- MIND THE GAP! -

GEOGRAPHIC TRANSFERABILITY OF ECONOMIC EVALUATION IN HEALTH

A thesis submitted for the degree of Doctor of Philosophy

By

Christian Ernst Heinrich Boehler

Health Economics Research Group (HERG) and

Multidisciplinary Assessment of Technology Centre for Healthcare (MATCH)

Brunel University

January, 2013

Abstract

Background: Transferring cost-effectiveness information between geographic domains offers the potential for more efficient use of analytical resources. However, it is difficult for decision-makers to know when they can rely on cost-effectiveness evidence produced for another context. **Objectives:** This thesis explores the transferability of economic evaluation results produced for one geographic area to another location of interest, and develops an approach to identify factors to predict when this is appropriate. **Methods:** Multilevel statistical models were developed for the integration of published international cost-effectiveness data to assess the impact of contextual effects on country-level; whilst controlling for baseline characteristics within, and across, a set of economic evaluation studies. Explanatory variables were derived from a list of factors suggested in the literature as possible constraints on the transferability of cost-effectiveness evidence. The approach was illustrated using published estimates of the cost-effectiveness of statins for the primary and secondary prevention of cardiovascular disease from 67 studies and related to 23 geographic domains, together with covariates on data, study and country-level. **Results:** The proportion of variation at the country-level observed depends on the appropriate multilevel model structure and never exceeds 15% for incremental effects and 21% for incremental cost. Key sources of variability are patient and disease characteristics, intervention cost and a number of methodological characteristics defined on the data-level. There were fewer significant covariates on the study and country-levels. **Conclusions:** Analysis suggests that variability in cost-effectiveness data is primarily due to differences between studies, not countries. Further, comparing different models suggests that data from multinational studies severely underestimates country-level variability. Additional research is needed to test the robustness of these conclusions on other sets of cost-effectiveness data, to further explore the appropriate set of covariates, and to foster the development of multilevel statistical modelling for economic evaluation data in health.

Contents

Tables

Figures

Abbreviations

ΔC	Incremental Cost
ΔE	Incremental Effect
CAD	Coronary Artery Disease
CD	cerebrovascular disease
CMM	Centre for Multilevel Modelling
CVD	Cardiovascular Disease
DAM	Decision Analytic Model
DBP	Diastolic Blood Pressure
DIC	Deviance Information Criterion
EPSRC	Engineering and Physics Science Research Council
ESRA	European Survey Research Society
ESRC	Economics and Social Science Research Council
FA	Factor Analysis
GDP	Gross Domestic Product
HDL	High Density Cholesterol Level
HESG	Health Economists Study Group
HTA	Health Technology Assessment
ICER	Incremental Cost-Effectiveness Ratio
IGLS	Iterative Generalised Least Squares
INMB	Incremental Net Monetary Benefit
IPD	Individual Patient Data
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
LDL	Low Density Cholesterol Level
LE	Life Expectancy
LYS	Life Year Saved
MAR	Missing At Random
MCA	Multiple Correspondence Analysis
MCAR	Missing Completely At Random
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
MLM	Multilevel Modelling
MNAR	Missing Not At Random
NHS	National Health Service
NICE	National Institute for Health and Clinical Excellence
NMB	Net Monetary Benefit
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Squares
PAD	Peripheral Arterial Disease
PCA	Principal Components Analysis
PCF	Principal Components Factor Analysis
PPP	Purchasing Power Parity

QALY	Quality adjusted Life Year
QHES	Quality of Health Economic Studies instrument
RCT	Randomised Controlled Trial
SBP	Systolic Blood Pressure
TCL	Total Cholesterol Level
VPC	Variance Partitioning Coefficient
WHO	World Health Organisation
WTP	Willingness to Pay

Acknowledgements

I am extremely grateful to my supervisors, Dr. Joanne Lord and Professor Martin J. Buxton for their excellent supervision throughout the development of this thesis. Only their guidance, advice and encouragement ensured the successful completion of this project.

Further, I would like to express my appreciation to my colleagues at the Health Economics Research Group (HERG) and the Multidisciplinary Assessment of Technology Centre for Healthcare (MATCH), especially Nicky Dunne, Eszter Nagy and Elisabeth Deadman for their continuing support throughout the course of this project. I would also like to thank my colleagues Catriona, Edit, Jeshika, Marta, Matthew, Nana, Subhash, as well as the entire HERG and MATCH team for their helpful comments and support at various stages of this project.

I thank the participants of the 2011 Health Economists' Study Group (HESG) meeting in York, participants of a multilevel session at the 2011 European Survey Research Society (ESRA) conference in Lausanne, Switzerland, fellow PhD-students participating at conferences of the German Academic Merit Foundation, as well as members of the Institute for Quality and Efficiency in Healthcare (IQWIG) in Cologne, the Department of Healthcare Management at Berlin Technical University and the Institut d'Economie et de Management de la Sante' (IEMS) at the University of Lausanne for their invaluable feedback from presenting this work. I take full responsibility of all the remaining errors and shortcomings of this thesis.

I wish to acknowledge the receipt of funding from MATCH, the Engineering and Physical Science Research Council (EPSRC), and the German Academic Merit Foundation.

All my love goes to my wife Timea and my son Benedek. Your patience and continuous encouragement over the years were so important for me. I would also like to express my unlimited gratitude for the forbearance and support I received from my mother and from my brother Martin. My thoughts are with my father in law, Dr. Lajos Helter, who gave me so much in the few years I had the chance to meet him.

Dedication

To my wife and son, Timea and Benedek

1. Introduction

To get the most benefit from healthcare resources available, we ought to know which technologies provide the best value given a restricted healthcare budget. Therefore, health economists conduct economic evaluation studies in order to compare different options to spend the limited resources available. Such economic evaluation studies may be defined as '*comparative analyses of alternative courses of action in terms of both their costs and consequences*' (Drummond et al., 2005a). Hence, '*for a meaningful comparison, it is necessary to examine the additional costs that one health care intervention imposes over another, compared to the additional benefits, or utilities it delivers'* (Drummond et al., 2005a).

Decision makers from an increasing number of countries require such cost-effectiveness data to inform the provision and reimbursement of new health technologies (Drummond et al., 2009). If a new health technology is to be launched in a specific country, manufacturers may therefore need to provide evidence not just of safety and clinical efficiency, but also of cost-effectiveness in the context of a particular healthcare market. However, this absorbs analytical resources, which are scarce and expensive. If economic evaluation results could be reliably transferred from one geographic domain to another, this would free analysts to study other important questions (Steuten et al., 2008). Therefore, it would be helpful to '*provide evidence for decision makers to establish the relevance or to adjust the results of a specific study to their location of interest'* (Sculpher et al., 2004).

A key barrier to transferability of economic evaluation results is a lack of understanding on the causes of variability in measures of cost-effectiveness. There is a plethora of literature discussing potential causes of variability (Sculpher et al., 2004; Goeree et al., 2007), but many of the suggested variability factors are fuzzy, hard or impossible to measure, and dependent on the health technology under consideration. Little is known about the relative impact of such

variability factors on measures of cost-effectiveness. For instance, Goeree et al. (2007) state that '*there is a lack of empirical studies which prevents stronger conclusions regarding which transferability factors are most important to consider and under which circumstances.'* In addition to that, Sculpher et al. (2004) highlight that '*research is required to identify (higher-level) covariates which are empirically useful [...] in terms of explaining differences in efficiency between locations and useful for policy-making purposes'*. This opinion is also being shared by Drummond et al. (2009) who suggest that 'm*ore research should be undertaken into those sources of local differences that affect economic data transferability. This would help justify jurisdiction-specific data requirements and inform the selection of jurisdiction-level covariates in statistical models.'*

1.1. Aims and objectives

To fill this research gap, this thesis aims to assess the impact of variability factors within and between economic evaluation studies and, ultimately, between geographic domains. In order to do so, a quantitative method is required to integrate international cost-effectiveness data elicited from published economic evaluation studies, which allows for the inclusion of covariates encoding variability factors working within and between studies and, ultimately, between countries represented in the data. Identifying such a method constitutes the first objective of this exercise, and it is argued in this thesis that multilevel statistical modelling (MLM) provides this methodological framework.

Having identified MLM as an appropriate method for the integration of international cost-effectiveness data from published economic evaluation studies, the second objective is to develop models which appropriately reflect the complex structures which are likely to be present in the data. These models need to take into account that cost-effectiveness data is grouped both in the studies which it was elicited from, and the countries which this data applies to. In addition, some studies provide data for more than one country, and sensible

assumptions regarding dependencies within this data are required to develop models which are suitable for the empirical analysis.

Once the appropriate multilevel model structure has been determined, the third objective is to control for variability factors working within and between studies represented in the data, which may also disclose further variability on country-level. For this reason, covariates which were drawn from a long list of variability factors as previously discussed in the relevant literature (Sculpher et al., 2004; Goeree et al., 2007) are being tested systematically within the multilevel models developed for the purposes of this project.

Only after controlling for factors causing variability in measures of cost-effectiveness within and between studies, can we appropriately assess the impact of country-level covariates. The reason is that variability caused by lower-level factors may feed through to higher levels. Hence, after controlling for such '*lower-level*' covariates, the fourth objective of this project is to investigate causes of country-level variability within the MLM framework. As a result, a number of country-level covariates may be identified which explain differences in incremental net monetary benefits (INMBs) as well as the stochastic components of the INMB statistic, i.e. incremental cost (ΔC) and incremental effects (ΔE) respectively. If variability in measures of cost-effectiveness turns out to be low between countries, an alternative objective is to assess reasons for a lack of country-level variation in international cost-effectiveness data.

Finally, the MLM framework offers a number of interesting analytical features, which have not yet been applied to the academic domain of economic evaluation in health. In particular, MLM allows modelling variation in the response variable directly as a function of explanatory variables (e.g. Steele, 2008). This concept of the '*variance function*' could be useful to show how variability in measures of cost-effectiveness between studies, or countries, changes as a function of explanatory variables. This may help identifying key areas within which disagreement between studies, or countries, is particularly high, and hence, the transfer of existing data to other geographic domains discouraged. As a result,

research resources may be focused on areas of high variability in existing international cost-effectiveness data, whilst the transfer of information may be rather indicated for situations where variability in existing cost-effectiveness evidence is low. For this reason, the final objective of this thesis is to explore analytic features of the MLM framework which may allow further insights into the geographic transferability of economic evaluation data in health.

The remainder of this introductory section provides a brief overview of the individual chapters of this thesis.

1.2. Overview of this thesis

The primary aim of this project is to address the transferability problem of economic evaluation in health by analysing what causes variability in measures of cost-effectiveness within and between studies, and ultimately, between geographic domains. For this reason, the literature concerned with the transferability and generalisability of economic evaluation in health is reviewed first in Section 2.1, confirming that '*the methods that have been proposed to address the transferability issue have often been relatively ad hoc, with the obvious consequence that the methodological literature in this area has evolved somewhat nonlinearly over time*' (Manca, 2009). The available literature also confirms that '*there is a lack of empirical studies which prevents stronger conclusions regarding which transferability factors are most important to consider and under which circumstances' (*Goeree et al., 2007). As a consequence, an analytic strategy for the purposes of this project is developed in Chapter 2.2, identifying MLM as a promising method for analysing factors causing variability in measures of cost-effectiveness.

The use of MLM for analysing variability factors for measures of cost-effectiveness is also in accord with the theoretical framework for the transfer of evidence between geographic domains developed in Chapter 2.3. In this section,

the transferability problem is described as an '*analogical inference*', where a mapping of relevant attributes between a source domain, about which more is apparently known, and a, less studied, target domain is produced to infer whether the information of interest may also hold in the target setting (Gentner & Markman, 1997; Forbus, 2001; French, 2002). This theoretical framework is then linked to the statistical concept of '*exchangeability*', which forms the conceptual basis of MLM. Multilevel modelling makes explicit the exchangeability assumption and allows for the assessment of variability factors for cost-effectiveness data within studies, between studies, and ultimately, between countries through the assumption of conditional independence (Drummond et al., 2009). Chapter 2 therefore concludes with systematically reviewing the use and applications of MLM in the area of economic evaluation in health, and shows that all applications of MLM in this area focus on the analysis of individual patient data from multicentre trials or observational studies, within which a strict two-level hierarchical data structure is commonly assumed.

Hence, using MLM as a mode for meta-regressing secondary cost-effectiveness data from published economic evaluation studies, as it is aimed in this project, is a novelty in this area, which is why Chapter 3 is dedicated to developing and testing methods to integrate cost-effectiveness data from different studies and applicable to different geographic domains. Starting with a simple ordinary least squares (OLS) regression equation, this chapter introduces, step by step, the features required to model complex data structures and (partial) exchangeability not just between studies, but also the geographic domains represented in the dataset. Within this process, a number of strictly hierarchical as well as cross-classified models are developed both within a univariate framework, with INMBs as single response variable, and a bivariate framework, with the stochastic components of the INMB statistics (ΔC and ΔE) as a vector of response variables.

Models are subsequently tested within a pilot study reported in Chapter 3.4 using a subset of cost-effectiveness data on statins for the primary and secondary prevention of cardiovascular disease (CVD). This intervention was chosen as it has been extensively researched in the past, suggesting that data from a sufficient number of includable studies and geographic locations will

justify the assumption of random parameters on study and country-level (Snijders, 2005). Results from the pilot show that the use of MLM for secondary data integration is promising and in line with Gelman et al. (2004), who state that '*the valid concern is not about exchangeability, but encoding relevant knowledge as explanatory variables where possible.'* However, the pilot study also shows the importance of making appropriate assumptions about (in-) dependencies in the data, especially with respect to measures of cost-effectiveness on country-level elicited from multinational economic evaluation studies.

A systematic literature review and data abstraction exercise is then reported in Chapter 4, which has the aim of populating a dataset with incremental cost, incremental effects and INMBs on statins for the primary and secondary prevention of CVD; together with additional data encoding potential variability factors for measures of cost-effectiveness. In total, 67 relevant studies were includable in this empirical exercise, reporting 2094 cost-effectiveness estimates applicable to 23 geographic domains. Covariates were derived from a long list of potential variability factors as previously reported in the literature (Sculpher et al., 2004; Goeree et al., 2007). Results of most studies refer to one geographic domain only, whilst six studies are multinational in nature.

When carrying out the systematic literature review and abstracting data from studies includable in this empirical exercise, it was apparent that some studies related to each other, for instance, through common authorship, the use of identical data sources, reuse of a previously published decision analytic model (DAM), or simply a common source of funding. This may violate the independence assumption between studies, which is necessary to fit the MLMs developed in Chapter 3. Therefore, Chapter 4 also looks into the '*genealogy*' of economic evaluation studies on the cost-effectiveness of statins in the primary and secondary prevention of CVD. Multiple correspondence analysis is used to ascertain whether studies are similar with respect to key study characteristics, and once a '*phenotypic*' similarity is disclosed, a '*genotypic*' relationship between studies is aimed to be established. This exercise resulted in some relationships being disclosed amongst the studies included in the dataset; however, the method does not (yet) prove sensitive or specific enough to justify alternative

MLM structures. Rather, a number of explanatory variables are derived with the aim of encoding existing relationships between studies, and it is concluded that this exploratory task into the genealogy of economic evaluation studies should be followed up further in future research.

Chapter 5 is concerned with the main empirical analysis of this project, and this analysis is partitioned in accord with the objectives outlined above. The first objective, assessed in Section 5.1, is to determine the appropriate MLM structure for this empirical analysis. This assessment is not just concerned with testing which MLM previously developed works well on the data collected, but also whether assumptions made in these models are justified for the data. In particular, this section shows that appropriate assumptions regarding (in-) dependencies are crucial for making correct inferences when analysing secondary cost-effectiveness data. For instance, if data from multinational studies shows much lower country-level variability, the independence assumption between countries may not be justified for this data.

The analysis in Section 5.1 also demonstrates the benefits of decomposing the INMB statistic into its components ΔC and ΔE in the bivariate framework and shows that part of the variability in international cost-effectiveness data on statins '*disappears'* when combining ΔC and ΔE to the INMB statistic. This very interesting finding is also subject to further analysis in Section 5.3, which is concerned with country-level variability, or the lack thereof, both within the univariate and bivariate MLM framework. Finally, Section 5.1 assesses, in depth, whether '*empirical Bayes shrinkage estimation*' may be regarded as appropriate in a model which attempts to integrate secondary data from published economic evaluation studies, where the weight of a particular study does not depend on individual patients considered, but rather on the extent to which subgroup and sensitivity analyses have been reported. It is shown that, due to high between group variability in the data, shrinkage factors are very high, which means that shrinkage is, at most, moderate. More importantly, however, this section argues that the impact of shrinkage on study means in this exercise depends not just on the respective number of data points from each study, but also on the within and

between group variability and the location of each study mean relative to the overall regression mean.

Section 5.2 is concerned with covariate adjustment on data and study-level, and to disclose the maximum amount of variability on country-level through controlling for multiple variability factors working within and between studies in the dataset. Covariates are drawn from a long list of variability factors as obtained from the literature (Sculpher et al., 2004; Goeree et al., 2007) and abstracted from the studies included in the systematic literature review reported in Chapter 4. Results show, for instance, that the effect of lower-level variability factors feeds through to higher levels so that the actual amount of country-level variability may only be unravelled by including relevant covariates both on data and study-level. The analysis also shows that country-level variability is increasing with the inclusion of lower-level covariates in the bivariate model, which may allow assessment of covariates on country-level in the bivariate framework. However, a different conclusion applies to the univariate framework, where country-level variability remains negligible throughout the course of this exercise. Section 5.2 also provides a number of interesting findings with respect to individual covariates tested on data and study-level, which are also discussed in depth in Chapter 6.

Section 5.3 is concerned with variability in measures of cost-effectiveness between countries and this analysis consists of two parts. Part one analyses potential causes for a lack of country-level variability in the univariate MLM with INMBs as single response variable. In part two, country-level covariates are tested in the bivariate framework, within which considerably more country-level variability was identified. To analyse potential causes for a lack of country-level variation in the univariate MLM, forest plots with country means and their respective confidence intervals are presented for each response variable (INMB, ΔC and ΔE). In addition, Pearson correlations for mean ΔC and ΔE are highly significant and close to unity, indicating that the lack of country-level variability in INMBs results from combining ΔC's and ΔE's which have similar patterns of variability - meaning that variability in one component of the INMB statistic is partly being offset by variability in the other component. Testing country-level

covariates in the bivariate model show small but significant coefficients for a number of explanatory variables and results are subsequently discussed in Chapter 6.

The final Section 5.4 of the empirical chapter is concerned with additional methodological features of the MLM framework which may be beneficial for addressing the transferability problem of economic evaluation data in health. Random slopes are fitted to covariates in the model and variation in international cost-effectiveness data is modelled directly as a function of explanatory variables. In this section, it is argued that the concept of the '*variance function*' relates to the transferability problem as it may be used to determine which of the available cost-effectiveness information is rather transferable to the target country, and where to prioritize research resources to generate new target specific cost-effectiveness evidence. Results show, for instance, that variability between studies is constantly increasing for the relationship between INMBs and total cholesterol (TCL) and ΔE's and TCL respectively; so that results may be less transferable the higher the total cholesterol level of the target population.

A number of issues are identified with respect to the variance function, which are discussed in far more detail in the discussion Chapter 6. For instance, can we determine a '*threshold value*' for study-level variability which may be helpful to guide the decision on whether or not to transfer existing evidence to the target country? Also, are there additional application areas where modelling the variance function may be useful, for instance within the context of international multicentre trials. Apart from that, Chapter 6 provides a thorough discussion on other findings of this project, policy implications, strength and weaknesses of the empirical analysis, and suggested areas for further research.

2. Background

This chapter provides the necessary background and develops a research strategy for this project. The economic evaluation literature on the transferability /generalisability of measures of cost-effectiveness is reviewed, identifying a number of general '*research themes'* within this topic. Based on this summary of the relevant literature, a research strategy is developed. Third, to provide a solid theoretical basis for this thesis, a theoretical framework is developed for the transfer of evidence between geographic domains. The transferability problem may best be described as an '*analogical inference*', where a mapping of relevant attributes between a source domain, about which more is apparently known, and a, less studied, target domain is produced to infer whether the information of interest may also hold in the target setting (Gentner & Markman, 1997; Forbus, 2001; French, 2002). This theoretical framework of analogical reasoning is then linked to the statistical concept of '*exchangeability*', which forms the conceptual basis of the empirical work of this project.

As a result, multilevel statistical modelling, as previously identified by Rice & Jones (1997), Sculpher et al. (2004), Drummond et al. (2009) and others is used in this thesis to analyse factors causing variability in measures of cost-effectiveness. MLM explicitly models the exchangeability assumption and, by relaxing this assumption, allows for the assessment of variability factors of measures of cost-effectiveness within studies, between studies, and ultimately, between geographic domains through the assumption of conditional independence (Drummond et al., 2009). The final section of this chapter therefore reviews the use and applications of MLM in the area of economic evaluation in health, before Chapter 3 reports in detail on the MLM methods developed for the purposes of this project and a pilot study to test these methods.

2.1. Existing literature on the transferability/ generalisability of economic evaluation in health

There is a considerable body of literature concerned with the transferability / generalisability of economic evaluation data and Manca (2009) states that *'the methods that have been proposed to address the transferability issue have often been relatively ad hoc, with the obvious consequence that the methodological literature in this area has evolved somewhat nonlinearly over time*'. This section therefore aims to bring some order into this body of literature. Though it was intended to thoroughly discuss relevant publications, this literature review may not be labelled as '*systematic*'. The '*nonlinear*' development of this field of research led to a large number of potential search words which also appear very regularly in unrelated publications. This makes it difficult defining a sufficiently '*sensitive'* search strategy which is also '*specific'* enough to obtain a manageable number of potentially relevant hits from searching scientific databases. On the other hand, however, a number of fairly recent key publications exist which draw together part of the relevant literature. Therefore, a rather '*organic*' search strategy was applied, starting off from some key publications in the area (for instance, Sculpher et al., 2004; Goeree et al., 2007 & 2011; Drummond et al., 2009) and then systematically following up papers which cited, or were cited, in these publications.

One of the first papers focussing on the transferability problem was written by Bernie O'Brien (1997). '*Because replication of trials is an expensive and inefficient undertaking*', he argued, '*analysts need to determine the validity of transferring cost-effectiveness data from one country to another*'. He further identified six '*threats*' to the transferability of economic evaluation data, namely 1) demography and epidemiology, 2) clinical practice and conventions, 3) incentives and regulations for healthcare providers, 4) relative price levels, 5) consumer preferences and 6) opportunity cost of resources. Much of the subsequent work on the transferability problem in general, but also on factors causing variability in measures of cost-effectiveness in particular (e.g. Drummond & Pang, 2001; Sculpher et al., 2004; Goeree et al., 2007) builds upon O'Brien (1997). Barbieri et al. (2005), for instance, systematically reviewed the literature to identify

economic evaluation studies conducted for two or more countries in order to assess their level of variability in measures of cost-effectiveness and the main causes of this variation (i.e. variability factors). It was also assessed whether differences in results would lead to different decisions in different countries. Results suggested that the type of economic evaluation study (i.e. trial based or decision analytic modelling based) had some impact on variability, but that the '*most important factor was the extent of variation across countries in effectiveness, resource use or unit costs, allowed by the researchers chosen methodology*'. The authors also devised a classification of studies with respect to their likely degree of variability in measures of cost-effectiveness.

Another body of literature emerged with respect to critical appraisal methods for the transferability potential of economic evaluation in health. In 1999, Späth et al. published a paper which aimed to define a method for assessing the eligibility of published economic evaluation studies for transfer to various settings in a given healthcare system. They built up from work undertaken by Heyland et al. (1996), who developed a basic '*transferability checklist'* to critically appraise the potential of transferring economic evaluation data from one context to another. Subsequently, a whole body of literature emerged concerning transferability checklists, decision charts, or indices, some of the most prominent examples were provided by Welte et al. (2004), Boulenger et al. (2005) the EUnetHTA adaption toolkit (Turner et al., 2009), EURONHEED (Nixon et al., 2004; Nixon et al., 2009) or Antonanzas et al. (2009). Work was also undertaken to validate or empirically apply transferability checklists (e.g. Knies et al., 2009; Essers et al., 2010; Wolfenstetter & Wenig, 2010), and a systematic review on critical appraisal tools was recently published by Goeree et al. (2011).

Meanwhile, other researchers looked at the transferability problem from a different angle. Their approach was to increase the generalisability of economic evaluation data by harmonizing HTA and economic evaluation methods across HTA agencies and geographic jurisdictions (e.g. Hjelmgrien et al, 2001). Sculpher & Drummond (2006) state that '*decision makers and analysts need to work together and where possible harmonize guidelines on methods for economic evaluations whilst recognising legitimate variation in the needs of different*

healthcare systems.' However, Birch & Gafni argued already in 2003, that '*the ICER associated with maximising health benefits for the community cannot be determined in isolation of the community context.*' Further, '*the validity of the method of valuation cannot be established independent of the setting in which it is to be used*' and finally, even if the methods of valuation are valid in each setting, the authors pose the questions of whether this implies that '*numbers produced by application of the methods are generalisable across individuals and settings*.' Birch & Gafni (2003) conclude that the '*generalisability of the validity of a method of valuation does not imply generalisability of the resulting valuations*' and with respect to the transferability problem in general, they conclude that '*the economic question of whether an activity adds more to well-being than the alternative uses of the same resources in a particular community cannot be answered by reference to the costs and consequences of the same activity in a different community*'. A similar opinion is shared by Vale (2010), who argues that *'despite common principles, the process of HTA, and more particularly its economic evaluation component, needs to take a national approach toward evaluation.'* Nevertheless, significant research into the harmonization of HTA guidelines and methods for economic evaluation in health has been carried out, for instance within the EUnetHTA WP 4, developing the 'CORE HTA Model' (EUnetHTA, 2008) or WHO work on '*generalised cost-effectiveness analysis'*, which may also be subsumed under this category (Murray et al., 2000; Tan-Torres Edejer, 2003).

Apart from the general research themes as outlined above, authors addressed transferability issues of particular relevance for decision analytic modelling studies on the one hand, and trial based analysis of individual patient data on the other. In 2004, Sculpher et al. published a landmark study on the generalisability of economic evaluation in health. In their work, which formed part of the NHS R&D HTA programme, they first conducted a number of systematic reviews on:

- Factors causing variability in economic evaluation studies
- Methods used to assess variability and enhance generalisability
    - in decision analytic modelling based economic evaluations
    - in trial based economic evaluations

- A systematic review on economic evaluations undertaken alongside multicentre randomised controlled trials and
- A structured review on model based economic evaluation studies in osteoporosis (results of this review were also published separately in *'Pharmacoeconomics'* (Urdahl et al., 2006)).

Further to that, Sculpher et al. (2004) produced two case studies, one on assessing generalisability in trial based economic evaluations using MLM, and another one on making economic evaluation more location specific using decision analytic modelling. With respect to DAM, Sculpher et al. (2004) mention that '*it facilitates the synthesis of data from several sources'*. As such, decision analytic models may be used for instance to:

- adjust trial results to reflect routine practice
- extend trial results to non-trial locations
- substitute input parameters with location specific data or
- build generic models which are then tailored to specific locations.

With respect to trial based economic evaluations, Sculpher et al. (2004) identify methods for increasing generalisabiliy relating to a) study design and b) data analysis. In terms of the design of studies, methods identified relate to cost estimation, currency conversion, centre selection, randomisation, data collection and adjustments to bridge data generated under artificial trial conditions to routine practice. With respect to the analysis of IPD from trial based economic evaluation studies, Sculpher et al. (2004) highlight the paramount importance of making explicit assumptions of (partial) exchangeability of economic evaluation data between centres and countries represented in multinational RCTs. Assuming exchangeability means that there are no a priori reasons to expect more or less favourable estimates of cost-effectiveness between centres or locations represented in the data (Drummond et al., 2009). According to Sculpher et al. (2004), one method to explore issues of (partial) exchangeability in economic evaluations alongside RCTs is the use of MLM to '*analyse data that fall naturally into hierarchical structures consisting of multiple macro units (contexts) and multiple micro units within each macro unit*' (Rice & Jones, 1997 cited from Sculpher et al., 2004)

With respect to MLM, Sculpher et al. (2004) conclude that further research should relate to 1) the overall specification of models, 2) the selection of patient and location specific covariates and the specification of their interaction with treatment, 3) the appropriate MLM approach when there are a number of levels in the data hierarchy (e.g. patients, surgeons, centres, countries), 4) appropriate methods when there are few locations in a trial, and 5) the use of Bayesian approaches to MLM. Furthermore, the '*assessment of alternative approaches to specifying multilevel models to the analysis of cost-effectiveness data alongside multilocation randomised trials*' and the '*identification of a range of appropriate covariates relating to locations (e.g. hospitals) in multilevel models'* were identified as overall priorities for further research with respect to the transferability / generalisability of economic evaluation in health (Note that the MLM literature related to health economic evaluation is also subject of a systematic literature review further below in this chapter).

The recommendations of Sculpher et al. (2004) with respect to both trial based and DAM based economic evaluations were subsequently drawn together and published by Drummond et al. (2005). Meanwhile, the transferability literature on both trial based and model based economic evaluations developed further. Manca & Willan (2006) proposed an algorithm to assist the choice of the appropriate analytical strategy when facing transferability issues in practice. Different scenarios were considered based on a) whether a country of interest participated in a trial and b) whether individual patient level data is available for that trial. This work differs from transferability checklists as it assists in determining the appropriate method to assess or enhance transferability, not to assess the transferability potential of the data itself. Again, the use of MLM was proposed as it mediates between the two extreme assumptions of either '*pooling*' or '*splitting*' the data from multinational trials. In addition, the potential of assessing both patient and country-level covariates within the MLM framework was highlighted as a particular advantage of this analytic approach (Manca & Willan, 2006). In the meantime, Mason & Mason (2006) reviewed the literature on the generalisability and transferability of economic evaluation in health and identified current issues within this area of research, one of which being, again, the use of MLM to assess the transferability of findings from trial based economic evaluations.

Another key paper was recently published by Drummond et al. (2009), which addresses issues falling into almost all of the research themes outlined above. Based on work by the ISPOR Good Research Practices Task Force, Drummond et al. first review what national economic evaluation guidelines suggest with respect to the transferability problem, discuss elements which may be deemed transferable, and make good research practices recommendations for both trial based and decision analytic modelling based economic evaluations. According to Drummond et al. (2009), most guidelines recognise the potential for differences in effectiveness data and suggest the transferability of relative risk reduction, whilst baseline risks should be context specific. With respect to health state valuations, only 8 of 21 methods guidelines make any recommendations in terms of the transferability of such data. For resource use and unit cost estimates, most guidelines agree that both should be reported separately, and all guidelines specify that unit cost shall be location specific.

Secondly, Drummond et al. (2009) propose an algorithm to determine whether simple or more elaborate methods are required for adjusting study results to the location of interest. This algorithm consists of four general steps and is partly based on Welte's decision chart for the transferability of economic evaluation data. The first step determines whether cost-effectiveness information is available, the second step specifies whether this data may be relevant to the decision problem and whether the methodology is deemed appropriate (this step is based on criteria proposed by Welte et al., 2004). The third step considers whether treatment patterns are comparable between the existing data and the location of interest, and the final step considers whether the cost-effectiveness data is based on a multilocation trial which includes the location of interest.

The remainder of the paper published by Drummond et al. (2009) is concerned with specific transferability/ generalisability issues with respect to either trial based or model based economic evaluation studies. In terms of DAM based studies, the authors identify situations when this may be the preferred vehicle for economic evaluation in health. Precisely, if:

- trials were undertaken wholly outside the jurisdiction of interest and one or more components of evidence cannot be generalised across jurisdictions
- to synthesize data from multiple sources of evidence relating to any aspect of the analysis
- To adjust aspects of the trial (e.g. time horizon) to what is considered relevant and appropriate in the location of interest
- To adapt a DAM developed for another country to the country of interest

Further issues relating to the geographic transferability of DAM based economic evaluation were mentioned, for instance determining the model structure, parameter estimation, or analysis as '*different jurisdictions require different analytical methods'*.

Finally, with respect to IPD analysis from randomised controlled trials, Drummond et al. (2009), state that '*with respect to transferability, analytic approaches address two sets of objectives. The first is to evaluate whether there is evidence of heterogeneity in patterns of resource use, costs, survival, and / or utilities and to explore potential sources of heterogeneity. The second objective is to obtain estimates of incremental resource use, cost, and/ or cost-effectiveness that are appropriate for decision making within particular jurisdictions that may or may not have been included in the trial*.' Accordingly, Drummond et al. (2009) identify three categories of statistical methods for analysis of IPD data, namely 1) the detection of heterogeneity, 2) fixed effects models and 3) multilevel or hierarchical models. With respect to MLM, the authors state that they:

- '*can appropriately handle the hierarchical nature of the data that manifests itself as a lack of independence of the errors between the observations'*
- '*provide the formal means of estimating jurisdiction specific measures of cost-effectiveness'*
- 'p*rovide a pooled, random effects estimate across all jurisdictions, equivalent to a random effects summary estimate from meta-analysis'*

- model partial exchangeability through the inclusion of covariates on each level of the modelled data hierarchy.

With respect to covariates, the authors mention that '*currently, evidence is lacking as to what types of higher-level covariates may be useful in this regard, but candidate variables may include those that are indicative of macroeconomic characteristics (e.g. gross national product), capacity constraints (e.g. limited availability of intensive care beds), economic incentives (e.g. pressure to minimise length of stay), or financing characteristics (e.g. global budgets vs. fee for service reimbursement).*' Finally, the authors highlight that the body of literature on MLM within economic evaluation is currently evolving so that it may be premature to comment on best practices or to suggest situations where multilevel modelling may offer advantages over simpler analytic approaches.

Table 2.1 General research themes emerging from the available literature

1. Assessing factors causing variability in measures of cost-effectiveness
2. Standardising economic evaluation methodology across HTA bodies and countries to increase the generalisability of findings
3. Developing or testing methods for the critical appraisal of economic evaluation in health resulting in transferability checklists, decision charts and indices
4. Using decision analytic modelling
    a. to adjust trial results to reflect routine practice (though this particular topic relates to another *'facet'* of generalisability which is not the focus here)
    b. to extend trial results to non-trial locations
    c. To apply decision models for one particular location to another or
    d. to build generic models which may then be tailored to specific locations of interest
5. Issues relating to economic evaluation alongside randomised controlled trials
    a. with respect to the design of economic evaluations conducted alongside multinational randomised controlled trials and
    b. with respect to developing methods of data analysis which appropriately reflect the assumption of (partial) exchangeability between locations, in particular the use of multilevel statistical modelling

This section aimed to provide an overview of the available literature on the transferability / generalisability of economic evaluation in health and identified a number of research *'themes'* to be followed up further within this thesis These

themes are also summarized in Table 2.1 above. In addition, MLM has already been identified as a promising method to address the transferability problem by a number of authors. The following section aims to describe the process within which the actual research strategy for this project evolved out of the themes identified above.

2.2. Development of a research strategy based on the available literature

Some of the research themes mentioned in Table 2.1 were dropped early from further consideration within the process of determining an appropriate research strategy for this project. This holds specifically for research into the standardisation of international HTA and economic evaluation methods. For instance, at the time the author of this thesis looked into potential areas for further research within this project (October 2008 to June 2009) considerable work was already in progress with respect to the CORE-HTA model (EunetHTA, 2008). This model is '*an attempt to define and standardise elements of an HTA'*. (EUnetHTA, 2008). It does so by dividing relevant information on the technology under assessment *'into standardised pieces, each of which describing one or more aspects of the technology that is likely to be useful when considering the adoption or rejection of the technology. (…) The elements that are most likely useful for international sharing of information are defined as core elements*.' (EUnetHTA, 2008). The CORE HTA model has since been adapted to two clinical intervention areas (one for medical and surgical interventions and one for diagnostic technologies) and research into the CORE-HTA model is still going on to date (EUnetHTA-web-link). As a result, trying to contribute to this growing research area would also mean to compete with at least one large scale multinational research collaboration which already builds up from years of experience. Findings within this thesis would probably quickly become outdated, or even contradicted from newly emerging evidence within this subject.

Conducting empirical work concerned with critical appraisal methods for the transferability potential of economic evaluation data (i.e. checklists, decision charts and indices) was rejected for another reason. Over the years, a number of tools emerged within this area (Goeree et al., 2011), and any work related to this subject would most likely end up with either empirically testing developed methods or contributing towards the development of another transferability checklist. Some work has been carried out already to test or empirically apply transferability checklists (e.g. Knies et al., 2009; Essers et al., 2010; Wolfenstetter & Wenig, 2010), and in the light of a vast number of checklists existing already, it was not deemed a high research priority to dedicate this project to the development of (yet another) tool for critically appraising the transferability potential of economic evaluation in health.

With respect to factors causing variability in measures of cost-effectiveness, a different conclusion was reached. Sculpher et al. (2004) systematically reviewed the economic evaluation literature to ascertain such variability factors. Both conceptual papers discussing variability factors as well as empirical papers trying to estimate the variability in cost-effectiveness results were considered. As a result, they compiled a list of 27 unique factors suspected to cause variability in measures of cost-effectiveness (Table 2.2), and these factors were grouped in characteristics of a) the patient, b) the clinician, c) the healthcare system and d) wider socioeconomic factors. Sculpher et al. (2004) also reviewed studies which empirically tried to assess sources of variation in cost-effectiveness data. According to the authors, both decision analytic models and studies based on individual patient data have been used to estimate variability between geographic locations. In both cases, the most common way to estimate variability by location was to substitute location specific unit cost data and thereby increasing the applicability of results to the respective target location. Apart from unit cost estimates, variation with respect to resource use patterns was quite commonly subject of analysis. However, Sculpher et al. (2004) state that most studies reviewed were standard economic evaluations with the aim to provide results for a number of geographic locations. Only very few studies set out with the specific aim to measure variation between locations, and these studies were *'rather descriptive than evaluative*' in nature.

Table 2.2: Variability factors reported by Sculpher et al (2004)

Factor	Definition
Absolute / relative costs	Unit costs/prices of inputs into healthcare
Artificial study conditions	Research environment versus routine practice
Capacity utilisation	Level of utilisation of inputs into healthcare
Case mix	Clinical and socio-demographic characteristics of patients undergoing treatment
Clinical practice variation	Variation in how healthcare is delivered
Compliance	Adherence to treatment regimen
Culture / attitudes	As affecting clinical practice
Demography	Patient non-clinical characteristics, e.g. sex, age
Disease Interaction	Association of primary disease with risk factors, other morbidity/mortality
Economies of Scale	Greater levels of 'production' leads to lower costs
Epidemiology	Incidence/prevalence of disease
Exchange Rates	Conversion rate of different currencies
Geographical Setting	Location such as country, type of facility
Health State valuations	Individuals' preferences for particular levels of health
Healthcare resources	Inputs into health delivery, e.g. personnel, equipment
Healthcare system	Regulatory and organisational infrastructure
Historical differences	History of organisation/practice
Incentives	Financial and other factors which affect individuals and organisational behaviour
Industry-related bias	Sponsor influence on study results
Joint production	Inputs into healthcare delivery are shared between different units/departments
Opportunity cost	Health benefits forgone by use of a resource in a particular way
Perspective	Viewpoint of economic analysis
Skills / Experience	Level of training and experience of health professional
Technological Innovation	Advancement of technology/practice
Timing of evaluation	Stage of conduct of study in the development of the technology
Treatment comparators	Available treatment options

Following the work of Sculpher et al. (2004), only one further study was found which aimed to assess variability in economic evaluation data (Barbieri et al., 2005). In their work, the authors systematically reviewed the literature to identify economic evaluation studies conducted for two or more countries in order to assess their level of variability in measures of cost-effectiveness and the main causes of this variation (i.e. variability factors). It was also assessed whether differences in results would lead to different decisions in different countries. Results suggested that the type of economic evaluation study (i.e. trial based or DAM based) had some impact on variability, but that the '*most important factor was the extent of variation across countries in effectiveness, resource use or unit costs, allowed by the researchers chosen methodology*'. The authors rightly claim

that - up to this point - their work constitutes '*the most comprehensive analysis of the variation in the results of cost-effectiveness studies, of drugs, in Europe*'.

Table 2.3: Variability factors reported by Goeree et al. (2007)

Patient Characteristics
- Demographics (age, gender, race), education, socio-economic status
- Risk factors, medical history, genetic factors
- Lifestyle, environmental factors
- Mortality rates, life expectancy
- Attitudes toward treatment, culture, religion, hygiene, nutrition
- Compliance and adherence rates, ethical standards
- Population values (utilities)
- Population density, immigration, emigration, travelling patterns
- Income, employment rates, productivity, work loss time, friction time
- Type of insurance coverage, user fees, co-payments, deductables
- Incentives for patients

Disease Characteristics
- Epidemiology (incidence/prevalence, disease progression, spread)
- Disease severity, case mix
- Disease interaction, co-morbidity, concurrent medications
- Mortality due to disease

Provider Characteristics
- Clinical practice, conventions, guidelines, norms
- Experience, education, training, skills, learning curve position
- quality of care provided
- Method of remuneration (supplier induced demand)
- Patient identification
- Cultural attitudes
- Incentives for providers, liability

Health care system characteristics
- Absolute or relative prices
- Available resources (staff, facilities, equipment), programs, services
- Organisation of delivery system, structure, level of competition
- Level of technological advancement, innovation and availability
- Available treatment options (comparators)
- Capacity utilisation, economies of scale, technical efficiency
- Input mix (personnel, equipment), specialisation of labour, joint production
- Access to programs and services, gatekeepers, historical differences
- Waiting lists, referral patterns
- Regulatory and organisational infrastructure, licensing of products
- Availability of generics or substitutes
- Market form of suppliers, payment of suppliers, supplier incentives
- Incentives from institutions

Methodological characteristics
- Costing methodology, estimation procedures (e.g. productivity cost)
- Study perspective
- Study factors (artificial trial conditions, industry related bias)
- Timing of the economic evaluation
- Clinical endpoints/outcome measures
- Discount rates
- Exchange rates, purchasing power parities
- Opportunity cost (foregone benefits)
- Affordability (CE thresholds)

Later, Goeree et al. (2007) provided a second systematic review on factors causing variability in measures of cost-effectiveness. With their work, they basically confirm the results previously published by Sculpher et al. (2004) and expand upon the list of factors potentially causing variability in measures of cost-effectiveness. In total, 77 unique transferability factors were found which were grouped by the authors in characteristics of a) the patient, b) the disease, c) the provider, d) the healthcare system and e) methodological factors causing variability in measures of cost-effectiveness (Table 2.3 above). Most importantly, Goeree et al. (2007) confirm what has been identified as a potential gap in research before. In their review, the authors clearly state that '*there is a lack of empirical studies which prevents stronger conclusions regarding which transferability factors are most important to consider and under which circumstances*'

Hence, for this thesis, it was concluded that a systematic assessment of factors causing variability in measures of cost-effectiveness constitutes an important research theme to consider. At this point, however, there was no clear strategy in terms of how to conceptualise such an assessment. On the highest level, this issue relates to a '*false dichotomy*' (Drummond et al., 2009) namely the question of whether to work within the context of individual patient data analysis from randomised controlled trials, or within the framework of decision analytic modelling to assess the cost-effectiveness of healthcare technologies.

With respect to the transferability problem, DAM has been previously used mainly to a) adjust trial results to reflect routine practice, b) to extend trial results to non-trial locations, c) to adapt decision models for one particular country to another or d) to build generic models which are then tailored to specific locations of interest (Sculpher et al, 2004; Drummond et al, 2009). Clearly, one may reasonably argue that adjusting trial results to routine practice (and with it the assessment of relating variability factors) does not closely relate to the aim of this project, which is rather concerned with the geographic transferability of economic evaluation data. Secondly, it was deemed highly unlikely to obtain access to IPD from one or more randomized controlled trials, which would be a pre-requisite to carry out work in this area going beyond

purely theoretical or conceptual/methodological contributions. Though the second subtheme (extending trial results to non-trial locations) falls much more into the scope of this thesis, the same problem regarding data availability applies, which ruled out this option for the purposes of this project.

As a result, considerable time was devoted to explore strategies to analyse variability factors using DAM either in terms of making a general model location specific or in terms of adapting a model from one jurisdiction to another. With each step of adjustment, one may measure the variability in cost-effectiveness estimates, thereby analysing the factors which cause this variability in international cost-effectiveness data. Though this idea looked appealing at first, serious limitations led to the decision of not following up this strategy any further. In particular, both the '*baseline model'* as well as adjustments to the model to transfer it to another jurisdiction would be subject to choices and assumptions made by the author of this thesis. This obviously bears the risk of bias as having influence on the study protocol at any time of the process might influence the results towards '*what someone seeks to show'*. This limitation was perceived as too strong so that it constituted a knock out criterion for this analysis strategy.

Having ruled out DAM, multilevel modelling was considered for the assessment of factors causing variability in measures of cost-effectiveness. MLM allows reflecting complex data structures which arise from non-independence of error terms within groups like centres or countries reflected in the dataset (e.g. Sculpher et al., 2004), it also allows for the inclusion of covariates at any level of the data hierarchy through the assumption of conditional independence (Drummond et al., 2009). Variability factors may hence be assessed as covariates on different levels of the data hierarchy. However, this method has been identified within the context of IPD analysis, and it was deemed unlikely to obtain access to such data within this project. Therefore, the work by Barbieri et al. (2005), which was summarised above, demonstrated a promising alternative solution. Instead of IPD from multinational trials, one may use secondary data from published economic evaluation studies to populate a dataset for the

assessment of factors causing variability in measures of cost-effectiveness using MLM. This approach offers several advantages, for instance:

- Access to secondary cost-effectiveness data from published economic evaluation studies is unproblematic. Hence, a systematic literature review on a particular healthcare intervention could be carried out to populate a dataset with secondary cost-effectiveness data as well as additional variables for covariate adjustment on each hierarchical level within the MLM framework

- Secondly, despite technical challenges which are addressed in more detail in Chapters 3 and 5, the approach of integrating secondary data from published studies constitutes a more appropriate basis for the assessment of variability factors than the use of IPD from multinational trials. Trials usually implement strict protocols which may be identical across centres and geographic domains (Ramsey et al., 2005). These protocols, though crucial to ensure internal validity of the trial results, artificially reduce the variability which is likely to exist between centres and countries under real world conditions (Ramsey et al., 2005). Secondary data from published economic evaluation studies, which are usually designed to inform decisions under real world conditions, may better reflect this variability and therefore constitute a more appropriate basis for the systematic assessment of variability factors for measures of cost-effectiveness. Furthermore, utilising IPD from one trial may not allow for the assessment of methodological factors causing variability in measures of cost-effectiveness between studies, which has been identified as an important source of variability in cost-effectiveness data by Barbieri et al (2005).

- With respect to the analytic technique, MLM makes explicit the exchangeability assumption, which '*mediates*' between assumptions of either identical (pooling) or independent (splitting) parameters (e.g. Spiegelhalter et al., 2000 & 2004). This assumption of exchangeability (between studies and countries) may allow for the integration of secondary cost-effectiveness data from different studies and different geographic domains without ignoring the fact that study and country

residuals are not independent. Subsequent chapters may show whether this hypothesized advantage may also hold in practice.

- Further, through the assumption of conditional independence (or partial exchangeability) one may assess the impact of variability factors modelled as covariates on each level of the data hierarchy. As Gelman et al. (2004) put it *'In this way exchangeable models become almost universally applicable, because any information to distinguish different units should be encoded.'* (Gelman et al., 2004 cited from Manca et al., 2007).

- If *'the analyst has identified the appropriate set of covariates for the exchangeability assumption to hold; and the characteristics of the country of interest are represented appropriately by countries in the dataset'* (Drummond et al., 2009), one may then extrapolate from existing data to domains for which cost-effectiveness information is currently missing.

- By expansion, one may turn the MLM into a bivariate framework, hence allowing for the simultaneous assessment of costs and effects as a vector of response variables (Bartholomew, 2008).

- Specific features of the MLM framework may even allow to explicitly model the variability in measures of cost-effectiveness as a function of explanatory variables (e.g. Rasbash et al., 2009).

As a result, it was decided to dedicate the empirical work within this project to the analysis of factors causing variability in international cost-effectiveness data from published economic evaluation studies using MLM. The following section aims to provide a theoretical basis for this research strategy.

2.3. Theoretical framework

This section aims to provide a theoretical basis for the use of MLM for the analysis of factors causing variability in international cost-effectiveness data from published economic evaluation studies. In addition, a working definition for the '*geographic transferability*' of economic evaluation data is provided.

As with any representation of the real world, a health economic evaluation, whether trial-based or model-based, is only useful to *'the degree to which it captures the reality as observed within the original setting'* (Sleigh, 1997). Therefore, establishing the relevance of economic evaluation results to any setting - including the one for which it was originally designed for - is a process within which we need to test whether the characteristics of that setting are appropriately reflected in the evaluation. We thus need to establish a correspondence between characteristics of the economic evaluation and the characteristics of the setting of interest. This is an argument by analogy (Juthe, 2005; Steel, 2008).

Analogical reasoning involves a mapping of attributes between a base domain, about which more is apparently known, and a, less studied, target domain (Gentner & Markman, 1997; Forbus, 2001; French, 2002). Then, the additional information about the base domain is hypothesized to hold in the target by virtue of the correspondences of those attributes which determine the information of interest (Juthe, 2005; Forbus, 2001; Klix, 2004). Hence, through the mapping of attributes we explicitly model our *a priori* belief that domains are similar in aspects which determine the information to be transferred. To estimate the attributes of interest for this mapping we need to know a) what causes variability in cost-effectiveness data and b) how to quantify the relative impact of such factors on measures of cost-effectiveness. There are numerous publications speculating about possible variability factors (Sculpher et al., 2004; Goeree et al., 2007), but little is known about the quantitative impact of such factors. This requires a simultaneous analysis of a number of cost-effectiveness studies across geographic domains. However, variability factors do not only

impact between geographic domains– there is also variation between studies within domains. When focussing on the *'higher-level variability*', it is hence necessary to control for any variability introduced on lower levels.

In conclusion, we need a statistical approach which explicitly and simultaneously models and tests the *a priori* belief of similarity between a) cost-effectiveness studies, and b) geographic domains. To identify such a quantitative technique, we need to define more precisely what we mean by an *'a priori belief of similarity'*. This definition is provided by the statistical concept of *'exchangeability'* (Spiegelhalter et al., 2000 & 2004).

The concept of exchangeability goes back to Bruno de Finetti and has been extensively discussed in Bernardo and Smith (1994). It means, in our context, that the joint probability distribution of the output parameters for each cost-effectiveness study is the same for all studies (Bernardo & Smith, 1994; Jeffrey, 2002). Likewise, without having any additional information on each geographic setting, we would not have any expectation of more, or less, favourable estimates of cost-effectiveness (Drummond et al., 2009). Hence, we assume that each set of cost-effectiveness measures available represents a *'random sample'* of some (hypothetical) population of measures of cost-effectiveness for that technology. This is also the standard assumption for random-effects meta-analysis (Greenland, 2000; Spiegelhalter et al., 2004). Spiegelhalter et al. (2004) further state that '*if a prior assumption of exchangeability is considered reasonable, a Bayesian approach to multiplicity is thus to integrate all the units into a single model, in which it is assumed that study parameters are drawn from some common prior distribution whose parameters are unknown: this is known as a hierarchical or multilevel model'* (Spiegelhalter et al., 2004).

This class of models is referred to as *'multilevel',* as allowing study and country parameters to vary randomly enables us to fit models for these parameters *'above'* the model for the actual cost-effectiveness data (Jackman, 2008). Hence, the overall model structure can follow the way the data is *'clustered'* within studies and geographic domains. Note that this, as this thesis will show,

constitutes a key challenge when fitting multilevel models to existing cost-effectiveness data from different geographic domains as we cannot establish a strict hierarchical structure between the study and the country-levels in the presence of data from multinational studies. This issue of '*cross-classified data structures'* is addressed in more detail in Chapters 3 and 5.

Within this multilevel framework, we can model exchangeability whilst also controlling for variability factors working within studies, between studies and between geographic domains, by allowing for covariate adjustment through the assumption of conditional independence (Gelman et al., 2004; Manca et al., 2007; Drummond et al, 2009). This should enable us to integrate cost-effectiveness data, and to quantify the impact of variability factors on data, study, and country-level. If we are successful in this endeavour, we could then extrapolate from existing data to domains for which cost-effectiveness information is currently missing. This can be achieved within the proposed framework and in accord with the principles of analogical reasoning if and only if *'the analyst has identified the appropriate set of covariates for the exchangeability assumption to hold; and that the characteristics of the country of interest are represented appropriately by countries in the dataset'* (Drummond et al., 2009). In this way, as Gelman et al. (2004) points out, '*exchangeable models become almost universally applicable, because any information to distinguish different units should be encoded.'* (Gelman et al., 2004, cited from Manca et al., 2007)

Before concluding this chapter with a review on the use of multilevel statistical modelling within the area of economic evaluation in health, it may be useful to provide a working definition for the '*geographic transferability*' of economic evaluation data, which, so far, has been used more or less interchangeably with the term '*generalisability*'. In fact, there may be some confusion around these terms, and different authors may provide differing definitions for both concepts. This is also summarized in Table 2.4 below. Therefore, based on the theoretical framework of analogical reasoning, an alternative working definition for the '*geographic transferability'* of economic evaluation data is provided below.

Table 2.4: 'Generalisability' and 'transferability' as defined by health economists

Author (year)	Term defined	Definition
Willke et al., 2003	Generalisability and transferability	*'generalisability' has often been used to refer to the problem of whether one can apply or extrapolate results obtained in one setting or population to another. Though this includes geographic settings, it can include demographic groups and various treatment settings, as well as extrapolating from trials to general populations. The term 'transferability has been used more specifically in reference to comparing results across countries*'
Drummond &Pang (2001)	Generalisability	*generalisability refers to the extent to which the results of a study, as they apply to a particular patient population and/or a specific context, hold true for another population and/or in a different context*
Sculpher et al. (2004)	Generalisability	*Adapt the same definition as Drummond and Pang (2001): 'generalisability refers to the extent to which the results of a study, as they apply to a particular patient population and/or a specific context, hold true for another population and/or in a different context*'
Drummond et al. (2005)	Generalisability	*Are the results of an HTA undertaken in one country relevant to another? Also, within a large country, do the results of a given HTA apply in all regions?*
Boulenger et al. (2005)	Transferability and generalisability	*Generalisability is de fined as: the degree to which the results of an observation hold true in other settings ('will a specific treatment produce the same results in a different location?').* *By adapting a definition by Späth et al (1999), transferability is defined as: the data, methods and results of a given study are transfer able if (a) potential users can assess their applicability to their setting and (b) they are applicable to that setting.*
Mason & Mason (2006)	Transferability and generalisability	*Mason and Mason (2006) define the term 'generalisability' as consisting of three elements: 1) technical quality, which refers to the robustness of the study methodology, 2) applicability, which refers to the extrapolation of the original setting of the clinical trial, i.e. whether the trial has adequate pragmatic qualities to be useful in its original setting and 3) transferability, which is the capacity to directly use the complete results of the economic evaluation in a setting different from the original one in which the technology was assessed.*
Manca et al. (2007)	Transferability and generalisability	*The authors state that the terms transferability and generalisability are used interchangeably. Strictly speaking, the former should be interpreted as ''the extent to which results from a given setting also apply to other settings,'' whereas the term generalisability should be used to indicate the extent to which results can be adapted to apply in other settings or can be interpreted for other settings*
Drummond et al (2009)	Transferability and generalisability	*The Task Force's working definitions were that economic evaluations were generalisable if they applied, without adjustment, to other settings. On the other hand, data were transferable if they could be adapted to apply to other settings*

The use of the term 'generalisability' for applying economic evaluation evidence from one particular context to another is being avoided in this thesis as this term may be confused with the statistical concept of '*generalisability'*, which is entirely based on inductive reasoning. Induction, in its strict logical sense, is '*an argument that employs premises containing information about some members of a class in order to support a generalisation about the whole class*' (Blaug, 1980). Hence it is a reasoning process from particular instances, say the participants of a trial, to general conclusions, as for example the cost-effectiveness of a new health technology compared to current practice in the population from which trial participants were originally recruited from. Though inductive reasoning is an inference from observed instances to unobserved instances, it is hence essential that all instances, the observed ones and the unobserved ones, belong to the same class (Steel, 2008). This definition would be very restrictive and only allows applying economic evaluation evidence to settings where a relationship between the studied sample and the target population can be established.

A more practicable definition for the purposes of this thesis would also incorporate situations where decisions need to be based on evidence which stems from an entirely different context. In the absence of country-specific cost-effectiveness data, it is common practice that decision makers in one country may allocate healthcare resources based on economic evaluation evidence originally generated in other countries (Goeree et al., 2007), and inductive reasoning, which assumes relatedness between the observed instances and the unobserved ones, does obviously not apply in this situation (Steel, 2008). Rather, the problem of '*transferring*' evidence from other contexts may best be described as an '*analogical inference'*, where a mapping of relevant attributes between a source domain, about which more is apparently known and a, less studied, target domain is produced to infer whether the information of interest may also hold in the target setting (Gentner & Markman, 1997; Forbus, 2001; French, 2002). In order to make inferences about the target setting, researchers may therefore look at characteristics of the existing evidence and the target context and thereby establish a relationship between the data and the context of interest (Juthe, 2005; Forbus, 2001; Klix, 2004, Steel, 2008). If this '*context*' is a geographic entity (for instance another country) we may refer to this process as

the assessment of the '*geographic transferability*' of economic evaluation in health.

The final section of this chapter reports on a systematic literature review on the use of MLM within economic evaluation in health.

2.4. Systematic literature review on the use of multilevel statistical modelling in economic evaluation in health

This systematic literature review on the use of MLM in economic evaluation in health was undertaken to inform the current state of research, major application areas of the method within this scientific domain, technical features which may be of interest for the empirical exercise within this project, as well as trends for further research. As the relevant body of literature relates strongly to general research on the transferability of health economic evaluation, is relatively small and very well cross-referenced, the author of this thesis was aware of many of the relevant papers even before the conduct of this systematic review exercise. However, to ensure that no important studies are missing from review, systematic database searches were conducted both in August 2010 (when developing and testing multilevel methods for the purposes of this project), as well as in May 2012 (when writing up for submitting this thesis). The following subsection reports on the search methods, inclusion and exclusion criteria as well as search results. Findings from analysing relevant papers are reported thereafter, with particular emphasis on conceptual or methodological features which may be relevant for the integration of international cost-effectiveness data from existing economic evaluation studies and the analysis of factors causing variability in this data.

2.4.1. Review methodology

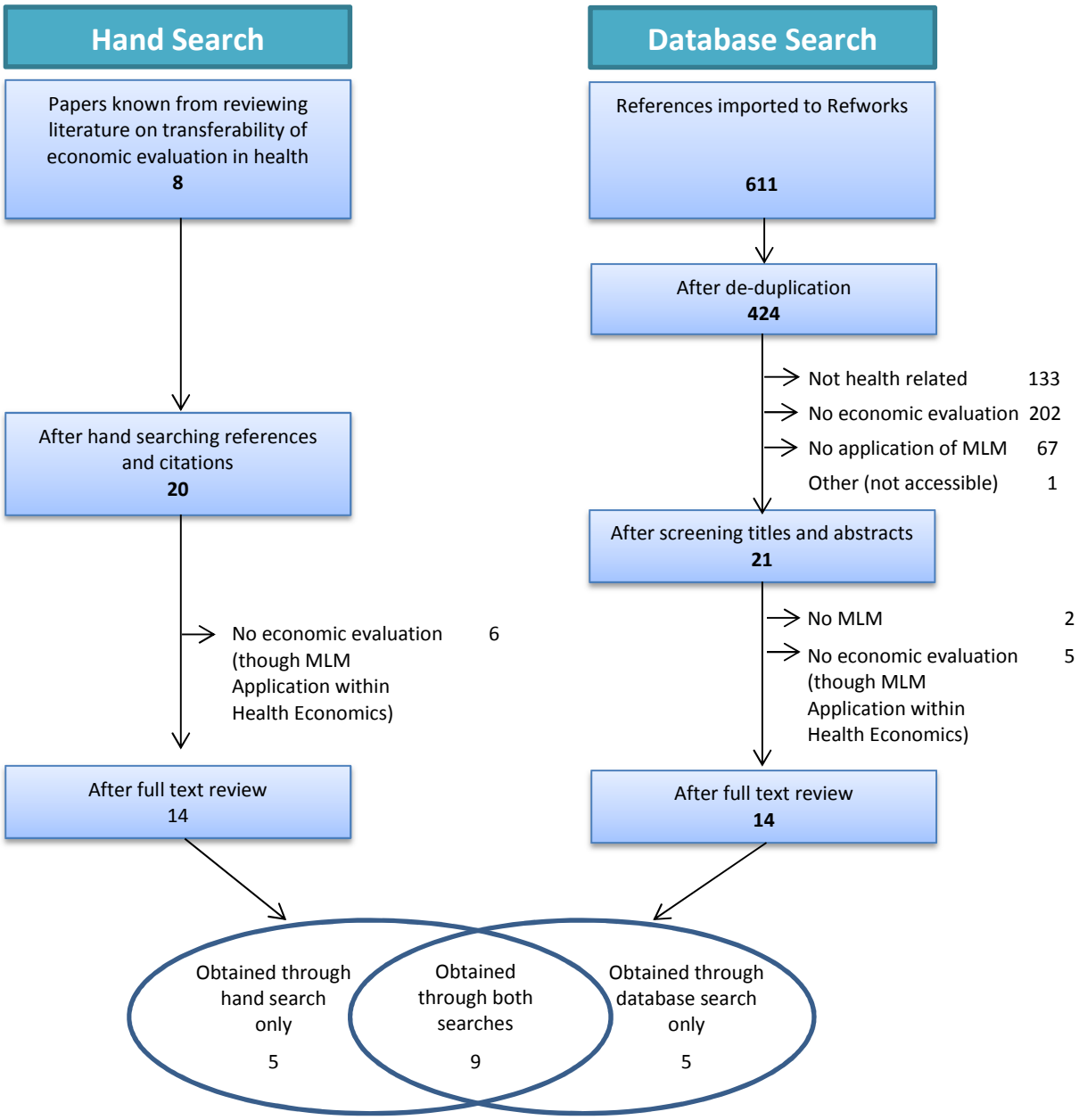
As mentioned, prior to this systematic review, the author was already aware of a large proportion of relevant studies. In particular, important papers were already under review when looking into the transferability of economic evaluation data in general (e.g. Rice & Jones, 1997; Sculpher et al., 2005; Manca et al., 2005; Drummond et al., 2005; Willan et al., 2005; Grieve et al., 2007; Manca et al., 2007; Drummond et al., 2009). Even before searching databases, these studies were systematically hand searched for further relevant references and SCOPUS was used to check whether subsequent studies citing these papers may be includable in this review. However, to ensure that no important papers are missing from this review, a systematic database search was also conducted. The databases searched were SCOPUS (Medline, Embase and Science direct), Web of Knowledge (Web of Science, Biosis Previews) and HEED, and search strategies for individual databases may be obtainable from Appendix 2.

Literature searches were performed in August 2010 (when the multilevel methodology for this thesis was developed and tested), and repeated in May 2012 (when writing up the results of this thesis for submission). No country, time, or language restriction was initially applied to the literature search. Studies concerned with economic evaluation in health which conceptualized, developed, empirically tested, or discussed the use of MLM were considered. In turn, this means that studies were not includable if they were a) not health related, b) not concerned with economic evaluation, and c) multilevel modelling methods were neither conceptualized, developed, applied, or discussed.

Figure 2.1 below shows the search algorithm for this systematic review. Eight papers were initially available from reviewing the general literature on the transferability of economic evaluation in health. These papers were first hand-searched for relevant references and SCOPUS citations, resulting in 20 papers eligible for full text review. With respect to database searches, 611 references were initially imported to the reference managing software Refworks. Subsequently, search results were de-duplicated, resulting in 425 hits remaining

in the database. These references were screened by titles and abstracts, resulting in 21 papers eligible for full text review.

Figure 2.1: Search Algorithm



Of all papers reviewed in full text, a further 13 were dropped. Five papers obtained from searching databases were not concerned with multilevel modelling. A further eight studies (two from database searches and six from hand search) relate to multilevel modelling within health economics in general, but did not relate to health economic evaluation defined as comparative analysis

of competing healthcare interventions in particular (Duncan et al., 1996; Duncan et al., 1998, Scott & Shiell, 1998; Burgess et al., 2000; Carey, 2000; Or et al., 2005; Morelle et al., 2009; Lee et al., 2010). Nine papers meeting all inclusion criteria were identified both from hand searching references as well as searching databases. A further five papers were only obtained through hand search, and five papers were only identified from searching databases respectively. Hence, 19 papers met all inclusion criteria and are analysed below with respect to concepts, applications, and / or discussions of MLM in the context of economic evaluation in health.

2.4.2. Results

Three of 19 papers meeting the final inclusion criteria were conceptual in nature (Rice & Jones, 1997; Spiegelhalter et al., 2000; Drummond et al., 2009), and one very recent paper critically appraises the use of hierarchical modelling techniques to analyse multinational cost-effectiveness data (Manca et al., 2010). An '*introductory account*' of multilevel models within the area of health economics has been provided by Rice & Jones (1997), who describe areas within this field of research that may benefit from the use of MLM. The authors explicitly discuss the use of MLM in economic evaluation alongside multinational RCTs and mention that '*the approach allows exploration of variation arising at different levels of the hierarchy and modelling of the correlation structure inherent in such data sets and leads to efficient parameter estimates.*'

In 2000, Spiegelhalter et al. published a review on Bayesian methods in health technology assessment within the NHS R&D HTA programme. Within this study, Spiegelhalter et al. (2000) explain the concept of exchangeability in the context of hierarchical models and state that '*the general Bayesian approach to multiplicity involves specifying a common prior distribution for the substudies that expresses a belief in the expected 'similarity' of all the individual unknown quantities being estimated. This produces a degree of pooling, in which an individual study's results tend to be 'shrunk' towards the average result by an amount depending on the variability between studies and the precision of the*

individual study. (….) This is essentially a random effect approach, often labelled as 'empirical Bayes' or 'multilevel' modelling' The authors further state that '*If there are known reasons to suspect specific units are systematically different, then those reasons need to be modelled*.'

About the exchangeability assumption between geographic domains in the context of cost-effectiveness analysis, Drummond et al. (2009) state *'the term "exchangeable" means that there are no other a priori reasons why one jurisdiction may have more or less favourable measures of cost-effectiveness than another.*' However, in accord with what has been mentioned by Spiegelhalter et al. (2000) above, the authors further state that, '*making an a priori assumption that one does not expect differences in jurisdiction-specific measures of cost-effectiveness may be unreasonable when the question we are trying to answer is whether or not such differences exist*.' To address this issue, Drummond et al. (2009) suggest the use of covariates on centre and / or country-level and state that '*currently, evidence is lacking as to what types of higher-level covariates may be useful in this regard*'. As potential candidates they suggest macroeconomic characteristics (e.g. gross national product), capacity constraints (e.g. limited availability of intensive care beds), economic incentives (e.g. pressure to minimise length of stay), or financing characteristics (e.g. global budgets vs. fee for service reimbursement). Drummond et al. (2009) conclude that '*if the analyst has identified the appropriate set of higher-level covariates for the exchangeability assumption to hold and that the characteristics of the country of interest are reflected appropriately by countries participating in a trial*' one may even transfer cost-effectiveness results to centres or countries that did not participate in the trial.

The vast majority of papers includable in this systematic review exercise develop a MLM for the analysis of individual patient cost-effectiveness data and test their approach within a case study (for a summary of review results and references of respective studies the reader is refereed to Table 2.5 below).

Table 2.5: Review Results

General			
Aim	Explaining a concept	3	1, 2, 16
	Developing a method + case study	13	3, 4 – 12, 14, 15, 18
	Empirical application	2	13, 19
	Critical appraisal	1	17
Data source	RCT	8	3, 5 - 8, 10, 14, 15,
	Observational data	4	4, 9, 11, 19
	Cluster randomized trial	3	12, 13, 18
	n.a. (conceptual, review paper)	4	1, 2, 16, 17
Multinational data	Yes	8	4, 5, 8 – 11, 14, 19
	No	7	3, 6, 7, 12, 13, 15, 18
	n.a. (conceptual, review paper)	4	1, 2, 16, 17
Transferability explicitly addressed through MLM	Yes	12	1, 3 – 11, 16, 17
	No	7	2, 12 – 15, 18, 19
Model specifications in methodological / empirical papers			
Nr of levels	Two	15	3 – 15, 18, 19
Hierarchy modelled	Patients in centres	8	3, 4, 6, 7, 9, 11, 15, 19
	Patients in clusters (CRTs design)	3	12, 13, 18
	Patients in countries	4	5, 8, 10, 14
Vector of response variables	Univariate	6	3 – 6, 9, 15
	Bivariate	5	7, 8, 10, 14, 18
	Both univariate and bivariate	2	11, 12
	Unclear	2	13, 19
Response variables modelled	NMB	4	3, 6, 11, 12
	Patient cost	4	4, 5, 9, 15
	Resource use	1	4
	Bivariate: patient cost / effects	7	7, 8, 10 – 12, 14, 18
	Unclear	2	13, 19
Other model features in methodological / empirical papers			
Random slopes	Yes	4	3, 6, 9, 12
	No	11	4, 5, 7-8, 10-11, 13-15, 18, 19
Random slope on which variables?	Treatm. effect varies across centres	3	3, 6, 12
	Incontinence status	1	9
	n.a.	11	4-5, 7-8, 10-11, 13-15, 18-19
Variance function modelled?	No	15	3 – 15, 18, 19
Gamma distributions for cost data	Yes	7	4, 7, 9, 12, 14, 15, 18
	No	8	3, 5, 6, 8, 10, 11, 13, 19
Multiplicative effects of covariates on outcomes	Yes	1	9
	No	14	3-8, 10-15, 18, 19
	n.a. (conceptual, review paper)	4	1, 2, 16, 17
Model implementation in methodological / empirical papers			
Software used	MLwiN	5	3, 4, 6, 11, 13,
	WinBugs	8	4, 7, 9, 10-12, 14, 18
	Stata	3	12, 15, 19
	R	1	15
	Unclear	2	5, 8
Estimation method	IGLS / RIGLS	1	4,
	MCMC	9	3, 4, 6, 7, 10-12, 14, 18
	Unclear	6	5, 8, 9, 13, 15, 19
Covariates suggested / tested			
Patient covariates *(covariates tested in bold/ italic with reference in brackets)*	baseline clinical and socio-demographic characteristics / ***age (4, 15, 19***) / ***gender (4, 15)*** / smoking status / ***pre-stroke living conditions (4)*** / ***stroke severity measures (4) / stroke subtype (4)*** / ***incontinence status (4, 9, 11) / paralysis at admission (4, 9)*** / ***presence of diabetes (15) / hypertension (15) / previous occurrence of AMI (15) / 'Karnofsky score' (19) / predicted mortality (19) / referral site (19) / Type of referral (19) indication for referral (19)***		
Centre covariates *(covariates tested in bold/ italic with reference in brackets)*	patient throughput / experience of clinical staff / hospital type (teaching vs. non-teaching) / financing characteristics (e.g. ***global budgets vs. fee for service reimbursement (4)*** / ***coronary angioplasty unit (15) / number of MIs (15) number of beds in cardiology (15***)		
Country covariates *(covariates tested in bold/ italic with reference in brackets)*	GDP / ***% GDP spent on HC (4, 9, 10, 11)*** / capacity constraints (e.g. limited availability of intensive care beds) / economic incentives (e.g. pressure to minimise length of stay) / dummy for low, middle, high income countries / ***patient co-payment for acute care (4)*** / ***country mean life expectancy at birth (10)*** /		

(1) Rice & Jones (1997)
(2) Spiegelhalter et al. (2000)
(3) Sculpher et al. (2004)
(4) Grieve et al. (2005)
(5) Willan et al. (2005)
(6) Manca et al. (2005)
(7) Nixon et al. (2005)
(8) Pinto et al. (2005)
(9) Thompson et al. (2006)
(10) Manca et al. (2007)
(11) Grieve et al. (2007)
(12) Bachmann et al. (2007)
(13) Coupe et al. (2007)
(14) Willan et al. (2008)
(15) Petrinco et al. (2009)
(16) Drummond et al. (2009)
(17) Manca et al. (2010)
(18) Grieve et al. (2010)
(19) Edbrooke et al. (2011)

Most case studies utilise data from RCTs, whilst three papers focus on the analysis of observational data, and two studies develop MLM methods for cost-effectiveness analysis alongside cluster-randomised trials. Further to that, data from a cluster randomized trial is also being analysed in one paper which empirically applies MLM for cost-effectiveness analysis, whilst another empirical application utilises data from a multinational observational study.

Of the 15 papers which either develop and test or empirically apply MLM methods, seven studies utilise data collected within one or maximum two countries, whilst multinational trials provide the data basis for the remaining eight studies under review. The transferability problem of economic evaluation data (either between centres, populations, settings or regions within one country or between countries) was specifically mentioned as one (though potentially not the sole) motivation for considering MLM techniques in 12 of the 19 papers, whilst *'accounting for clustering'*, '*appropriately reflecting the hierarchical data structure'* and thereby '*obtaining more appropriate regression outcomes'* appeared to be the main motivation for the use of MLM in the remaining papers under review.

When moving on to model specifications within methodological or empirical papers, it gets obvious that all studies without exception apply a strictly hierarchical two-level structure to their respective datasets. Within this two-level structure, individual patients are always modelled at level one, whilst level two may represent centres (eight studies), clusters (which may coincide with centres) as in cluster randomised trials (three studies), or countries (four studies). With respect to the response variable, six studies develop a univariate model, whilst a bivariate framework was developed in five studies. Two studies apply both, a univariate and a bivariate model. The two empirical applications (Coupe et al., 2007; Edbrooke et al., 2011) are ambiguous in most respects of the MLM specification, including the vector of response variables. If a univariate framework was applied, the response variable was patient cost (four studies), resource use (one study), or net monetary benefit (four studies). If net monetary benefit was the response variable, the respective MLM may be classified as a multilevel application of Hoch's net benefit regression framework which models

NMB as response variable and uses a treatment dummy to discriminate between patients in the respective trial arms, thereby modelling incremental net monetary benefit (Hoch et al., 2002)

The bivariate models which were developed in seven of the 19 papers under review represent a further expansion from this net benefit regression framework, within which patient cost and outcomes are decomposed and modelled as a vector of response variables. Again, a treatment dummy may discriminate between patients in the respective trial arms and thereby allow estimation of incremental cost and incremental effects within one model. With respect to such bivariate models, Manca et al. (2007) state that '*this (bivariate) approach has three main advantages. First, it facilitates explicit modelling of both costs and effects while allowing the inclusion of a set of covariates. Second, it exploits the existence of correlation, at the patient level, between costs and effects, thereby improving the efficiency of the estimation process when this correlation is different from zero. Third, unlike the standard (univariate) net benefit regression it does not require a new regression to be estimated for every value of the cost-effectiveness threshold*.'

Further features of the multilevel models developed in the relevant body of literature include random slopes (four studies), gamma distributions for cost data (seven studies) and multiplicative effects of covariates on outcomes. With respect to random slopes, three studies allowed the treatment effect to be different across centres, whilst one study (Thompson et al., 2006), allowed for heterogeneity of the patient-level variable '*incontinence status*' on centre level, thereby allowing mean costs to vary across centres with respect to whether patients are incontinent following the experience of a stroke or not. Fitting a random slope to this variable on centre level significantly improved the fit of the model. Even though some studies allowed for random slopes of covariates, none of them considered the explicit modelling of variation in the outcome variable as a function of explanatory variables, hence estimating a '*variance function*' within the multilevel modelling framework.

(Thompson et al., 2006) also modelled a multiplicative rather than an additive effect of covariates on outcomes. Again focussing on incontinence status as a patient covariate, the assumption was that the effect of incontinence may multiply the costs by a certain factor rather than adding to costs in each centre. Again, model fit improved with this assumption. Furthermore, to better reflect the fact that cost data is usually skewed to the right, seven papers modelled gamma distributions instead of normal distributions to cost data, and all studies observed an improved model fit from doing so.

Moving on to covariates assessed within the MLM studies under review, it gets apparent that, though it has been suggested as an important area for further research already in 2004 (Sculpher et al., 2004), no systematic assessment has yet been carried out to analyse covariates on each level of the data hierarchy. The inclusion of covariates served, at most, the purpose of testing a concept. Further, patient level variables under assessment are usually very disease and intervention specific, whilst only very few centre and country variables were considered at all; amongst them the percentage of GDP spent on healthcare, patient co-payments for acute care and life expectancy at birth within countries. As all multilevel models were applied within the context of IPD analysis from one single trial or observational study, methodological study characteristics, which were previously highlighted as important variability factors (e.g. Barbieri et al., 2005; Goeree et al., 2007), are identical across observations within trials and were therefore never subject to analysis within any of the studies under review.

2.5. Summary and conclusions

This chapter was concerned with providing the background and developing a research strategy for this project. Reviewing the economic evaluation literature on the transferability / generalisability of measures of cost-effectiveness showed that there is a need to systematically assess factors causing variability in international cost-effectiveness data. To do so, the analysis of secondary data abstracted from published economic evaluation studies has been proposed. This is justified from a pragmatic point of view as it would be highly unlikely to obtain

access to IPD within the scope of this project. However, integrating secondary data from published economic evaluation studies also allows the assessment of variability factors which may not be present in individual patient data from single trials or observational studies, especially with respect to methodological characteristics of health economic evaluation studies.

To facilitate this assessment, MLM was identified as a promising method. A systematic literature review on economic evaluation studies proposing or applying MLM methods showed that this technique has been considered with the particular aim to address the transferability problem of measures of cost-effectiveness. However, all applications of MLM in economic evaluation in health thus far relate to IPD from multicentre studies which may or may not be multinational in nature. Accordingly, model hierarchies always consist of two levels; more complex data structures, as they may appear when integrating secondary cost-effectiveness data (for instance cross-classifications), have not yet been proposed. Furthermore, existing studies consider covariates on different hierarchical levels only to demonstrate the concept. Accordingly, there is still a need to systematically assess covariates on all hierarchical levels encoding variability factors for measures of cost-effectiveness.

The following Chapter 3 aims to develop MLM methods appropriate for the integration of secondary cost-effectiveness data abstracted from international economic evaluation studies. This model development builds up from a basic ordinary least squares (OLS) regression model and, step by step, introduces the features necessary to integrate cost-effectiveness data from different studies which are also applicable to different geographic domains. The MLM methods developed are subsequently tested in a pilot study utilizing data from 16 international cost-effectiveness studies on statins in the primary and secondary prevention of CVD, before Chapter 4 reports on a systematic literature review and data abstraction exercise to populate a dataset for the main empirical analysis reported in Chapter 5.

3. Multilevel model methodology

The previous chapter began by summarizing the literature on the transferability and generalisability of economic evaluation in health and a research strategy was developed for the purposes of this project. Subsequently, a theoretical framework for the use of MLM to analyse variability factors for measures of cost-effectiveness was also developed, and the health economic evaluation literature concerned with MLM was systematically reviewed. As a result, MLM was identified as a promising strategy for the integration of international cost-effectiveness data from existing economic evaluation studies and the assessment of factors causing variability in measures of cost-effectiveness.

This chapter is concerned with the development of a number of MLMs for the purpose of secondary data-integration and the analysis of variability factors within this framework. It starts off with a simple OLS regression equation and then, step by step, introduces additional features to model complex data structures and (partial) exchangeability not just between studies, but also geographic domains. The final Section 3.4 of this chapter reports on a pilot study using a set of data collected from the health economic literature which entails cost-effectiveness information on one health intervention which was measured in different studies within, and across, geographic domains. The area of statins for the primary and secondary prevention of CVD was chosen for this pilot.

The chapter is organised as follows: It first elaborates on the exchangeability assumption of international cost-effectiveness data to derive a model suitable for secondary data integration in Section 3.1. It then shows how the assumption of exchangeability of cost-effectiveness data allows modelling complex data structures in a MLM framework. Next, this section expands upon this framework to include explanatory variables on each level of the data hierarchy. It then elaborates on even more complex data structures, including cross-classification of higher levels. Finally, slopes of regression coefficients are allowed to vary randomly across higher-level units.

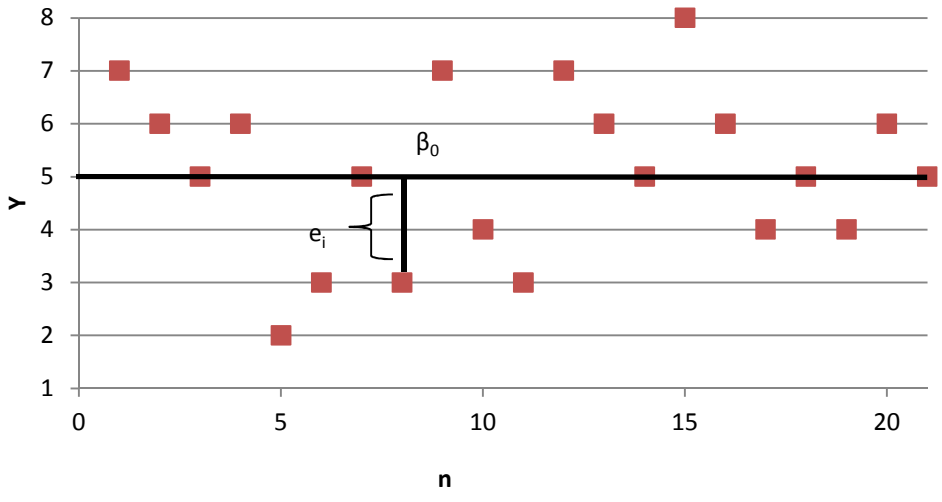
As Section 3.1 predominantly concentrates on the right hand side of the regression equation and the appropriate hierarchical structure of the model, Section 3.2 elaborates on the question of how the dependent variable, i.e. '*measures of cost-effectiveness'* of a healthcare technology, ought to be expressed. The advantages and disadvantages of the incremental cost-effectiveness ratio (ICER) and the incremental net monetary benefit (INMB) approaches are discussed. Following from that, the net benefit regression framework proposed by Hoch et al (2002) is considered for the purposes of this exercise. Subsequently, this framework, which was originally developed for the analysis of RCT data, is developed into a model which is more suitable for the integration of secondary cost-effectiveness data. Net monetary benefits are being decomposed and, in accordance with what has been proposed by other researchers in the field (Nixon et al., 2005; Pinto et al., 2005; Manca et al., 2007; Grieve et al., 2007; Bachmann et al., 2007; Willan et al., 2008; Grieve et al., 2010), a bivariate model is developed which enables inclusion of incremental costs and incremental effects separately as response variables in one regression equation.

In Section 3.3, the models derived in Sections 3.1 and 3.2 are combined, leading to a *'bivariate multilevel model for secondary data integration'*. This model is tested in the pilot study in Section 3.4 of this chapter, together with a number of two-level specifications and, for comparative purposes, an OLS regression model.

3.1. Modelling complex data structures within secondary cost-effectiveness data

Suppose we are interested in the cost-effectiveness of a certain healthcare technology and we find that this was measured in a number of cost-effectiveness studies available from the literature. There is an explicit theory on how the cost-effectiveness of a healthcare technology ought to be measured, and several authors suggest the use of incremental net monetary benefits (INMB's) as most suitable for econometric analysis (e.g. Claxton & Posnett, 1996; Stinnet & Mullahy, 1998; Briggs & Fenn, 1998; Tambour et al., 1998; Hoch et al., 2002). This theory is addressed in Section 3.2. However, let us leave this theory aside for the moment to concentrate on the complex data structures in international cost-effectiveness data only. To do so, denote a single cost-effectiveness estimate, however it may be measured, with 'Y_i'(with i=1,....,n). This could be, for example, a cost-effectiveness estimate applicable to one particular subgroup of patients. In addition, economic evaluation studies frequently report more than one single estimate of cost-effectiveness of a particular healthcare technology; usually, results are reported for different patient subgroups, and different assumptions tested in sensitivity or scenario analyses.

Figure 3.1: Example dataset and single level OLS regression without explanatory variables



Now, consider the following hypothetical situation: after conducting a literature review on the cost-effectiveness of a certain healthcare technology, we found four includable papers, reporting n= 21 cost-effectiveness estimates in total. The resulting dataset with sample data is plotted above in Figure 3.1. The easiest solution in this case is to treat parameters from all four studies as *'identical'* (Spiegelhalter et al., 2000, Spiegelhalter et al., 2004; Willan et al. 2005). In this situation, all observations of 'Y' may be pooled and we completely ignore that they stem from different studies available in the literature. Hence, we do not reflect the fact that cost-effectiveness estimates are *'nested'* within studies and regard all 'Y_i' as if they were obtained from the same source, irrespective from the fact that different studies may have employed different methods, assumptions, patient groups etc., which most likely causes dependency of cost-effectiveness results within each study. A very basic econometric model, which is also referred to as ordinary least squares (OLS) regression model, simply draws a line 'β_0' through the dataset which minimises its squared distance to each cost-effectiveness estimate 'y_i' (e.g. Maddala, 2001; Grieve et al., 2005). This may be expressed as:

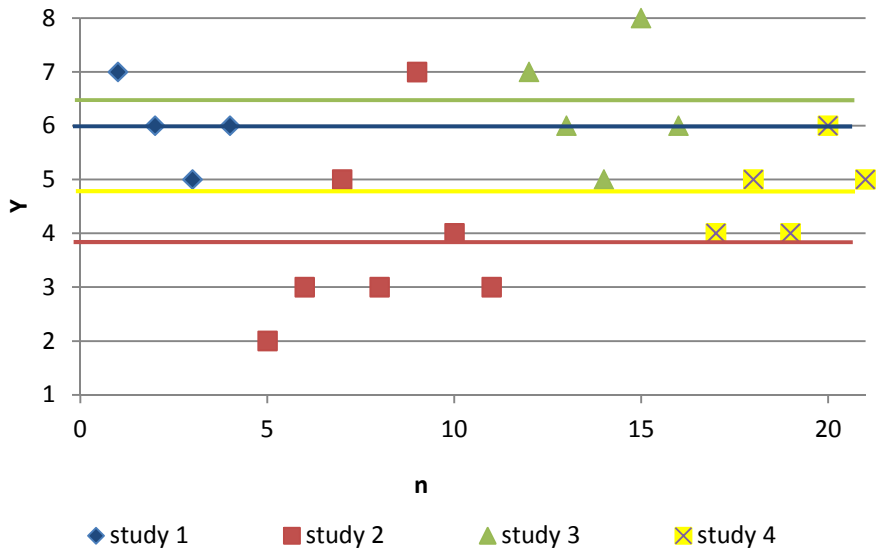
$$y_i = \beta_0 + e_i \qquad \text{with} \qquad e_i \sim N(0, \sigma^2) \qquad (1)$$

Model (1) is also called an '*empty',* or *'intercept only model'* as it does not carry an explanatory variable which may explain variation in cost-effectiveness estimates (Hox, 2010). Such an explanatory variable will be introduced later. In the case of this empty model shown in equation (1), the estimator of 'β_0' is simply the mean cost-effectiveness observed across all four studies (Maddala, 2001). Whereas 'β_0' is the so called deterministic component of model (1), 'e_i' is also referred to as the 'random', or 'stochastic' component of the regression equation (Maddala, 2001). The assumptions about 'e_i' are (Maddala, 2001):

- a mean of zero, i.e. $E(e_i) = 0$
- a common variance (homoscedasticity), i.e. $Var(e_i) = \sigma^2$
- the errors, or residuals 'e_i' are mutually independent and
- errors are normally distributed

As mentioned, whilst assuming that parameters of the four studies are *'identical'*, we can simply apply equation (1) to the data pooled across all four studies. Needless to say that complex data structures cannot simply be ignored without risking to overestimate accuracy, or to make plainly wrong inferences (Manca et al., 2005, Rasbash, 2008, Bartholomew et al, 2008; Hox, 2010).

Figure 3.2: Example dataset of 21 cost- effectiveness estimates, grouped in studies



Now, rather than simply assuming that studies are *'identical'* in order to pool the data, we may have another look at this dataset first. Figure 3.2 shows the same dataset as before, but cost-effectiveness estimates from each study are now shown in different colours and shapes. From this illustration it gets immediately clear that studies may not be regarded as identical. Studies employ different methods and assumptions, rely on different patient groups and apply to different geographic settings, etc. Therefore, cost-effectiveness estimates obtained from one study may be more similar to each other than they are to estimates from other studies. This, however, violates one of the assumptions introduced above, namely that the errors e_i are '*mutually independent*' (e.g. Rasbash, 2008). As a result, we cannot simply pool the data, but may have to fit a model for each of the resulting groups of cost-effectiveness-data separately. Hence, we treat our observations of 'Y' within each study as completely unrelated from the other studies available. In other words, we assume '*independence'* between studies (Spiegelhalter et al., 2000; Spiegelhalter et al, 2004; Willan et al. 2005). To

estimate a model for each study, equation (1) may be re-written after introducing additional subscripts to make clear that the dataset consists of cost-effectiveness estimates obtained from more than one single source. Instead of using 'i' as subscript for all cost-effectiveness estimates in the dataset as before, 'i' is now used as subscript for cost-effectiveness estimates within one single study with ($i=1,....,n_i$) and j as subscript for studies with ($j=1,...,J$), so that Y_{ij} denotes the i^{th} cost-effectiveness estimate in the j^{th} study. Rewriting equation (1) leads to:

$$y_{ij} = \beta_0 + e_{ij} \qquad \text{with} \qquad e_{ij} \sim N(0, \sigma^2) \qquad (2)$$

The corresponding least squares lines are illustrated in Figure 3.2. Of course, assuming *'independence'* may not be correct either as studies often do have things in common. For instance, they often rely on the same sources of effectiveness data, or they share methodological standards, etc. Most importantly, however, this assumption would preclude the chance of integrating secondary cost-effectiveness data from different published studies, which is the aim of this exercise. Hence, just like assuming study parameters to be *'identical'* to allow complete pooling may be regarded as somewhat too crude, one can conversely argue that assuming complete *'independence'* is too restrictive, as there must be some similarities between studies which, in theory, allows at least some sort of data integration (e.g. Spiegelhalter, 2000)

In conclusion, assuming either *'identical'* or *'independent'* parameters are two extreme assumptions about the data, and it would be much more sensible to assume that the data from different sources can be somehow pooled together, but without completely ignoring that cost-effectiveness estimates do stem from different studies indeed. This may be achieved through the assumption of *'exchangeability'* (e.g. Spiegelhalter et al., 2000, Spiegelhalter et al., 2004, Sculpher et al., 2004; Manca et al., 2005; Drummond et al., 2009). Exchangeability is a concept which goes back to Bruno de Finetti and it has been extensively discussed in Bernardo & Smith (1994). Essentially, it means that the joint probability distribution of the parameters 'Θ' (i.e. study mean and variance) observed in each of the 'j' studies available is the same for any permutation of

the index of studies 'j' (Bernardo & Smith, 1994; Jeffrey, 2002). In other words, without having any additional information on each of the 'j' studies, we would not have any expectation of more, or less, favourable estimates of cost-effectiveness to be found in each of the studies (Drummond et al., 2009). In this sense, exchangeability is, like Greenland (2000) puts it, a '*much weaker assumption than that studies really are equal; it only says that, without seeing the data, we can't yet tell how they might differ*'. Further, exchangeability is *'essentially an uncertain and qualitative prior guess about the similarity'* of the studies available (Greenland, 2000)

Spiegelhalter et al. (2000 & 2004) state that '*if a prior assumption of exchangeability is considered reasonable, a Bayesian approach to multiplicity is thus to integrate all the units into a single model, in which it is assumed that study parameters* $\Theta_1, \ldots, \Theta_j$ *are drawn from some common prior distribution whose parameters are unknown: this is known as a hierarchical, or multilevel, model'.* In other words, we assume that each of the 'j' studies available represents a '*random sample'* of some (hypothetical) population of cost-effectiveness studies for that technology. This is also the standard assumption in a traditional random-effects meta-analysis (Spiegelhalter et al., 2000 & 2004). The class of models which may be built upon this assumption is called hierarchical, or multilevel, because assuming exchangeability enables study parameters to vary randomly, and this allows building a model for the study parameters on top of the model for the actual cost-effectiveness data (Jackman, 2008).

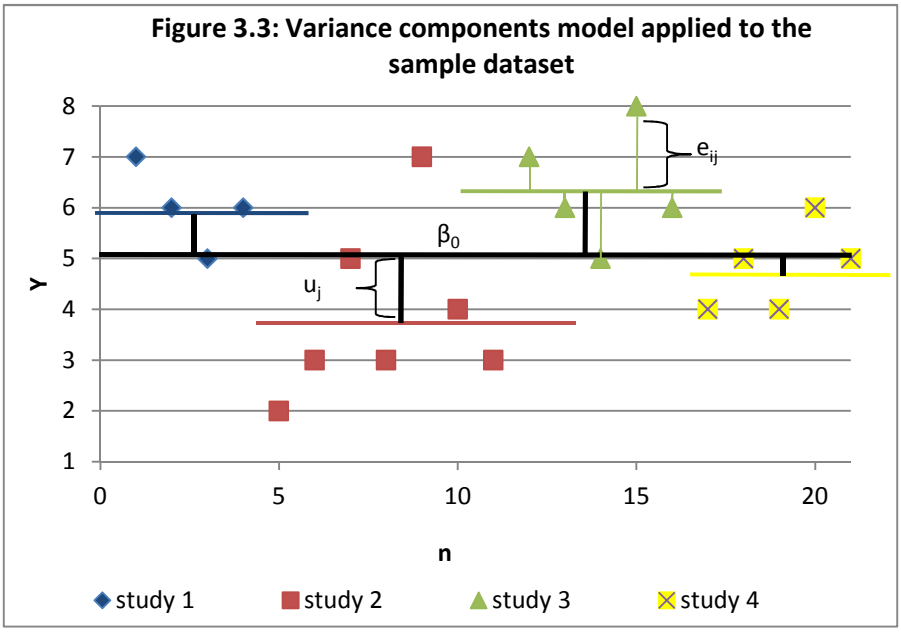
To sum up, through the assumption of exchangeability we may allow study parameters to vary randomly between studies, and this allows building a hierarchical model which mediates between the two extreme assumptions of either independent or identical parameters. What is important now is to translate these findings into a quantitative model for data integration which makes explicit the assumption of exchangeability of cost-effectiveness data across studies and geographic domains.

3.1.1. Modelling exchangeability – 'empirical Bayes shrinkage estimation'

Recall model (2)

$y_{ij} = \beta_0 + e_{ij}$ with $e_{ij} \sim N(0, \sigma^2)$ (2; repeated)

This model is a single level-model of cost-effectiveness estimates obtained from each of the 'j' studies. It is also an *'empty model'* as it does not carry explanatory variables on the right hand side of the equation. The subscripts indicate that a separate model may be fitted to each study independently. The stochastic component 'e_{ij}' allows the cost-effectiveness estimates 'y_{ij}' within each study to vary randomly, assuming a normal distribution of the error term. This is illustrated in Figure 3.2 above, where one regression line, representing mean cost-effectiveness, is fitted for each study separately assuming independence of data across studies.



However, we learned that modelling exchangeability allows study parameters (i.e. study mean and variance) to vary randomly between studies too. We also learned that this allows placing a model for the study parameters above the

model for the cost-effectiveness data (Jackman, 2008); which is also illustrated in Figure 3.3 above. The black line represents the overall mean regression line; one could say the pooled effect over all four studies in the dataset (e.g. Drummond et al., 2009). The distance between this pooled effect and the study regression lines represents the random variation of the study means from the overall mean effect just as the distance between the data and the study regression lines represent the random variation of cost-effectiveness data from the individual study means (Bickel, 2007; Bartholomew et al., 2008; Steele, 2008; Hox, 2010). As this random variation of the data within each study is captured with the stochastic error term' e_{ij}', a second error term is needed which encapsulates the random variation between the overall mean cost-effectiveness, captured by 'β_0', and the corresponding study means (Bickel, 2007; Bartholomew et al., 2008; Steele, 2008; Hox, 2010). This error term is labelled with 'u_j'. Just like it was assumed that the 'e_{ij}' are normally distributed with zero mean and variance 'σ_e^2', an analogous assumption can be made for 'u_j' (Bartholomew et al., 2008; Steele, 2008; Hox, 2010). Hence, through the inclusion of this second error term 'u_j', model (2) may be turned into a basic MLM of the form:

$$y_{ij} \sim N(XB, \Omega) \qquad (3)$$
$$y_{ij} = \beta_0 + u_j + e_{ij} \text{ with}$$
$$u_j \sim N(0, \sigma_u^2)$$
$$e_{ij} \sim N(0, \sigma_e^2)$$

Model (3) is a *'variance components model',* the simplest form of MLM (Bartholomew et al., 2008; Steele, 2008; Hox, 2010). As mentioned, 'β_0' is the overall mean for 'y_{ij}' (across all studies), '$\beta_0 + u_j$' is the mean cost-effectiveness within each study, 'u_j' is the difference between a studies mean cost-effectiveness and the overall mean, and 'e_{ij}' is the difference between the 'y_{ij}' for the i[th] measurement and the group mean of that study, i.e. '$e_{ij} = y_{ij} - (\beta_0 + u_j)$' (Steele, 2008). The statement '$y_{ij} \sim N(XB, \Omega)$' makes clear that we assume the response variable 'y_{ij}' to be normally distributed (Rasbash et al., 2009). 'XB' is the fixed part of the model, whereas 'Ω' denotes the random part over all the levels of the data (Rasbash et al., 2009). In the case of model (3), it simply

denotes the variances at both the data and the study-level (Rasbash et al., 2009). Through the assumption that both error terms 'u_j' and 'e_{ij}' are normally distributed with zero mean and variance 'σ_u^2' and 'σ_e^2' respectively, the total variance can be written as (Steele, 2008; Rasbash et al., 2009):

$$total\ variance = \sigma_u^2 + \sigma_e^2 \quad (4)$$

This also allows calculating the proportion of the total variance which can be attributed to differences between studies, which is referred to as the *'variance partition coefficient'* (VPC) (Steele, 2008; Rasbash et al., 2009)

$$VPC = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2} \quad (5)$$

From equation (5) it can be seen that the VPC tends towards zero if $'\sigma_e^2'$ is large in comparison to $'\sigma_u^2'$, meaning that there would be little variation from '*study effects'* in the data (Steele, 2008; Rasbash et al., 2009). On the other hand, if $'\sigma_u^2'$ is large in comparison to $'\sigma_e^2'$, then the VPC tends towards one; and we learn that there is little variation from '*within-study differences'* in the data as most variation stems from differences between studies (Steele, 2008; Rasbash et al., 2009). A further assumption of model (3) is that the residuals at the same level are uncorrelated with one another, that is: $'\mathrm{cov}(u_{j1}, u_{j2}) = 0'$ for different studies and $'\mathrm{cov}(e_{j1k1}, e_{j2k2}) = 0'$ for different cost-effectiveness values obtained from different studies (Rasbash et al., 2009). Finally, it is assumed that the residuals at different levels are uncorrelated with one another, that is: $'\mathrm{cov}(e_{jk1}, u_{k2}) = 0'$ for the same or different groups (Rasbash et al., 2009).

To sum up, due to assuming exchangeability we may regard the study parameters to be random draws from some prior distribution which allows fitting a model for the overall mean cost-effectiveness above the models for the study means (e.g. Spiegelhalter et al, 2000 & 2004; Jackman, 2008). We also assumed error terms $'u_j'$ and $'e_{ij}'$ to be normally distributed with zero mean and variance $'\sigma_u^2'$ and $'\sigma_e^2'$ respectively (Steele, 2008; Rasbash et al., 2009). The VPC is a measure which tells us how much of the total variation in the data may be

attributed to the data-level or to the study-level respectively (Steele, 2008; Rasbash et al., 2009). However, model (3) needs to be specified further. Assuming exchangeability is supposed to '*mediate*' between the two extreme viewpoints of either identical or independent study parameters in the sense that it allows for some sort of pooling, but without ignoring the fact that the data stems from different studies indeed. As Jackman (2008) puts it, hierarchical models help us to find the *'sweet spot'* between the assumptions of identical parameters on the one hand and independent parameters on the other. To be able to make comparisons between studies, the study residual 'u_j' needs to be estimated (Rasbash et al., 2009). An estimate of 'u_j' may be derived by calculating the *'mean raw residual',* that is: (Rasbash et al., 2009).

$$\bar{r}_j = \bar{y}_j - \hat{\beta}_0 \tag{6}$$

Where '$\bar{y}_j$' is the mean of cost-effectiveness in study 'j', and '$\widehat{\beta}_0$' is an estimator of the overall mean cost-effectiveness (Rasbash et al., 2009). This raw residual is then multiplied by a so called shrinkage factor 'S' (Steele, 2008; Rasbash et al., 2009):

$$\hat{u}_j = S\bar{r}_j \quad \text{where} \quad S = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + (\frac{\hat{\sigma}_e^2}{n_j})} \tag{7}$$

'n_j' is the sample size in study 'j', hence the number of cost-effectiveness estimates reported in, and abstracted from, that study. '$\widehat{\sigma}_e^2$' and '$\widehat{\sigma}_u^2$' are estimates of the variances of the within-study and between-study error terms respectively (Steele, 2008; Rasbash et al., 2009)

Now it can be illustrated why the assumption of exchangeability of study parameters mediates between the two extremes of either identical or independent parameters obtained from different cost-effectiveness studies. If the between study variance 'σ_u^2' is assumed to be zero, then this is equal to say that all variation in the reported cost-effectiveness measures stems from '*within-*

study' variability, hence, the mean study effects are '*identical'* between studies (Steele, 2008; Rasbash et al., 2009). This would mean that there are no differences between studies and all cost-effectiveness estimates may be safely pooled together. If, on the other hand, $'\sigma_u^2 \to \infty'$, then the study effects are regarded to be independent, meaning that data from different sources may not be pooled together (Steele, 2008; Rasbash et al., 2009). Assuming exchangeability allows to model that the *'reality'* might be somewhere in between those extreme viewpoints (e.g. Spiegelhalter et al., 2000 & 2004; Willan et al., 2005; Manca et al., 2007; Drummond et al., 2009)

From what was stated above, it can be seen that $'\sigma_u^2 \to 0'$ is the special case where the shrinkage factor $'S'$ tends towards zero and the study effects are completely shrunken towards the overall mean cost-effectiveness estimate $'\beta_0'$ (Steele, 2008; Rasbash et al., 2009). If, however, $'\sigma_u^{2\prime}$ is high, then the shrinkage factor $'S'$ tends towards one, meaning that the between study variance is high and shrinkage of the study effects towards the overall mean $'\beta_0'$ is small (Steele, 2008; Rasbash et al., 2009). It is important to note, however, that the extent of shrinkage does not only depend on the amount of between study variability $'\sigma_u^{2\prime}$. The higher the number of cost-effectiveness estimates provided by one single study (n_j), the more information is provided by that particular study to the overall model, and the less will the study mean be shrunken towards the overall mean (e.g. Steele, 2008; Rasbash et al., 2009; Drummond et al., 2009). On the other hand, if a study provides only few estimates of cost-effectiveness, then shrinkage is high as this study '*borrows*' a lot of information from all other studies available (Steele, 2008; Rasbash et al., 2009; Drummond et al., 2009).

The mechanism explained above is what the literature refers to as '*empirical Bayes shrinkage estimation*', or *'empirical Bayes'* (e.g. Spiegelhalter, 2000; Willan et al., 2005; Steele, 2008). This process makes explicit the assumption of exchangeability between studies, and, by extending the basic model shown in equation (3), will also provide a means to model exchangeability between geographic domains and to assess variability factors on each level of the data hierarchy. The next section expands upon model (3) to:

- introduce slope parameters to model dependency of 'Y' upon some vector of explanatory variables (random intercepts model)
- adding a third level to the model to simultaneously assess variability in cost-effectiveness estimates on data, study and country-level whilst dealing with the problem of *'cross-classification'* of data from multinational studies, and finally,
- assuming both slopes and intercepts of the regression line to vary randomly (random-slopes model)

In Section 3.2, the question of how to measure the cost-effectiveness of a healthcare technology as a dependent variable within the MLM framework is discussed. Finally, before actually testing models in a pilot study which is reported on in Section 3.4 of this chapter, the MLMs derived in this section are combined with the regression model derived in Section 3.2, resulting in a *'bivariate multilevel model for secondary data integration'* with incremental costs (ΔC) and incremental effects (ΔE) as response variables nested in studies and geographic domains. However, for now, let us turn back to the variance components model illustrated in equation (3) and extend upon it to make it more suitable for the purposes of this thesis.

3.1.2. Extending the basic multilevel model

This section expands upon the variance components model shown in equation (3) to make it more suitable for the assessment of variability in secondary cost-effectiveness data between studies and geographic domains. The first step is to introduce '*slope parameters'* to model dependency of 'Y' upon some vector of explanatory variables. This results in a so called '*random intercepts model'*.

3.1.2.1. The random intercepts model

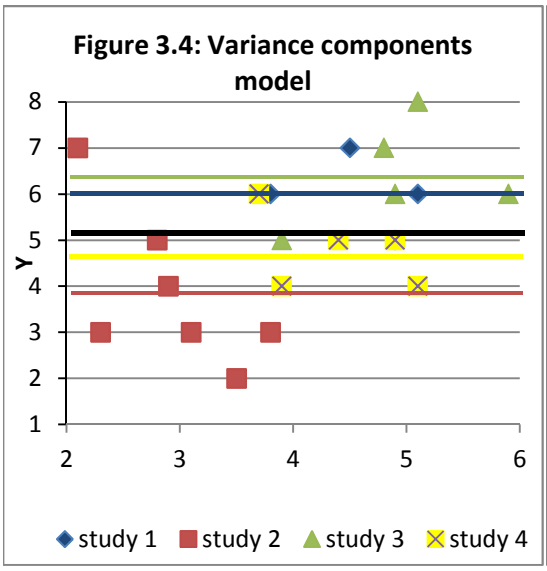
Recall model (3):

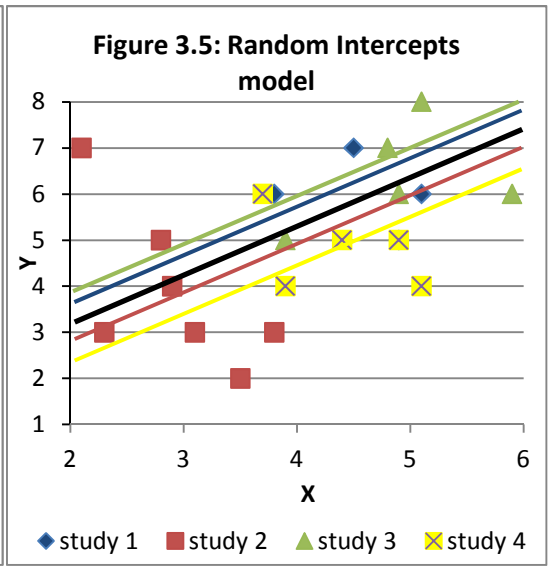
$y_{ij} \sim N(XB, \Omega)$ (3; repeated)

$y_{ij} = \beta_0 + u_j + e_{ij}$ with

$u_j \sim N(0, \sigma_u^2)$
$e_{ij} \sim N(0, \sigma_e^2)$

This *'variance components model'* essentially tells whether differences in cost-effectiveness data exist between studies, which studies have higher, or lower, mean cost-effectiveness estimates, and how much of the variation in cost-effectiveness is attributable to each level of the data hierarchy. Here, a *'slope parameter'*, or *'explanatory variable'*, on data-level is being introduced. Consider Figure 3.5 below which displays the same sample dataset as before; however, the ordering of the data has changed. In the previous section, cost-effectiveness estimates from different studies were in no particular order, hence, leaving the horizontal axis in the diagram without a meaning. Now, the horizontal axis captures additional information, for instance patient characteristics like mean age or mean body mass index (BMI). As the variance components model does not carry an explanatory variable, we would fail to capture this additional information when using model (3). This situation is also illustrated in Figure 3.4 below, where the slopes of the regression lines are 'flat'. Now, to capture this additional information on the x-axis, we need to introduce a parameter which captures the slope of the regression lines (Bickel, 2007; Steele, 2008, Rasbash et al., 2009, Hox, 2010). This will then allow explaining some of the variability between cost-effectiveness estimates. It will also tell us how much variability will remain on both levels (data and study) after controlling for a particular covariate (Steele, 2008; Bartholomew et al., 2008, Rasbash et al., 2009, Hox, 2010).





Introducing an explanatory variable on the data-level turns model (3) into (Steele, 2008, Rasbash et al., 2009):

$$y_{ij} \sim N(XB, \Omega) \tag{8}$$

$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$ with
$u_j \sim N(0, \sigma_u^2)$
$e_{ij} \sim N(0, \sigma_e^2)$

Unlike the variance components model shown in equation (3), this *'random intercepts model'* allows taking a look at, and control for, characteristics of cost-effectiveness estimates collected from the available literature. The fixed part of the model is now given by, $'\beta_0 + \beta_1 x_{ij}'$ whereas the random part $'u_j + e_{ij}'$ remains unchanged (Steele, 2008, Rasbash et al., 2009). $'\beta_0 + \beta_1 x_{ij}'$ is also the equation for the pooled regression line, with $'\beta_0'$ being the intercept and $'\beta_1 x_{ij}'$ being the slope of that regression line (Steele, 2008, Rasbash et al., 2009). Analogously to the variance components model, the intercepts of the individual study regression lines are given by $'\beta_0 + u_j'$, so that the individual lines are still located parallel around the pooled regression line (Steele, 2008, Rasbash et al., 2009). As the distance between the intercepts of each individual study regression line to the pooled regression line is encapsulated in the random component $'u_j'$, this class of models is referred to as *'random intercepts models'* (Steele, 2008, Bartholomew et al., 2008, Rasbash et al., 2009)

As within the variance components model above, a) residuals for different levels, b) residuals at level two for different groups, and c) level one residuals for different observations, are uncorrelated, that is (Goldstein, 1999; Steele, 2008; Rasbash et al.; 2008, Hox; 2010; CMM-workshops/random intercepts):

a) $\mathrm{cov}(u_{i1}, e_{i1j1}) = 0$ b) $\mathrm{cov}(u_{i1}, e_{i2}) = 0$ c) $\mathrm{cov}(e_{i1j1}, e_{j2j1}) = 0$

$\mathrm{cov}(u_{i1}, e_{i1j2}) = 0$ $\mathrm{cov}(e_{i1j1}, e_{i2j2}) = 0$

In addition, we assume the residuals and the covariates to be uncorrelated within the random intercepts model, that is (Goldstein, 1999; Steele, 2008; Rasbash et al.; 2008, Hox; 2010; CMM-workshops/random intercepts):

$$\mathrm{cov}(u_j, x_{ij}) = 0$$
$$\mathrm{cov}(e_{ij}, x_{ij}) = 0$$

This leaves us with a '*block-diagonal*' correlation matrix as shown in Figure 3.6 below, where all cost-effectiveness measurements within the same study are correlated, but all measurements from different studies are uncorrelated (Goldstein, 1999; CMM-workshops/random intercepts). 'P' is simply the variance partitioning coefficient shown in equation (5) above (Goldstein, 1999; CMM-workshops/random intercepts).

Figure 3.6: Covariance matrix in a two-level hierarchical model

Study		1	1	1	2	2	2	2	3	3
	data	1	2	3	1	2	3	4	1	2
1	1	1	p	p	0	0	0	0	0	0
1	2	p	1	p	0	0	0	0	0	0
1	3	p	p	1	0	0	0	0	0	0
2	1	0	0	0	1	p	p	p	0	0
2	2	0	0	0	p	1	p	p	0	0
2	3	0	0	0	p	p	1	p	0	0
2	4	0	0	0	p	p	p	1	0	0
3	1	0	0	0	0	0	0	0	1	p
3	2	0	0	0	0	0	0	0	P	1

Of course, we may add more than just one explanatory variable, as well as interaction terms. However, thus far, the class of explanatory variables

introduced relates to the data-level only. Hence, this model may help explaining differences which exist, for instance, between different patient subgroups assessed within different studies. However, it may not yet be able to capture differences between studies, like the timing of an economic evaluation study, or the general study design (RCT or decision analytic model). These differences relate to level-two of the data-hierarchy. Furthermore, once adding a third level to the model which captures geographic domains, we may want to assess what explains differences in cost-effectiveness data between countries. This requires the inclusion of country-level covariates; hence, we need to add higher-level covariates to the model.

To provide another justification, Drummond et al. (2009) state: '*the term 'exchangeable' means that there are no other a priori reasons why one jurisdiction may have more or less favourable measures of costs or cost-effectiveness than another. Making an a priori assumption that one does not expect differences in jurisdiction-specific measures of cost-effectiveness may be unreasonable when the question we are trying to answer is whether or not such differences exist.'* (Drummond et al., 2009). Hence, it is necessary to relax the initial assumption of exchangeability of study and country parameters to allow for some sort of differences between studies, or countries, which may account for more, or less favourable cost-effectiveness estimates of a healthcare technology.

There are strong reasons to suggest that the exchangeability assumption does, in fact, not hold between geographic domains. For example, as Grieve et al. (2007) puts it: *'in multinational CEA a priori reasoning would suggest that if the countries included [...] are at different stages of economic development then systematic variations in the relative cost-effectiveness of health care interventions would be anticipated. In this context the exchangeability assumption would be implausible and a multilevel model that assumed exchangeability would be inappropriate.'* Analogously, the same holds between studies within one geographic domain. For example, if two studies were exchangeable in terms of the health technology under assessment, the comparator, and the general study design, but would differ with respect to the

elicitation of utility estimates and timing, then the exchangeability assumption may be violated in these aspects. Hence, what is needed is a model which *'avoids making such stringent exchangeability assumptions'* (Grieve et al., 2007), and rather permits the adjustment of the dependent variable with respect to factors which are anticipated to be responsible for more or less favourable cost-effectiveness estimates between studies and countries. Such models rely on the assumption of '*partial exchangeability'*, that is, we assume that the '*mean estimates of cost-effectiveness are exchangeable, but only after covariate adjustment*.' (Grieve et al., 2007). Hence, including covariates on both levels of the model developed thus far will turn the random intercepts model in equation (8) into:

$$y_{ij} \sim N(XB, \Omega) \tag{9}$$

$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_j + u_j + e_{ij}$ with
$u_j \sim N(0, \sigma_u^2)$
$e_{ij} \sim N(0, \sigma_e^2)$

Note that this model only consists of a data and a study-level, and therefore does not yet carry country-level covariates. Country-level covariates may be considered further below once a country-level has been added to the model. It is also important to note that the study-level covariate in equation (9) does not carry a 'i'- subscript as it does not vary within, but only between studies in the dataset (e.g. Steele, 2008; Rasbash et al; 2009). This specification enables exploring the impact of study-level effects while simultaneously controlling for factors on data-level and allowing for the fact that the cost-effectiveness of a healthcare technology measured in different studies may be influenced by variability factors within and between studies in the dataset. Further, as Steele (2008) states, if contextual effects are of interest, then a multilevel approach is vital as '*the standard errors of coefficients of higher-level variables may be severely underestimated when a single level model is used*'.

Analogously to single level multiple regression analysis, we may also model that the effect of one explanatory variable on cost-effectiveness depends on the

value of another explanatory variable through interaction effects and interaction effects may be included between any set of explanatory variables (Steele, 2008). If explanatory variables belong to different levels, corresponding interaction terms are referred to as *'cross-level interactions'* (Steele, 2008). With such cross-level interactions, we could examine, for instance, the relationship between timing of an economic evaluation on study-level and the cost of the intervention on data-level. The idea is that intervention cost would change over time and this should be reflected in the data elicited from different economic evaluation studies which differ, amongst other things, by their respective timing. A random intercepts model with a cross-level interaction between data and study-level variables is presented in equation (10):

$$y_{ij} \sim N(XB, \Omega) \tag{10}$$

$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_j + \beta_3 x_{ij} x_j + u_j + e_{ij}$ with
$u_j \sim N(0, \sigma_u^2)$
$e_{ij} \sim N(0, \sigma_e^2)$

Again, everything mentioned in this section also holds for covariates relating to differences between countries which are considered below once a country-level has been added to the model. This will ultimately allow simultaneous assessment of variability factors on data-, study- and country-level. As Gelman (2004) puts it, '*the valid concern is not about exchangeability, but encoding relevant knowledge as explanatory variables where possible. In essence, 'the usual way to model exchangeability with covariates is through conditional independence. In this way exchangeable models become almost universally applicable, because any information to distinguish different units should be encoded.'* (Gelman (2004), cited from Manca et al., 2007). With respect to differences between geographic domains, Drummond et al. (2009) state, assuming that *'if the analyst has identified the appropriate set of higher-level covariates for the exchangeability assumption to hold; and that the characteristics of the country of interest are represented appropriately by countries in the dataset'*, the inclusion of higher-level covariates allows estimation of cost-effectiveness for geographic domains which are not included in the dataset. The next section is concerned with adding a country-level to the model, which also allows including such higher-level covariates encoding differences between geographic domains.

However, before moving on to model the country-level within the multilevel framework, an important question remains: If our interest is to analyse contextual effects between countries, why do we control for variables on lower levels of the data hierarchy. The answer is that the variation on higher levels depends on what happens on lower levels of the model (Hox, 2010). In a single level model, the introduction of additional explanatory variables always decreases the variance of the error term or, at least, leaves it unchanged; but it never increases (Steele, 2008, Rasbash et al., 2009). However, in a MLM, the variance of the error term of the next level might stay unchanged, decrease, or even increase with the introduction of a lower-level covariate (Steele, 2008, Rasbash et al., 2009). Hence, without controlling for lower-level variables, we might get a distorted picture on what happens on study or country-level. Some contextual effects might be overestimated; others might be disguised by confounders on lower levels of the data hierarchy. For instance, in the case of statins for the primary and secondary prevention of CVD, not controlling for the annual drug cost on data-level may distort, or disguise the impact of GDP per capita on measures of cost-effectiveness on the country-level. To provide another example, Sculpher et al. (2004) state that patient factors constitute an important source of lower-level variability which potentially '*feeds through to centre or country variations in cost-effectiveness if these subgroups of patients are not evenly distributed between locations'.*

In conclusion, we cannot simply leave all the variation on lower levels of the data hierarchy to the relating error term and then concentrate on contextual variables on country-level only; we need to investigate what explains variation on lower levels first before moving to the next higher level of the data hierarchy (e.g. Hox, 2010). This has important implications for the design of the empirical analysis. With respect to data abstraction, for instance, it makes it necessary to put considerable effort into the collection of additional data from includable cost-effectiveness studies to control for effects on data, and study-level; and this before even considering the analysis of contextual effects on country-level. In addition, data analysis within the empirical chapter may proceed from the lower level to the higher level, as the analysis of country-level covariates may only be valid once we appropriately controlled for variability factors on data and study-level (Steele, 2008; Rasbash et al., 2009; Hox, 2010). Hence, within the empirical

exercise there is a strong focus on factors which may introduce variability in measures of cost-effectiveness on all levels of the data hierarchy. Fortunately, there is extensive literature on what may cause variability in cost-effectiveness estimates within and across studies as well as between geographic domains, which may also help determining a strategy for this empirical exercise (Sculper et al., 2004; Goeree et al., 2007).

3.1.2.2. Adding a country-level to the multilevel model

To this point, the methodological discussion was limited to a hierarchical model with two levels, namely, the data-level, with measures of cost-effectiveness collected within each study, and above that, the study-level, with studies included in the dataset. The hierarchical structure arises from cost-effectiveness estimates being nested within studies, and each estimate of cost-effectiveness nests in one study, the higher-level unit, only. The corresponding unit and classification diagrams for this situation are illustrated in Figure 3.7 (Rasbash, 2008). If the problem was to integrate secondary cost-effectiveness data from different economic evaluation studies *within* one geographic domain, then this two-level model structure would probably be sufficient for that purpose.

Figure 3.7: Unit and classification diagrams of a two-level hierarchical data structure

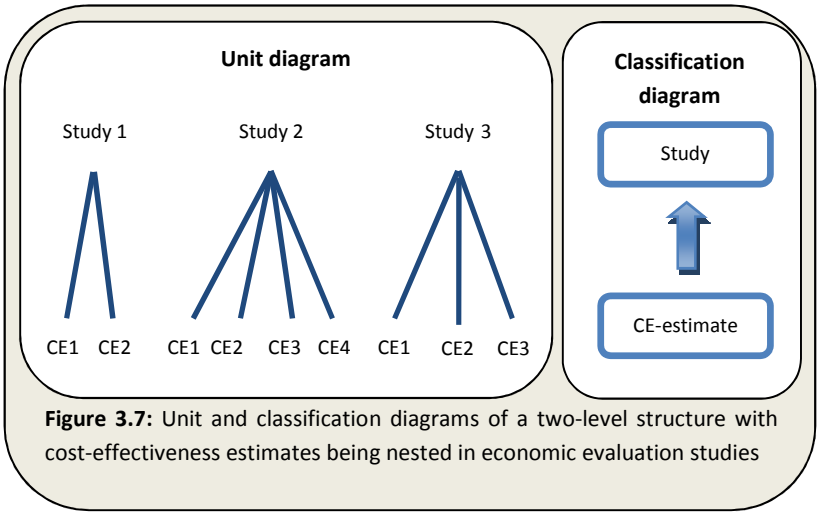


Figure 3.7: Unit and classification diagrams of a two-level structure with cost-effectiveness estimates being nested in economic evaluation studies

However, as the aim of this thesis is to assess the '*geographic transferability*' of health economic evaluation data, there must be yet another hierarchical layer above the study-level (Figure 3.8), meaning that the two- level model discussed thus far may not be suitable to capture the full structure of our data. Hence, the first extension to the model to capture this additional layer is to add another level analogously to what has been done to move from the single level regression framework to the two-level model in Section 3.1.1.

Figure 3.8: Unit and classification diagrams of a three-level hierarchical data structure

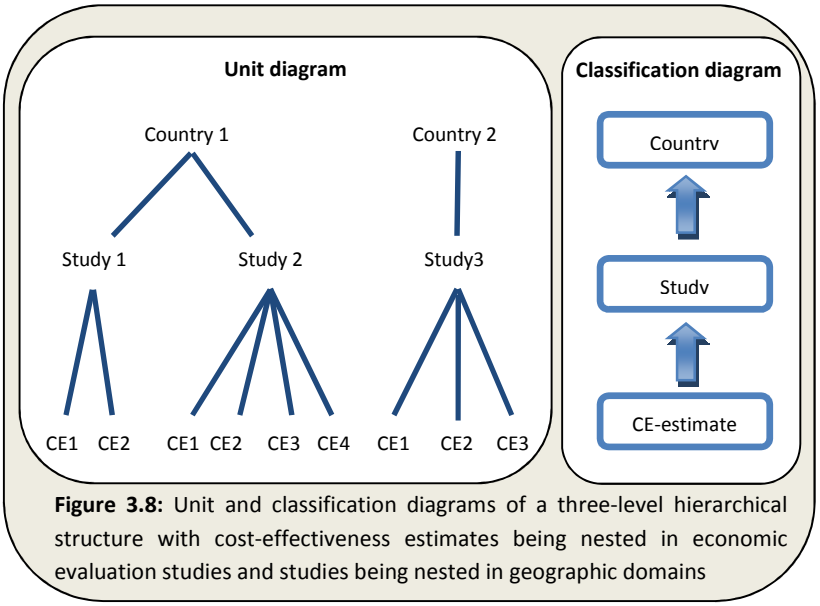


Figure 3.8: Unit and classification diagrams of a three-level hierarchical structure with cost-effectiveness estimates being nested in economic evaluation studies and studies being nested in geographic domains

To do so, reconsider the random intercepts model in equation (9):

$$y_{ij} \sim N(XB, \Omega) \qquad \text{(9, repeated)}$$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_j + u_j + e_{ij} \text{ with}$$
$$u_j \sim N(0, \sigma_u^2)$$
$$e_{ij} \sim N(0, \sigma_e^2)$$

Extending this model to turn it into a three-level hierarchical model is straightforward. We add another error term 'v_{0k}' for the geographic location, and locations receive the subscript 'k' (with k=1,….,K countries) (e.g. Rasbash et al., 2009).

$$y_{ijk} \sim N(XB, \Omega) \tag{11}$$

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + \beta_2 x_{jk} + \beta_3 x_k + v_{0k} + u_{0jk} + e_{0ijk} \quad \text{with}$$

$$v_{0k} \sim N(0, \sigma_{v0}^2)$$
$$u_{0jk} \sim N(0, \sigma_{u0}^2)$$
$$e_{0ijk} \sim N(0, \sigma_{e0}^2)$$

With model (11) we have specified a full three-level hierarchical model with covariates on data-, study- and country-level. Unfortunately, however, this model does still not capture the full complexity of the data-structure which we aim to reflect. An additional problem arises as adding the country-level to the current two-level structure may not necessarily result in a strict hierarchy where each cost-effectiveness estimate belongs to one study only and each study nests in one particular geographic location. Sometimes, a study produces estimates of cost-effectiveness for more than one geographic location. For example, an economic evaluation conducted alongside a multinational RCT usually reports results for each participating jurisdiction. As a result, cost-effectiveness estimates for these '*multinational studies*' are nested within studies, and they are also nested within geographic domains. However, studies are not nested within geographic domains, and geographic domains are not nested within studies. Rather, studies and geographic domains are *'cross-classified'*. To illustrate the problem of cross-classified data, consider Figure 3.9. The unit diagram shows how cost-effectiveness estimates obtained from specific studies may belong to different geographic domains. Figure 3.9 also contains the related classification diagram for this situation, which can be regarded as a more general case as the hierarchical structures in Figurers 3.7 and 3.8 (Rasbash, 2008). In fact, the hierarchical structure is only a special case of the cross-classified model, and the existence of these non-hierarchical structures justifies the use of the term '*multilevel models'* in preference to the more specific term '*hierarchical models*' (Rasbash, 2008).

Figure 3.9: Unit and classification diagrams of a cross-classified data structure

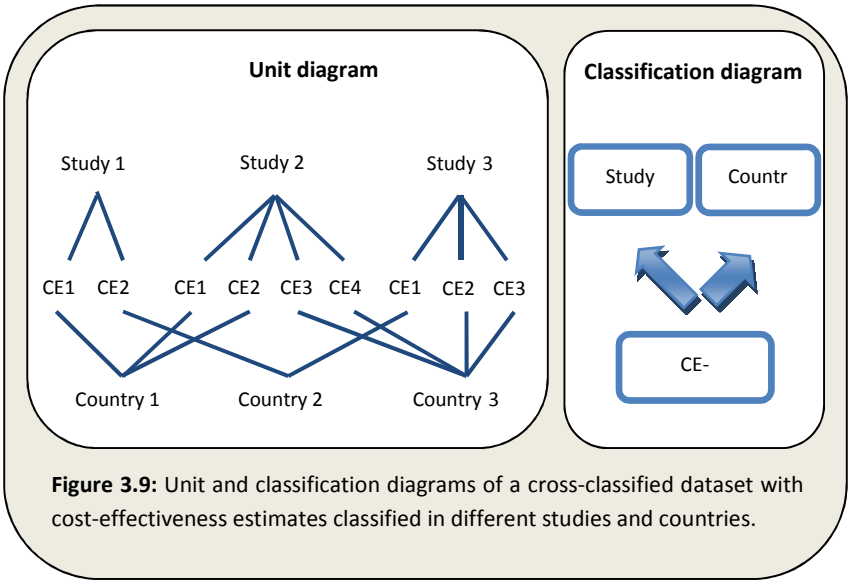


Figure 3.9: Unit and classification diagrams of a cross-classified dataset with cost-effectiveness estimates classified in different studies and countries.

The question is how to expand upon the strictly hierarchical model introduced above to allow for cross-classified data. First, we do not longer talk about three '*levels*', as studies and countries become classifications on the same level (Rasbash et al., 2009). Hence, the model conceptually turns into a two-level, cross-classified structure (Goldstein & Sammons, 1997; Rasbash et al., 2009; Hox, 2010). Next, unique classification identifiers are needed (Rasbash et al., 2009). Thus far, within the strictly hierarchical model, one subscript was used for each data-level with 'y_{ijk}' denoting the i^{th} cost-effectiveness estimate from the j^{th} study and k^{th} country. As models with many classifications are possible, the use of subscripts may become bulky, which is why other notations exist (Rasbash et al., 2009). However, in our model with two levels and only two classifications, we can comfortably carry on with using subscripts 'i', 'j' and 'k' and simply use parentheses to group together the subscripts 'j' and 'k' to express that studies and countries are cross-classified (Hox, 2010). Hence, the three-level random intercepts model introduced in equation (11) turns into a two-level cross-classified model of the form:

$$y_{i(jk)} \sim N(XB, \Omega) \quad (12)$$

$$y_{i(jk)} = \beta_0 + \beta_1 x_{i(jk)} + \beta_2 x_j + \beta_3 x_k + u_{0k} + u_{0j} + e_{0i(jk)} \quad \text{with}$$

$$u_{0k} \sim N(0, \sigma^2_{u0k})$$
$$u_{0j} \sim N(0, \sigma^2_{u0j})$$
$$e_{0i(jk)} \sim N(0, \sigma^2_{e0})$$

Note that the notation for the country and study error terms also changed (Rasbash et al., 2009). As we have now specified that both studies and countries are cross-classified on level two, we denote the error terms with 'u_k' for country and 'u_j' for study rather than 'v_k' and 'u_{jk}' in the strictly hierarchical model.

The estimation procedure is more complicated for the cross-classified model compared to its strictly hierarchical counterpart (Rasbash et al.; 2009, Hox, 2010). Thus far, the hierarchical structure of the data permitted the structure of the covariance matrix to be '*block-diagonal*' as it was shown in Figure 3.6 in Section 3.1.2.1. However, as Rasbash et al. (2009) make clear, the cross-classified model requires a non-block diagonal covariance structure. For this reason, to be able to estimate cross-classified effects between study and country using iterative generalised least squares (IGLS), a third '*dummy-level*' needs to be introduced with one unit that spans the entire dataset (Rasbash et al.; 2009, Hox, 2010). Then, dummy variables are created on the cost-effectiveness level with one dummy variable for each country (Rasbash et al.; 2009, Hox, 2010). Finally, coefficients of the dummy variables created are permitted to vary randomly at the country-level and, whilst covariances between dummies are assumed to be zero, their variances are assumed to be equal (Rasbash et al.; 2009, Hox, 2010). Hence, according to Hox (2010), we estimate one variance component for the countries to which cost-effectiveness estimates apply to, and by creating a separate level for countries, we make sure that the covariance between studies and countries is zero. Again, although technically this model uses three data-levels, it is conceptually a two-level cross-classified model with studies and countries being cross-classified on the same level (Rasbash et al.; 2009, Hox, 2010). As Hox (2010) further states, the third level is '*just a computational device to allow estimation using standard multilevel software*'. Therefore, we refer to it as *'dummy-level'* (Hox, 2010). Note that an alternative procedure to estimate

cross-classified models in MLwiN, which is also used for the purposes of this thesis, is Markov Chain Monte Carlo (MCMC) estimation (Browne, 2012).

Acknowledging cross-classification within multilevel models is not new, though, to the knowledge of the author, there has not yet been an application of a cross-classified multilevel-model within the area of economic evaluation in health. Several authors used two-level models for the analysis of individual patient data collected alongside randomised controlled trials, cluster randomised trials, or observational studies (e.g. Grive et al., 2005; Willan et al., 2005; Manca et al., 2005; Nixon et al., 2005; Pinto et al., 2005; Thompson et al., 2006; Manca et al., 2007; Grieve et al., 2007; Bachmann et al., 2007; Coupe et al., 2007; Willan et al., 2008; Petrinco et al., 2009; Grieve et al., 2010; Edbrooke et al., 2011) but datasets analysed in these studies were strictly hierarchical; hence, cross-classification was not an issue. However, acknowledging cross-classification allows integrating secondary data of both single-country studies and multi-country studies, as the latter cause the strict hierarchy between studies and countries to break down. The underlying assumption is that data within individual studies and countries is dependent, whilst independence is assumed between studies and between countries reflected in the dataset. However, there are reasons to be critical about the independency assumption of data for countries considered within multinational studies. Multinational studies may underestimate variability between countries because of, for instance, standardised trial protocols (e.g. Ramsey et al., 2005), the use of pooled effectiveness or resource use data (e.g. Barbieri et al., 2005), or methodological characteristics of the study which may cause cost-effectiveness estimates between countries to be more in the same range compared to evidence from single country-studies. As a result, underestimated country-level variability in multinational studies may affect the overall variability observed between countries in the cross-classified model.

For this reason, Section 3.4 reports on a pilot study in which MLMs were tested before the main empirical analysis was carried out. If country-level variability is observed to be suspiciously low in the cross-classified model in the pilot study, this may be a reason to consider alternative assumptions regarding the

independence of data between countries from multinational studies. One alternative may be to fall back to a strictly hierarchical three-level model as reported in equation (11), where data from multinational studies are simply clustered in a separate group on country-level. However, for the purposes of the pilot, the cross-classified model (12) is the main model of interest, which is why the remainder of this chapter further elaborates on this particular framework. Nevertheless, everything mentioned in the remainder of this chapter holds just as much for a hierarchical three-level model. The pilot study in Section 3.4 and Chapter 5.1, which is concerned with determining the appropriate MLM structure for the main empirical analysis, further elaborate on this matter.

Before moving on to the question of how the dependent variable may be defined within the MLM framework in Section 3.2 of this chapter, there is a third extension to the model developed thus far. The MLM framework offers the unique opportunity to model the variation in the response variable as a function of explanatory variables through the inclusion of random slopes to the model. This is considered next in Section 3.1.2.3.

3.1.2.3. Random slopes and the variance function

In the variance components model, the effect of explanatory variables on the cost-effectiveness of a healthcare technology is not considered. Therefore, the random intercepts model with explanatory variables on each level of the data hierarchy was introduced in Section 3.1.2.1. However, within that section, the effect of explanatory variables on cost-effectiveness estimates was assumed to be the same across all studies (and countries). Hence, all regression lines were placed parallel to each other with $'\beta_0'$ being the intercept and $'\beta_1 x_{i(jk)}'$ being the slope of the overall regression line and $'\beta_0 + u_j'$ being the intercepts of the individual study regression lines. This was illustrated in Figure 3.5. However, the effect of explanatory variables may be different, for instance, for different studies in the dataset. In this case, the random intercepts model does not adequately capture the reality of such relationships. Therefore, a new class of

models is now being introduced, the *'random slopes model' (Steele, 2008; Bartholomew et al., 2008; Rasbash et al., 2009; Hox, 2010)*.

In contrast to the random intercepts model, each regression line in the random slopes model has its individual intercept and slope (*Steele, 2008; Bartholomew et al., 2008; Rasbash et al., 2009; Hox, 2010*). This allows each explanatory variable to have an individual effect, for instance, on each study in the dataset, which is also illustrated in Figure 3.10. The question is how to expand the random intercepts specification to turn it into a random slopes model?

Recall equation (12):

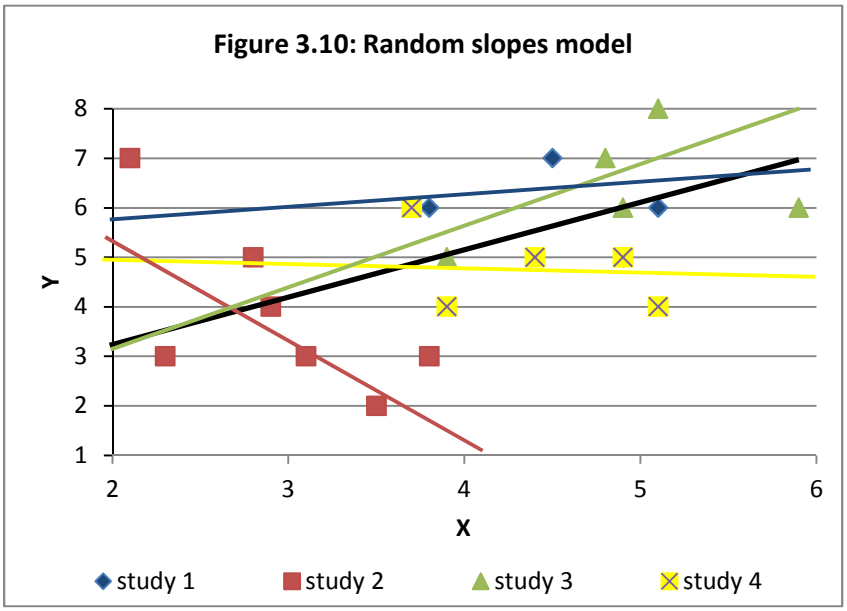
$$y_{i(jk)} \sim N(XB, \Omega) \qquad \text{(12, repeated)}$$

$$y_{i(jk)} = \beta_0 + \beta_1 x_{i(jk)} + \beta_2 x_j + \beta_3 x_k + u_{0k} + u_{0j} + e_{0i(jk)} \qquad \text{with}$$

$$u_{0k} \sim N(0, \sigma^2_{u0k})$$
$$u_{0j} \sim N(0, \sigma^2_{u0j})$$
$$e_{0i(jk)} \sim N(0, \sigma^2_{e0})$$

We can turn this model into a random slopes model by adding a random term to any of the slope parameters on any level of the model (*Steele, 2008; Bartholomew et al., 2008; Rasbash et al., 2009; Hox, 2010*). For instance, if we would like the model to capture that the relationship between patient

Figure 3.10: Random slopes model



characteristics on data-level and measures of cost-effectiveness may be different for different studies in the dataset, we add a random slope on level two for the respective patient covariate. The idea is that, just like the random term $'u_{0j}'$ allows the intercepts of individual study regression lines to vary randomly, this additional random component $'u_{1j}'$ attached to the regression coefficient will allow its respective slope to vary randomly too (*Steele, 2008; Bartholomew et al., 2008; Rasbash et al., 2009; Hox, 2010*).

This turns model (12) into:

$$y_{i(jk)} = \beta_0 + (\beta_1 x_{i(jk)} + u_{1j}) + \beta_2 x_j + \beta_3 x_k + u_{0k} + u_{0j} + e_{0i(jk)} \quad (13)$$

which can be rearranged so that:

$$y_{i(jk)} \sim N(XB, \Omega) \quad (14)$$

$y_{i(jk)} = \beta_0 + \beta_1 x_{i(jk)} + \beta_2 x_j + \beta_3 x_k + u_{0k} + u_{0j} + u_{1j} + e_{0i(jk)}$ with

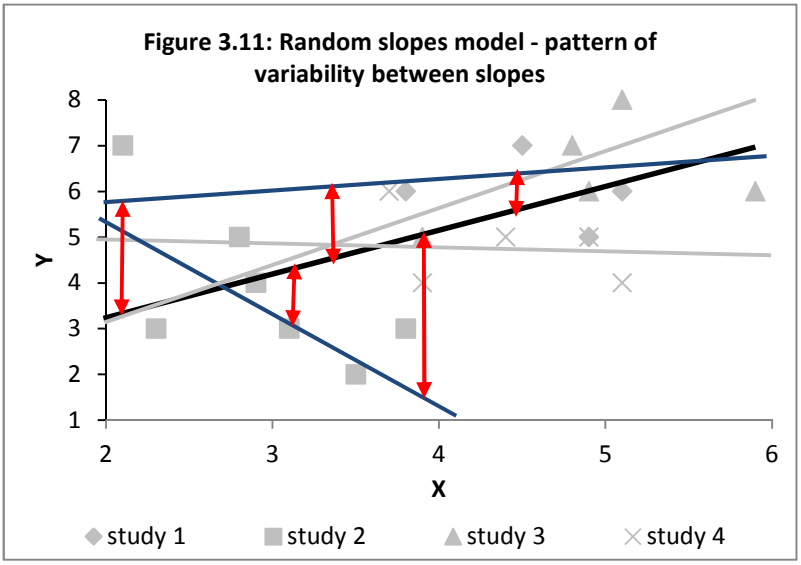
$[u_{0k}] \sim N(0, \Omega_{uk})$ where $\Omega_{uk} = [\sigma^2_{u0k}]$

$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_{uj})$ where $\Omega_{uj} = \begin{bmatrix} \sigma^2_{u0j} & \\ \sigma_{u01j} & \sigma^2_{u1j} \end{bmatrix}$

$[e_{0i(jk)}] \sim N(0, \Omega_e)$ where $\Omega_e = [\sigma^2_{e0}]$

As before in the random intercepts model, we can interpret the parameter $'\beta_0'$ as the intercept and $'\beta_1 x_{1i(jk)}'$ as the slope of the pooled regression line (Steele, 2008; Rasbash et al., 2009, CMM-workshops/random slopes). Also analogously to the random intercepts model, $'\sigma^2_{e0}'$ is the variance for the within-study error term $'e_{0i(jk)}'$ (Steele, 2008). However, in contrast to the random intercepts model, the slope is no longer identical between the pooled regression line and the individual study regression lines (*Steele, 2008; Bartholomew et al., 2008; Rasbash et al., 2009; Hox, 2010*). Hence, special attention needs to be placed on

the interpretation of $'\Omega_{uj} = \begin{bmatrix} \sigma^2_{u0j} & \\ \sigma_{u01j} & \sigma^2_{u1j} \end{bmatrix}'$ in which $'\sigma^2_{u1j}'$ is the variance in slopes between studies, $'\sigma^2_{u0j}'$ is the variance in intercepts between studies, and $'\sigma_{u01j}'$ is the covariance between intercepts and slopes between studies (Steele, 2008; CMM-workshops/random slopes).



Specifically, as there is no variation in a random intercepts model between slopes of the individual regression lines, the covariance for slopes and intercepts $'\sigma_{u01j}'$ is not defined (Steele, 2008; CMM-workshops/random slopes). However, in a random slopes model, both intercepts and slopes may vary randomly between studies included in the dataset, making $'\sigma_{u01j}'$ an important parameter (Steele, 2008). If the covariance between slopes and intercepts is positive, that means there is a pattern of regression lines over the range of the explanatory variable which is '*fanning out*' (Steele, 2008; CMM-workshops/random slopes). This is, for example, the case in the sample dataset illustrated in Figure 3.10. Conversely, if the covariance $'\sigma_{u01j}'$ is negative, then the regression lines are '*fanning in*' (Steele, 2008; CMM-workshops/random slopes). Finally, if the covariance is zero, then there is no pattern in the variation of slopes over the range of the explanatory variable x (Steele, 2008).

Hence, the matrix '$\Omega_{uj} = \begin{bmatrix} \sigma^2_{u0j} & \\ \sigma_{01j} & {}^{2}_{u1j} \end{bmatrix}$' defines the relationship between the error terms of the intercepts 'u_{0j}' and the error term of the slopes 'u_{1j}' (Steele, 2008; CMM-workshops/ random slopes). It also turns out that through introducing random slopes, the value of the intercept depends on where we define the explanatory variable 'x' to be zero (Steele, 2008; CMM-workshops/ random slopes). Consider Figure 3.11: depending on where the y-axis cuts the x-axis, we observe different values of the between group error term 'u_{0j}', the variance of the intercepts between studies 'σ^2_{u0j}' and the covariance between slopes and intercepts 'σ_{u01j}' (Steele, 2008; CMM-workshops/ random slopes). Specifically, whilst the data-level and country-level variance terms remain unchanged with 'σ^2_{e0}' and 'σ^2_{u0k}' respectively, we now have two random terms at the study-level, and the total level two variance may be expressed as a quadratic function of the explanatory variable (Steele, 2008; Rasbash, 2009; CMM-workshops/ variance function):

$$Var(u_{0j} + u_{1j}x_{1ij}) = \sigma^2_{u0j} + 2\sigma_{u01j}x_{1ij} + \sigma^2_{u1j}x^2_{1ij} \quad (15)$$

Accordingly, the VPC may be expressed as (e.g. Steele, 2008):

$$VPC = \frac{\sigma^2_{u0}+2\sigma_{u01}x_{1ij}+\sigma^2_{u1}x^2_{1ij}}{\sigma^2_{u0}+2\sigma_{u01}x_{1ij}+\sigma^2_{u1}x^2_{1ij}+\sigma^2_{e0}} \quad (16)$$

Now that a model has been specified which allows both intercepts and slopes to vary randomly across studies, the question is why this concept may be valuable for addressing the transferability problem of economic evaluation data. The answer is that equation (15) represents a '*variance function*' which shows how study or country-level variance changes over the range of an explanatory variable (Steele, 2008; Rasbash et al., 2009; CMM-workshops/ variance function). Hence, random slopes models provide us with the opportunity to explicitly model variability in measures of cost-effectiveness across studies, or geographic domains, as a function of explanatory variables. This is also shown in Figures 3.12 and 3.13 below. Figure 3.12 shows how the variance in cost-effectiveness data between study regression lines depends on the value of the explanatory variable,

and this variance between studies has been plotted in Figure 3.13. Being able to model means and variances simultaneously offers unique opportunities for addressing the transferability problem of health economic evaluation data. For instance, Sculpher et al. (2004) mentioned that patient characteristics in economic evaluation studies (like age, gender, risk factors, etc.) constitute critical variability factors which may feed through to higher levels of the model hierarchy.

Figure 3.12: Random slopes model - repeated

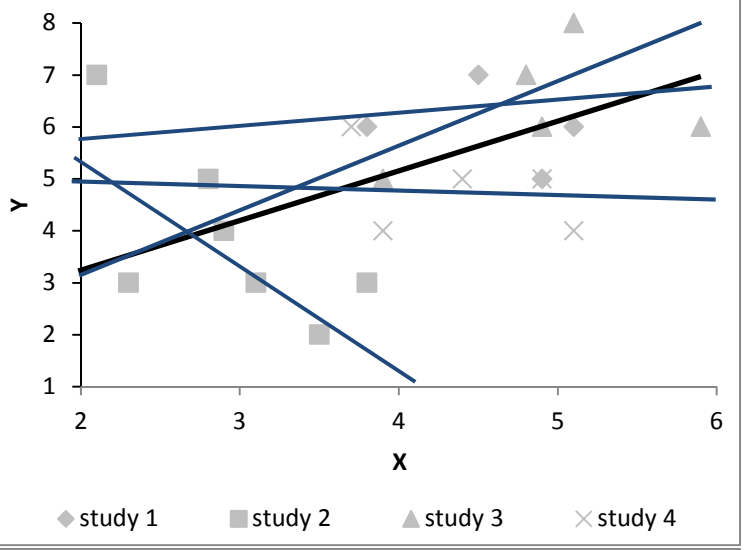
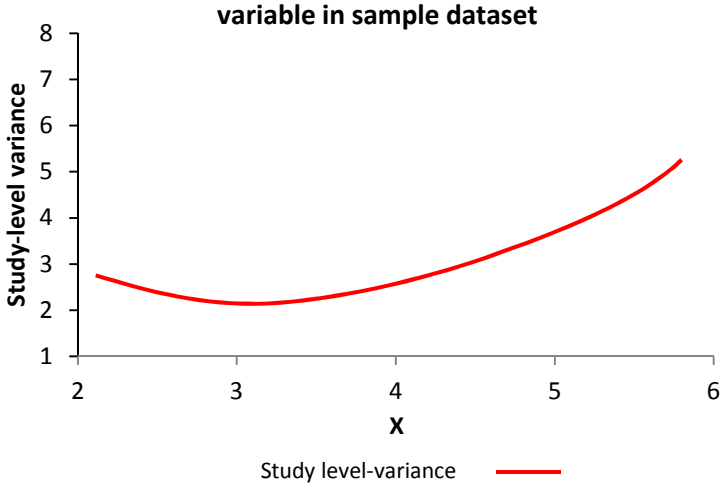


Figure 3.13: Variance function for explanatory variable in sample dataset



Modelling variability in measures of cost-effectiveness between economic evaluation studies from different geographic domains as a function of (lower-level) patient characteristics, shows for which range of the respective patient

factor variability between studies and/or countries is lowest. In other words, we infer the range of values for the explanatory variable for which agreement in results between studies, or countries, is highest, and transferring evidence to the target domain may be most indicated. Conversely, for ranges of the explanatory variable for which variability between studies / countries is high, transferring evidence to the target domain may not be indicated and one may rather conclude commissioning a new study for the target country. In this respect, the variance function may be utilised to target research resources more efficiently to those study questions for which disagreement in existing international economic evaluation data is highest.

For example, suppose the explanatory variable modelled in the sample dataset is the mean low density cholesterol level (LDL) of patient subgroups in each study and a random slope is specified on study-level, i.e. allowing that the effect of LDL on the cost-effectiveness of statins is different between studies. What we learn from Figure 3.13 is that the variability in cost-effectiveness estimates increases between studies with increasing LDL. Hence, we may be more inclined to transfer findings to a new setting for lower LDL patients, whilst suggesting additional evidence for the target domain for higher LDL patients. Although we do not have clear thresholds to either accept, or decline a given level of variability within the data, this approach could help setting more specific research priorities in the light of limited research resources; i.e., decision makers could invest research resources within one intervention area more targeted to cases of high variation in existing cost-effectiveness data, whilst relying on transferred evidence for cases which show comparatively low variation in cost-effectiveness estimates between studies and / or countries.

Thus far, exchangeability of study parameters was assumed to allow for the integration of secondary cost-effectiveness data. This resulted in the variance components model introduced in Section 3.1.1. Subsequently, this model was extended to include explanatory variables on each level of the data hierarchy in Section 3.1.2.1, and more complex data structures were modelled including the cross-classification between studies and countries due to multinational study data in Section 3.1.2.2. In this particular section, random slopes and the

variability in measures of cost-effectiveness were modelled as a function of explanatory variables. However, the question of how to define the dependent variable in the model, i.e. how to express the cost-effectiveness of a healthcare technology as a response variable in a MLM for secondary data integration, needs to be addressed next. Section 3.2 elaborates on this question. Finally, in Section 3.3, models of both previous sections are combined, resulting in a '*bivariate multilevel model for secondary data integration'.*

3.2. Dependent variable in a model for secondary economic evaluation data integration

As mentioned above, in order to develop a multilevel model for secondary data integration from international health economic evaluation studies, one has determine how to define the dependent variable in the model, i.e. how to express the '*cost-effectiveness'* of a healthcare technology as a response variable in a MLM framework. There is an explicit theory on regression analytic modelling within health economic evaluation, and analysts suggest the use of the *'incremental net monetary benefit framework'* (INMB) as this has more favourable statistical properties than other, non-linear, measures of cost-effectiveness (Claxton & Posnett, 1996; Stinnet & Mullahy, 1998, Briggs & Fenn, 1998; Tambour et al., 1998). In addition, bivariate modelling has been suggested to include measures of incremental costs and incremental effects separately in a vector of response variables in one regression framework (Nixon et al., 2005; Pinto et al., 2005; Manca et al., 2007; Grieve et al., 2007; Bachmann et al., 2007; Willan et al., 2008; Grieve et al., 2010). This section elaborates on how to define the response variable, or a vector thereof, within the MLM for secondary data integration.

3.2.1. Incremental Cost-Effectiveness Ratio (ICER) or Incremental Net Monetary Benefits (INMB)

From standard health economics textbooks, we can learn that *'for a meaningful comparison, it is necessary to examine the additional costs that one health care intervention imposes over another, compared to the additional benefits, or utilities it delivers'* (Drummond et al., 2005a). A commonly used summary measure for this *'incremental'* approach is the so called '*Incremental Cost-Effectiveness Ratio'* (ICER), that is:

$$ICER = \frac{C_i - C_a}{E_i - E_a} = \frac{\Delta C}{\Delta E} \quad (17)$$

C_i= (mean) cost of intervention
C_a= (mean) cost of alternative
E_i= (mean) effectiveness of intervention
E_a= (mean) effectiveness of alternative

The ratio can be interpreted as the additional financial resources which need to be invested in order to achieve a unit of health gain compared to an alternative strategy (Stinnet & Mullahy, 1998). However, there are several problems associated with the ICER as a summary statistic for the cost-effectiveness of a healthcare intervention which are well discussed in the relevant literature (Drummond et al., 2005a; Claxton & Posnett, 1996; Stinnet & Mullahy, 1998; Briggs & Fenn, 1998):

- the ICER may entail information on the relative cost-effectiveness of a healthcare intervention, but it is not sufficient to decide upon whether or not to implement that healthcare technology. To make this decision, we need an external criterion, i.e. a threshold which represents the willingness to pay for the additional health outcome achieved. (Drummond et al., 2005a) This threshold is denoted by (λ) and the decision rule is to adopt the technology if:

$$ICER = \frac{C_i - C_a}{E_i - E_a} < \lambda \quad (18)$$

- The interpretation of ICERs, is not unambiguous without having additional information on the location of the new technology on a cost-effectiveness plane. In other words, the ICER of an intervention may be identical if it is less costly and more effective, or more costly and less effective. Without having additional information on the quadrant of the CE plane in which the ICER falls, there is no meaningful interpretation of this ratio-statistic (Stinnet & Mullahy, 1998). In addition, negative ICER's are not defined, and hence, are not informative for decision making (Briggs & Fenn, 1998)

- A statistical problem that follows from this ambiguity relates to the calculation of confidence intervals (CIs) for ICERs if there is a non-negligible chance that the ratio may be negative. As the probability distribution of the ICER does not carry any information regarding its location on the cost-effectiveness plane, its distribution and hence the construction of CIs is ambiguous (Stinnet & Mullahy, 1998).

- Also, if there is a non-negligible probability of the denominator taking values close to zero, then the quantification of sampling uncertainty becomes an issue as the moments of the sampling distribution may be undefined (Briggs & Fenn, 1998)

Hence, though the ICER follows an incremental approach to cost-effectiveness analysis, which is advocated by the relevant literature (Drummond et al., 2005a), it possesses some qualities which cast into doubt its usefulness as a dependent variable within a regression analytic framework (Claxton & Posnett, 1996; Stinnet & Mullahy, 1998; Briggs & Fenn, 1998). To overcome the problems associated with ICERs, several health economists advocated the *'incremental net monetary benefit approach'* (INMB) as an alternative (e.g. Stinnet & Mullahy, 1998; Briggs & Fenn, 1998; Tambour et al., 1998). INMBs may be calculated simply by rearranging equation (18). Recall that we decide to adopt a health technology if:

$$ICER = \frac{C_i - C_a}{E_i - E_a} < \lambda \qquad \text{(18, repeated)}$$

Rearranging this equation leads to:

$$\lambda * (E_i - E_a) > C_i - C_a \qquad (19)$$

And finally

$$\lambda * (E_i - E_a) - C_i - C_a > 0 \quad (20)$$

or

$$\Delta E * \lambda - \Delta C > 0 \qquad \text{(INMB)} \quad (21)$$

As the INMB is not a ratio-statistic, it essentially solves the problems mentioned above. The INMB statistic is defined both for negative and positive values and there is no ambiguity in its interpretation with respect to its location on the cost-effectiveness plane; i.e. a positive (negative) ICER always means that the new intervention is more (less) favourable than the comparator technology and, the higher the INMB value, the more favourable is the intervention under investigation (Stinnet & Mullahy). In other words, whereas ICERs do not obey to the law of transitivity, preferences are monotonic in INMBs (Stinnet & Mullahy, Hoch et al., 2002). In addition, the lineal form of the INMB statistic solves the statistical shortcomings of the ICER, making it more suitable when employing regression analytic methods (Hoch et al., 2002), which is the aim of this exercise.

3.2.2. Net benefit regression for RCT data

So far we have learned that the cost-effectiveness of a health care technology should be measured in terms of both its incremental costs and incremental health effects of one intervention compared to another. Secondly, because of its linear form, the INMB approach may be preferred over to the ICER statistic, which suffers from several shortcomings in terms of its interpretation and in terms of its statistical properties. However, the next task is to employ the INMB statistic within a regression analytic framework. We previously defined that:

$$\Delta E * \lambda - \Delta C > 0 \qquad \text{(INMB)} \quad (21, \text{repeated})$$

Hoch et al. (2002) transferred the INMB framework into a linear model suitable for net benefit regression in a randomised controlled trial. First, they estimated a net benefit value for each subject in a trial setting in the form of.

$$NMB_i = \lambda * E_i - C_i \tag{22}$$

where $'E_i'$and $'C_i'$ are the observed effects and cost for subject 'i' in the trial (Hoch et al., 2002). A simple linear regression model was then defined in the form of:

$$NMB_i = \beta_0 + \delta t_i + \varepsilon_i \tag{23}$$

where, 'β_0' represents an intercept term, 't' a treatment dummy taking the value zero for the standard treatment and the value one for the treatment under consideration, and $'\varepsilon'$ is the stochastic error term (Hoch et al., 2002). The coefficient $'\delta'$ provides the estimate of the incremental net monetary benefit ($INMB_i = NMB_{1i} - NMB_{0i}$) In an extension of model (23), Hoch et al. (2002) showed how this regression framework can be exploited to add explanatory variables to the regression equation, which then takes the following form:

$$NMB_i = \beta_0 + \sum_{p=1}^{P} \beta_p x_{ip} + \delta t_i + \varepsilon_i \tag{24}$$

In this model, there are 'p' covariates and the INMB provided by $'\delta t_i'$ is estimated whilst taking into account the impact of these explanatory variables, which could be, for example, age, gender, or severity of disease of patients participating in the trial (Hoch et al., 2002).

Hoch et al. (2002) made a very important contribution as they transferred the INMB-statistic into a regression framework. However, three important issues of dependency within (secondary) cost-effectiveness data need to be addressed to derive a model suitable for the purposes of this thesis. The three dependencies to be addressed are:

- no independence between intervention and control groups within a study if data is combined from different studies
- no independence of costs and effects within each measure of cost-effectiveness, and
- Dependence of the INMB statistic on a (country specific) threshold value λ

The following section addresses each of these issues showing that all of them can be considered appropriately within a *'multivariate regression framework'.*

3.2.3. Decomposing the INMB statistic

As mentioned above, there are several dependencies within secondary cost-effectiveness data from different economic evaluation studies which the standard net benefit framework proposed by Hoch et al. (2002) does not adequately capture. First of all, as the net benefit regression framework was purposely developed for economic evaluations '*within*' RCTs, it treats both costs and effects in the treatment and the control arm as independent. As shown in equation (24), the treatment dummy $'\delta t_i'$ is '1' for the intervention and '0' for the control group, assuming zero covariance between the interventions and the controls. Within the boundaries of a single RCT (for which the net benefit regression framework was developed), this is a perfectly reasonable assumption, as the trial protocol is supposed to eliminate dependencies between both groups. However, when the intension is to integrate evidence from different studies, then estimates from the same study suddenly become highly dependent.

The second issue relates to the use of NMBs as dependent variable in the net benefit regression framework. In particular, using NMBs on the left hand side of the equation means combining estimates on both costs and effects in the dependent variable, so that the existence of correlation between the two components of the NMB statistic is not being made explicit (note again that in Hoch's net benefit regression framework, NMBs constitute the dependent

variable, the incremental approach is introduced through the inclusion of the treatment dummy $'\delta t_i'$ (Hoch et al (2002)). However, each component of the NMB statistic is likely to contribute in different amounts to the overall variability in measures of cost-effectiveness, and explanatory variables are likely to have a differential effect on each of these components (e.g. Manca et al., 2007). Using NMB as outcome measure would hence preclude the chance of estimating the differential impact of covariates on both costs and effects of the healthcare intervention and its respective comparator. One could argue that fitting separate models may solve the problem. However, doing so would mean to model the opposite extreme, where the existence of correlation between the components of net monetary benefits is being ignored altogether. Therefore, several authors already suggested the use of bivariate models to overcome the problem of dependency between cost and effects of a healthcare technology (Nixon et al., 2005; Pinto et al., 2005; Manca et al., 2007; Grieve et al., 2007; Bachmann et al., 2007; Willan et al., 2008; Grieve et al., 2010).

In addition to all of the above, the standard net benefit regression framework requires a new regression to be estimated for each value of $'\lambda'$ as different cost-effectiveness thresholds might be applicable to different jurisdictions (e.g. Manca et al., 2007). A solution to all of these problems may exist in the form of multivariate models as they allow for correlation amongst a set of response variables. How to derive such a multivariate model is the question for the remainder of this section.

First of all, instead of calculating the NMB for interventions and controls and using $'\delta t_i'$ as a treatment dummy, we may first decompose the INMB framework, which provides four potential response variables for a multivariate model. These four response variables are the costs and effects of the intervention, and the costs and effects of the comparator respectively.

Denote by 'Y_{di}' the response on variable 'd' with (d=C_{int}, E_{int}, C_{comp}, E_{comp},) for each estimate of INMB 'I' with (i=1,....,n). Modelling costs and effects of both intervention and comparator simultaneously as response variables within one

regression equation makes explicit the correlation between all components of the INMB statistic. However, having more than one dependent variable on the left hand side of the equation requires making further adjustments to the overall regression equation, which is why response indicators 'r_d' are introduced (Bartholomew et al., 2008). These response indicators are dummies being coded '1' for each observation on variable 'd' and '0' otherwise (Bartholomew et al., 2008). In the current example case, we have a dataset with 'i=1,...,n' observations of INMB, and 'd=4' response variables '$d_1 = C_{int}$' , '$d_2 = E_{int}$', '$d_3 = C_{comp}$' and '$d_4 = E_{comp}$'. The respective dataset may look like it is shown below in Table 3.1 (Bartholomew et al., 2008):

Table 3.1: Multivariate data structure for two data-points

i	d	Y_{di}	r_1	r_2	r_3	r_4
1	C_{int}	Y_{11}	**1**	0	0	0
1	E_{int}	Y_{21}	0	**1**	0	0
1	C_{comp}	Y_{31}	0	0	**1**	0
1	E_{comp}	Y_{41}	0	0	0	**1**
2	C_{int}	Y_{12}	**1**	0	0	0
2	E_{int}	Y_{22}	0	**1**	0	0
2	C_{comp}	Y_{32}	0	0	**1**	0
2	E_{comp}	Y_{42}	0	0	0	**1**
...	...	...	...	...		

For instance, as can be seen from Table 3.1, the response indicator 'r_1' is equal to '1' if the row in the dataset refers to the cost measured for the intervention, and '0' otherwise, and 'r_4' is '1' if the row in the dataset refers to the effect of the comparator, and '0' otherwise. We can now rewrite our model as:

$$y_{di} \sim MVN(XB, \Omega) \qquad [25]$$
$$Y_{di} = \beta_{0d} * r_d + u_{di} * r_d$$

with

$$[u_{di}] \sim MVN\ (0, \Omega_u) \quad \text{where } \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & & & \\ \sigma_{u01} & \sigma_{u1}^2 & & \\ \sigma_{u02} & \sigma_{u12} & \sigma_{u2}^2 & \\ \sigma_{u03} & \sigma_{u13} & \sigma_{u23} & \sigma_{u3}^2 \end{bmatrix}$$

In a sense, response indicators take over the role of the treatment dummy $'\delta t_i\ '$ in Hoch et al's (2002) net benefit regression framework. However, In model (25), an error variance is assigned to each response variable (Bartholomew et al., 2008) and the model allows the covariances between response variables within each measurement of INMB being different from zero as it is shown by the covariance matrix $'\Omega_u'$ (Bartholomew et al., 2008). Hence, model (25) represents a very basic multivariate model, which takes into account the correlation between different components of the INMB statistic. An important advantage of this model is that by adding explanatory variables to the equation, one can infer differential effects of covariates on each response variable (Bartholomew et al., 2008). Model (25) turns into:

$$y_{di} \sim MVN(XB, \Omega) \qquad [26]$$
$$Y_{di} = \beta_{0d} r_d + \beta_{1d} x_{1i} r_d + u_{di} r_d$$

with

$$[u_{di}] \sim MVN\ (0, \Omega_u) \quad \text{where } \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & & & \\ \sigma_{u01} & \sigma_{u1}^2 & & \\ \sigma_{u02} & \sigma_{u12} & \sigma_{u2}^2 & \\ \sigma_{u03} & \sigma_{u13} & \sigma_{u23} & \sigma_{u3}^2 \end{bmatrix}$$

Because of the response indicator $'r_d'$ and the covariance matrix $'\Omega_u'$, which entails a variance term for each response variable and also allows covariances between response variables to be different from zero, we can now fit a different intercept to each response variable, while the interactions between the explanatory variable $'x_i'$ and the response indicator $'r_d'$ allow the effect of the

explanatory variable $'x_i'$ to differ between response variables (Bartholomew et al., 2008). The response variable specific residuals are fitted by permitting the intercept of the response indicator $'r_d'$ to vary randomly across the measurements of INMB (Bartholomew et al., 2008). Note that models (25) and (26) technically belong to the family of MLMs as the response variables are being '*nested*' within each estimate of INMB. However, there are important differences to standard multilevel models, which is why these models are being referred to as *'simple multivariate models'* (Bartholomew et al., 2008). Acording to Bartholomew et al. (2008), these characteristics are:

- the intercept term $'\beta_{0d}'$ allows for different means of each response variable
- As there is only one error term 'u_{di}' which is modelled on the level of the INMB-measure 'u_{di}', this may be viewed as the INMB-specific residuals that have different variances for each response variable. In a standard MLM, there would be one error term on each level.
- Finally, the model allows the error term to be correlated across responses.

By modelling a vector of response variables for each component of the INMB statistic, the net benefit regression framework has been turned into a multivariate model which takes into account correlations between the components of INMBs. If published studies on the cost-effectiveness of a healthcare intervention would explicitly report data on costs and effects of both the intervention and the comparator, then this would be the appropriate model. However, a practical limitation lies in the reporting of cost-effectiveness evidence. Unfortunately, many publications report their cost-effectiveness results only in terms of ICERs or INMBs, without decomposing the measure of cost-effectiveness into its components. If this is the case, the study may not be includable in this empirical exercise. However, whilst information on cost and effects of both the intervention and comparator are almost never reported explicitly, quite a number of studies decompose measures of cost-effectiveness into incremental cost (ΔC) and incremental effects (ΔE) respectively.

The implication is that the multivariate model with four response variables may not be practicable as only very few studies, if any at all, make explicit the

required information on all response variables to populate the dataset. If this is the case, the next best alternative is to include ΔC and ΔE as a vector of response variables in a bivariate regression framework. Though this does not take into account the differential impact of intervention and comparator on INMBs, it still decomposes the INMB statistic so that we may assess the differential impact of explanatory variables on incremental cost and incremental effects whilst taking into account the correlation between the two components of the INMB statistic. As a result, a simplified, bivariate model is proposed, in which the two response variables are ΔC and ΔE of a healthcare intervention. Hence, the vector of response variables (d=C_{int}, E_{int}, C_{comp}, E_{comp},) is replaced by (d= $\Delta C, \Delta E$) and model (26) turns into:

$$\begin{bmatrix} Y_{1i} \\ Y_{2i} \end{bmatrix} \sim BVN(XB, \Omega) \qquad [27]$$

$$Y_{di} = \beta_{0d} r_d + \beta_{1d} x_{1i} r_d + u_{di} r_d$$

$$r_1 = \begin{Bmatrix} 1 \; if \; \Delta \, cost \\ 0 \; if \; \Delta \, effects \end{Bmatrix} \qquad r_2 = 1 - r_1 \qquad \text{with}$$

$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim BVN\ (0, \Omega_u) \quad \text{where } \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

In this section, a bivariate model was developed which makes explicit the correlation between the two components of the INMB statistic ΔC and ΔE and also allows for differential effects of explanatory variables on each response variable. In addition, a MLM with data, study, and country-levels was developed in Section 3.1 of this chapter. The next section combines both models, ultimately deriving a multilevel model suitable for secondary data integration of international health economic evaluation studies.

3.3. Bivariate multilevel model for secondary data integration

In Section 3.1, a MLM was developed which accommodates both a study and a country-level to allow integrating data from different studies applicable to different geographic domains. In addition, cross-classification was modelled as multinational studies provide cost-effectiveness data for more than one country, causing the strict hierarchy between data, study and country-level to break down. In Section 3.2, the measure of cost-effectiveness suitable for econometric analysis was determined. Following from Hoch's net benefit regression framework, a bivariate model was developed with the two components of the INMB statistic (ΔC and ΔE) as a vector of response variables. In this section, both models previously developed are being combined, resulting in a bivariate MLM suitable for secondary data integration of economic evaluation studies applicable to different geographic domains. To do so, reconsider the random intercepts model which makes explicit that data is grouped in studies and countries and also allows for cross-classified data structures due to multinational study data:

$$y_{i(jk)} \sim N(XB, \Omega) \qquad \text{(12, repeated)}$$

$$y_{i(jk)} = \beta_0 + \beta_1 x_{i(jk)} + \beta_2 x_j + \beta_3 x_k + u_{0k} + u_{0j} + e_{0i(jk)} \qquad \text{with}$$

$$u_{0k} \sim N(0, \sigma^2_{u0k})$$

$$u_{0j} \sim N(0, \sigma^2_{u0j})$$

$$e_{0i(jk)} \sim N(0, \sigma^2_{e0})$$

Also, recall the bivariate model shown above in equation (27):

$$\begin{bmatrix} Y_{1i} \\ Y_{2i} \end{bmatrix} \sim BVN(XB, \Omega) \qquad \text{(27, repeated)}$$

$$Y_{di} = \beta_{0d} r_d + \beta_{1d} x_{1i} r_d + u_{di} r_d$$

$$r_1 = \begin{Bmatrix} 1 \; if \; \Delta \, cost \\ 0 \; if \; \Delta \, effects \end{Bmatrix} \qquad r_2 = 1 - r_1 \qquad \text{with}$$

$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim BVN\,(0, \Omega_u) \quad \text{where } \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

To combine both models, we posit model (12) above model (27). As model (27) may be regarded as a two-level hierarchy with responses 'ΔC' and 'ΔE' grouped in individual observations, the resulting model technically consists of four levels. As mentioned, responses are modelled on the lowest level. Above that is the level of individual observations. The third and fourth levels are those of the study and the country. However, conceptually, we may still speak of a two-level model (Bartholomew et al., 2008). The reasons are that the lowest level is only introduced to model a vector of response variables and it does not carry its own residual term (Bartholomew et al., 2008). Secondly, studies and countries are assumed to be cross-classified on level two of the data hierarchy (Hox, 2010). To combine both models, each response variable has its own error variance on each level, and the model also allows the covariance between error terms of each response indicator to be different from zero on each level of the data hierarchy (Bartholomew et al., 2008). This results in the following model:

Bivariate random intercepts model for secondary data integration

$$\begin{bmatrix} Y_{1,i(jk)} \\ Y_{2,i(jk)} \end{bmatrix} \sim BVN(XB, \Omega) \tag{28}$$

$$y_{d,i(jk)} = (\beta_{0d} + \beta_{1d}x_{i(jk)} + \beta_{2d}x_j + \beta_{3d}x_k + u_{0dk} + u_{0dj} + e_{0di(jk)}) * r_{d,i(jk)}$$

$$r_{1,i(jk)} = \begin{Bmatrix} 1\ if\ \Delta\ cost \\ 0\ if\ \Delta\ effects \end{Bmatrix} \quad r_{2,i(jk)} = 1 - r_1$$

with:

$$\begin{bmatrix} u_{0,0,k} \\ u_{0,1,k} \end{bmatrix} \sim BVN\ (0, \Omega_u) \quad \text{where } \Omega_u = \begin{bmatrix} \sigma^2_{u0,0k} & \\ \sigma_{u0,01k} & \sigma^2_{u0,1k} \end{bmatrix}$$

$$\begin{bmatrix} u_{0,0,j} \\ u_{0,1,j} \end{bmatrix} \sim BVN\ (0, \Omega_u) \quad \text{where } \Omega_u = \begin{bmatrix} \sigma^2_{u0,0j} & \\ \sigma_{u0,01j} & \sigma^2_{u0,1j} \end{bmatrix}$$

$$\begin{bmatrix} e_{0,0,i(jk)} \\ e_{0,1,i(jk)} \end{bmatrix} \sim BVN\ (0, \Omega_u) \quad \text{where } \Omega_e = \begin{bmatrix} \sigma^2_{e0,0} & \\ \sigma_{e0,01} & \sigma^2_{e0,1} \end{bmatrix}$$

Now, consider we introduce a random slope on study-level to allow the effect of explanatory variables on each response variable to be different between

different studies included in the dataset. This means including an additional error term for the slope variance and that for each response variable, one has to consider not just their slope and intercept variances, but also their respective covariances. This results in a 4x4 covariance matrix on level two of the bivariate model. The respective model is shown below in equation (29).

Bivariate random slopes model for secondary data integration

$$\begin{bmatrix} Y_{1,i(jk)} \\ Y_{2,i(jk)} \end{bmatrix} \sim BVN(XB, \Omega) \tag{29}$$

$$y_{d,i(jk)} = (\beta_{0d} + \beta_{1d}x_{i(jk)} + \beta_{2d}x_j + \beta_{3d}x_k + u_{0dk} + u_{0dj} + u_{1dj} + e_{0di(jk)}) * r_{d,i(jk)}$$

$$r_{1,i(jk)} = \begin{Bmatrix} 1\ if\ \Delta\ cost \\ 0\ if\ \Delta\ effects \end{Bmatrix} \quad r_{2,i(jk)} = 1 - r_1$$

$$\begin{bmatrix} u_{0,0,k} \\ u_{0,1,k} \end{bmatrix} \sim BVN\ (0, \Omega_u) \quad \text{where } \Omega_u = \begin{bmatrix} \sigma^2_{u0,0k} & \\ \sigma_{u0,01k} & \sigma^2_{u0,1k} \end{bmatrix}$$

$$\begin{bmatrix} u_{0,0,j} \\ u_{0,1,j} \\ u_{1,0,j} \\ u_{1,1,j} \end{bmatrix} \sim BVN\ (0, \Omega_u) \quad \text{where } \Omega_u = \begin{bmatrix} \sigma^2_{u0,0j} & & & \\ \sigma_{u00,01j} & \sigma^2_{u0,1j} & & \\ \sigma_{u01,00j} & \sigma_{u01,10j} & \sigma^2_{u1,0j} & \\ \sigma_{u01,01j} & \sigma_{u01,11j} & \sigma_{u11,01j} & \sigma^2_{u1,1j} \end{bmatrix}$$

$$\begin{bmatrix} e_{0,0,i(jk)} \\ e_{0,1,i(jk)} \end{bmatrix} \sim BVN\ (0, \Omega_u) \quad \text{where } \Omega_u = \begin{bmatrix} \sigma^2_{e0,0} & \\ \sigma_{e0,01} & \sigma^2_{e0,1} \end{bmatrix}$$

The multilevel models for the purposes of this thesis have now been developed. However, before applying these models to the data in the main empirical exercise, it was decided to carry out a pilot study. The final Section 3.4 of this chapter reports on this pilot study and discusses its results in depth and with particular emphasis on the implications for the design and execution of the main empirical analysis, which is reported in Chapter 5.

3.4. Pilot Study

In this section, the models developed above are applied to a set of secondary cost-effectiveness data from 16 international economic evaluation studies on the primary and secondary prevention of CVD. Starting off from an OLS regression equation, multilevel model features, as elaborated on in this chapter, are introduced step by step. This demonstrates the relative merits of the MLM framework for the integration of international economic evaluation data and also gives valuable insights for the design and execution of the main empirical analysis. This main empirical analysis, which is reported in Chapter 5, also begins with an exercise to determine the appropriate MLM structure for analysing data abstracted from 67 international cost-effectiveness studies on statins for the primary and secondary prevention of CVD. The design of this exercise was based to a large extend on the experiences from this pilot study.

The pilot study is organised as follows: Section 3.4.1 states the aim and objectives of this pilot. Then, in Section 3.4.2, the methods of analysis are explained before the data is introduced in Section 3.4.3. Results are reported in Section 3.4.4. Finally, the discussion, which focuses on the implications of pilot study findings on the design and execution of the main empirical analysis, are reported in Section 3.4.5.

3.4.1. Aim and objectives of the pilot study

The primary aim of this pilot is to inform the design and execution of the main empirical analysis, which is reported in Chapter 5. This empirical analysis may be described as a secondary data integration exercise, based on a systematic literature review on the cost-effectiveness of statins in the primary and secondary prevention of CVD (details on the systematic literature review and data abstraction are available from Chapter 4). Within this exercise, it is aimed to apply MLM methods to analyse factors causing variability in international cost-effectiveness data on each level of the data hierarchy. For this purpose, a

multilevel model structure needs to be developed which is suitable for the data available. The first part of the main empirical analysis reported in Chapter 5.1 is therefore concerned with determining this multilevel model structure. However, for the appropriate design of this exercise it was crucial to carry out this pilot study. Accordingly, three main objectives are specified for this pilot.

- First, it is important to test the models developed theoretically in Sections 3.1 to 3.3 of this chapter using real world data. For this reason, the pilot study builds up a MLM in the same sequence as it was developed above, starting off from an OLS regression equation and subsequently elaborating on this model.

- Secondly, as the literature on secondary data integration of economic evaluation data is scarce, and some concerns exists over the overall validity of such an exercise due to enormous variability, especially in socioeconomic data (e.g. Jefferson et al, 1996), it was intended to obtain input from the wider health economics and multilevel modelling communities. For this reason, a paper on this pilot study has been presented at various conferences, including the winter meeting of the Health Economists Study Group (HESG) in January 2011 in York, UK, and a MLM session at the conference of the European Society of Survey Research (ESRA) in July 2011 in Lausanne, Switzerland. The feedback obtained from these conferences significantly influenced the design of the main empirical analysis and is discussed later in this section.

- Finally, the importance of appropriately controlling for variability on each level of the MLM has been stressed already. What follows is that data for covariates need to be obtained on each level of the model hierarchy. Therefore, a data abstraction form was developed based on the existing health economic literature on factors causing variability in economic evaluation data. This data abstraction form was tested, revised, and improved upon within this pilot study. Note, however, that specifics on this aspect of the pilot are reported in full detail in the subsequent Chapter 4, which is concerned with the systematic literature review and data abstraction exercise to obtain data for the main empirical analysis.

The next section introduces the methods of analysis to carry out the pilot study. The data for the pilot is summarized thereafter in Section 3.4.3.

3.4.2. Methods of analysis

The example of statins for the primary and secondary prevention of CVD was chosen as this is an extensively studied area, with many cost-effectiveness studies across many countries, hence allowing for the assumption of random parameters on study and country-level. At the time of the pilot study, a systematic literature review was not yet completed, so that the pilot uses a dataset with studies previously identified by Franco et al. (2005) (details about search strategy, study selection process, etc. can be obtained from this source). Only studies comparing statins with '*doing nothing*' were included. Although some studies estimate QALYs, most do not, so for this exercise, life years saved (LYS) were used as the measure of effect. Studies not reporting incremental costs and incremental effects separately were also excluded, as INMB could not be calculated for these studies.

As a result, from the 24 papers identified by Franco et al. (2005), a further 7 were dropped because of the additional exclusion criteria defined above (Goldman et al., 1991; Hay et al., 1991; Goldman et al., 1993; Riviere et al., 1997; Huse et al., 1998; Pickin et al., 1999; Russel et al., 2001). One further study was not obtainable at the time the pilot study was carried out (Grover et al., 2001). From the remaining 16 papers, cost-effectiveness estimates were collected for the base case, as well as for subgroup analysis, and sensitivity analyses exploring variation by: efficacy of intervention, baseline risk, annual drug costs of statin therapy, duration of statin therapy in years, costs of CVD related events, and discount rates for costs and effects. These sensitivity analyses results were considered as they were usually reported in most studies included in this exercise, which allowed definition of covariates encoding differences between data points for multilevel analyses. However, a number of abstracted cost-effectiveness estimates had to be dropped from the dataset as they referred to less frequently reported forms of sensitivity analyses, and could not be included in the multilevel analysis without losing large numbers of studies. Excluding

these sensitivity analyses reduced the pilot study dataset by around 160 datapoints (25% of all cost-effectiveness estimates abstracted from pilot study papers).

If a data-point was considered includable, data on incremental cost and incremental effects were abstracted. If only ΔC or ΔE was explicitly reported in conjunction with a cost-effectiveness estimate, the missing component of the ICER or INMB statistic was calculated by rearranging the formula for ICERs or INMBs respectively. For instance, if a study reported an ICER of £16,000 and ΔE of 0.5 LYS, ΔC was calculated as $\Delta C = ICER * \Delta E = £8{,}000$. Or if an INMB of say £7,000 was reported at a given threshold value (say £30,000) and ΔE of 0.5, ΔC was calculated as $\Delta C = \Delta E * \lambda - INMB = £8{,}000$. Problems did arise, for instance, if a study reported cost-effectiveness as ICERS, and the use of statins resulted in cost-savings. As studies rightly omitted negative ICERs, this meant, however, that ΔC and ΔE could not be calculated and the respective data point had to be dropped. This resulted in a loss of a further 36 data points within the pilot study dataset. The discussion in Chapter 6 will explicitly discuss potential reasons for bias that might arise from excluding certain data-points and studies from the dataset. In particular, there is a risk that such exclusions may result in a violation of the assumption of random parameters which is essential for fitting multilevel models.

Abstracting data for the pilot study in this way resulted in a total of 464 data points, clustered in 16 studies, and applicable to 16 geographic domains. The fact that there were 16 studies and 16 countries included in the dataset was a coincidence as some studies were multinational in nature, hence, reporting data on more than one country and thereby introducing the cross-classification problem. Local currencies were transferred to Pound Sterling using Purchasing Power Parities (PPP) and updated to 2009 using country specific GDP deflators (Shemilt et al., 2008; OECD, 2010). For countries which adopted the Euro, historic currencies were first converted to Euros using irrevocable Euro conversion rates as adopted by the Council of the European Union on January 1, 1999 (IMF, 2010).

As a first step, models were fitted to the data without any covariates. Accordingly, the most fundamental model in this pilot study was an empty OLS regression equation. Building up from that, two-level hierarchical models were fitted with a) data clustered in countries only and b) data clustered in studies only. Finally, a cross-classified model, which accommodates both a study and country-level, was specified and fit to the pilot study data. All models were implemented in a univariate version with INMB as the only response variable and a bivariate version with ΔC and ΔE as a vector of response variables. For univariate models, a cost-effectiveness threshold level (λ) of £30,000 per LYS was assumed. This does not equate to £30,000 per QALY gained, as most additional years of life will be lived in less than perfect health states. Random slopes were not tested within this pilot study as, at the time it was carried out, the author was not yet successful in fitting them to the models of interest. Table 3.2 provides an overview of the models tested within this pilot study.

Table 3.2: Model specifications applied to the data within the pilot study

	Univariate model with INMB as response variable	Bivariate model with ΔC and ΔE as vector of response variables	Variance components model without covariates	Random intercepts model with covariates	Random slopes model
OLS regression model	√	√	√	√	x
Two-level hierarchical model with data and country-level	√	√	√	√	x
Two-level hierarchical model with data and study-level	√	√	√	√	x
Cross-classified model with data, study and country-level	√	√	√	√	x

After implementing models as variance components models without covariates, a set of explanatory variables was tested in each model. All continuous variables were centred around their mean values, which has the advantage that the intercept is easier to interpret as it represents the predicted INMB (ΔC, ΔE) for average values for each explanatory variable (Steele, 2008; Rasbash et al., 2009; Hox, 2010). The set of potential explanatory variables within this pilot study

analysis was drawn from a long list of variability factors suggested by previous authors (Sculpher et al., 2004; Goeree et al., 2007). For this purpose, a data abstraction form was developed, tested and improved upon within this pilot study. However, the details on the development of the data abstraction form, which was one of the key objectives of this pilot study, are presented in Chapter 4, which is concerned with a systematic literature review, data abstraction exercise, and genealogy study of cost-effectiveness papers on statins in the primary and secondary prevention of CVD.

Within the boundaries of this pilot study, a much reduced dataset, with only a few covariates found to be significant when fitting multilevel models, was used. As a result, All models include three variables on data-level (mean pre-treatment total cholesterol level (TCL), age and percentage of females in the sample), and two variables on the country-level (GDP per capita and total life expectancy at birth). Although a range of study-level covariates was tested (e.g. general study design, timing, or the method of effect calculation, to name a few), the pilot study did not result in a set of study-level covariates which captured significant variation across the studies included in the dataset. Observations with missing values were generally rare for the explanatory variables included in this pilot, and listwise deletion was applied as a simple ad hoc strategy for dealing with missing values within this pilot study (note, however, that missing values were more of a concern in the main empirical analysis which is reported in Chapter 5. Therefore, to minimise the potential for bias and overestimated precision, a much more elaborated approach for dealing with missing values is applied in Section 5.2 of the empirical chapter, where the purpose is to test covariates within the MLM framework).

All models were implemented in MLwiN using Markov Chain Monte Carlo (MCMC) estimation (Rasbash et al., 2009a; Browne, 2012). Though models could also be implemented using iterative generalised least squares (IGLS), it is a more complex procedure, which is why it has been strongly advocated to use MCMC for cross-classified and especially for bivariate models in MLwiN (Browne, 2012; and personal communication with Prof William Browne, CMM, Bristol). A detailed step by step guide on how to implement models in MLwiN can be found

in Appendix 3. The statistical software package MLwiN was chosen for this exercise as it is special MLM software with unique capacities to specify complex data structures (Rasbash et al., 2009a). Though it is acknowledged that there are other software applications allowing to fit multilevel models (e.g. HLM, STATA or R), the author found the features offered by MLwiN particularly useful. In addition, the Centre for Multilevel Modelling in Bristol, which developed the software package MLwiN, is sponsored through the Economics and Social Science Research Council (ESRC) which enables free access to an unrestricted version of MLwiN to all researchers based in a UK academic institution.

Before reporting on the results of the pilot study, the dataset is introduced next in more detail, including descriptive statistics for the response variables and explanatory variables tested within this pilot.

3.4.3. Pilot study data

Table 3.3 provides details on the studies included in the dataset. Further information on the studies included may also be obtainable from Franco et al. (2005). Most studies were based on decision analytic modelling (DAM), although three (Caro et al., 1997; Jonsson et al., 1996; Jonsson et al., 1999) were directly based on clinical trial data. Studies included populations without any known history of CVD (primary prevention), those with at least one prior event of CVD (secondary prevention), or both. Subgroup analysis was usually performed with respect to age, gender, and pre-treatment cholesterol level. As can be seen in Table 3.3, when assuming a threshold value (λ) of £30,000, mean INMBs within each study range from £-2,670 (Pharoa et al., 1996) to over £48,660 (Grover et al., 1999), with an overall mean INMB of £11,632 across all studies included in the dataset.

Table 3.3: Study characteristics and descriptive statistics

Study	Nj*	K**	Study type+	Prev. cat.++	Age range	Mean TCL mmol/L	Mean ΔC 2009 GBP	Mean ΔE LYS	Mean INMB λ= £30.000, 2009 GBP
Ashraf et al., 1996	9	1	2	2	60	6.00	1228	0.157	3472
Caro et al., 1997	1	1	1	1	45-64	7.01	2758	0.098	182
Grover et al., 1999	48	1	2	2	40-70	6.50	16387	2.169	48669
Grover et al., 2000	52	1	2	3	40-70	6.51	25874	1.73	26021
Hamilton et al., 1995	20	1	2	1	30-70	7.22	19734	0.97	9396
Johannesson et al.,1997	39	1	2	2	35-70	6.75	1120	0.22	5433
Muls et al., 1998	15	1	2	2	60	6.00	1885	0.16	2815
Perreault et al., 1998	36	1	2	1	44-57	8.14	18698	0.64	452
Pharoah et al., 1996	46	1	2	3	45-64	7.10	5452	0.09	-2671
Szucs et al., 1998	42	1	2	2	45-65	5.41	3575	0.29	5041
Szucs et al., 2000	8	1	2	2	45-65	5.62	3588	0.41	8562
Van Hout et al., 2001	5	1	2	3	25-75	?	11524	0.93	16436
Jonsson et al., 1996	18	10	1	2	35-70	6.74	1791	0.24	5409
Jonsson et al., 1999	109	11	1	2	35-70	6.74	1342	0.30	7723
Ganz et al., 2000	12	1	2	2	75-84	?	3856	0.26	3794
Martens et al., 1994	4	1	2	1	45	5.56	7223	0.19	-1485
Sum:	**464**	**16**	**Mean:**		**57.06**	**6.66**	**8678**	**0.677**	**11632**

* Number of INMB estimates clustered within a study

** Number of countries included in that study

\+ 1 = Primary modelling (directly based on observations from trial data) / 2 = Secondary modelling (studies based on any form of decision analytic model (DAM))

++ Prevention category: 1 = primary prevention / 2 = secondary prevention /3 = both

See also Franco et al (2005) for additional information on study characteristics [49]

Table 3.4: Country characteristics

Country	Nk*	J**	GDP***+ per capita	Health care spending***+ % of GDP	Life exp. at birth***+ years	Mean ΔC 2009 GBP	Mean ΔE LYS	Mean INMB λ=30.000, 2009 GBP
Australia	1	1	21541	8.0	77.83	1967	0.240	5233
Belgium	23	3	22556	7.8	76.84	1855	0.201	4180
Canada	160	5	22696	9.5	78.05	20179	1.483	24297
Denmark	11	1	25259	8.0	75.95	854	0.305	8290
Finland	11	1	20968	7.4	76.88	1687	0.305	7458
France	8	2	21556	9.6	78.28	1101	0.285	7437
Germany	62	4	23289	10.8	76.84	3137	0.305	6006
Italy	12	2	22460	7.6	78.58	1092	0.299	7891
Netherlands	5	1	26933	8.2	77.83	11524	0.932	16436
New Zealand	1	1	17143	7.3	76.73	3012	0.240	4188
Norway	8	2	27416	7.6	78.09	1371	0.285	7168
Portugal	12	2	14324	7.9	75.39	2013	0.299	6970
Spain	12	2	17554	7.4	78.55	1542	0.299	7441
Sweden	58	3	22161	8.5	78.83	1127	0.238	6007
UK	59	4	20241	6.9	76.91	4607	0.135	-562
USA	21	2	29919	14.0	76.17	2730	0.213	3656
Sum / mean	464	16	22473	9.05	77.58	8678	0.677	11632

* Number of INMB estimates clustered within a country

** Number of studies providing INMB estimates for that country

**** As INMB measurements within one country may stem from several studies with differential timing, the values in these columns represent means across the years for which INMB values are reported

\+ Source: OECD Health Data, 1999 [69]

Table 3.4 provides some characteristics and summary statistics of the countries included. As INMB measurements within one country may stem from several studies with differential timing, values elicited on GDP per capita, healthcare spending as percentage of GDP and life expectancy at birth are presented as mean values across the years for which INMB measures were reported. Again, there is considerable spread in mean INMB's across countries, ranging from £-552 (UK) to £24,296 (Canada), assuming λ= £30,000. Looking at the third column in both tables, we can see that most studies (14 out of 16) report data applicable to one geographic domain only. However, two studies (Jonsson et al., 1996; Jonsson et al., 1999) report data which applies to several geographic domains, resulting in studies and countries being cross-classified. Further descriptive statistics of response variables and explanatory variables tested within this pilot study are available from Tables 3.5 and 3.6 below.

Table 3.5: Descriptive statistics of continuous variables tested in the pilot

Variable (in 2009 £-Sterling)	Level	Obs.	missing	in %	Min	Max	Mean	Std. Dev.
incr_cost	response	464	0	0%	10.44	64259	8678	11231
incr_effect	response	464	0	0%	.00885	5.4	0.677	0.944
inmb	response	464	0	0%	-42959	140518	11632	23407
age	Data	448	16	3.45%	30	70	57.06	7.57
gender	Data	449	15	3.23%	0	1	32.03%	40.20%
tcl	Data	447	17	3.66%	5.4	9.9	6.66	0.942
dr_cost	Data	464	0	0%	0	10%	3.55%	1.77%
dr_effect	Data	464	0	0%	0	10%	3.17%	2.10%
GDP/capita	Country	16	0	0%	13071	31653	22473	2729
total_hcs	Country	16	0	0%	6.8	14.1	9.03	1.62
public_hcs	Country	16	0	0%	4.7	8.5	6.72	0.85
le_birth	Country	16	0	0%	75.06	79.20	77.57	0.92

The mean age across all patient subgroups considered in the studies included was 57.06 (SD 7.57), and 32.03% (SD: 40.2%) of all patients considered were female. The overall mean TCL was 6.66 mmol/L (SD: 0.942), and the majority of patients already experienced at least one CVD event in the past (65.73%). The most common statin under assessment was simvastatin (57.76%), followed by pravastatin (19.83%) and lovastatin (12.28%). Moving on to methods applied in the studies, the mean discount rates were 3.55% for costs and 3.17% for effects

respectively. The time horizon was most commonly between 6 to 10 years (40.30%) or above 20 years (46.55%). Most studies were based on DAM (13 out of 16), and CVD risk reduction was the most common way to capture treatment effectiveness (11 out of 16 studies), whilst 5 studies measured the impact of statins on cholesterol levels and then extrapolated findings to life years saved using risk equations. Finally, 50% of all studies were industry funded, whilst the general funding source was unclear in 25% of the studies included in the dataset.

Table 3.6: Descriptive statistics of categorical variables tested in the pilot

Variable category	Level	Frequency	In %	cumulative
Intervention				
- Pravastatin	data	92	19.83%	19.83%
- Simvastatin		268	57.76%	77.59%
- Lovastatin		57	12.28%	89.87%
- Fluvastatin		1	0.22%	90.09%
- unclear		46	9.91%	100%
CHD history				
- no	data	117	25.22%	25.22%
- yes		305	65.73%	90.95%
- both		42	9.05%	100%
Time horizon				
- 0 to 5 years	data	43	9.27%	9.27%
- 6 to 10 years		187	40.30%	49.57%
- 11 to 20 years		18	3.88%	53.45%
- > 20 years		216	46.55%	100%
Funding				
- Government	Study	1	6.25%	6.25%
- Industry		8	50%	56.25%
- University		1	6.25%	62.50%
- Industry + Government		1	6.25%	68.75%
- No funding source		1	6.25%	75.00%
- unclear		4	25%	100%
General study design				
- IPD	Study	3	18.75%	18.75%
- DAM		13	81.25%	100%
Method of effect calculation				
- CVD risk reduction	Study	11	68.75%	68.75%
- Cholesterol reduction		5	31.25%	100%
Timig				
- 1992	Study	1	6.25%	6.25%
- 1993		1	6.25%	12.50%
- 1995		6	37.50%	50.00%
- 1996		4	25.00%	75.00%
- 1997		1	6.25%	81.25%
- 1998		2	12.50%	93.75%
- 1999		1	6.25%	100%

3.4.4. Results

Table 3.7 shows results for running variance components models without explanatory variables in a univariate framework with INMB as response variable; whilst Table 3.8 contains results of running the same models as bivariate models with ΔC and ΔE as a vector of response variables. Model 1 in each table represents an empty OLS regression model, where the intercept is simply the pooled mean INMB (ΔC, ΔE) across all studies and countries. Models 2a and 2b are variance components models which take into account 2-level hierarchical data structures with data clustered in countries (model 2.a) or studies (model 2.b) respectively. Model 3 is a cross-classified model which considers that data is grouped both in studies and countries simultaneously. Though theoretically irrelevant, MLwiN encounters problems if the discrepancy in the error variance of the response variables in the bivariate model is high, which is why ΔC was linearly transformed by dividing it by 100 (personal communication with R. Pillinger, CMM Bristol).

Table 3.7: Univariate variance components models run on pilot study data

	Model 1 (single level OLS)	**Model 2.a** (2-level model with data clustered in countries only)	**Model 2.b** (2-level model with data clustered in studies only)	**Model 3** (Cross-classified model with data clustered in studies and countries)
Fixed part:				
Intercept (λ=£30.000)	£11629	£8224	£10209	£9581
Random part:				
σ^2_{u0j} (Study)	--	--	198216240	193738848
σ^2_{u0k} (Country)	--	46423764	--	341225
σ^2_{e0} (Data)	549323520	467745120	343737600	343655648
VPC - study VPC – country VPC - data	-- -- 100%	-- 9.03% 90.97%	36.57% -- 63.43%	36.03% 0.06% 63.91%
DIC	10654	10579	10436	10436

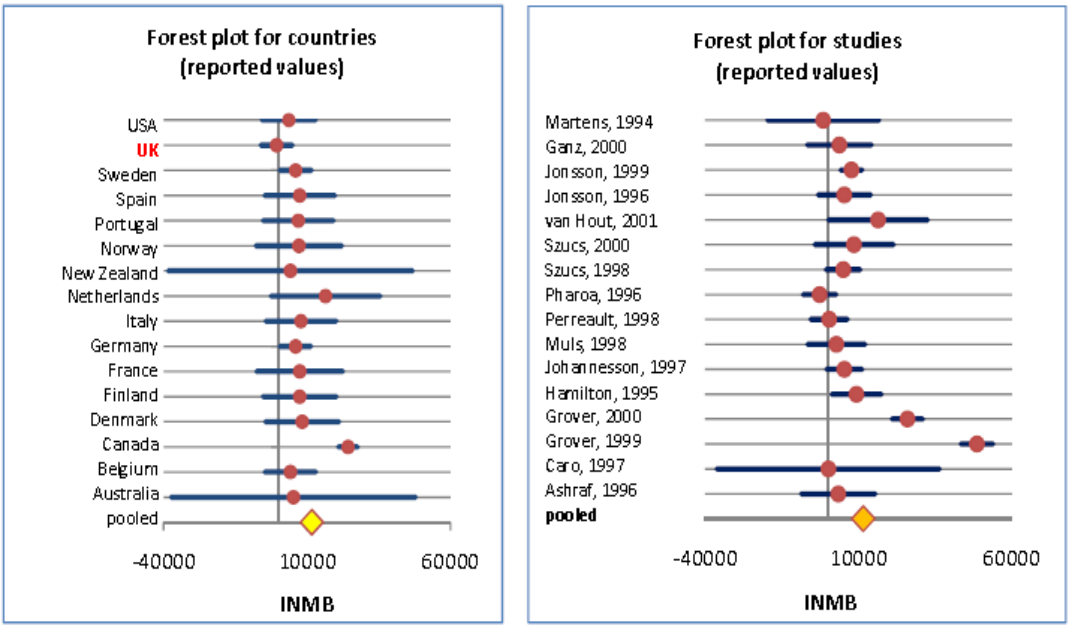
Table 3.8: Bivariate variance components models run on pilot study data

	Model 1 (single level OLS)		**Model 2.a** (2-level model with data clustered in countries)		**Model 2.b** (2-level model with data clustered in studies)		**Model 3** (Cross-classified model with data clustered in studies and countries)	
	ΔC/100	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**
Fixed part:								
Intercept	£87.66	0.685	£36.03	0.372	£73.34	0.510	£83.69	0.597
Random part:								
σ^2_{u0j} (Study)	--	--	--	--	10185	0.632	8629	0.558
σ^2_{u0k} (Country)	--	--	3289	0.320	--	--	43.37	0.005
σ^2_{e0} (Data)	13049	0.921	5824	0.575	5028	0.427	5057	0.429
VPC - study	--	--	--	--	66.95%	59.68%	62.85%	56.25%
VPC – country	--	--	36.09%	35.75%	--	--	0.32%	0.5%
VPC - data	100%	100%	63.91%	64.25%	33.05%	40.32%	36.83%	43.25%
DIC	6538		6130		5927		5928	

Model 2a in Table 3.7 shows that 9.03% of the total variation is attributable to differences between countries, and comparing models 1 and 2a using the deviance information criterion (DIC) (Browne, 2012) shows that model 2a is a better fit. However, moving on to model 2b, which also assumes a two-level hierarchy, but with INMB values clustered in studies rather than countries, we observe that there is much stronger variation between studies than between countries (36.57% as opposed to 9.03%), and that this model fits the data even better than model 2a. Moving on to the cross-classified model 3, where INMB values are clustered in both studies and countries, we see that the variation attributable to the country-level virtually disappears. However, the DIC indicates that model 3 is not a better fit than model 2b. Finally, moving on to the bivariate framework in Table 3.8, we observe a drastic improvement in the DIC statistic for all model specifications. As within the univariate framework, the study-level in model 2.b shows much more variability than the country-level in model 2.a. Also in accord with the univariate model in Table 3.7, we observe that the country-level variability virtually disappears in the cross-classified framework. Moreover, the VPC for studies drastically increases, with a variance component of 62.85% and 56.25% on study-level for ΔC and ΔE respectively.

Figure 3.14 confirms the findings reported in Table 3.7. The two forest plots provide mean INMB values (assuming λ=£30,000) and their respective 95% confidence intervals. The pooled estimate presented at the bottom of Figure 3.14 is the result obtained from the empty OLS-regression model in Table 3.7. The figure shows considerably more variation when sorting the data by studies rather than countries.

Figure 3.14: Variation in INMB values on country and study-level



Finally, a set of explanatory variables was tested in each model (Tables 3.9 and 3.10). All models include three variables on the data-level (mean pre-treatment total cholesterol level (TCL), age and percentage of females in the sample), and two variables on the country-level (GDP per capita and total life expectancy at birth). Although a range of study-level variables were tested (e.g. general study design, timing, funding source, or the method of effect calculation), this pilot study did not result in a set of study-level covariates which captures some of the variation across the studies included in the dataset.

Table 3.9: Univariate random intercepts models run on pilot study data

	Model 1 (single level OLS)	**Model 2.a** (2-level model with data clustered in countries)	**Model 2.b** (2-level model with data clustered in studies)	**Model 3** (Cross-classified model with data clustered in studies and countries)
Fixed part:				
Intercept	£12107	£9762	£8396	£8066
TCL	-1145	-1675*	4013***	4031***
(SE)	(1139)	(1137)	(1119)	(1114)
Age	-1045***	-807***	-974***	-976***
(SE)	(138)	(138)	(116)	(118)
% women	-11159***	-13346***	-13876***	-13904***
(SE)	(2587)	(2509)	(2112)	(2106)
GDP	1.48***	1.63***	0.12	0.13
(SE)	(0.46)	(0.66)	(0.45)	(0.46)
Life exp. at birth	5358***	648	364	356
(SE)	(1155)	(2034)	(1278)	(1285)
Random part:				
σ^2_{u0j} (Study)	--		210684928	279913696
σ^2_{u0k} (Country)	--	55614192		413570
σ^2_{e0} (Data)	449880704	400137568	269516096	273600160
VPC - study	--	--	43.87%	50.05%
VPC – country	--	12.2%	--	0.07%
VPC - data	--	87.80%	56.13%	49.39%
DIC	10175	10141	9986	9985

* Significant at the 10%-level
** Significant at the 5%-level
*** Significant at the 1%-level

Table 3.10: Bivariate random intercepts models run on pilot study data

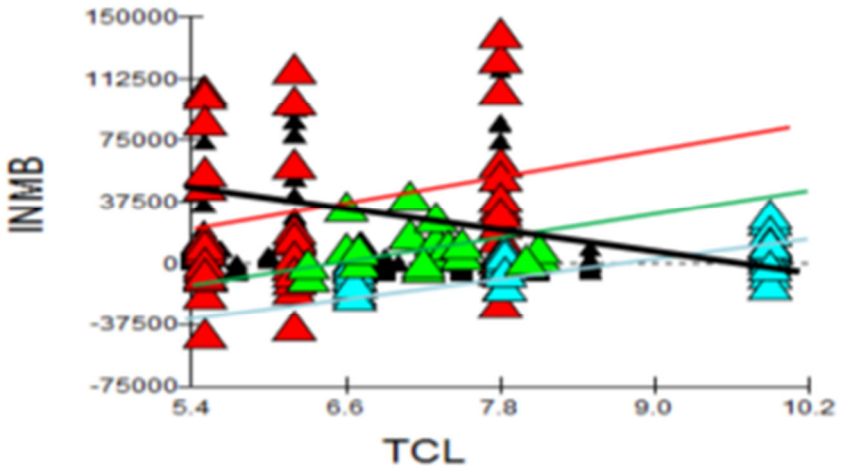
	Model 1 (single level OLS)		**Model 2.a** (2-level model with data clustered in countries)		**Model 2.b** (2-level model with data clustered in studies)		**Model 3** (Cross-classified model with data clustered in studies and countries)	
	ΔC/100	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**
Fixed part:								
Intercept	£87.75	0.685	£52.18	£0.482	£66.44	0.451	£67.75	0.465
TCL	-7.63*	-0.063	-11.68***	-0.091***	-8.50***	0.095***	-8.70***	0.097***
(SE)	(4.26)	(0.041)	(3.15)	(0.035)	(3.16)	(0.034)	(3.20)	(0.034)
Age	-9.70***	-0.067***	-7.40***	-0.052***	-7.82***	-0.058***	-7.77***	-0.058***
(SE)	(0.51)	(0.005)	(0.39)	(0.004)	(0.33)	(0.004)	(0.33)	(0.003)
% women	58.77***	-0.176*	35.69***	-0.328***	29.16***	-0.377***	28.97***	-0.376***
(SE)	(9.70)	(0.093)	(7.07)	(0.079)	(5.99)	(0.063)	(5.92)	(0.063)
GDP	0.01***	0.000	0.009***	0.000	-0.001	0.000	-0.001	0.000
(SE)	(0.002)	(0.000)	(0.003)	(0.000)	(0.001)	0.000	(0.001)	(0.000)
Life exp. at birth	12.77***	0.220	11.29	0.311	-1.08	0.004	-1.02	0.001
(SE)	(3.75)	(0.042)	(10.65)	(0.131)	(3.61)	(0.038)	(3.84)	(0.041)
Random part:								
σ^2_{u0j} (Study)	--	--	--	--	6319	0.557	6859	0.628
σ^2_{u0k} (Country)	--	--	3503	0.502	--	--	12.24	0.003
σ^2_{e0} (Data)	6375	0.592	3082	0.391	2181	0.246	2174	0.245
VPC - study	--	--	--	--	74.34%	69.37%	75.82%	71.69%
VPC – country	--	--	53.19%	56.22%	--	--	0.14%	0.35%
VPC - data	100%	100%	46.81%	43.78%	25.66%	30.63%	24.04%	27.96%
DIC	6148.773		5706.022		5320.925		5321.919	

* Significant at the 10%-level
** Significant at the 5%-level
*** Significant at the 1%-level

As TCL is a risk factor for CVD, and as statins reduce TCL, one would expect a positive relationship between TCL and INMB as well as TCL and ΔE; and a negative relationship between TCL and ΔC due to cost of future CVD events avoided. For age, prediction is less straightforward. One may expect a positive relationship between age and INMB as older people have a higher risk of a CVD event. However, this relationship may also be negative beyond a certain age, as older patients may have less to gain from cholesterol reduction (Ward et al., 2007). The effect of gender is clearer, as men are generally at higher risk and therefore likely to benefit more from statin therapy, especially in primary prevention (Ward et al., 2007). A positive relationship between the cost-effectiveness of statins and GDP per capita was hypothesized, as a higher level of economic attainment may correlate with higher future healthcare costs potentially avoided through statin therapy. Likewise, a positive relationship between INMB and life expectancy at birth was expected, as avoiding a CVD event in societies with a higher life expectancy leads to more life years saved.

We can see from Table 3.9 and 3.10 that the prediction for age and gender has been confirmed by each model. However, Models 1 and 2a in Table 3.9 show a non-significant negative relationship between the INMB for statins and TCL. Accordingly, the relationship between TCL and ΔE is negative for the bivariate versions of models 1 and 2.a in Table 3.10, and this negative relationship is even highly significant in model 2.a. The reason may be that, within the data, TCL has both an effect within, as well as between studies. Whilst the relationship between INMB and TCL may be positive within each study, it may be negative between studies. This could be the case, for instance, as studies which focus on high risk groups generally report lower INMB-values; although, within that study, the anticipated positive relationship between TCL and INMB still applies. Figure 3.15 below illustrates what happens within and between studies by plotting INMB against TCL and highlighting the predictions for some of the studies in the dataset. Whilst the between study regression line is negative, the positive within study effect still applies. In conclusion, using a model which does not acknowledge that data is clustered in studies (models 1 and 2a) may simply lead to wrong inferences as these models do not make explicit the distinction between within and between study effects (Steele, 2008; Rasbash et al., 2009).

Figure 3.15: Within-study and between-study effect of TCL on INMB



Coloured lines: **Individual study regression lines** (Within Group Effect)
Black line: **Overall regression line** (Between Group Effect)

Another interesting finding relates to the country-level variables in the dataset. Using a single level OLS, one would infer that both GDP per capita and life expectancy at birth are positive and highly significant for INMB and ΔC, which accords expectations. However, in models 2b, and 3, neither GDP nor life expectancy at birth is significant. The reason may be that treating country-level variables as if they referred to the data-level spuriously inflates the amount of information they provide, and hence, overestimates precision (e.g. Steele, 2008). To obtain correct standard errors, one therefore needs to take into account the correct structure of the data.

Finally, it can be seen that after covariate adjustment on data and country-level, about 50% of the overall variation in the data refers to the study-level in the univariate model 3 and country-level variation remains negligible. Even higher study-level variation is observed in the bivariate cross-classified model, with 75.8% and 71.7% of the total variation attributable to differences between studies. Also, this model provides further insights into the differential impact of covariates on ΔC and ΔE. For instance, within models 2.b and 3, TCL is negatively related with ΔC, whilst the relationship to ΔE is positive. This accords expectations, as people with higher cholesterol levels have generally more to

gain from statin therapy which reduces ΔC through potential cost avoided and also increases life years saved. However, the DIC statistic does not improve between model 2b and 3 both in the univariate and bivariate framework.

3.4.5. Discussion

The primary aim of this pilot was to inform the design and execution of the main empirical analysis, which is reported in Chapter 5. The three main objectives of this pilot were i) to test the MLMs developed in this chapter, ii) to obtain feedback from the wider health economists and multilevel modeller's communities, and iii) to inform the design of a data abstraction form which is used in a systematic literature review reported in the subsequent Chapter 4. Accordingly, this discussion starts off with general issues regarding the performance of the MLMs tested in this pilot study. Thereafter, the discussion addresses issues raised by participants of various conferences where this work was presented. This feedback proofed particularly helpful for the design and execution of the main empirical analysis, which is reported in Chapter 5.

3.4.5.1. Discussing the results of this pilot study

The analysis within this pilot study seems to confirm three key points. First, ignoring the clustering which occurs naturally when integrating secondary cost-effectiveness data, induces the risk of overestimated precision and, potentially, wrong inferences about factors causing variability in measures of cost-effectiveness (Steele, 2008; Rasbash et al., 2009, Hox, 2010). This can be seen by comparing the single level specification (model 1) with the more sophisticated multilevel models. Secondly, when the aim is to examine the exchangeability assumption between geographic domains by assessing country-level covariates, one cannot simply ignore the grouping which occurs naturally on study-level. This gets apparent from a comparison of models 2a and 2b. Finally, multinational

studies provide data on more than one geographic domain, which causes the strict hierarchy between studies and countries to break down. The cross-classified model allows acknowledging that data is clustered in studies and countries simultaneously. However, the results of the two-level hierarchical model with measures of cost-effectiveness clustered in studies (model 2b) are similar to those of the cross-classified model 3, where negligible country-level variation and no improvement in the DIC statistic were observed. This indicates that the more elaborated cross-classified model is not an improvement in fit. Hence, particular emphasis needs to be placed on this matter to inform appropriate MLM structures for the main empirical analysis.

There may be several potential causes for the observation that the cross-classified model did not perform better than the two-level hierarchical model with data clustered in studies only. Some of those causes may be related to the pilot study data itself, others however, concern the assumptions made to fit a cross-classified model and may therefore lead to alternative model structures to be tested in the main empirical analysis. The key question is whether the explicit modelling of a country-level itself may proof redundant as differences between studies constitute an overriding source of variability in measures of cost-effectiveness, or whether the issue of cross-classification and related assumptions about (in-)dependencies within the data may be responsible for the fact that model 3 failed to be an improvement in fit.

A reason to defend the cross-classified model relates to the identification and use of appropriate covariates, especially on study-level. Whilst this pilot aimed to test the feasibility and potential of the analytical framework, there is considerable work to be done to find this set of covariates to control for variability in measures of cost-effectiveness. This holds especially true for study-level covariates, as the pilot study clearly showed that this is a dominating source of variability in measures of cost-effectiveness. Furthermore, within the multilevel framework, controlling for variability on data and study-level may disclose further variability between countries, and hence, may constitute a pre-requisite for making correct inferences for higher-level covariates (e.g. Sculpher et al., 2004; Steele, 2008; Hox, 2010). Therefore, within the main empirical

exercise, a much larger set of potential covariates on all levels may be tested, drawn from a long list of variability factors published in the relevant literature (Sculpher et al., 2004; Goeree et al., 2007). However, operationalizing these variability factors is not straightforward. There are challenges around defining, measuring and selecting covariates. These issues are considered in far more detail in the subsequent Chapter 4, which is concerned with a systematic literature review and data abstraction exercise to generate a dataset on the cost-effectiveness of statins in the primary and secondary prevention of CVD.

Secondly, as shown in Table 3.3, 14 out of 16 studies included in the pilot study are strictly hierarchical as they provide data on one geographic domain only. Only two studies (Jonsson et al., 1996; Jonsson et al., 1999) introduce the problem of cross-classification between studies and countries. As these two studies provide 127 data points (27.4% of the whole dataset), this provides strong justification for not simply dropping these studies and loosing valuable data. A second reason to defend model 3 originates from the two included cross-classified studies themselves; though it also discloses a problem which needs to be addressed within the main empirical analysis. Having a closer look at these two studies (Jonsson et al., 1996; Jonsson et al., 1999) we learn that both originate from the same country, and that they use the same set of effectiveness and resource use data for all countries. Hence, adaptation of the cost-effectiveness results to each jurisdiction was achieved only through the use of country-specific unit cost estimates. It is not uncommon in economic evaluations, both trial based and model based, to apply data collected in one country to other locations of interest, without appropriately recognizing or exploring issues of transferability (Barbieri et al., 2005). It is suggested that this may explain the lack of additional geographic variation found in the cross-classified model. One way to address this problem in the main empirical analysis may be through a categorical variable capturing the degree of '*context specificity*' of measures of cost-effectiveness, and the author identified a potentially appropriate system to classify data which has been previously developed by Barbieri et al. (2005). Again, further details on this matter are obtainable from the subsequent Chapter 4 as well as Section 5.2 of the empirical exercise, which is concerned with covariate adjustment on data and study-level.

However, if the same problem relating to the two affected studies in the pilot study dataset generally applies to multinational economic evaluation studies, this may cast into doubt key assumptions necessary to fit the cross-classified model. Precisely, the cross-classified model assumes dependency of data within studies and also within countries, whilst independence is assumed between studies and between countries represented in the dataset. The two multinational studies included in the pilot both originate from the same country, and adaptation of cost-effectiveness results to each jurisdiction was achieved only through the use of country-specific unit cost estimates; whilst all other input parameters were transferred from the primary target country. This practice, which has not been without criticism within the health economics literature (Barbieri et al., 2005) may lead to drastically underestimated variability in measures of cost-effectiveness between countries included in multinational studies, hence, casting into doubt the independence assumption of data for the countries modelled in those studies. This data may then *'infect'* individual country-parameters in the cross-classified model as it is assigned to their respective target domains. In other words, measurements from multinational studies are being spread across the countries which they refer to and independence of data is assumed between those countries. As this data is, in fact, not independent, this may cause drastically underestimated country-level variability, and therefore redundancy of the country-level in the cross-classified model.

In conclusion, an alternative model structure may be necessary which does not assume independence of data from multinational studies on country-level but still allows utilizing this data for the purposes of assessing variability factors on data and study-level. As mentioned, the cross-classified model assigns data from both single-country-studies and multinational studies to their respective study on study-level, hence assuming that this data is dependent within studies and independent between studies. Exactly the same is assumed on country-level though this assumption may not hold for multinational studies. Instead of simply dropping the affected data points which may result in a considerable loss of valuable data, the data from multinational studies may be '*pooled*' in a distinct group on country-level, thereby removing its influence on other country-parameters. Obviously, as some countries in the dataset are only considered

within those multinational studies, this would mean to lose some parameters on country-level. However, what is gained is the chance to analyse the full dataset with all studies originally includable in this exercise whilst still obtaining '*clean*' country-level estimates from all the other studies in the dataset. Secondly, the problem of cross-classification in the data does no longer exist so that a strictly hierarchical three-level model with data being clustered in studies and studies being clustered in countries (as shown in equation (11) in Section 3.1.2.2), applies. However, most importantly this strategy addresses what has been discussed in the literature before (Ramsey et al., 2005; Barbieri et al., 2005), namely that multinational studies might not appropriately reflect country-level variability, which may also be the cause for (severely) underestimated country-level variation observed in the cross-classified model in this pilot study.

In conclusion, the main empirical analysis, which is reported in Chapter 5, starts off with an exercise on the appropriate MLM structure for analysing the data obtained from the systematic literature review and data abstraction exercise. One key aspect of this exercise is to investigate the appropriateness of the independency assumption of multinational study data on country-level, which leads to a direct comparison of the cross-classified model, and an alternative three-level hierarchical model structure which clusters multinational study data in a separate group on country-level.

3.4.5.2. Feedback from presenting this pilot study to a wider health economics and multilevel modelling audience

Next to testing the MLMs theoretically developed in this chapter, a second objective of this pilot study was to generate feedback on the proposed method of secondary data integration from published economic evaluation studies to analyse factors causing variability in international cost-effectiveness data from a wider health economics and multilevel modelling audience. For this reason, the pilot study was presented at a number of seminars and conferences. The remainder of this section discusses the invaluable feedback obtained.

One of the most valuable comments received relates to the above discussion regarding the appropriate MLM structure. In particular, a three-level hierarchical model was proposed instead of the cross-classified data-structure because of the negligible country-level variation in the cross-classified model and the suspicion that this might be due to inappropriate assumptions regarding the independence of data from multinational studies on country-level. At the time of carrying out the pilot study, the author was not clear on how to fit a three-level hierarchical model to the data without dropping measurements from multinational studies. However, before completing the systematic literature review and data abstraction exercise for the main empirical analysis, alternative model structures were developed which acknowledge that data from multinational studies is not independent on country-level. These data structures were subsequently tested and their performance compared to the cross-classified framework. Results are reported in great detail in Section 5.1 of the empirical chapter.

The impact of shrinkage on study and country parameters may constitute an issue for the integration of secondary data from international economic evaluation studies. Shrinkage partly depends on the number of observations within a particular group, so that the '*gravity*' of a study increases with the number of cost-effectiveness estimates provided to the overall dataset. If the unit of observation is an individual within a randomized controlled trial, this makes perfect sense as the weight of a study is proportional to the number of patients under assessment. However, as the data for this exercise stems in part from decision analytic modelling studies, the respective number of cost-effectiveness estimates may only reflect the rigor with which subgroup and sensitivity analyses have been conducted and reported within a particular study. As a result, MLMs may bias towards those studies which report in greater detail on the results of subgroup or sensitivity analysis, irrespective of the quality of the particular study or the strength of evidence underlying the input data for decision analytic models. This important issue is addressed in great depth in Section 5.1.5 of the empirical chapter.

As a further development of the MLMs tested within the pilot study, it was suggested to consider fitting random slopes to acknowledge that the relationship

between measures of cost-effectiveness and explanatory variables may be different for different studies (or countries) in the dataset. Moreover, fitting random slopes to the data allows modelling the variation in international cost-effectiveness data directly as a function of explanatory variables. At the time of carrying out the pilot study, the author did not yet succeed in fitting random slopes models, however, this has been subsequently achieved within the main empirical analysis and details are reported in Section 5.4 of the empirical chapter.

Finally, a number of comments received relate to the impact of particular variability factors on the results of economic evaluation studies. For instance, the impact of 'time' on the cost and efficiency of the intervention, differential discount rates, and further methodological characteristics of studies under assessment were mentioned. For this reason, the following Chapter 4, which is concerned with a systematic literature review and data abstraction exercise, reports on the development of a data abstraction form, which is based on a long list of potential variability factors as obtained from the relevant economic evaluation literature (Sculpher et al., 2004; Goeree et al., 2007). A key objective of the main empirical analysis in Chapter 5 is then to test these variability factors within the MLM framework and to ascertain a set of covariates which controls for part of the variability in measures of cost-effectiveness on each level of the model hierarchy.

Further comments received from researchers involved with economic evaluation in health, MLM, or both, which are not particularly related to the design and execution of the main empirical analysis, may be discussed later in the overall discussion section in Chapter 6. The following chapter reports on a systematic literature review and data abstraction exercise to populate a dataset for the main empirical analysis. As the pilot study before, this exercise focusses on the cost-effectiveness of statins in the primary and secondary prevention of CVD.

4. Systematic literature review, data abstraction and 'genealogy' study

The previous chapter was concerned with the MLM methods relevant for the empirical exercise within this project. In Section 3.1 of Chapter 3, a number of MLM structures were theoretically developed, including a three-level hierarchical model with measures of cost-effectiveness clustered in studies and countries, as well as a two-level cross-classified model where the hierarchy between studies and countries breaks down due to data from multinational studies. Then, in Section 3.2., the dependent variable, or a vector thereof, was specified. INMBs were identified as the appropriate response variable in a univariate model, and the INMB statistic was decomposed into its stochastic components ΔC and ΔE in a bivariate MLM. Finally, the models developed in Chapter 3 were tested in a pilot study utilizing data from 16 international economic evaluation studies on the cost-effectiveness of statins in the primary and secondary prevention of CVD, and this was reported and discussed with particular emphasis on the design of the main empirical exercise in Section 3.4 of the previous chapter.

In this chapter, the primary aim is to develop a dataset to carry out this empirical exercise. Specifically, the chapter begins with a systematic literature review on the cost-effectiveness of statins in the primary and secondary prevention of CVD. This intervention was chosen as it has been extensively researched in the past, meaning that a sufficient number of includable studies and geographic locations is hypothesized to be present in the data to justify the assumption of random parameters on study and country-level (Snijders, 2005). This assumption of random parameters on higher levels is crucial for fitting multilevel models (Snijders, 2005).

After carrying out a systematic literature review, the second task to populate a dataset for the main empirical analysis is to develop a data abstraction form to collect data on the response variables as well as covariates to be tested on data and study-level. This process starts with reviewing the relevant literature to obtain a long list of factors potentially causing variability in international cost-

effectiveness data. Subsequently, variability factors are operationalized and a data abstraction form is developed. This form, which was also extensively tested and improved upon during the pilot study reported in the previous chapter, is used to obtain data from the studies includable in this empirical exercise. Details on this process are reported in Section 4.2.

Finally, the systematic literature review, which resulted in 67 studies includable in this empirical exercise, showed that there may be a number of potential relationships between published economic evaluation studies. Such relationships may relate to common authorship, partly or fully recycled models, or the use of identical data sources, to name a few. As a result, the independency assumption of data between studies may be violated analogously to the problems introduced by multinational study data on country-level as discussed within the pilot study. For this reason, an explorative analysis of the '*genealogy*' of economic evaluation studies is reported to either inform alternative data structures, or at least, to derive further covariates encoding potential relationships between economic evaluation studies. This genealogy study differs from the other two sections of this chapter as it addresses an important research question in its own right rather than being a pure prerequisite for the main empirical analysis. Results of this explorative exercise into the genealogy of economic evaluation studies are reported in Section 4.3 of this chapter.

4.1. Systematic literature review on the cost-effectiveness of statins for the primary and secondary prevention of CVD

The primary aim of this systematic review is to populate a dataset for the empirical analysis. As the MLM methodology applied in the empirical exercise relies on the assumption of random parameters on study and country-level (e.g. Spiegelhalter et al., 2004), decisions were made in order to achieve sufficient numbers of higher-level units for this assumption to hold. First, with statins for

the primary and secondary prevention of CVD, the author chose an extensively studied intervention area for the purposes of this project. Secondly, a highly sensitive search strategy was developed to ensure sufficient numbers of includable studies applicable to as many countries as possible. However, this does not mean that any study concerned with the cost-effectiveness of statins in the primary and secondary prevention of CVD is includable in this systematic review. Rather, a number of strict inclusion and exclusion criteria were defined to ensure that only studies providing information in a format that is suitable for secondary data integration as laid out in the MLM methods Chapter 3 would enter this exercise. The following Section 4.1.1 outlines the search strategy as well as inclusion and exclusion criteria for this systematic review. Subsequently, search results are presented in Section 4.1.2. The second part of this chapter is concerned with the development of a data abstraction form to populate a dataset for the main empirical analysis within this project. As part of this section, descriptive statistics of studies included in this exercise are reported. The final Section 4.3 of this chapter focusses on the genealogy of economic evaluation studies.

4.1.1. Review Methodology

To devise a highly sensitive search strategy which identifies a large number of includable studies on the cost-effectiveness of statins in the primary and secondary prevention of CVD, a number of existing systematic review papers on this intervention area were consulted to learn about respective search strategies applied. Probably one of the most rigorous reviews within this area was undertaken by Ward et al. (2007), who aimed to '*identify and evaluate studies exploring the cost-effectiveness of statins in primary and secondary prevention of CHD and CVD in the UK*.' Though the literature review in this thesis is not limited to a UK setting, some principles from Ward et al. (2007) were adapted and developed further for the purposes of this project.

Table 4.1: Databases searched

Search Engine	Databases searched	Comments
Ovid	- British Nursing Index - Medline	- Search performed April 15th, 2011
SCOPUS	- Embase - Medline - Science Direct	- Search performed April 16th, 2011 - Medline was not dropped from this search as the SCOPUS search engine differs from Ovid which may have led to differences in search results
Ebsco Host	- Academic Search Complete - Busines Source Premier - CINAHL	- Search performed April 16th, 2011
HEED		- Search performed April 19th, 2011
CRD	- Database of Abstracts of Reviews of Effects (DARE) - Health Technology Assessment Database (HTA) - NHS Economic Evaluation Database (NHS EED)	- Search performed April 19th, 2011
Cochrane Library	- Cochrane Database of Systematic Reviews (CDSR) - Cochrane Central Register of Controlled Trials - Cochrane Methodology Register - Database of Abstracts of Reviews of Effects (DARE) - Health Technology Assessment Database (HTA) - NHS Economic Evaluation Database (NHS EED)	- Search performed April 21th, 2011 - Cochrane database of systematic reviews was searched as review papers were hand searched for relevant references - DARE, HTA and NHS EED were included as the Cochrane search engine differs from CRD, which may have led to differences in search results
Pubmed		- Search performed April 19th, 2011 - Though Medline and Pubmed are essentially identical, a separate search was performed in PubMed as a different search engine applies which may have led to differences in search results
Web of knowledge	- Web of Science - Biosis Reviews	- Search performed April 19th, 2011
JStor		- Search performed April 19th, 2011 - Search results subsequently dropped from further assessment
Wiley		- Search performed April 20th, 2011 - Search results subsequently dropped from further assessment

Literature searches were performed between April 15th and April 21st 2011 using the databases listed in Table 4.1 above. Some databases may have been

searched more than once as they form part of several search engines used. However, as search engines may differ, so may search results, which is why it was decided not to drop the respective search results but rather to deal with potential duplicates after exporting results to the reference managing software RefWorks. Two databases (Jstor and Wiley) were initially searched but search results were subsequently dropped from further analysis (more details on this matter are available in the next section). For Medline searched via OVID, the same search strategy as developed by Ward et al. (2007) was applied, though results were not limited by geographic setting or publication year. For other search engines, the search strategies are reported in Appendices 4.1 to 4.10.

As mentioned above, no country or language restriction was initially applied to the literature search as it was aimed to represent as many geographic domains as possible in the resulting dataset. However, at a later stage, it was decided to drop papers which were not written in either English or German language as resources to translate studies written in other languages were not available. As with geographic origin, no time restriction was placed on the literature search as it was aimed to reflect the whole continuum of statin related cost-effectiveness literature in the dataset and as '*timing*' may be one potential explanatory within the main empirical analysis.

Only studies following an incremental approach were includable in this review exercise, and as comparator technology, only i) '*doing nothing,* ii) '*other statin*' or iii) '*same statin in different dosage'* were considered. Studies which compared the intervention to any combination of statins and other technologies (e.g. statin vs. statin plus dietary advice, statin vs. statin plus ACC, or statin vs. statin plus antihypertensive drug), were not includable as differences in combinations of therapies potentially introduce further variability in measures of cost-effectiveness which cannot be easily controlled for through explanatory variables within the MLM framework.

Even If the appropriate intervention and comparator technologies were considered, studies were only includable if the ICER or INMB statistic was

decomposed into its stochastic components ΔC and ΔE. Otherwise, it may not be possible to either re-combine ΔC and ΔE to INMBs with a common threshold value λ for the univariate multilevel model or to run a bivariate model with ΔC and ΔE as a vector of response variables. Further, studies were only includable if effectiveness was measured either in life years saved (LYS) or quality adjusted life years (QALYs). Hence, studies reporting intermediate outcomes such as cholesterol reduction or change in cardiovascular events due to statin prevention were not includable in this review exercise.

Studies utilizing individual patient data and decision analytic modelling studies were suitable for this exercise. In addition, studies which adapted published results to other geographic settings were includable if results were not simply 'currency adjusted copies' of the original data. Finally, only adult populations were considered for this systematic review.

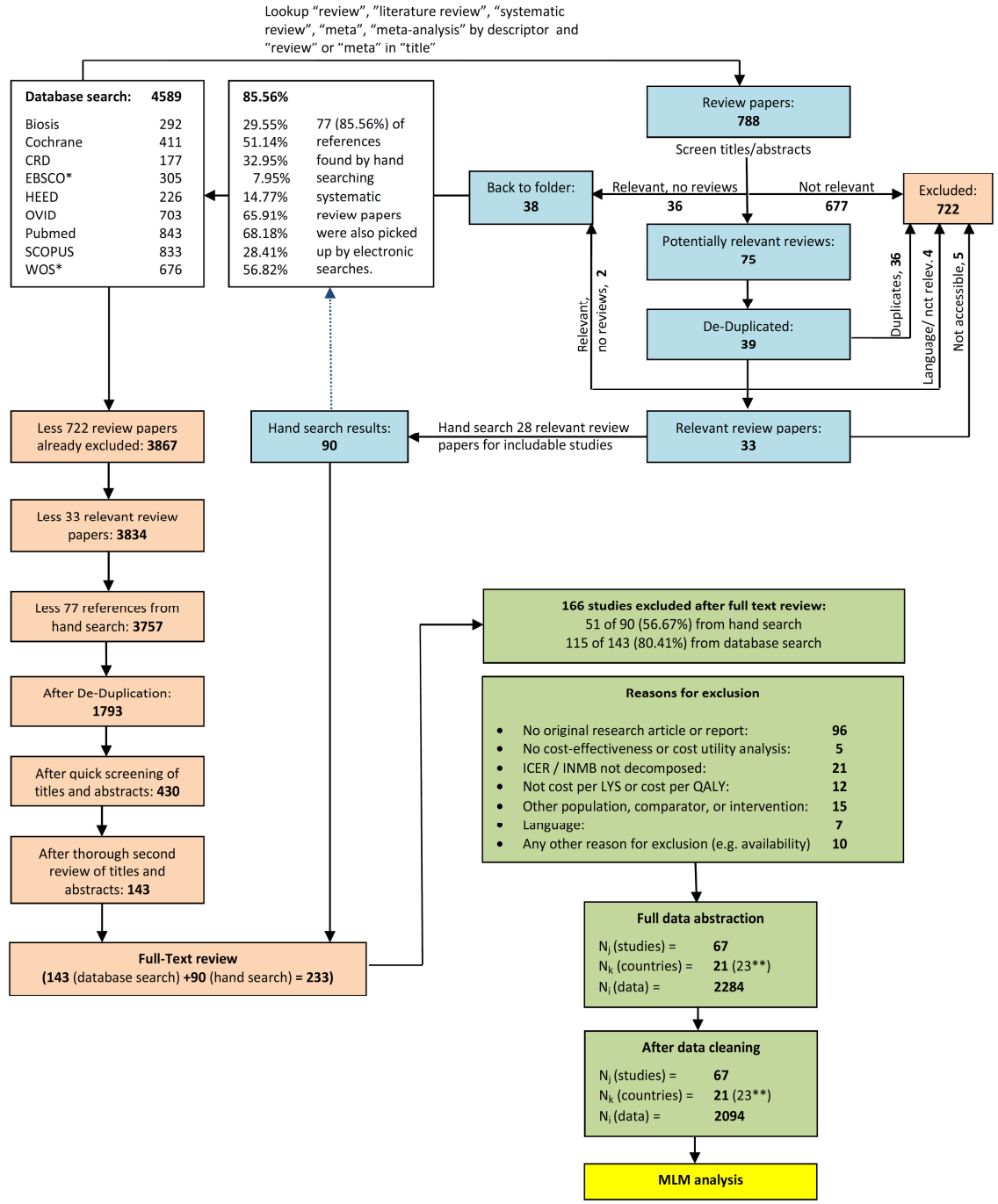
In contrast to other systematic reviews, this exercise did not a priori define a minimum set of methodological requirements. There were several reasons for this. First, methods standards may differ between geographic domains, so that it may be difficult to define a particular set of minimum requirements without biasing search results towards particular settings. Secondly, the main empirical exercise aims to control for variability factors on different levels of the data hierarchy through the assumption of conditional independence (e.g. Drummond et al., 2009). Hence, differences in study methods may be partly controlled for through the inclusion of appropriate covariates on data and study-level. Finally, it was aimed to explicitly rate study quality through the use of 'QHES', which is a validated quality checklist for economic evaluation studies in health (Ofman et al., 2003). The results of this quality assessment may then be used as a further explanatory in the MLM framework. More details on the use of QHES within this empirical exercise are available from Section 4.2 of this chapter, which is concerned with the development of an abstraction form to populate a dataset for the main empirical exercise of this project.

4.1.2. Search Results

Figure 4.1 below shows the search algorithm applied in this systematic literature review. After searching individual databases, 4589 search results were exported to the reference managing software 'RefWorks'. Though review papers, meta-analyses, opinion pieces etc. are not includable in the empirical, they were not a priori excluded from electronic searches. Rather, the RefWorks search facility was used after importing results to filter out relating references. This resulted in 788 hits. Secondly, titles and abstracts of these 788 references were screened and de-duplicated, resulting in a list of 33 highly relevant review papers or meta-analyses. Finally, 28 review papers which were accessible via Brunel University subscriptions or inter library loans, were hand searched for potentially relevant references of original research articles on the cost-effectiveness of statins in the primary and secondary prevention of CVD (references of these 28 papers are available from Appendix 4.11). This resulted in a list of 90 references.

One advantage of first hand searching systematic review papers for potentially relevant references is that one may use results to obtain an indication of *'sensitivity'* of the initial electronic search performed. If an electronic search picks up a high number of references also obtained from hand searching review papers, this search may be deemed as fairly sensitive. In total, 77 of 90 (85.56%) references obtained from hand searching relevant review papers were also picked up by electronic searches. However, some databases provided much better results than others, with the most successful searches performed in PubMed and OVID. Conversely, two databases (Jstor and Wiley) picked up less than 2% of references obtained from hand searching review papers. Hence, these databases may not focus on research in the relevant area, which is why the respective search results were dropped from further assessment and omitted in the search algorithm below.

Figure 4.1: Search Algorithm



* Without Medline

** England/Wales, Scotland, and UK as separate geographic entities

After screening out and hand searching relevant review papers, 3757 hits remained in the database. These search results were first de-duplicated, resulting in 1793 references. Secondly, titles and abstracts of these 1793 references were screened twice. A first screening round reduced references to 430, and a second - more thorough - screening of titles and abstracts reduced potentially relevant references to 143. Subsequently, a full text review was conducted of the 143 references from electronic searches plus the 90 references previously obtained from hand searching relevant systematic review papers, resulting in 233 references eligible for full text review. Of these 233 references, a further 166 were excluded. 96 papers were excluded as they did not constitute an original research article or report. Five papers did not perform a cost-effectiveness or cost-utility analysis. A further 21 papers did not decompose the ICER or INMB into its components ΔC and ΔE, and twelve papers did not report health outcomes in terms of LYS or QALYs. 15 papers focused on the wrong population, intervention or comparator and seven papers were written in any other language than English or German. Finally, ten papers were excluded for other reasons, most commonly the fact that access was neither provided through Brunel University subscriptions nor inter library loans from the British Library. As a result, 67 studies were eligible for this empirical exercise, of which 39 studies were initially obtained from hand searching systematic review papers for potentially relevant references, and a further 28 references from electronic searches. References of includable studies are available from Appendix 4.12.

The 67 studies obtained were subsequently abstracted to populate a dataset for the main empirical analysis, which is reported in Chapter 5. To do so, a data abstraction form needed to be developed and tested which ascertains information on the response variables as well as potential variability factors for measures of cost-effectiveness on data and study-level. The development of this data abstraction form is reported next in Section 4.2 of this chapter, which also includes some key descriptive statistics of the studies included in this systematic review exercise. Finally, Section 4.3 of this chapter is concerned with an exercise into the 'genealogy' of international economic evaluation studies on the cost-effectiveness of statins in the primary and secondary prevention of CVD.

4.2. Developing a data abstraction form and populating a dataset for the main empirical analysis

The previous section described how studies includable in the empirical exercise were identified through a systematic literature review on the cost-effectiveness of statins in the primary and secondary prevention of CVD. This section describes how a data abstraction form to populate a dataset for the empirical analysis was developed. In general, this exercise began with screening the relevant literature for potential variability factors for measures of cost-effectiveness to obtain a long list of potential variables on the abstraction form. Secondly, the variability factors mentioned in the literature were operationalized so that data could be abstracted in a format which is suitable for quantitative analysis in the MLM framework. The third step was to test the resulting data abstraction form within the pilot study which was previously reported in Chapter 3.4. Experiences from the pilot study were then used to improve upon the data abstraction form. In some instances that meant to change the scale of a continuous variable, or to add or delete categories of categorical variables. In other cases, experiences from the pilot study led to dropping variables either because studies failed to report the relevant data or because responses would not vary within and between studies in the dataset. Finally, the resulting abstraction form was used to obtain data from the 67 studies includable in this empirical exercise.

The following Section 4.2.1 reports on the development and use of the data abstraction form as it was outlined above. Subsequently, descriptive statistics of studies included in this empirical exercise are reported in Section 4.2.2. Finally, in Section 4.3, the data obtained is used within an empirical exercise on the 'genealogy' of economic evaluation studies on the cost-effectiveness of statins in the primary and secondary prevention of CVD.

4.2.1. Development and use of a data abstraction form

The development of a data abstraction form started off from a long list of factors potentially causing variability in measures of cost-effectiveness as previously discussed within the relevant literature. Two papers were of utmost relevance, namely Sculpher et al. (2004) and Goeree et al. (2007). Both papers systematically reviewed the economic evaluation literature in health to obtain factors which may be responsible for variation in international cost-effectiveness data. Sculpher et al. (2004) obtained a list of 27 variability factors relating to a) the patient, b) the clinician, c) the healthcare system or d) wider socioeconomic factors. Goeree et al. (2007) confirmed and updated the list of Sculpher et al. (2004). Their systematic literature review resulted in 77 unique variability factors based on characteristics of a) the patient, b) the disease, c) the provider, d) the healthcare system and e) methodology used in the analysis. The results of Goeree et al. (2007) constitute the *'long list'* of potential variability factors on which the development of the data abstraction form is based upon. It is acknowledged that the '*space of variability factors'* is generally unlimited, so that factors neither mentioned by Sculpher et al. (2004) nor by Goeree et al. (2007) may also be important. However, as their work is supposed to represent a comprehensive list of those factors which other health economists previously suspected to be responsible for variability in cost-effectiveness data, their results may be regarded as the most appropriate starting point for this analysis.

Though Sculpher et al. (2004) and Goeree et al. (2007) made an invaluable contribution to the field, the variability factors mentioned in their papers needed further refinement to be practicable within this data abstraction exercise. Some factors mentioned are somewhat '*fuzzy*', not measurable without further adjustment or simply not applicable to the intervention area of statins in the primary and secondary prevention of CVD. In addition, a number of variability factors specifically relate to variation between geographic domains. These factors are considered explicitly within Section 5.3 of the empirical chapter where country-level variability is assessed within the MLM framework. However, for the purposes of developing a data abstraction form for the studies included in this empirical exercise, country-level characteristics were not considered as the

relevant data may not be obtainable from the majority of studies but rather from alternative data sources like WHO, World Bank or OECD databases. In terms of the variability factors reviewed by Goeree et al. (2007), this relates specifically to the group of 'healthcare system characteristics' (see Figure 4.2 below).

Figure 4.2: Flow chart showing how the data abstraction form was developed

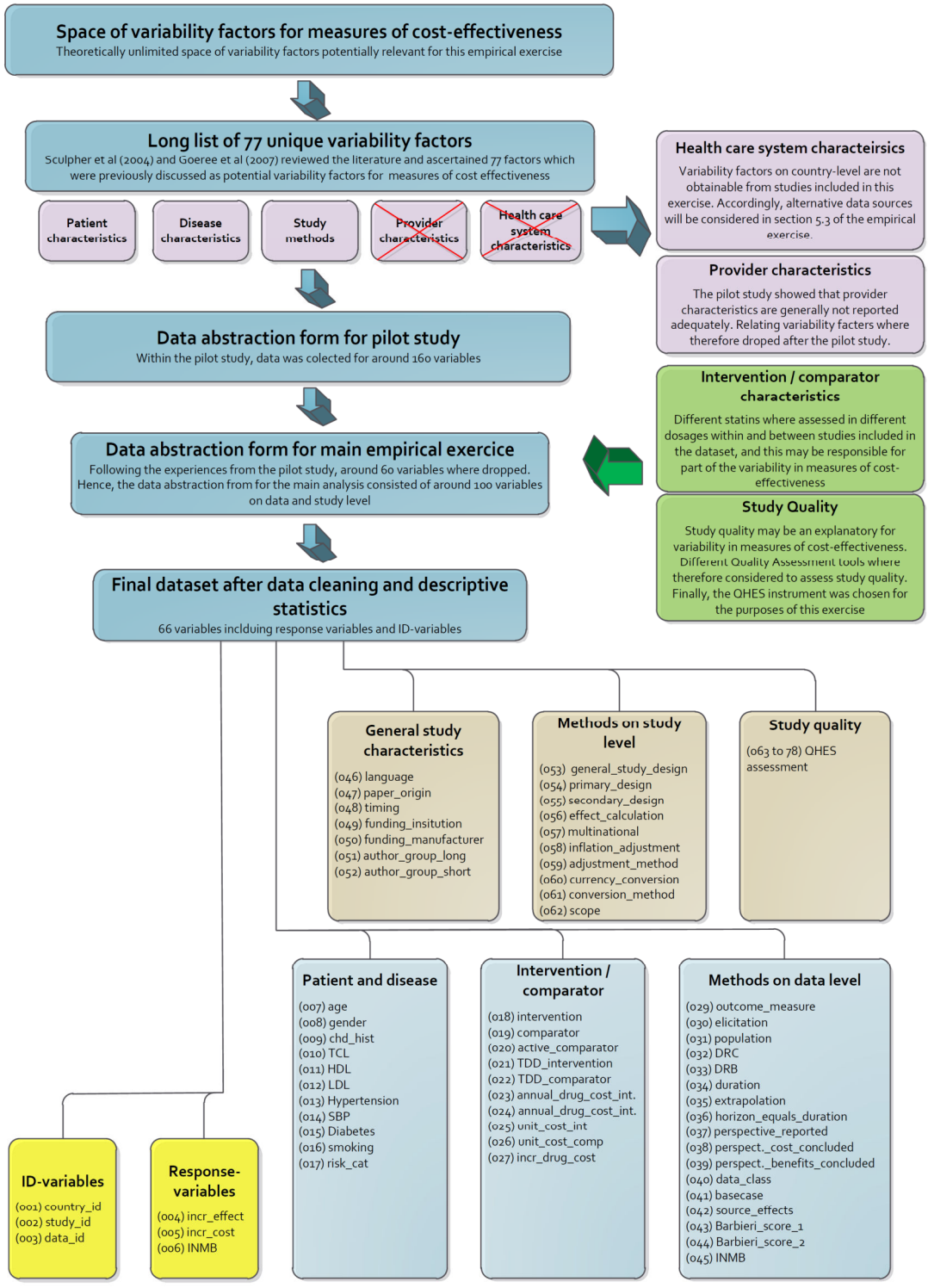


Table 4.2: Response variables, ID-variables and covariates on data-level

Variable name	Description	Level	Nature of variable
	Response variables		
Incr_effect	*incremental effectiveness of intervention*	*Dependent variable*	*continuous*
Incr_cost	*incremental cost of intervention in 2010 £-Sterling*	*Dependent variable*	*continuous*
INMB	*INMB in 2010 £-Sterling*	*Dependent variable*	*continuous*
	Level-IDs		
country_id	*Which country does the CE-estimate refer to?*	*Level 3 ID (country)*	*Level ID no covariate*
Study_id	*Which study does the CE-estimate refer to?*	*Level 2 ID (study)*	*Level ID no covariate*
Data_id	*Data-level identifier*	*Level 1 ID (data)*	*Level ID no covariate*
	Group 1a covariates: Patient and disease characteristics		
age_cat	*What was the age of the sub-population modelled*	*Level 1 (data)*	*ordered, categorical*
gender_cat	*What was the gender of the population (percentage of men in population)*	*Level 1 (data)*	*ordered, categorical*
CVD_hist	*What was the CHD related medical history*	*Level 1 (data)*	*unordered, categorical*
Tcl	*What was the total cholesterol level baseline*	*Level 1 (data)*	*Continuous*
Hdl	*What was the high density lipoprotein level at baseline*	*Level 1 (data)*	*Continuous*
Ldl	*What was the low density lipoprotein level*	*Level 1 (data)*	*Continuous*
Hypert	*What was the percentage of hypertensive people in the subsample*	*Level 1 (data)*	*Continuous*
Sbp	*What was the mean systolic blood pressure at baseline*	*Level 1 (data)*	*Continuous*
Diab	*What was the percentage of diabetic patients at baseline*	*Level 1 (data)*	*Continuous*
smokers	*What was the percentage of smokers at baseline*	*Level 1 (data)*	*Continuous*
risk_cat	*What was the risk category of the subsample*	*Level 1 (data)*	*ordered, categorical*
	Group 1b covariates: Intervention and comparator		
intervention	*What was the brand name of the intervention drug?*	*Level 1 (data)*	*unordered, categorical*
comparator	*What was the brand name of the comparator drug?*	*Level 1 (data)*	*unordered, categorical*
act_comp	*Was the comparator no active intervention?*	*Level 1 (data)*	*unordered, binary*
tdd_int	*What was the total daily dose of the intervention*	*Level 1 (data)*	*ordered, categorical*
tdd_comp	*What was the total daily dose of the comparator*	*Level 1 (data)*	*ordered, categorical*
Cost-int	*What are the annual drug cost of the intervention in 2010 £-Sterling*	*Level 1 (data)*	*Continuous*
Unitcost_int	*What was the unit cost of the intervention*	*Level 1 (data)*	*Continuous*
Cost_comp	*What are the annual drug cost of the comparator in 2010 £-Sterling*	*Level 1 (data)*	*Continuous*
Unitcost_comp	*What was the unit cost of the comparator*	*Level 1 (data)*	*Continuous*
Incr_cost	*What was the incremental annual drug cost of the intervention*	*Level 1 (data)*	*Continuous*
	Group 1c covariates: Methodological characteristics on data-level		
outc_measure	*How was health outcome reported in the study*	*Level 1 (data)*	*unordered, binary*
elicitation	*If QALYS were used, what was the method of preference elicitation?*	*Level 1 (data)*	*unordered, categorical*
population	*If QALYS were used, what do the utility values reflect (patient / population values)*	*Level 1 (data)*	*unordered, categorical*
DRC	*What was the discount rate on costs*	*Level 1 (data)*	*Continuous*
DRB	*What was the discount rate on benefits*	*Level 1 (data)*	*Continuous*
duration	*What was the treatment duration modelled*	*Level 1 (data)*	*ordered, categorical*
extrapol	*Was there any extrapolation beyond the latest follow up?*	*Level 1 (data)*	*unordered, binary*
horizon	*What was the time horizon?*	*Level 1 (data)*	*ordered, categorical*
hor_eq_dur	*Does the time horizon equal the treatment duration?*	*Level 1 (data)*	*unordered, binary*
Persp_rep	*What was the study perspective as reported by the authors of the article*	*Level 1 (data)*	*unordered, categorical*
Persp_cost_concl	*What was the study perspective on costs as concluded by the reviewer*	*Level 1 (data)*	*unordered, categorical*
Persp_ben_concl	*What was the study perspective on outcomes as concluded by the reviewer*	*Level 1 (data)*	*unordered, categorical*
data_class	*How was the datapoint classified*	*Level 1 (data)*	*unordered, categorical*
basecase	*Was the data point result of a base case or sensitivity analysis?*	*Level 1 (data)*	*unordered, binary*
source_effects	*From which source (trial, meta-analysis) was effectiveness data taken from*	*Level 1 (data)*	*unordered, categorical*
Barbieri_score_1	*How context specific is the CE estimate judged from the input parameters*	*Level 1 (data)*	*unordered, categorical*
Barbieri_score_2	*How context specific is the CE estimate judged from the input parameters*	*Level 1 (data)*	*unordered, categorical*

Table 4.3: Covariates on study-level

Variable name	Description	Level	Nature of variable
Group 2a covariates: General Study characteristics			
language	*In which language was the paper written?*	*Level 2 (study)*	*unordered, binary*
paper_origin	*In which country was the paper written (if authors from several jurisdictions were involved, where is the lead author based?*	*Level 2 (study)*	*unordered, categorical*
Timing	*What is the timing of the economic evaluation*	*Level 2 (study)*	*unordered, categorical*
fund_inst	*What was the primary source of funding (institution)*	*Level 2 (study)*	*unordered, categorical*
fund_man	*If funding source was private, which manufacturer was involved?*	*Level 2 (study)*	*unordered, categorical*
Author_group _long	*Variable which encodes relationships between published papers in terms of common authorship*	*Level 2 (study)*	*unordered, categorical*
Author_group _short	*Variable which encodes relationships between published papers in terms of common authorship*	*Level 2 (study)*	*unordered, categorical*
Group 2b covariates: Methodological characteristics on study-level			
gen_des	*What was the general study design?*	*Level 2 (study)*	*unordered, categorical*
prim_des	*If primary modelling, what was the specific study design?*	*Level 2 (study)*	*unordered, categorical*
sec_des	*If secondary modelling, what was the specific study design*	*Level 2 (study)*	*unordered, categorical*
effect_calc	*Method of effect calculation*	*Level 2 (study)*	*unordered, categorical*
multinational	*Was the study multinational*	*Level 2 (study)*	*unordered, binary*
infl_adj	*Were cost estimates in the model adjusted for inflation?*	*Level 2 (study)*	*unordered, categorical*
adj_method	*If cost estimates were adjusted for inflation, what was the adjustment method*	*Level 2 (study)*	*unordered, categorical*
cur_conv	*Was currency converted*	*Level 2 (study)*	*unordered, binary*
conv_method	*If currency was converted, what was the conversion method used by the authors?*	*Level 2 (study)*	*unordered, categorical*
scope	*What was the scope of assessment*	*Level 2 (study)*	*unordered, categorical*
Group 2c covariates: Study Quality indicators			
qhes_cata	*What was the overall QHES category given a strict application of the QHES criteria*	*Level 2 (study)*	*ordered, categorical*
qhes_catb	*What was the overall QHES category given a pragmatic application of the QHES criteria*	*Level 2 (study)*	*ordered, categorical*
Qhes_conta	*What was the overall QHES score given a strict application of the QHES criteria?*	*Level 2 (study)*	*continuous*
Qhes_contb	*What was the overall QHES score given a practicable application of the QHES criteria?*	*Level 2 (study)*	*continuous*

Further, experiences from the pilot study showed that studies on the cost-effectiveness of statins in the primary and secondary prevention of CVD usually do not explicitly report information on healthcare providers involved in delivering the intervention. At most, studies considered a fixed cost-component per annum for GP involvement in screening appropriate patients, titration and monitoring. In addition, some studies considered a lump sum for distributing statins through pharmacies. However, in both cases, this forms part of the annual cost of intervention, and is therefore considered as an '*intervention characteristic*'. Beyond this, provider characteristics were generally not reported and respective variability factors were therefore dropped from further consideration after carrying out the pilot study.

On the other hand, neither Sculpher et al. (2004) nor Goeree et al. (2007) consider a separate group of variability factors relating to the intervention and the comparator under assessment. However, even for a relative homogeneous group of agents as it may be the case for statins in the primary and secondary prevention of CVD, differences may exist which feed through to the cost and / or effectiveness of treatment. Additionally, total daily dosages may differ between patients, which again may cause variability in measures of cost-effectiveness. For these reasons, variables entered the abstraction form relating to the statins under assessment, total daily dosages, as well as unit cost and annual drug cost of both intervention and comparator.

For the remaining variability factors mentioned by Goeree et al. (2007) relating to a) the patient, b) the disease and c) methods used in health economic evaluations, literature was consulted to define variables for the data abstraction form which are suitable for subsequent quantitative analysis within the MLM framework. Patient and disease characteristics potentially causing variability in the cost-effectiveness of statins were operationalized by consulting literature on CVD risk factors and risk estimation. There is a vast literature available on CVD risk factors and a number of validated tools are available to estimate patients CVD related risk over a certain time period. Probably the most prominent CVD risk estimation tool is the Framingham risk equation (Anderson et al., 1991). However, in recent years the Q-Risk tool (Hippisley-Cox et al., 2007; Hippisley-Cox et al., 2008) has gained popularity (Cooper et al., 2008). Hence, risk equations were used to ascertain a long list of patient and disease characteristics potentially relevant for the cost-effectiveness of statins in the primary and secondary prevention of CVD. This long list of patient and disease characteristics was then tested in the pilot study, where data was abstracted from 16 studies previously identified by Franco et al. (2005). It turned out that some covariates included in risk equations are generally not (or at least not frequently) reported in the statins related cost-effectiveness literature. This holds, for instance, for left ventricular hypertrophy (Anderson et al., 1991), BMI (Hippisley-Cox et al., 2008), Family history of CVD (Hippisley-Cox et al., 2007; Hippisley-Cox et al., 2008), social status, social deprivation (Hippisley-Cox et al., 2007; Hippisley-Cox et al., 2008), a townsend deprivation score (Hippisley-Cox et al., 2008), ethnicity (Hippisley-Cox et al., 2007; Hippisley-Cox et al., 2008), as well as rheumatoid

arthritis, chronic renal disease and atrial fibrillation (Hippisley-Cox et al., 2008). As a result, only those patient and disease characteristics listed in Figure 4.2 as well as Table 4.2 above remained part of the data abstraction form, whilst variables less frequently reported were dropped from further analysis. The final data abstraction form for this empirical exercise is provided in Appendix 4.13.

The data collected on individual patient and disease characteristics of patient subgroups subsequently allowed applying the original Framingham risk equation (Anderson et al., 1991) to obtain an overall score on 10 year CVD related risk. Framingham was chosen as the risk equation is - in contrast to QRISK I and II - freely available so that a model could be set up in MS Excel and, secondly, as the pilot study showed good availability of data for most risk factors considered in this risk equation. In addition, Framingham is probably still the most prominent tool in the literature, and even though alternatives exist, its use is still recommended by NICE (Cooper et al., 2008). Initially, it was intended to use the resulting estimate of 10 year CVD risk as a further continuous explanatory within subsequent multilevel analysis. However, there were several obstacles:

First, not all studies included in the systematic review report the data required to populate the Framingham equation. Hence, if data was not obtainable from a cost-effectiveness study included in the systematic review exercise, alternative data sources were considered to fill gaps in the dataset. For instance, literature on RCTs providing data to populate the economic model proved particularly helpful to obtain missing information on patient subgroups considered. However, if neither the study included in this review exercise nor accompanying literature provided the data required to apply the Framingham risk equation, assumptions had to be made about the data. For instance, if data on TCL, HDL and triglycerides was available, but authors did not report an estimate of pre-treatment LDL, the 'Friedewald Function' (Friedewald et al., 1972) was used to fill this gap. Likewise, if only diastolic blood pressure (DBP) was reported, assumptions were made about systolic blood pressure (SBP). In particular, if DBP was elevated, it was assumed that the same holds for SBP and vice versa. If neither DBP nor SBP was reported, but a patient's hypertension status was positive, an elevated SBP was assumed. Finally, diabetes, smoking and

hypertension status were considered as continuous variables capturing the proportion of patients being affected by any of the stated conditions. Further details on the variables considered in the data abstraction form and the way they were operationalized for the purposes of this empirical analysis are available from Tables 4.2 and 4.3 above and Appendix 4.13.

After filling gaps in the data for patient and disease factors, the Framingham risk equation was used to estimate patient subgroups 10-year CVD related risk. However, the resulting score was not applicable to data points which referred to secondary prevention with statins as Framingham is only valid for patients who have not experienced a CVD event (Anderson et al., 1991). A previous CVD event drives the subsequent CVD risk so that results from the Framingham risk equation, which does not take into account CVD history, are invalid (Anderson et al., 1991). Unfortunately, this precludes the chance to use the resulting risk estimate as a continuous variable in the MLM framework. As an alternative, a categorical variable was defined with categories relating to a 10 year CVD risk of <10% (very low), 10% to <20% (low), 20% to <30% (medium), 30% to <40% (high), >40% (very high) and 'secondary prevention'. The resulting variable 'RISK-CAT' is tested as an explanatory in the MLM framework in Chapter 5.2.

Finally, methodological characteristics as mentioned by Goeree et al. (2007) were operationalized using standard health economics textbooks and key publications on economic evaluation in health. One source which proved particularly useful was the ISPOR online tool which summarizes HTA guidelines around the world (ISPOR, 2011). Five key challenges relate to the variability in cost-effectiveness data as a result of differences in study methods. First, some methods may only change between studies included in the dataset (for instance the general study design), whilst other methodological characteristics may also change within one study (for instance the discount rate on costs and effects). The MLM framework offers an excellent opportunity to assess the impact of different methods on data and study-level on measures of cost-effectiveness as it allows inclusion of covariates on each hierarchical level. As a result, the data abstraction form contains methodological characteristics on data and study-level respectively (Figure 4.2 above and Appendix 4.13).

Secondly, it is perhaps impossible to encode all study characteristics potentially causing variability in measures of cost-effectiveness within and between studies in terms of covariates for quantitative analysis. As a result, the data abstraction form may be regarded as an attempt to capture the most important methodological characteristics based upon those variability factors which have been previously discussed by other researchers. Accordingly, the reader may think of further variability factors relating to study methodology for future analysis. Presenting pilot study results at various conferences and seminars helped to expand upon the list of potential explanatory variables to consider in the main empirical analysis. Hence, this exercise aims to reflect both those factors previously mentioned in the literature as well as the feedback received from presenting the pilot study at various conferences and seminars.

The third challenge relates to the aim of incorporating study quality into the analysis. Study quality was frequently mentioned as a potential variability factor for measures of cost-effectiveness and fellow researchers highlighted that this exercise should not simply assign corresponding variability to the respective error term. Rather, one should aim to capture study quality somehow and include this information as a further covariate in the model. There are a number of quality checklists available within the economic evaluation literature and in a first attempt to capture study quality, the author considered the use of a checklist developed by Drummond & Jefferson (1996), which was developed with the aim to *'improve the quality of submitted and published economic articles'* (Drummond & Jefferson, 1996). Hence, the pilot study data abstraction form initially included 35 yes/no items referring to the above named checklist. However, the checklist developed by Drummond & Jefferson (1996) proved to be rather related to the quality of reporting economic evaluation results, and not the methodological rigour with which the study was initially conducted. Though this critique may apply to some extend to any quality checklist for economic evaluations, it was decided to consider alternatives for the empirical analysis.

The '*Quality of health Economics Studies'* (QHES) instrument, developed by Ofman et al. (2003), was chosen for a number of reasons: First, the QHES instrument is the only quality checklist which provides individual scores for each

dimension considered. These scores were generated using '*random-effects general least-squares regression based on a conjoint analysis of survey results from 120 international health economists*' (Ofman et al., 2003). Adding together individual scores for each item allows assigning an overall quality score to each study, with a perfect quality score of 100. This score may be used as an explanatory variable in the MLM analysis. Secondly, the QHES instrument was validated in a survey by 60 experts (30 clinicians and 30 health economists) in six disease categories. Third, QHES proved relatively straightforward to apply in this empirical exercise, with only 16 criteria to consider. However, there were also some problems associated with QHES. In particular, Ofman et al. (2003) designed some dimensions on their checklist with multiple subcategories. Hence, an individual study may score 'yes' in one subcategory but 'no' in others. For instance, consider question eight of the QHES checklist:

'Did the analytic horizon allow time for all relevant and important outcomes? Were benefits and costs that went beyond 1 year discounted (3% to 5%) and justification given for the discount rate?'

Unfortunately, the developers of QHES did not specify what to do if a study does, for instance, consider a sufficient analytic horizon, but discounted costs with 6% and effects with 1.5%. Or alternatively, if both analytic horizon and discount rate are appropriate, but no justification is given for the choice of the discount rate. As a result, some assumptions were made to apply QHES to the 67 studies included in this systematic review exercise and two QHES-scores were calculated for each study considered. First, criteria were applied in their strictest sense, meaning that a score of zero was assigned to a QHES dimension if any of the subcategories was answered with 'no'. Secondly, however, the item score was divided by the number of subcategories, and a partial score was assigned for each of the subcategories answered with 'yes'. Obviously, this less stringent application of the QHES instrument led to generally higher overall QHES scores.

Nevertheless, the results of applying quality checklists should always be considered with highest caution. All checklists considered for this exercise rely on judgements which may sometimes bias results. In addition, there may always be the concern that not all important quality dimensions were considered or, in the

particular case of QHES, the weights reflected in individual item scores may not perfectly reflect the relative importance of individual criteria on the checklist. To take into account the resulting uncertainty in results, continuous QHES scores were also converted into categorical variables in this thesis, with five categories with increments of 20 QHES points ranging from '0' to '100'.

The fourth challenge with respect to study methods relates to the geographic source of input parameters. Barbieri et al. (2005) grouped studies with respect to this source of variability. Accordingly, there may be differing degrees of '*context specificity*' in measures of cost-effectiveness. To assess this potential source of variability, information was first collected from studies to record the geographic origin of the data used to populate the economic model. Hence, data was collected to record the geographic origin of a) resource use data, b) unit cost estimates, c) effectiveness data and d) utility values. Secondly, covariates were derived which group data with respect to its '*geographic specificity*' of input parameters. For clarification, Table 4.4 relates the geographic source of the input data to the target country considered in the economic analysis. If all input parameters are taken from the target country, the highest context specificity is assumed. This context specificity decreases with an increasing number of input parameters obtained from other geographic domains.

Table 4.4: Context specificity of Input parameters
(adapted from Barbieri et al, 2005)

	Context specificity of input parameters			
Group	Utility weights	Effectiveness	Resource Use	Unit cost
Type 4	Yes	Yes	Yes	Yes
Type 3	No	Yes	Yes	Yes
Type 2	No	No	Yes	Yes
Type 1	No	No	No	Yes

As a result, variables entered the dataset which aim to encode the context specificity of input parameters. One variable (Barbierie_score_1) aims to reflect different combinations of input parameters being drawn directly from the target country, whilst another variable (Barbieri_score_2) counts the number of context

specific categories of input parameters, resulting in 4 categories according to what has been presented in Table 4.4 above. The raw data to derive such covariates has been omitted from further analysis.

The final challenge with respect to methodological characteristics of studies included in this empirical exercise relates to the potential number of relevant methods characteristics to consider in the MLM framework. In other words, if a large number of methodological characteristics are abstracted from the relevant literature, it is questionable whether one can simultaneously accommodate all relevant variables in the MLM framework. Hence, what would be desirable is a summary measure which captures a '*methodological profile'* of studies included in the dataset. For this reason, the use of multiple correspondence analysis was considered – the '*categorical equivalent to principal component analysis'* (Le Roux & Rouanet, 2010) - to group together studies of similar methodological profile. Those studies may then be assigned a score which can be used as an explanatory variable in the MLM. This idea was subsequently developed further into a study on the '*genealogy*' of economic evaluation studies, The underlying assumption is that studies which are '*phenotypically*' similar may share a '*genotypic*' relationship, for instance in terms of common authorship, common funding source or reuse of a previously developed DAM. The importance of this question arises out of the assumption of independence between studies included in the dataset. This assumption, which is necessary between groups on study-level in the multilevel framework, may be violated if studies are very similar, and this similarity is not simply a coincidence but rather the result of an existing relationship between the studies in question. Section 4.2.3 is concerned with this genealogy study.

After deciding on the variables to collect data for, an abstraction form was implemented in MS Access for the purposes of the pilot study. However, one key advantage of MS Access, which is the possibility to link tables with subtables to create a data hierarchy, hence avoiding redundancies in datasets, turned out to be a disadvantage for this particular exercise. The reason is that MLwiN requires data for higher-level variables assigned to each data point, so that the above mentioned redundancies occur (Rasbash et al., 2009). Hence, for the main

empirical analysis, a data abstraction form was implemented in MS Excel so that data for higher-level covariates could simply be copied to each respective data point. This format allowed instant exporting of the resulting dataset to STATA 12 and MLwiN. A word document containing the final data abstraction form is also available from Appendix 4.13.

The following section summarizes some key descriptive statistics based on the data abstracted from the studies included in this empirical exercise. Subsequently, the genealogy study is reported in Section 4.3 of this chapter.

4.2.2. Descriptive statistics of studies included in the empirical exercise

The first section of this chapter reported on a systematic literature review to identify studies on the cost-effectiveness of statins in the primary and secondary prevention of CVD which meet the inclusion criteria for this empirical exercise. The second section was concerned with the development of a data abstraction form which was based on a long list of potential variability factors for measures of cost-effectiveness as previously discussed in the relevant literature. This section reports some descriptive statistics of studies included in this empirical exercise which are also relevant for the genealogy study reported further below. Note that Section 5.2.3 of the empirical chapter also reports on descriptive statistics, missing values and correlations between explanatory variables for all covariates in the dataset. Further, descriptive statistics for all covariates are available from Appendix 5.

67 studies were includable in this systematic literature review. References of these 67 studies can be found in Appendix 4.12. In addition, Table 4.6 at the end of this section lists all 67 studies with key characteristics discussed in this section. 61 studies are '*single country*' studies, providing measures of cost-effectiveness for one geographic domain only. The remaining six studies (8.96%) are

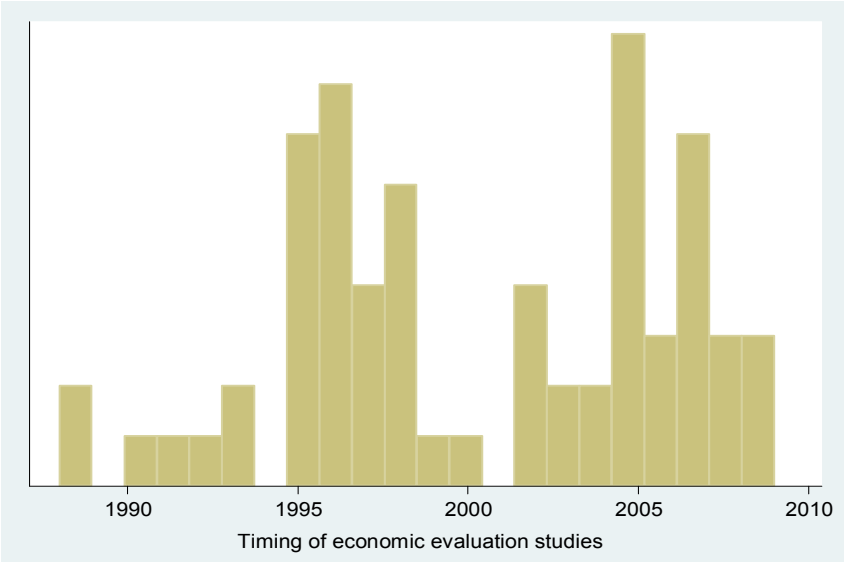
multinational in nature, providing data on more than one geographic domain. The countries considered in these six multinational studies are Denmark, France, Italy, New Zealand, Norway and Portugal. Table 4.5 below shows the distribution of studies per country.

Table 4.5: Distribution of studies per target country

Country	Frequency	Percent	Cummulative
Australia	1	1.49%	1.49%
Belgium	2	2.99%	4.48%
Brazil	1	1.49%	5.97%
Canada	12	17.91%	23.88%
Finland	2	2.99%	26.87%
Germany	5	7.46%	34.33%
Hong Kong	1	1.49%	35.82%
Hungary	1	1.49%	37.31%
Japan	1	1.49%	38.81%
Netherlands	3	4.48%	43.28%
Spain	1	1.49%	44.78%
Sweden	4	5.97%	50.75%
Switzerland	1	1.49%	52.24%
UK	7	10.45%	62.69%
UK (Engl./Wales)	7	10.45%	73.13%
UK (Scotland)	1	1.49%	74.63%
USA	11	16.42%	91.04%
Multinational*	6	8.96%	100%
Total	**67**	**100%**	--
*Denmark, France, Italy, New Zealand, Norway and Portugal			

Note that some studies provided data applicable to 'England/Wales', others provided data for 'Scotland' and finally, some studies referred to the 'UK' as a whole. Accordingly, three distinct geographic entities were defined. 61 out of 67 studies were written in English language. The timing of studies (which is not the year of publication) ranges from 1988 to 2009, with two peaks between 1995 to 1998 and 2005 to 2007 respectively (see Figure 4.3 below). Industry was involved in the funding of 39 publications (58.21%), whilst funding was unclear for a further 17 studies (25.37%). If industry funding was available, the manufacturers most commonly involved were Pfizer with 13 studies and Merck with 12 studies.

Figure 4.3: Timing of economic evaluation studies included in this empirical exercise



Simvastatin was by far the most commonly assessed intervention, followed by pravastatin, atorvastatin, lovastatin, rosuvastatin and fluvastatin. In most studies, the intervention was compared to '*doing nothing'* whilst 17 studies also considered other statins as comparator. The mean annual drug cost of the intervention (converted to £-Sterling using Purchasing Power Parities (PPP) and updated to 2010 using country specific GDP deflators (Shemilt et al., 2008; OECD, 2010) is £521.59. The annual drug cost is highest for lovastatin at £932.14 (SD: 515.14), followed by pravastatin at £858.00 (SD: 236.67), atorvastatin at £503.89 (SD: 232.19), simvastatin at £477.70 (SD: 312.81), rosuvastatin at £337.28 (SD: 247.16) and fluvastatin at £ 293.08 (SD: 103.10) respectively.

Methods on study-level show that most studies (61; 91.04%) rely on secondary modelling, whilst only six studies (8.96%) made direct use of individual patient data. For secondary modelling, the most common model used was a Markov state transition model (41; 61.19%). Seven studies (10.45%) were based on decision trees, and other modelling approaches involved life tables, or discrete event simulation. An important question is how effectiveness was measured and modelled within a study, and there are two general approaches. Most studies (61.19%) modelled the reduction in risk of experiencing a CVD event in the future to estimate incremental effectiveness. 26 studies (38.81%), however, used the intermediate outcome of cholesterol reduction to approximate its impact on CVD

risk, which then resulted in an estimation of life years or QALYs saved. Moving on to the outcome measure itself, QALYs were considered in 32 (47.76%) out of the 67 studies. If QALYs were considered, the utility weights represent population values in 15 studies (22.39%), patient values in 13 studies (19.40%), and in 4 studies (5.97%) it was unclear whether utility weights represent patient or population preferences.

The majority of 35 studies (52.24%) explicitly looked into the effect of statins on coronary heart disease (CHD) and cerebrovascular disease (CD), whilst 18 studies (26.87%) looked at CHD only. 11 studies (16.42%) looked at CHD, CD and peripheral arterial disease (PAD). Inflation adjustment of cost estimates to a common baseline year was explicitly reported in 18 studies (26.87%), whilst this was unclear in 35 cases (52.24%), If inflation adjustment was applied, the most common method was to use the healthcare component of the target countries consumer price index with 10 studies (14.93%). Currency conversion of any kind was applied in 15 studies (22.39%), and the most common method of currency conversion was the use of real exchange rates (11, 16.42%). In terms of context specificity, by far the most measures of cost-effectiveness (49.33%) where generated with unit cost and resource use data from the target country, whilst effectiveness estimates and utility weights were transferred from another jurisdiction. Only 56 data points (2.67%) were generated with all data sources from the target domain.

Table 4.6 below summarizes some key characteristics of all 67 studies included in this empirical exercise. Full descriptive statistics as well as missing values analysis and analysis of correlations between explanatory variables are provided in Section 5.2.3 of the empirical chapter as well as in Appendix 5.

Table 4.6: Key study characteristics – summary table

ID	Authors (pub. year)	Timing	Multi-country	Target country	Primary Modelling	Outcome measure	Industry Funding
1	Ashraf et al. (1996)	1995	No	USA	No	LYS	Yes
2	Caro et al. (1997)	1996	No	UK (Scotland)	Yes	LYS	Yes
3	Grover et al. (1999)	1996	No	Canada	No	LYS	Unclear
4	Grover et al. (2000)	1996	No	Canada	No	LYS	Yes
5	Hamilton et al. (1995)	1992	No	Canada	No	LYS	Yes
6	Johanneson et al. (1997)	1995	No	Sweden	No	LYS	Yes
7	Muls et al. (1998)	1995	Yes	Belgium	No	LYS	Yes
8	Perreault et al. (1998)	1995	No	Canada	No	LYS	No
9	Pharoah et al. (1996)	1995	No	UK (England/Wales)	No	LYS	No
10	Szucs et al. (1998)	1996	No	Germany	No	LYS	Unclear
11	Szucs et al. (2000a)	1998	No	Germany	No	LYS	Unclear
12	van Hout et al. (2001)	1999	No	Netherlands	No	LYS	Unclear
13	Jönsson et al. (1996)	1995	Yes	Multi-country	Yes	LYS	Yes
14	Jönsson et al. (1999)	1997	Yes	Multi-country	Yes	LYS	Yes
15	Ganz et al. (2000)	1998	No	USA	No	LYS / QALYs	No
16	Grover et al. (2001)	1998	Yes	Multi-country	No	LYS	Yes
17	Martens et al. (1994)	1993	No	Canada	No	LYS	Yes
18	Alonso et al. (2008)	2005	No	Spain	No	LYS	Yes
19	Annemans et al. (2010)	2009	No	Belgium	No	LYS / QALYs	Yes
20	Araujo et al. (2007)	2007	No	Brazil	No	LYS	Unclear
21	Lindgren et al. (2010)	2007	No	UK (England/Wales)	No	LYS / QALYs	Yes
22	Grover et al. (2008)	2002	No	Canada	No	LYS	Yes
23	Franco et al. (2007)	2003	No	Netherlands	No	LYS	No
24	Greving et al. (2011)	2008	No	Netherlands	No	QALYs	No
25	HPS Group (2009)	2006	No	USA	No	LYS / QALYs	Yes
26	Khoury et al. (2009)	2007	No	Canada	No	LYS / QALYs	Yes
27	Kongnakorn et al. (2009)	2005	No	USA	No	LYS / QALYs	Yes
28	Morris (1997)	1996	No	UK	No	LYS	Unclear
29	Morris & Godber (1999)	1997	No	Canada	No	LYS	Yes
30	Rosen (2010)	2007	No	USA	No	LYS / QALYs	Yes
31	Scuffham et al. (2005)	2002	No	UK	No	LYS / QALYs	Yes
32	Scuffham et al. (2006)	2005	No	Hungary	No	LYS / QALYs	Yes
33	Tailor et al. (2009)	2005	Yes	Multi-country	No	LYS / QALYs	Unclear
34	Tonkin et al. (2006)	1998	No	Australia	Yes	LYS	Yes
35	Wagner et al. (2009a)	2007	No	Canada	No	LYS / QALYs	Yes
36	Wagner et al. (2009b)	2006	No	Canada	No	LYS / QALYs	Yes
37	Berger et al. (1997)	1996	No	Germany	No	LYS	Unclear
38	Obermann et al. (1997)	1993	No	Germany	No	LYS	Yes
39	Davies et al. (2006)	2005	No	UK	No	QALYs	Yes
40	Spaans et al. (2003)	1996	No	Canada	No	LYS	Unclear
41	Soini et al. (2010)	2007	No	Finland	No	LYS / QALYs	Yes
42	Peura et al. (2008)	2006	No	Finland	No	LYS / QALYs	Yes
43	Slejko et al. (2010)	2008	No	USA	No	QALYs	Unclear
44	Nherera et al. (2010)	2009	No	UK (England/Wales)	No	QALYs	Unclear
45	Szucs et al. (2000b)	1997	No	Switzerland	No	LYS	Unclear
46	Sigvant et al. (2011)	2009	No	Sweden	No	LYS / QALYs	Yes
47	Johannesson et al (1996)	1991	No	Sweden	Yes	LYS	Yes
48	Troche et al. (1998)	1995	No	Germany	No	LYS	Unclear
49	Szucs et al. (2004)	2003	Yes	Multi-country	No	LYS	Unclear
50	Nagata et al. (2005)	2002	No	Japan	No	QALYs	Unclear
51	Lindgren et al. (2007)	2005	Yes	Multi-country	No	LYS / QALYs	Yes
52	HPS Group (2006)	2005	No	UK	No	LYS / QALYs	Yes
53	Tsevat et al. (2001)	1996	No	USA	No	QALYs	Yes
54	Raikou et al. (2007)	2004	No	UK	Yes	LYS / QALYs	Yes
55	Ramsey et al. (2008)	2005	No	USA	No	LYS / QALYs	Yes
56	Scuffham et al. (2004)	2002	No	UK (England/Wales)	No	LYS / QALYs	Yes
57	Hjialte et al. (1989)	1988	No	Sweden	No	LYS	Unclear
58	Caro et al. (2003)	1998	No	USA	No	LYYS	Yes
59	CDC Group. (2002)	1997	No	USA	No	LYS / QALYs	No
60	Chau et al. (2001)	1998	No	Hong Kong	No	QALYs	Yes
61	Grover et al. (2003)	2000	No	Canada	No	LYS	Yes
62	Glick et al. (1992)	1988	No	UK	No	LYS	Yes
63	NICE (2008)	2007	No	UK (England/Wales)	No	QALYs	No
64	Drummond et al. (1993)	1990	No	UK	No	LYS	No
65	Chan et al. (2007)	2005	No	USA	No	LYS / QALYs	No
66	Ward et al. (2007)	2004	No	UK (England/Wales)	No	QALYs	No
67	Ara et al. (2009)	2008	No	UK (England/Wales)	No	QALYs	No

4.3. ‘Genealogy’ study of the statin related cost-effectiveness literature

When carrying out the systematic literature review and abstracting data from studies, it became apparent that some studies are related to each other, for instance, through common authorship, the use of identical data sources, reuse of a previously published DAM, or simply a common source of funding. This, however, may violate the independence assumption between studies, which is necessary to fit the MLMs developed in Chapter 3. If two studies are related, one may consider pooling them in one group on study-level rather than defining separate groups for each of the 67 studies included. Though this may reduce parameters on study-level, it may also be regarded as more appropriate to fit MLMs to the data, especially when two related studies are very similar in many aspects. Therefore, the aim of this final section of Chapter 4 is to look into the ‘*genealogy*’ of economic evaluation studies on the cost-effectiveness of statins in the primary and secondary prevention of CVD. Though the motivation of this study arises out of the aim of critically appraising the independency assumption between studies in the MLM framework, this exploratory task into relationships and potential similarities between studies as a consequence of such relationships constitutes an interesting piece of original research in its own right, which could also be followed up further in future work.

The aim of this exercise is to look into the genealogy of economic evaluation studies includable in this empirical exercise. There are two primary objectives: the first is to assess whether there may be alternative hierarchical structures to those assumed in the multilevel methods chapter and pilot Study which may fit the data better and hence would need to be acknowledged in the MLM. In other words, it is to consider whether modelling data clustered in published papers is the most reasonable hierarchical structure, or whether it is indicated to look into alternative data structures based on the relationships between published papers. This could lead, for instance, to studies which belong to one group (however this group may be defined) being aggregated on level two of the data hierarchy so that data is no longer clustered in papers, but rather in groups of papers. If it is

concluded that modelling data clustered in single papers remains the most appropriate hierarchical structure, the second objective is to derive covariates to capture some of the observed relationships between papers. Hence, instead of informing alternative model hierarchies, this exercise could lead to a number of covariates which encode relationships between papers and thereby allow their influence on the variability in measures of cost-effectiveness to be assessed.

Both objectives relate back to an important issue in MLM, namely whether to capture something as a level or as a covariate (Rasbash et al., 2009). In conclusion, looking into the genealogy of studies, i.e. the relationships between papers which exist within the sample of papers included in this exercise, is a necessary step towards the theoretical validation of the model structure assumed in this study. This assessment, however, is explorative in nature, meaning that there is neither a strong body of literature to build upon nor is it clear whether this exercise may actually lead to results robust enough to inform alternative model structures. Nevertheless, beyond the primary objectives of this exercise, the issue of study genealogy constitutes an important research question in its own right, which deserved little attention thus far and should therefore also be considered in future research.

4.3.1. Problem

Relationships between published economic evaluation studies may exist on several dimensions like common authorship, recycling previously published DAMs, or using identical data sources, just to name a few. It is one thing to test the influence of such relationships on the response variable by including respective covariates into a MLM, but to inform an entire alternative model structure, one would have to establish not just an existing relationship, but also show that studies are similar in terms of characteristics determining the response variable so that one may group studies together on level two. For this reason, it is imperative not just to look into the '*genotypic*' existence of relationships between published economic evaluation papers, but also to look at

the '*phenotypic*' similarity in terms of study characteristics influencing the results of the affected studies. In very simple terms, things which are related (on one level) ought to look similar (on that level), if they don't, we shall continue to treat them as independent!

Only looking at differences in the response variable is not sufficient as results may appear unrelated due to differences on other levels, which would then draw a curtain over actual similarities of the affected studies. For instance, two studies could be absolutely identical in terms of the economic model applied and a number of key study methods which determine measures of cost-effectiveness, but differ with respect to the patient population under assessment - a data level characteristic. This patient characteristic, for example mean TCL, may cause substantial variability in measures of cost-effectiveness even though the two studies are almost identical. In conclusion, after controlling for TCL on data-level, one may regard study parameters to be identical, which would then justify pooling both studies in one group on study-level. Only looking at the response variable, i.e. the cost-effectiveness of the intervention, may not be sufficient as INMBs may differ sharply due to the difference in mean TCL amongst patient subgroups considered in both studies.

Two papers in the dataset (Grover et al., 1999 and Grover et al., 2000) illustrate the above mentioned example. Though both papers are related through common authorship, and remarkably similar in terms of model structure, data sources and a number of additional study characteristics, their respective mean INMBs differ considerably (£47803 vs. £26230). A key difference likely to be responsible for this strong variation in mean INMBs between both papers is that one paper considers patients suffering from diabetes whilst the other paper does only consider patients without that condition. This difference, however, may not be captured on study-level, but rather on data-level so that one could regard parameters of both studies to be identical after controlling for diabetes status on data-level. Hence, one may argue in favour of pooling both studies in one group on study-level even though their respective mean INMBs differ considerably.

4.3.2. Method

The question is then how to establish this similarity between related studies. It is difficult enough ascertaining simple relationships between published papers for covariate adjustment. However, performing a detailed assessment of similarity for 67 studies included in the dataset and also determining a '*threshold*' above which we shall no longer consider two studies as independent but rather pool them on study-level, would lead into a study of considerable complexity and uncertain outcome in terms of its usefulness for informing alternative MLM structures. Though this appears to be a very important research question, it is questionable whether this project, which focusses on a different matter, can accommodate a case study of this magnitude with uncertain outcome.

As an alternative to a full '*qualitative*' assessment of relationships and resulting similarities between published economic evaluation studies, this exercise therefore looks into methods which may be used to build upon the existing dataset. In other words, if a quantitative method existed which is capable of disclosing patterns in the existing data which may then show whether economic evaluation studies can be pooled on study-level, this could then be used as a less time consuming alternative. After studying techniques for the analysis of multivariate data, several candidate methods were considered.

One multivariate method often used to assess patterns in data is factor analysis (FA). According to Acock (2011), FA is a collection of methods that does exploratory analysis to ascertain whether there are items that may be clustered in particular groups. In this form of analysis, the observed variables are presented in terms of '*linear combinations of a few random variables, called factors*' (Rencher, 2002). As Rencher (2002) further states, '*the goal of factor analysis is to reduce the redundancy among the variables by using a smaller number of factors*.' These factors are, unlike the observations, unobserved and therefore also referred to as '*latent variables*'. Factor analysis may be used to assess patterns in the observed data and to reduce the complexity in a dataset by presenting groups of correlated variables in a smaller number of uncorrelated

factors (Acook, 2011). A related method to this is principal component analysis (PCA), which is also used to assess patterns in the data and to reduce the complexity of the dataset (Rencher, 2002). However, there are differences between both methods: Most importantly, '*principal components are defined as linear combinations of the original variables, whilst in factor analysis, original variables are expressed as linear combinations of the factors'* and secondly '*PCA explains large part of the total variance of the variables whilst in FA, we seek to account for the covariances or correlations among the variables.'* (Rencher, 2002).

It is important to note that there is considerable confusion around the terminology of these methods. What is referred to as PCA in one textbook, may be labelled as principal components factor analysis (PCF) in another one. Accordingly, software packages are not consistently using the same terminology for the same methods as STATA, for example, labels as PCF what SPSS refers to as PCA (Acook, 2011). Though this may further add to the confusion, one thing is common to all methods introduced above. Whilst they are, in theory, very appropriate for assessing patterns in the data in the way required for this genealogy study, they are only defined for continuous variables (Rencher, 2002; Kolenikov & Angeles, 2004; Acook, 2011), whilst the vast majority of variables to consider for this assignment is categorical in nature. For this reason, a package developed by Kolenikov & Angeles (2004) was considered as an alternative, which implements '*polychoric principal components analysis'* into the software environment STATA. However, as Kolenikov & Angeles (2004) state, this method may only be valid for continuous as well as ordered categorical variables, which again leaves out the majority of variables relevant for this assessment.

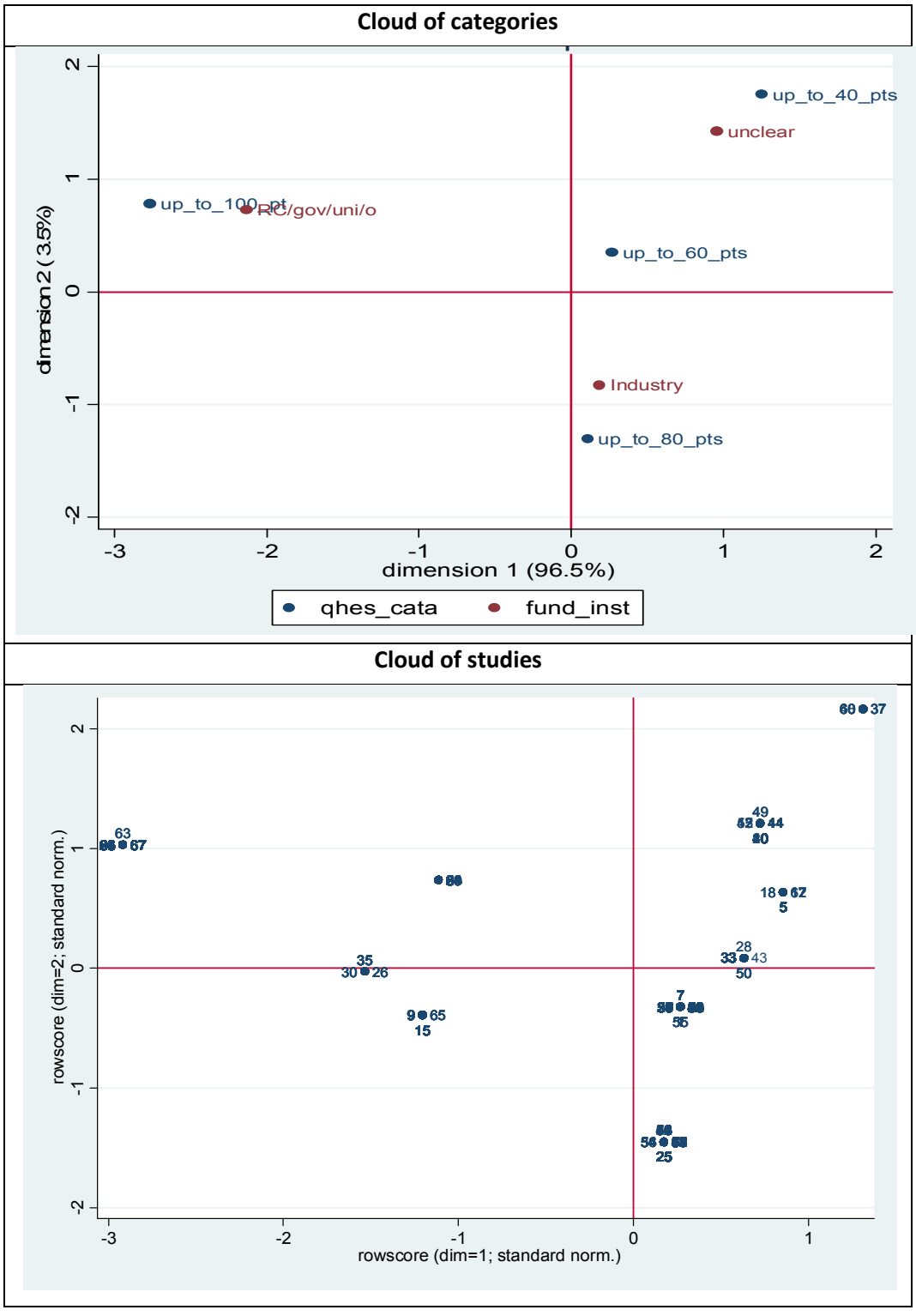
In conclusion, as neither FA, PCA nor polychoric PCA appeared to be applicable to unordered categorical data, an alternative methodology was considered for this exercise, commonly referred to as multiple correspondence analysis (MCA). The term '*correspondence analysis'* stems from the French term '*Analyse Factorielle de Correspondence*' which was defined by Benzéri et al. (1973) and the technique is also referred to as the '*categorical equivalent to PCA'* (Le Roux & Rouanet, 2010). The aim of this form of analysis is to visualise the raw data in a low-

dimensional space (usually two dimensions) which then helps to identify patterns in this data (Bartholomew et al., 2008). It does so by converting categories of variables into points on a plane (the biplot), and the researcher may then analyse the resulting cloud and sub-clouds of points in this geometric space (Le Roux & Rouanet, 2010).

There are actually two clouds to consider. The '*cloud of categories*' shows the chi-square distance between categories of variables, the '*cloud of individuals'* reflects dissimilarities in response patterns of 'individuals' - or rather 'studies' in the case of this genealogy exercise (Le Roux & Rouanet, 2010). Hence, the cloud of categories would tell us whether certain combinations of categories of variables are more common than others. If this is the case, then the respective categories would appear close to each other on the biplot. Though this is an important aim of correspondence analysis, it is not what we are primarily interested in within this genealogy study. Rather, the primary aim is to analyse the cloud of individuals which shows which studies share common response patterns in terms of those categories.

Consider Figure 4.4 below. The biplot on the top shows the cloud of categories for the two variables 'Funding institution' and 'QHES'. From this biplot, we can learn that the lowest QHES scores most commonly appear in studies which did not disclose their funding sources. Conversely, government/ research council funded studies appear close to a QHES score of 80 to 100. In between are the industry funded studies, which most commonly score between 60 and 80 QHES points. The interesting question is now which studies actually have similar response patterns in terms of funding source and QHES score, and this is exactly what the cloud of studies on the bottom of Figure 5.4 shows. Hence, markers of studies which are absolutely identical in terms of the categories of variables under consideration, overlap on the biplot. The more differences between combinations of categories, the greater is the distance between studies on the biplot.

Figure 4.4: Example of a MCA on '*funding source*' and '*study quality*'



Hence, the idea of using this technique in this genealogy study is simply that studies which are very similar in terms of their response patterns (i.e. the categories of variables), appear in close proximity to each other on the cloud of

individuals (Bartholomew et al., 2008; Le Roux & Rouanet, 2010). If this is the case, the respective studies could be pooled within the same group on study-level within subsequent MLM analysis. Hence, MCA is not employed to ascertain combinations of categories that are more common than others (cloud of categories), but rather to assess which studies share common response patterns (cloud of individuals) even if the underlying data is diverse In nature.

Of course, the more variables we include in this assessment, the less likely it may be to find studies which are completely alike in terms of their response patterns. In other words, the diversity between pairs of studies is likely to increase with an increasing amount of individual characteristics to look at. On the other hand, with an increasing amount of study characteristics, similarity may be less likely a coincidence but rather the result of studies actually being related in some way. In conclusion, biplots will be produced with different subsets of variables, starting off with the complete set of variables presented in Table 4.7 below, and then subsequently dropping variable by variable in increasing order of relevance to the response patterns observed. The more variables we drop, the less diverse are the response patterns, and the more points on the cloud of individuals may coincide (Bartholomew et al., 2008; Le Roux & Rouanet, 2010). However, the more variables are being dropped, the more likely it gets that similarities between studies are not a result of actual relationships but rather a coincidence, as they are based on viewer study characteristics under consideration.

Accordingly, after disclosing similarities between studies in this way, a subsequent step is to validate findings by looking into the actual studies and trying to find the underlying relationship which may have caused the observed similarity. In a way, we confirm the '*genotypic*' relationship after disclosing '*phenotypic*' similarity. If there is no apparent relationship, then the observed similarity should not be considered when re-grouping data on study-level in subsequent multilevel analysis.

To sum up, MCA helps to assess which studies share common characteristics and should therefore be regarded as similar. It does so by utilising the data already

collected within the literature review and data-abstraction exercise reported earlier in this chapter and therefore constitutes a time-saving alternative to a full qualitative assessment of relationships between studies and their resulting similarities in terms of study-characteristics. Nevertheless, it needs to be emphasized that this whole exercise is explorative in nature, and what may work in theory, may not lead to unambiguous results in practice, e.g. due to noise in the data or other factors distorting the results.

4.3.3. Data

As detailed above, this genealogy study starts off from the data already available from the literature review and data abstraction exercise. This data was collected from 67 economic evaluation papers which were published on the cost-effectiveness of statins in the primary and secondary prevention of CVD. Only papers which decomposed the ICER or the INMB statistic were included as only this allows calculating as response variable an INMB with a common threshold value or running a bivariate MLM with incremental cost and incremental effects as a vector of response variables. Covariates were defined from a long list of potential variability factors which were previously reported in Sculpher et al. (2004) and Goeree et al. (2007). A number of variables entered the data abstraction exercise which may also be useful for the purposes of ascertaining relevant relationships between published papers. As a result, a very rich dataset was obtained with data collected on more than 80 distinct variables on data and study-level (details may also be obtained from previous sections of this chapter).

Certainly, not all of these variables may be relevant for this genealogy exercise. Rather, it is indicated to first hypothesize about potential relationships between published papers and then determine which study characteristics encoded as variables in the dataset may be affected by any relationship between papers. This is equivalent to choosing covariates for a regression analytic model, where one would only include candidate variables of which a relationship to the dependent variable is anticipated. Hence, as the basic principle '*garbage in,*

garbage out' applies just as much to MCA as it does to regression analytic modelling, it is indicated to first think about what may be affected if two studies are related in some way, and then to screen the dataset for the existence of appropriate candidate variables.

It is appreciated that, out of the potentially unlimited space of existing relationships and study characteristics affected by such relationships, this exercise may only look into a very limited number of obvious candidates. This underpins the character of this exercise as an explorative study into the genealogy of economic evaluation studies and one may think of a more systematic approach to the selection of variables for future research. Potential relationships which were considered are '*common authorship'*, '*recycled model*', '*common funding source'* and whether studies were based on the same '*methods guideline'*. The existing dataset was then screened for variables which are considered sensitive to the existence of any of these relationships between papers and therefore includable in the MCA. Results may be obtained from Table 4.7. Based on this data, biplots were produced with different subsets of variables, starting off with the complete set of variables presented in Table 4.7, and then subsequently dropping variable by variable in increasing order of relevance to the response patterns observed. The more variables we drop, the less diverse are the response patterns, and the more points on the cloud of individuals should coincide. However, the more variables we drop, the more likely it gets that similarities between studies are simply coincidence, as they are based on viewer study characteristics under consideration.

Table 4.7: Potential candidate variables for MCA

Variable	*Description*	*Level*	*Nature*
outc_measure	*How was health outcome reported in the study*	*Level 1 (data)*	*unordered, binary*
elicitation	*If QALYS were used, what was the method of preference elicitation?*	*Level 1 (data)*	*unordered, categorical*
population	*If QALYS were used, what do the utility values reflect (patient / population values)*	*Level 1 (data)*	*unordered, categorical*
DRC	*What was the discount rate on costs*	*Level 1 (data)*	*Continuous (converted to ordered, categorical)*
DRB	*What was the discount rate on benefits*	*Level 1 (data)*	*Continuous (converted to ordered, categorical)*
duration	*What was the treatment duration modelled*	*Level 1 (data)*	*ordered, categorical*
extrapol	*Was there any extrapolation beyond the latest follow up?*	*Level 1 (data)*	*unordered, binary*
horizon	*What is the time horizon?*	*Level 1 (data)*	*ordered, categorical*
Persp_rep	*What was the study perspective as reported by the authors of the article*	*Level 1 (data)*	*unordered, categorical*
Persp_cost_concl	*What was the study perspective on costs as concluded by the reviewer*	*Level 1 (data)*	*unordered, categorical*
intervention	*What was the brand name of the intervention drug?*	*Level 1 (data)*	*unordered, categorical*
comparator	*What was the brand name of the comparator drug?*	*Level 1 (data)*	*unordered, categorical*
source_effects	*From which source (trial, meta-analysis) was the effectiveness data taken from*	*Level 1 (data)*	*unordered, categorical*
paper_origin	*In which country was the paper written (if authors from several jurisdictions were involved, where is the lead author based?*	*Level 2 (study)*	*unordered, categorical*
multinational	*Was the study multinational*	*Level 2 (study)*	*unordered, binary*
fund_inst	*What was the primary source of funding (institution)*	*Level 2 (study)*	*unordered, categorical*
fund_man	*If funding source was private, which manufacturer was involved?*	*Level 2 (study)*	*unordered, categorical*
gen_des	*What was the general study design?*	*Level 2 (study)*	*unordered, categorical*
prim_des	*If primary modelling, what was the specific study design?*	*Level 2 (study)*	*unordered, categorical*
sec_des	*If secondary modelling, what was the specific study design*	*Level 2 (study)*	*unordered, categorical*
effect_calc	*Method of effect calculation*	*Level 2 (study)*	*unordered, categorical*
Timing	*What is the timing of the economic evaluation*	*Level 2 (study)*	*unordered, categorical*
cur_conv	*Was currency converted*	*Level 2 (study)*	*unordered, binary*
conv_method	*If currency was converted, what was the conversion method used by the authors?*	*Level 2 (study)*	*unordered, categorical*
scope	*What was the scope of assessment*	*Level 2 (study)*	*unordered, categorical*

Note that descriptive statistics of the variables considered in this exercise were already reported in the previous section of this chapter. In addition, full descriptive statistics are available from Chapter 5.2.3 and Appendix 5.

4.3.4. Results and discussion

Figure 4.5 shows the results of performing a MCA to ascertain similarities in study characteristics between published economic evaluation papers included in this empirical exercise. Each point on the biplot represents one study, and the distance between individual points is determined by differences in the response patterns of studies. In other words, studies which share common response patterns should be in very close proximity on the biplot. If studies are absolutely identical in their response patterns, their markers on the biplot should coincide.

Figure 4.5a shows a pattern which one could almost refer to as random. In other words, there are virtually no apparent clusters on the plot, meaning that the existence of common response patterns in studies is very low. However, with decreasing number of categorical variables included in the assessment, one can observe that points on the biplot increasingly group together in clusters. This is in line with expectations as the more study characteristics are compared, the less likely it should be that two studies are completely alike. Results therefore have face validity. However, what is worrying is that the combinations of coinciding points on the biplot change and therefore appear to be sensitive to the choice of covariates to include. In other words, if we observe two studies response patterns to be very similar in biplot 4.5a then ideally we should observe this result to be robust in all four biplots. In fact, we observe one such example with studies 15 and 65 coinciding exactly on all four biplots, but for other pairs of studies (e.g. 39 and 44 on biplot 4.5a) this close proximity disappears with dropping variables from the assessment.

After analysing biplots, the next step in this case study is to ascertain whether there are any apparent relationships between the studies with similar response patterns which coincide on the biplots presented in Figure 4.5. For this reason, a full text review was conducted for each pair (or group) of studies circled in red in Figure 4.5. This is a descriptive exercise and results can be obtained from Table 4.8 below.

Figure 4.5: Biplots showing clouds of studies with decreasing number of study characteristics under consideration

Figure 4.5a.

Inclusion of all variables from Table 8 considered sensitive to relationships between published studies

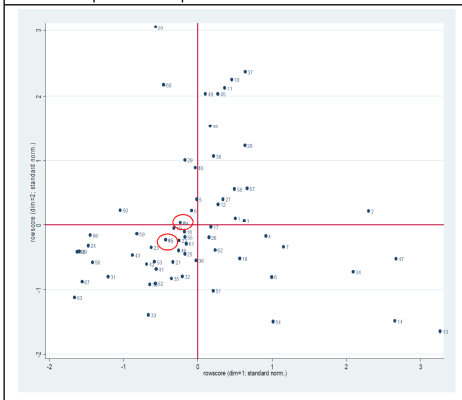


Figure 4.5b

Dropping scope, drc_cat and drb_cat

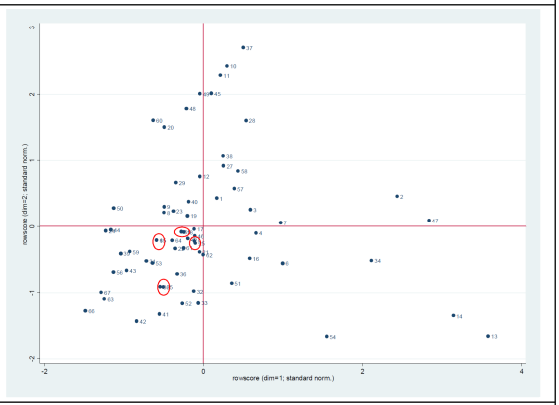


Figure 4.5c

Dropping scope, drc_cat, drb_cat, fund_inst and fund_man

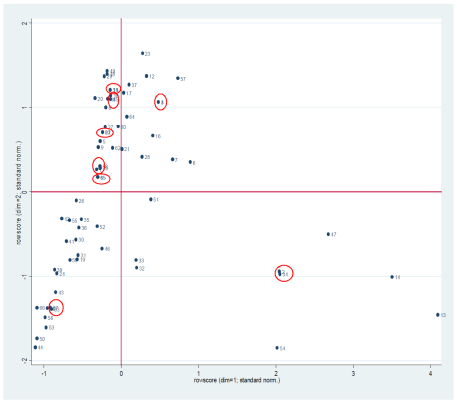
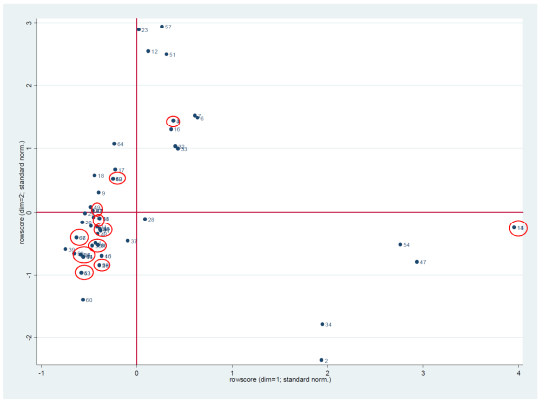


Figure 4.5d

Dropping scope, drc_cat, drb_cat, fund_inst, fund_man, duration, extrapol and paper_origin



Figue 4.5: clouds of individual studies included in this exercise. Each point on the biplot represents one study, and the distance between individual points is determined by differences in the response patterns. In other words, studies which share common response patterns, should be in very close proximity on the biplot. One can see that with a decreasing number of categorical variables included in the assessment, the number of coinciding points on the biplot increases.

As can be seen from Table 4.8, it was possible to establish a direct relationship between papers in almost 50% (10 out of 21) of the matches on the biplots shown in Figure 4.5. These relationships were mostly based on common authorship (7 of 10). In one case, the relationship between two papers, which were identical in terms of data sources and the model to estimate cost-effectiveness, was almost undetectable as the authors completely failed to cross-reference their papers. A link could only be established through the funding source, which was identical in both papers (Annemans et al., 2010; Khoury et al., 2009). The strong link between these two papers would still be hidden in the dataset if one would only look into genotypic relationships between studies and not the phenotypic similarities between those studies.

Table 4.8: Descriptive analysis of relationships between published papers appearing in close proximity on the cloud of studies

Papers coinciding on biplots	Biplots on which papers coincide	Description of the relationship between papers coinciding on biplots	Descriptive analysis accords MCA?
IDs: 1 / 25 / 58 Ashraf et al. (1996); HPS collaborative Group (2009); Caro et al. (2003)	c	There is no apparent relationship in terms of common authorship, use of the same model, funding source or identical data sources between the papers.	No
IDs: 2 / 34 Caro et al. (1997); Tonkin et al. (2006)	c	There is no apparent relationship in terms of common authorship, use of the same model, funding source or identical data sources between the papers	No
IDs: 3 / 4 Grover et al. (1999); Grover et al. (2000)	c / d	There is a strong link between both papers in terms of common authorship, as both papers were published by Grover et al. This may explain the use of the identical model (CVD life expectancy model) as well as similar data sources.	Yes
IDs: 5 / 25 Hamilton et al. (1995); HPS Collaborative Group (2009)	b	There is no apparent relationship in terms of common authorship, use of the same model, funding source or identical data sources between the papers	No
IDs: 5 / 41 Hamilton et al. (1995); Soini et al. (2010)	d	There is no apparent relationship in terms of common authorship, use of the same model, funding source or identical data sources between the papers	No
IDs: 9 / 64 Pharoah (1996); Drummond et al. (1993)	a	There is no apparent relationship in terms of common authorship, use of the same model, funding source or identical data sources between the papers	No
IDs: 10 / 45 Szucs et al. (1998); Szucs et al. (2000)	c / d	There is a strong relationship between published papers. Both papers share common authorship, and economic models are almost identical. Whilst data sources vary for effectiveness data, cost data is taken from the same sources. Both studies share the same origin, and papers are published in German.	Yes
IDs: 11 / 38 Szucs et al. (2000); Obermann et al. (1997)	c / d	Both papers were published in Germany, and hence, similar methods standards may apply. However, sources of effectiveness and cost data differ, and there is no apparent relationship in authorship.	Yes
IDs: 13 / 14 Jonsson et al. (1996); Jonsson et al. (1999)	d	Both studies share a strong relationship through common authorship and the fact that they rely on the analysis of RCT data from 4S. Whilst ID 13 (Jönssen et al (1996)) focusses on the cost-effectiveness of simvastatin as assessed within the whole 4S trial population, ID 17 (Jönsson et al (1999) focus on diabetic patients from that cohort only.	Yes
IDs: 15 / (36) / 65 Ganz et al. (2000); Wagner et al. (2009); Chan et al. (2007)	a / b / c / (d)	Two studies (Wagner et al (2009) and Chan et al (2007)) are partly based on the same effectiveness data from the IDEAL trial. All three studies rely on a Markov model to estimate cost-effectiveness of statins. However, there are no further apparent relationships between the three papers.	No
IDs: 18 / 55 Alonso et al. (2008); Ramsey et al. (2008)	b	There is no apparent relationship in terms of common authorship, use of the same model, funding source or identical data sources between the papers	No
IDs: 19 / 26 / 58 Annemans et al. (2010); Khoury et al. (2009); Caro et al. (2003)	d	Studies by Annemans (2010) and Khoury (2009) were both funded by Pfizer. Both studies rely on data from CARDS. The DAM used in both studies is identical. Interestingly, though both papers are very similar and obviously related, no reference has been made by their authors to each other. The third paper (Caro et al (2003)) does not have an apparent relationship to the other two papers mentioned.	Yes

IDs: 22 / 61 Grover et al. (2008); Grover et al. (2003)	c	There is a strong link between both papers in terms of common authorship. This is also reflected in the use of the identical DAM (CHD life expectancy model) and further similarities in terms of study methods and unit cost sources etc.	Yes
IDs: 24 / 56 Greving et al. (2011); Scuffham et al. (2004)	d	There is no apparent relationship in terms of common authorship, use of the same model, funding source or identical data sources between the papers	**No**
IDs: 35 / 36 Wagner et al. (2009a); Wagner et al. (2009b)	b	There is a very strong link between both papers in terms of authorship, methods and data sources	Yes
IDs: 40 / 52 / 62 Spaans et al. (2003); HPS coll. Group (2006); Glick et al. (1992)	d	There is no apparent relationship in terms of common authorship, use of the same model, funding source or identical data sources between the papers	**No**
IDs: 42 / 61 Grover et al. (2003); Peura et al. (2008)	d	There is no apparent relationship in terms of common authorship, use of the same model, funding source or identical data sources between the papers	**No**
IDs: 43 / 53 Slejko et al. (2010); Tsevat et al. (2001)	d	There is no apparent relationship in terms of common authorship, use of the same model, funding source or identical data sources between the papers	**No**
IDs: 44 / 63 Nherera et al. (2010); NICE lipid guideline 67 (2008)	d	Strong relationship in terms of authorship model structures, study methods and data sources	Yes
IDs: 63 / 67 NICE lipid guideline 67 (2008); Ara et al. (2009)	c	Transition probabilities in the HTA report from ARA et al partly taken from the study by Cooper et al, which led to the NICE guideline 67.	Yes
IDs: 66 / 67 Ward et al. (2007); Ara et al. (2009)	d	Both studies are related through common authorship and they were also both conducted as part of the HTA Programme. Both studies were conducted by SCHARR and the DAM underlying both studies is known as the SCHARR-model	Yes

Though the findings reported above indicate that the method applied to investigate the genealogy of economic evaluation studies is very promising, there is still concern in terms of basing an alternative MLM structure upon it. For instance, the method produced roughly 50% of '*false positive*' matches on the respective biplots, and one can almost be certain that a number of '*false negative*' relationships are still hidden in the dataset. This means that the results reported above are, at most, partially disclosing existing relationships within the data. This method may, with further refinement in a systematic exercise as it is suggested for future research, be very useful to investigate links between economic evaluation studies and the extent to which studies replicate each other. However, within this project, the method did not appear to be '*sensitive*' and '*specific*' enough to base alternative MLM structures upon its findings. Potential reasons may lie, for instance, in the choice of variables to enter the MCA. As these variables were drawn from existing data which was not purposely collected for this genealogy study, one may argue that important study

characteristics were missing from this assessment. Furthermore, this case study showed that MCA results are very sensitive with respect to the categorical variables considered, which most certainly asks for a much more systematic approach to variable selection when considering MCA to assess similarities between economic evaluation studies in future research.

On the other hand, it was not possible to allocate more time to this exercise which is, though related to the question of the appropriate MLM structure, not crucially relevant for addressing the transferability problem. Changing the MLM structure by pooling individual studies to groups of studies may have an effect on the results when running the respective models. However, there are also a number of valid reasons to retain the current multilevel structure even if some papers are found to be very similar in some aspects. For instance, pooling papers would mean that the number of level two units decreases, potentially casting into doubt the assumption of random parameters on that level (Snijders et al., 2005). Furthermore, even if studies are similar in most aspects, pooling them would obviously have an impact on those study-level covariates which still show differences between those studies. For instance, within the affected studies shown in Table 4.8, timing constitutes an important difference even if all other observed study-level characteristics are identical. When pooling those studies, differential timing may no longer be assessable as a study-level covariate within the multilevel framework.

It is also important to note that parameters of studies become, in theory, fully exchangeable after adjusting for the appropriate covariates (e.g. Gelman et al., 2004), so that the question of similarity between economic evaluation studies may be entirely shifted to the matter of covariate adjustment. Finally, the current model shows level two units in its most disaggregated form. One may argue that further aggregation may also lead to false inferences because of the ecological fallacy (e.g. Hox, 2010). This holds especially true in a situation when there is no clear cut between two studies which may be pooled because of their apparent similarity, and two studies which are similar in some aspects, but not '*similar enough'* to justify pooling in the researchers judgement.

Nevertheless, though it may not be indicated to alter the MLM structures theoretically developed and tested in Chapter 3 based on the results of this genealogy study, there are important findings which may impact on the remainder of this empirical exercise. Precisely, not considering existing relationships between published economic evaluation studies with respect to the MLM structure does not mean that this project ignores such relationships altogether. In line with the assumption of partial exchangeability, one may consider links between studies in terms of covariates on study-level. First, each of the variables considered above to assess similarity between economic evaluation studies will be tested individually for significance in the multilevel framework. This analysis of covariates is reported in Chapter 5.2.

In addition, one may further look into the issue of common authorship to group studies accordingly. For this matter, authors of all 67 studies included in this empirical exercise were first listed in a spreadsheet, resulting in a list of 351 authors. Next, authors of the respective studies were ordered alphabetically, so that it was relatively straightforward to group studies of common authorship together (note that this part of the assessment did purposely not distinguish between first and co-authorship). This way, it was possible to ascertain the most frequent authors in the dataset, namely S.A. Grover and B. Jonsson, each involved in seven studies, L. Coupal (six studies) as well as T.D. Szucs and H. Zowall, with five studies each in the dataset. Finally, authors and studies were grouped according to the most frequent relationships in terms of authorship, resulting in twelve groups of studies (Table 4.9 below). It turned out that only 18 studies in the dataset were not linked through common authorship to any of the other studies considered. Considering all links between studies results in one large group of related studies and a further six smaller groups. Only considering the strongest links between studies and ignoring some relationships between groups of studies results in the above mentioned 12 groups of papers related through common authorship. Finally, two categorical variables were generated for testing in the multilevel framework in Chapter 5.2. The hypothesis is that results from studies from one group of authors may be more in the same range compared to results from other groups of authors so that variability in measures of cost-effectiveness is lower for studies of common authorship.

Table 4.9: Relationships between papers in terms of common authorship

Nr	Authors	Studies of common authorship
0	No links through common authorship	1, 7, 9, 10, 12, 17, 18, 19, 20, 23, 24, 24, 38, 46, 48, 50, 60, 65
1	Caro, Shepherd, McGuire, Klittich	2, 27, 33, 54, 58, 64
2	Roberts, Merike, Wagner, Johnson, Goetghebeur, Sullivan	26, 27, 35, 36, 43, 55
3	Grover, Coupal, Zovall, Hamilton, Lavoie	3, 4, 5, 8, 16, 22, 40, 61
4	Pandya, Taylor, Weinstein, Thompson, Drummond	30, 33, 35, 53, 59, 62, 64
5	Jonnsson, Pedersen, Wedel, Johannesson, Olsson, Kjekhus, Lindgren	6, 13, 14, 21, 36, 47, 51, 57
6	Berger, Szucs, Maerz, Schaefer, Kuntz, Klose	11, 15, 37, 45, 49, 53
7	Davies, Martikanen, Niskanen, Soini	39, 41, 42
8	Neil, Calvert, Minhas, Nherera, Thorogod, Fuller	44, 54, 59, 63
9	Ward, Ara, Pandor	27, 66, 67
10	Scuffham, Chaplin	31, 32, 56
11	Mihaylova (HPS-Group)	25, 52
12	Morris	28, 29

4.3.5. Conclusion

This case study showed that MCA is a promising method to look into similarities between existing economic evaluation studies and thereby assessing the extent to which evidence replicates itself. Applied to an extensively studied area such as statins for the primary and secondary prevention of CVD, the author could show a fair degree of overlap between phenotypic similarities between studies detected by the method, and genotypic relationships between those studies assessed within a subsequent descriptive analysis. Future research may look into more systematic ways of variable selection and additional analytic options offered by the method to refine its results. In terms of this empirical analysis, however, it is not recommended to alter the multilevel model structure but rather to acknowledge existing links between economic valuation studies in terms of covariates on study-level.

5. Empirical Analysis

In Chapter 3, a number of MLMs were developed for the integration of secondary cost-effectiveness data measured as incremental net monetary benefits (INMB). These models were then developed further into a bivariate framework for the explicit and simultaneous assessment of incremental cost and incremental effects in one model. The models were then put to the test and compared to a standard OLS regression model fitted to a pilot dataset on the cost-effectives of statins in the primary and secondary prevention of CVD. It turned out that the MLM framework consistently outperformed the OLS framework, and that the hierarchical structure assumed in each model was a key factor for the overall fit of the model. Hence, from the pilot study reported in Chapter 3.4 it got apparent that a key question for the empirical work of this project is to test which MLM works best for the purposes of this exercise. This chapter brings together the MLM methods developed in Chapter 3, the findings from the pilot study reported in Chapter 3.4, and the data obtained from a systematic review and data abstraction exercise on the cost-effectiveness of statins in the primary and secondary prevention of CVD, which was the focus of the previous Chapter 4.

The primary aim of this chapter is to run the models developed on the data abstracted to address the transferability problem of economic evaluation data in health between geographic domains. The chapter first addresses the question of the most appropriate MLM structure before the focus is on covariate adjustment on data and study-level in Section 2 of this empirical exercise. Both issues need to be addressed prior to the assessment of country-level variability – or the lack thereof – which is assessed in detail in Section 3 of this chapter. Specifically, the potential to assess country-level covariates and implications of findings with respect to the transferability of economic evaluation data between countries are considered. Finally, a case study within which random slopes are added to the model demonstrates how the multilevel framework may be applied to explicitly model variation in the data as a function of explanatory variables. The case study

shows how this '*variance function'* addresses the core of the transferability problem as it makes explicit the variability in the relationship of measures of cost-effectiveness and explanatory variables. For values of the explanatory variable for which variability in measures of cost-effectiveness is low, transferring evidence to the target domain may be rather indicated compared to values of the explanatory variable for which variability in international cost-effectiveness data is particularly high.

Accordingly, four fundamental objectives are addressed and elaborated on throughout the course of this chapter. The first objective is to determine the appropriate model structure which best fits the data on the cost-effectiveness of statins in the primary and secondary prevention of CVD. This involves not just testing which multilevel structure previously developed works well on the data collected, but also whether assumptions made in these models are justified for the data. The starting point of this first exercise is to run the models developed in Chapter 3 and tested in the pilot study on the full dataset without the inclusion of any covariates. Thereafter, it is tested whether the assumption of independence between countries is also justified for the subset of data which stems from multinational studies as not just the relevant economic evaluation literature, but also findings from the pilot study indicate that this might not be the case. It turns out that this issue is intrinsically tied to the issue of cross-classification in the data. Finally, the issue of '*shrinkage*' within the multilevel framework is addressed further, and it is elaborated on the question whether this constitutes a potential source of bias when analysing not individual patient data but rather secondary data from economic evaluation studies. MLM features such as 'weighting' are considered in response to this issue. Note that links between economic evaluation studies, e.g. in terms of common authorship, and the question whether such links shall be considered in terms of model structure or covariates, was the focus of the genealogy study reported in Chapter 4. Hence, this question is not picked up again before section two, which is concerned with covariate adjustment on data and study-level.

All models are first run as variance components models without the inclusion of any slope parameters. This serves several objectives. First, it allows to assess

whether there is variation in INMBs (as well as ΔC and ΔE in the bivariate framework), within and between studies, and also between countries included in the dataset (Steele, 2008; Rasbash et al., 2009; Hox, 2010). It also allows quantifying this variability on each level modelled (Steele, 2008; Rasbash et al., 2009; Hox, 2010). Furthermore, it can be assessed which studies, and countries, are outliers in terms of cost-effectiveness of the technology under consideration before adjusting for any covariates (Rasbash et al., 2009, CMM workshops / variance components). Finally, the variance components model serves as a benchmark for further analysis (Rasbash et al., 2009; Hox, 2010). As a result of analysis one of this empirical exercise we determine a multilevel model structure which is then carried forward to the second part of this chapter.

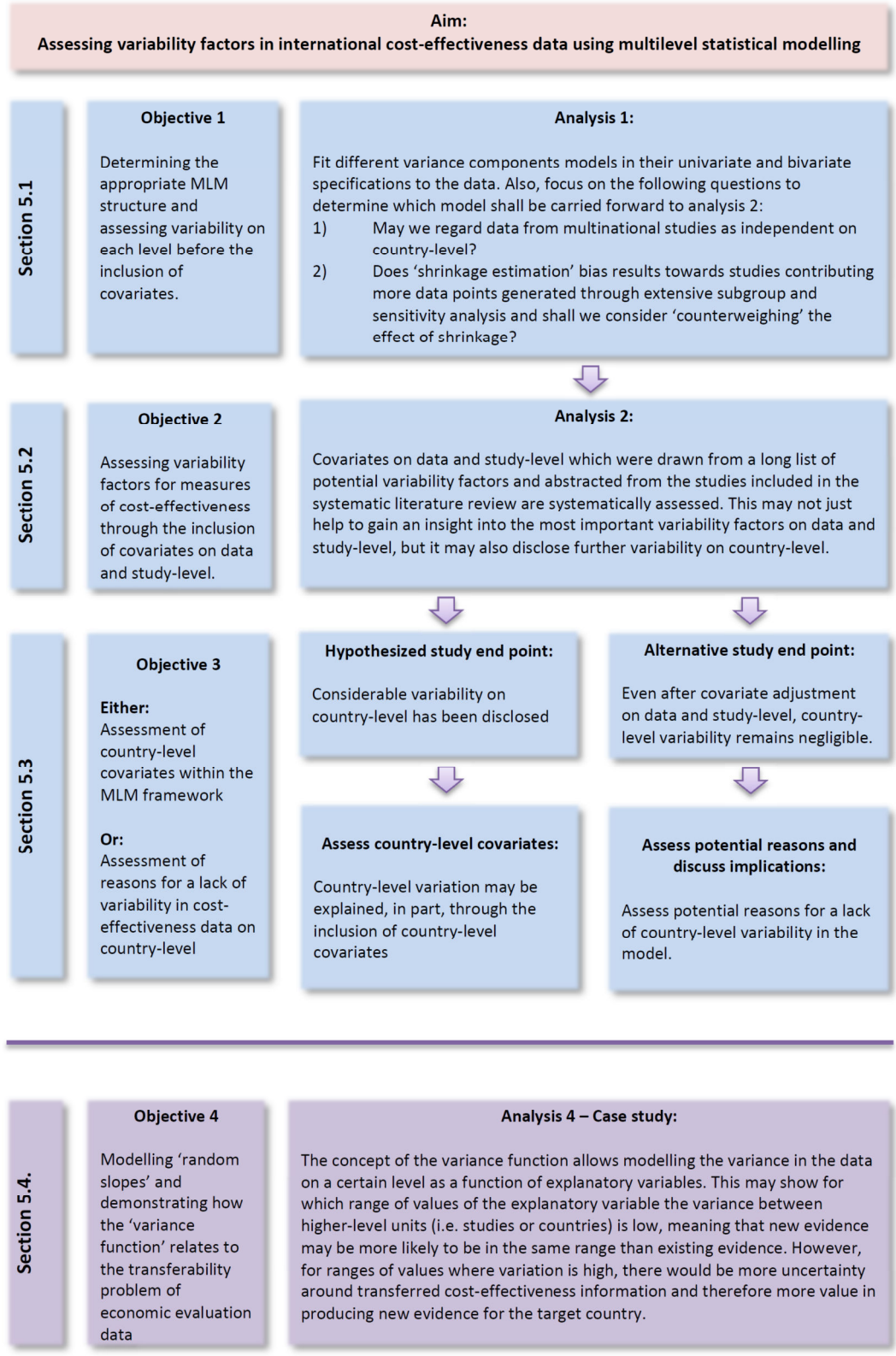
The objective for the second part of this empirical exercise is to systematically assess covariates on data and study-level which were drawn from a long list of variability factors as obtained from the relevant literature on the transferability of economic evaluation data (Sculpher et al., 2004, Goeree et al., 2007). This assessment does not just serve the aim to infer which covariates may be related to INMB, ΔC or ΔE of statins in the primary and secondary prevention of CVD. Rather, in a multilevel framework, changing anything on a lower level might have an impact on each subsequent level (e.g. Hox, 2010). Hence, the inclusion of lover level covariates in the multilevel framework could potentially disclose further variation between countries included in the dataset, so that controlling for lower-level covariates is imperative for the assessment of variability factors on country-level (Hox, 2010). Hence, the second objective is to assess whether the model carried forward from Section 5.1 of this exercise performs better in this sense after the inclusion of covariates on data and study-level.

After all the efforts taken to disclose country-level variability, it may be concluded that this source of variation is, in fact, present in the data, which then allows further analysis of potential causes of this variability between countries. This would be done by including country-level covariates in the model (Sculpher et al., 2004; Drummond et al., 2009). However, if country-level variability remains low throughout the course of this exercise, this could have a number of reasons. There may be, for instance, another model specification which fits the data better and which was not explicitly modelled in this study; or the

'*appropriate set of covariates to include'* was not yet found (Drummond et al., 2009). Finally, it could in fact mean that there is simply not much of the variability in international cost-effectiveness data due to differences between countries, as most of this variability relates to differences within and between studies included in the dataset. Assessing these questions depending on the actual amount of variability disclosed on country-level both in the univariate and bivariate framework is the focus of Section 5.3 of this chapter.

Finally, the fourth objective of this empirical exercise is to explore additional analytic features which are unique to the MLM framework. In particular, the final section of the empirical analysis looks into fitting random slopes to the model and, related to that, the concept of the '*variance function'*, where the variation in cost-effectiveness data may be explicitly modelled as a function of explanatory variables (Steele, 2008; Rasbash et al., 2009). When fitting covariates to the model in the second and third part of this exercise, it is assumed that intercepts of study regression lines vary randomly whereas the slopes of these regression lines are assumed to be fixed. This leads to parallel regression lines between studies included in the dataset (Steele, 2008; Rasbash et al., 2009). However, the relationship between measures of cost-effectiveness and explanatory variables may be different for different studies (or countries), which means that not just the intercepts, but also the regression slopes for these variables may differ between higher-level units. This has been previously acknowledged by some health economists who used MLM in the context of trial based economic evaluation (Sculpher et al., 2004; Manca et al., 2005; Thompson et al., 2006; Bachmann et al.; 2007). However, what is entirely new to the domain of economic evaluation in health is the concept of the variance function. Precisely, if one allows for random slopes, the variance on the level for which a random slope was modelled becomes a quadratic function of the explanatory variable (Steele, 2008) which may be used to express variation in measures of cost-effectiveness data as a function of covariates in the model.

Figure 5.1.: Overall analysis plan for the empirical exercise of this project



This concept relates to the transferability problem at its core as one may argue that for values of the explanatory variable for which variation in higher-level units (i.e. studies and/or countries) is low, additional evidence for the target country may also be more likely to be in the same range in terms of cost-effectiveness. In contrast, the higher this variation, the more indicated it may be to produce new target specific evidence. Hence, the variance function may be used to target research resources more specifically to those questions for which variance in international cost-effectiveness data is particularly high.

The analysis strategy as outlined above is also in accord with what is recommended in the MLM literature. A bottom up approach is usually advocated, starting with the simplest possible model and then including parameters which are tested for significance after they have been added (e.g. Hox, 2010). This procedure starts by building the fixed part of the model, and then continues to the random components (Rasbash et al., 2009; Hox, 2010). It ensures to keep the model as simple as possible and thereby complies with the 'l*ex parsimoniae*' (*Occam's Razor*). The simplest possible model is the variance components model without explanatory variables (Steele, 2008; Rasbash et al., 2009; Hox, 2010). It is therefore also referred to as '*intercepts only model'* as the regression lines in this model are parallel and '*flat*' (Hox, 2010). After having specified the variance components model, covariates are tested. This is done by proceeding from the lower level to the higher level as changes on the lower level might impact anything observed on subsequent levels (Steele, 2008; Rasbash et al., 2009; Hox, 2010). Hence, data-level covariates are tested first, and once the model is fully specified on data-level, study-level and country-level covariates may be added. Once a full random intercepts model has been specified in that way, the next step is to test the random part of the model, hence fitting random slopes of covariates. Testing random slopes is executed on a parameter by parameter basis, as not doing so might lead to an '*overparameterized model with serious estimation problems'* (Hox, 2010). It is acknowledged that covariates which were not significant with fixed slopes may be significant in a random slope specification (Rasbash et al., 2009; Hox, 2010). However, searching amongst the full dataset on a parameter by parameter basis (i.e. even separately for each category in a categorical variable) is not realistic in terms of the time effort involved and also not the purpose of the case study reported in Section 5.4, so that it was decided to scale down on this particular exercise.

The remainder of this chapter is organized alongside the four fundamental objectives defined above. Hence, the following Section 5.1. reports on various variance components models in the univariate and bivariate framework. Thereafter, covariates on data and study-level are assessed systematically for the inclusion in the multilevel model carried forward from part one of this empirical exercise. This is reported in Section 5.2. The covariates under consideration were drawn from a long list of variability factors as obtained from the relevant literature (Sculpher et al., 2004; Goeree et al., 2007) and further details about a systematic literature review and data abstraction exercise are also obtainable from the previous Chapter 4. Discussing the choice of covariates within each level of the model and each subgroup of variables constitutes a key challenge in this empirical exercise due to the theoretically unlimited space of variability factors and consequently the large number of covariates abstracted from the studies includable in this project. After having specified a full random intercepts model with data and study-level covariates in Section 5.2., the following Section 5.3 focusses on the country-level by either including covariates, or considering potential reasons for a lack of country-level variability on that level.

Independent of the outcome of the exercise as outlined above, the final Section 5.4 of this chapter demonstrates the value of fitting random slopes and to model variation in international cost-effectiveness data as a function of explanatory variables. It is argued that the concept of the '*variance function'* addresses the transferability problem as it may be applied to target research resources more specifically to those questions for which variation in measures of cost-effectiveness is particularly high.

All multilevel analyses are carried out in MLwiN (Rasbash et al., 2009a) and performed in a univariate framework with INMBs as response variable, and also in a bivariate framework with the two stochastic components of the INMB statistic (ΔC and ΔE) as a vector of response variables. Additional analyses (descriptive statistics, principal component factor analyses, multiple correspondence analyses etc.) were carried out in STATA 12.

5.1. Objective one: Determining the appropriate multilevel model structure

The aim of this empirical exercise is to address the transferability problem of health economic evaluations by fitting a multilevel model to international data on the cost-effectiveness of statins in the primary and secondary prevention of CVD. This allows the assessment of variability factors on each level of the modelled data hierarchy. In the MLM methods Chapter 3, a number of alternative model structures have been developed, starting off from a single level OLS regression and ending up with a cross-classified bivariate model which allows not just directly assessing a country-level in the presence of cross-classified data from multinational studies, but also decomposes the INMB statistic by modelling both ΔC and ΔE as a vector of response variables. These alternative model structures were then tested in a pilot study in Chapter 3.4, which was carried out on a reduced dataset on the cost-effectiveness of statins. This pilot study already showed that, due to ignoring that data in studies is not independent, the OLS regression model and the two-level model which clusters data in countries only, were clearly outperformed by those multilevel structures which did explicitly account for a study-level.

However, it was also observed that the cross-classified model, which takes into account that data is clustered in both studies and countries and also allows the inclusion of data from multinational studies which are responsible for the cross-classification problem, did not perform any better than the two-level hierarchical model which completely ignores the existence of a country-level. Not just was there no improvement in model fit due to the inclusion of a country-level, there was also a lack of noteworthy variability on the country-level itself. As a potential reason for this finding it was hypothesized that the assumption of independence between geographic domains may not be adequate for those studies in the dataset which are '*multinational*' in nature and thereby introduce the issue of cross-classification as their data may not appropriately reflect the variation in cost-effectiveness attributable to differences between the countries considered. In other words, if data from multinational studies is less affected by variability on country-level, this potentially '*lays a curtain*' over the overall country-level

variability present in the rest of the data from non-multinational studies and therefore casts into doubt whether the assumption of independence between countries actually holds for those studies. The suspicion that data from multinational studies may not be as context specific as generally assumed, e.g. due to standardised trial protocols (e.g. Ramsey et al., 2005), or the fact that not all data required to populate an economic model is country-specific (e.g. Barbieri et al., 2005), has been discussed before in the economic evaluation literature. Therefore, it is indicated to investigate whether the assumption of independence on country-level actually holds for data from multinational studies before settling on a MLM structure to be carried forward to the analysis of potential variability factors in the second part of this empirical exercise.

Finally, as this is also intrinsically tied to determining the appropriate model structure, additional assumptions required to fit a MLM to this dataset on the cost-effectiveness of statins are considered before adding covariates to the model. The fact that this exercise utilizes secondary data and not individual patient data requires further attention as the '*gravity*' of a study in a multilevel framework partly depends on the number of data points provided by that study, and this, in turn, affects the amount of shrinkage that study is subject to (Steele, 2008; Rasbash et al., 2009). Hence, as data points do not represent actual individual patients, a studies' '*weight*' in the overall model might only depend on the extensive reporting of subgroup and/or sensitivity analyses. Therefore, it is to consider whether this constitutes a potential thread to the validity of results and, if so, whether there are any strategies available, such as adding weights to individual studies or to bootstrap studies and resample data before fitting the MLM, which could efficiently counteract this potential source of bias.

The following Section 5.1.1 details the analysis strategy to investigate the appropriate multilevel structure for the remainder of this empirical exercise. Subsequently, the data and methods of analysis to carry out this assessment are explained in Sections 5.1.2 and 5.1.3. Section 5.1.4 reports and discusses the results of the running various variance components models before Section 5.1.5 picks up on the issue of shrinkage as a potential source of bias in the assessment of secondary data from international cost-effectiveness literature.

5.1.1. Plan of analysis for part one of this empirical exercise

The starting point of this analysis is the cross-classified model developed in the MLM methods Chapter 3 which was already applied to the pilot study dataset in Chapter 3.4. This model offers the most flexibility in terms of accommodating both a study and country-level and simultaneously allows for cross-classified data structures due to multinational studies being part of the dataset. However, this model also comes with considerable complexity, and its relative value depends upon its ability to capture variability on each level of the data hierarchy. Pilot study results indicate, however, that the cross-classified model may not capture a sufficient amount of country-level variability to permit the assessment of covariates on that level. Hence, if variability between countries is low, then we first need to rule out that model assumptions may cause this failure to capture variability in measures of cost-effectiveness on country-level.

As outlined in the introduction to this section, one reason may be the impact of multinational trial data on between-country variability as data from multinational trials is suspected to draw a curtain over potentially existing country-level variability in the cross-classified model. It is hypothesized that data from multinational trials shows much lower between-country variability, and this data '*infects*' country parameters in the cross-classified model and drags them towards each other in the MLM. Potential reasons for this lower country-level variability have been discussed in the literature (e.g. Ramsey et al., 2005). The easiest way to address this issue would obviously be to simply drop the data from multinational studies from any further analysis. This would result in a strictly hierarchical dataset which would allow assessing whether country-level variation would increase as compared to the cross-classified model. However, though this model is also run to receive some benchmark values on country-level variation without multinational studies in the dataset, dropping data means losing valuable observations, and especially in a multilevel context, where one needs a certain number of units on higher levels to assume random parameters (Snijders, 2005), this may not be the preferred strategy as the number of studies and countries in the dataset may be affected too. In addition, one would compare two different models being applied to two different subsets of the data,

meaning that differences in model fit are not directly comparable across both experiments.

There is an alternative analysis strategy which offers additional insights into the causes of country-level variability, or the lack thereof, in the dataset. It all comes back to the question of whether subsets of the data are actually independent. The cross-classified model assigns data from both single-country-studies and multinational studies to their respective study on study-level, hence assuming that this data is dependent within studies and independent between studies. Exactly the same is assumed on country-level though this assumption may not hold for multinational studies. Therefore, instead of dropping the affected data points, the data from multinational studies may simply be '*pooled*' in a distinct group on country-level, thereby removing its influence on other country-parameters. Obviously, as some countries in the dataset are only considered within those multinational studies, this would mean to lose some parameters on country-level. However, what is gained is the chance to analyse the full dataset with all studies originally includable in this exercise whilst still obtaining '*clean*' country-level estimates from all the other studies in the dataset. Secondly, the problem of cross-classification in the data does no longer exist so that a strictly hierarchical three-level model with data being clustered in studies and studies being clustered in countries, applies. However, most importantly this strategy addresses that multinational studies might not appropriately reflect country-level variability, which may also be the cause for (severely) underestimated country-level variation as observed in the pilot study.

In conclusion, if the suspicion that data from multinational studies may draw a curtain over existing country-level variation is supported by the data, much higher country-level variation would be observed when running such a three-level hierarchical model on a dataset where data from multinational studies is pooled in a separate group on country-level. To rule out that the model architecture itself conceals country-level variability, a cross-classified model is also run on an '*intermediate*' dataset, where some data from multinational studies has been pooled on country-level, and the rest of that data has been assigned to their respective target countries. If the problem lies in the data and not in the model architecture that allows for cross-classification, then country-

level variation would be somewhere in between the fully cross-classified model and the alternative three-level model which pools all data from multinational studies on country-level.

It is important to note that, instead of pooling data from multinational studies on country-level, it was also considered to do the opposite, namely splitting data from these studies on study-level. This would introduce a strict 1:1 relationship between '*hypothetical*' studies and the respective countries modelled in the affected studies and thereby also eliminate the cross-classification problem. However, though the number of countries would remain unaffected, this would spuriously inflate the number of studies on study-level. Doing so would have strong implications on the random part of the model, potentially resulting in severely underestimated study-level variability (e.g. Steele, 2008). To understand this issue it is useful to draw an analogy to the OLS regression model which overestimates precision for higher-level covariates if the underlying data is not independent (Steele, 2008; Rasbash et al., 2009; Hox, 2010). Furthermore, whilst the assumption of dependent country-level data within those multinational studies is supported by the relevant literature (e.g. Ramsey et al., 2005; Barbieri et al., 2005), the assumption of independent data on study-level with respect to those countries is most certainly not. On top of that, if country-level variability increased in such a model, this could simply be due to some of the study-level variability being '*dragged*' to the country-level because of the strict 1:1 relationship between hypothetical study-groups and countries within these multinational studies. Finally, comparing the cross-classified model to a three-level hierarchical model where data from multinational studies is pooled on country-level allows looking into the issue of whether multinational study data shows lower country-level variability, which constitutes in itself an important finding from this empirical exercise. On the other hand, splitting data on study-level would not permit any additional insights into this important issue.

In conclusion, models in this first exercise will be run on:

- the fully cross-classified dataset
- A cross-classified dataset where data from multinational studies for some countries is pooled in one group, whilst data for other countries has been assigned to their respective country groups (intermediate dataset)
- the full dataset where cross-classification has been completely eliminated due to pooling data from multinational studies on country-level (hierarchical dataset)
- a reduced dataset where all multinational studies were dropped for comparative purposes (reduced dataset)

The main models of interest are the cross-classified MLM, the three-level hierarchical model and, for comparative purposes following the results from the pilot study, the two-level hierarchical model which ignores the existence of a country-level. In addition, the two-level model which clusters data in countries only and the OLS regression model, which completely ignores complex data structures, are also implemented to confirm findings from the pilot study in terms of their worse fit compared to the above mentioned model specifications. Table 5.1 shows the analysis plan for this first exercise with the dataset being subject to different assumptions as introduced above on the vertical axis and the respective MLMs to fit on the horizontal axis.

Table 5.1: Analysis strategy for the first exercise to determine the appropriate multilevel model structure

Dataset / Models	Cross-classified dataset	Intermediate dataset	Hierarchical dataset	Reduced dataset
Cross-classified model	X	X	--	--
Three-level hierarchical model	--	--	X	X
Two-level model with data clustered in studies	X			X
Two-level model with data clustered in countries	X	X	X	X
OLS regression model	X			X

Because assumptions regarding independence on country-level as introduced above leave the data and study-levels unaffected, the two-level model which clusters data in studies only yields identical results for the cross-classified, the intermediate and the hierarchical dataset. Only the reduced dataset may yield different results due to all data points from multinational studies being dropped from this analysis. The same holds for the OLS regression model as this specification does not consider any complex data structures. Only the two-level hierarchical model, which clusters data in countries, may differ in its results between different assumptions regarding the data on country-level as these directly impact on the respective model output.

Before further specifying all models to be run in this exercise, the following section introduces the data which is used for this analysis in more detail.

5.1.2. Data for exercise one

Chapter 4 already reported on a systematic literature review on the cost-effectiveness of statins in the primary and secondary prevention of CVD and explained how the form for abstracting data from these studies was developed from a long list of potential variability factors drawn from the relevant economic evaluation literature and used to populate a dataset for this empirical exercise. This chapter also reports on the efforts taken to prepare the dataset for further analysis and the resulting dataset is now being used for the analysis as outlined above. The intervention of statins in the primary and secondary prevention of CVD was chosen as it has been extensively researched in the past, meaning that a sufficient number of includable studies and geographic locations was hypothesized to be present in the data to justify the assumption of random parameters on study and country-level which is crucial for fitting multilevel models (Spiegelhalter et al., 2000; Spiegelhalter et al., Rasbash et al., 2009; 2004; Hox, 2010).

During the systematic literature review, 67 studies were found to be includable in this exercise, providing in total 2094 estimates of incremental net monetary

benefit where the authors also decomposed the INMB statistic, hence explicitly reporting data on ΔC and ΔE of the healthcare intervention. This data is clustered in 23 different geographic locations. As some data is applicable to the UK as a whole, but other data only to England/Wales on the one hand or Scotland on the other, three distinct categories were introduced to the dataset to reflect these geographic entities. Table 5.2 provides an overview of the geographic locations represented in the data and Figure 5.2 is a Venn-diagram which groups data according to the respective studies being '*multinational*' or '*single country*' in nature. Note that Table 5.2 differs from Table 4.5 in Chapter 4.2.2 as it shows the distribution of data points – not studies – per country.

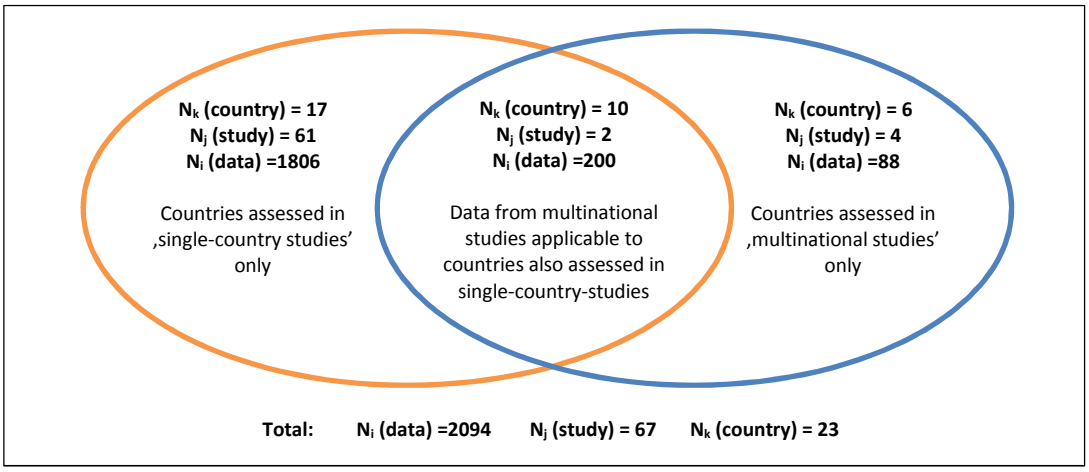
Table 5.2: Geographic locations represented in the dataset

Country	Frequency	In %	Cummulative	Data points for that country from multinational studies	% of data points for that country from multinational studies
Australia	13	0.62%	0.62%	1	7.69%
Belgium	30	1.43%	2.05%	8	26.67%
Brazil	2	0.1%	2.15%	--	--
Canada	422	20.15%	22.30%	16	3.79%
Denmark*	13	0.62%	22.92%	13	100%
Finland	39	1.86%	24.79%	13	33.33%
France*	24	1.15%	25.93%	24	100%
Germany	133	6.35%	32.28%	39	29.32%
Hong Kong	8	0.38%	32.66%	--	--
Hungary	4	0.19%	32.86%	--	--
Italy*	28	1.34%	34.19%	28	100%
Japan	6	0.29%	34.48%	--	--
Netherlands	70	3.34%	37.82%	--	--
New Zealand*	1	0.05%	37.87%	1	100%
Norway*	10	0.48%	38.35%	10	100%
Portugal*	12	0.57%	38.92%	12	100%
Spain	40	1.91%	40.83%	36	90%
Sweden	81	3.87%	44.70%	34	41.98%
Switzerland	3	0.14%	44.84%	1	33.33%
UK	408	19.48%	64.33%	36	8.82%
UK (Engl./Wales)	475	22.68%	87.01%	--	--
UK (Scotland)	11	0.53%	87.54%	--	--
USA	261	12.46%	100%	16	6.13%
Total	**2094**	**100%**	**100%**	**288**	13.75%
* Data for this country was only available from multinational studies					

As can be seem from Table 5.2 and Figure 5.2, the vast majority of data stems from single-country-studies with 1806 data points from 61 studies applicable to 17 geographic domains. On the other hand, 16 countries were considered in six multinational studies yielding a total of 288 data points. These 288 data points from multinational studies introduce the cross-classification problem to the dataset as it causes the strict hierarchical structure between studies and countries in the remaining data to break down. Of the 16 geographic domains

considered in these multinational studies, six only appear in those studies adding up to 88 data points, meaning that none of the single-country-studies provides data for any of those countries (Denmark, France, Italy, Netherlands, New Zealand and Norway). The remaining ten countries considered in multinational studies add up to 200 data points which, in the cross-classified model, would impact on country-parameters of those countries which were also considered in single-country-studies. Table 5.2 shows that this impact ranges between roughly 4% (Canada) to up to 90% (Spain) of all data points for those ten countries being drawn from multinational studies. It is hypothesized that this fact, plus the existence of countries only considered in multinational studies, significantly reduces country-level variation in the fully cross-classified model as data from multinational studies severely underestimates country-level variation in the cost-effectiveness of statins for the primary and secondary prevention of CVD.

Figure 5.2: Venn-diagram ordering country-specific data according to the nature of the underlying economic evaluation study



The easiest way to eliminate this impact is obviously be to drop all six studies from further analyses. To yield estimates of country-level variability without the influence of data from multinational studies, a three-level hierarchical model is implemented on this reduced dataset for comparative purposes. In a sense it constitutes the '*opposite extreme*' to the full, cross-classified dataset, resulting in 1806 data-points clustered in 61 studies and 17 countries. However, an alternative solution is to pool data from multinational studies on country-level. This allows keeping data from all multinational studies in the dataset and only reduces the number of parameters on country-level by the respective amount of

countries considered only in multinational studies. Hence, this dataset, which is also strictly hierarchical in nature, consists of 2094 data points from 67 studies and 17+1 countries (the additional country-group refers to the data from multinational studies being pooled in a separate group on country-level).

In addition, an '*intermediate solution*' is considered which also reduces the impact of data from multinational studies on country-parameters, but still requires fitting a cross-classified model to the data. All six countries which only appear in multinational studies (Denmark, France, Italy, New Zealand, Norway and Portugal) are still considered as separate groups on country-level. On the other hand, data from multinational studies which refers to one of the ten countries already existing in the dataset (i.e. the '*overlapping*' part of the Venn-diagram in Figure 5.2.) is pooled in a separate group on country-level. This means that 88 data points from multinational studies are assumed to be country specific, resulting in the maximal number of 23 country-specific groups on that level. The remaining 200 data points from multinational studies are pooled in a further group on country-level, so that this dataset consists of 2094 data points from 67 studies and 23+1 countries. The value of this additional analysis lies in the fact that it introduces cross-classification on a '*lower scale*', and if low country-level variability in the fully cross-classified model as observed in the pilot study is only due to the data from multinational studies showing much less variability then the rest of the data, then running a MLM on this intermediate dataset would result in country-level variation being somewhere in between the fully cross-classified dataset and the three-level hierarchical dataset where all data points from multinational studies are being pooled on country-level.

Table 5.3 below summarizes the four resulting datasets which are used to assess the appropriate model structure for further analysis. Further details on the studies included in this empirical exercise are also available from Chapter 4, which reports on the systematic literature review on the cost-effectiveness of statins for the primary and secondary prevention of CVD.

Table 5.3: Datasets for determining the appropriate multilevel model structure

Dataset Nr. and name	Assumptions regarding data structure	Data structure	N_i (data)	N_j (study)	N_k (country)	Data points from multinational studies pooled on country-level
1a cross-classified	data within multinational studies is independent on country-level ➔ All data from multinational studies grouped to respective target countries	Cross-classified	2094	67	23	0
1b inter-mediate	Part of the data from multinational studies is dependent on country-level. ➔ Data from multinational studies for ten countries pooled on country-level	Cross-classified	2094	67	23+1	200
1c hier-archical	All data from multinational studies dependent on country-level ➔ All data from multinational studies pooled on country-level	3-level hier-archical	2094	67	17+1	288
1d reduced	All data from multinational studies dependent on country-level ➔ All data from multinational studies dropped	3-level hier-archical	1806	61	17	0

Now that several variants of the dataset for assessing the appropriate MLM structure have been introduced (which all differ by the assumptions made regarding independence of cost-effectiveness data on country-level), the actual variables required for this analysis are introduced next. As only the response variables (INMB, ΔC and ΔE) are modelled as clustered in studies and geographic locations without the inclusion of any covariates, the few variables required are i) an ID-variable for the study the data point was drawn from, ii) an ID-variable for the country the data point refers to and iii) data on the response variables of interest (i.e. INMB, ΔC and ΔE). As all data points obviously refer to a specific study, the country of interest was always specified in each study, and as the reporting of INMBs, ΔC and ΔE was a study inclusion criterion, there is no data missing in any of these variables. Descriptive statistics for the three response variables can be found below in Table 5.4.

Table 5.4: Descriptive statistics of response variables for exercise one

Variable	Obs	Mean	Std. Dev.	Min	Max
ΔC	2094	£ 8871.77	£ 15061.75	£ -3688.47	£ 178653.1
ΔE	2094	0.636	0.908	-.03	5.4
INMB	2094	£ 10209.56	£25422.09	£ -151053.1	£ 150430.5

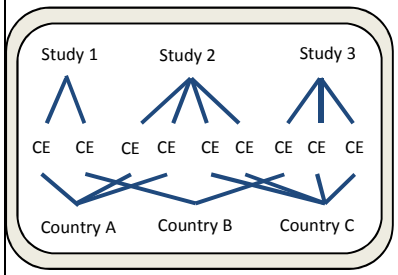
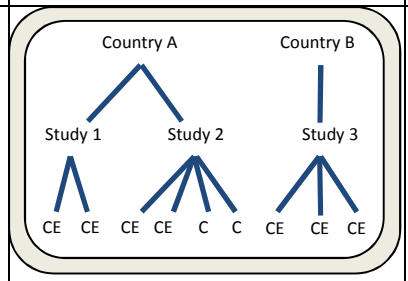
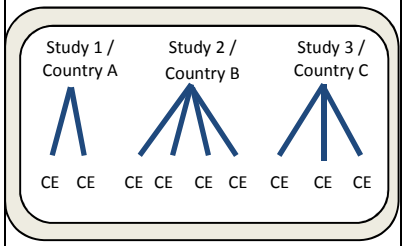
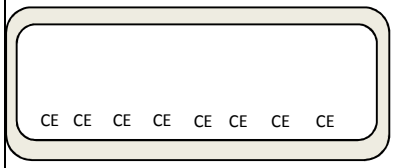
Obviously, when multiplying the raw mean ΔE of 0.636 with the threshold value λ of 30.000 assumed in this exercise and subtracting the raw mean ΔC of 8871.77, one arrives almost exactly at the mean INMB of £ 10209.56 measured in 2010 £-Sterling.

5.1.3. Methods of analysis for exercise one

As detailed in the plan of analysis, the models of interest in this first exercise are the cross-classified specification with data nested in studies and geographic locations, where studies and countries are cross-classified and, secondly, the three-level hierarchical model, where data is clustered in studies and studies are clustered in countries respectively. Both models are compared to a simpler two-level hierarchical model which groups data in i) studies only and ii) countries only. In addition, the respective OLS regression model, which does not take into account complex data structures, is implemented to confirm findings from the pilot study. All models are run in a univariate version with INMB as the only response variable and a bivariate version with ΔC and ΔE as a vector of response variables. Table 5.5 shows the unit diagrams as well as the respective algebraic forms of these models. Details on complex data structures, MLM methodology, and the algebra presented below are also provided in Chapter 3.

Table 5.5 summarizes all five models to be run in this first exercise to determine the appropriate multilevel model structure. As mentioned, models 1.a and 1.b both take into account that international cost-effectiveness data is nested in the studies it was drawn from and the countries it refers to. However, model 1.a. assumes that studies and countries are cross-classified due to the data drawn from multinational studies being grouped to their respective target countries whilst model 1.b assumes a strict hierarchical order between studies and countries as data from multinational studies may either be dropped or pooled in a separate group on country-level. Models 1.c and 1.d are both two-level models which either assume that data is only clustered in the studies it was drawn from (model 1.c) or the countries it refers to (model 1.d). Model 1.e. is an empty OLS regression model which ignores the existence of complex structures in the data.

Table 5.5. Multilevel Models for Exercise one

Model summary	Unit diagrams	Univariate model specification	Bivariate model specification
Model 1 a Cross-classified variance components model with cost-effectiveness data clustered in studies and countries where studies and countries are cross-classified.		$y_{i(jk)} \sim N(XB, \Omega)$ $y_{i(jk)} = \beta_0 + u_{0k} + u_{0j} + e_{0i(jk)}$ With $u_{0k} \sim N(0, \sigma^2_{u0})$ $u_{0j} \sim N(0, \sigma^2_{u0})$ $e_{0i(jk)} \sim N(0, \sigma^2_{e0})$	$\begin{bmatrix} Y_{0,i(jk)} \\ Y_{1,i(jk)} \end{bmatrix} \sim BVN(XB, \Omega)$ $y_{d,i(jk)} = (\beta_{0d} + u_{0dk} + u_{0dj} + e_{0di(jk)}) * r_{d,i(jk)}$ $r_{1,i(jk)} = \begin{Bmatrix} 1 \; if \; \Delta \; cost \\ 0 \; if \; \Delta \; effects \end{Bmatrix} \quad r_{2,i(jk)} = 1 - r_1$ With: $\begin{bmatrix} u_{0,0,k} \\ u_{0,1,k} \end{bmatrix} \sim BVN\,(0, \Omega_u)$ where $\Omega_u = \begin{bmatrix} \sigma^2_{u0,0} & \\ \sigma_{u0,01} & \sigma^2_{u0,1} \end{bmatrix}$ $\begin{bmatrix} u_{0,0,j} \\ u_{0,1,j} \end{bmatrix} \sim BVN\,(0, \Omega_u)$ where $\Omega_u = \begin{bmatrix} \sigma^2_{u0,0} & \\ \sigma_{u0,01} & \sigma^2_{u0,1} \end{bmatrix}$ $\begin{bmatrix} e_{0,0,i(jk)} \\ e_{0,1,i(jk)} \end{bmatrix} \sim BVN\,(0, \Omega_{ue})$ where $\Omega_e = \begin{bmatrix} \sigma^2_{e0,0} & \\ \sigma_{e0,01} & \sigma^2_{e0,1} \end{bmatrix}$
Model 1 b Three-level hierarchical variance components model with cost-effectiveness estimates being nested in economic evaluation studies and studies being nested in geographic domains.		$y_{ijk} \sim N(XB, \Omega)$ $y_{ijk} = \beta_0 + v_{0k} + u_{0jk} + e_{0ijk}$ With $v_{0k} \sim N(0, \sigma^2_{v0})$ $u_{0jk} \sim N(0, \sigma^2_{u0})$ $e_{0ijk} \sim N(0, \sigma^2_{e0})$	$\begin{bmatrix} Y_{0,ijk} \\ Y_{1,ijk} \end{bmatrix} \sim BVN(XB, \Omega)$ $y_{d,ijk} = (\beta_{0d} + v_{0dk} + u_{0djk} + e_{0dijk}) * r_{d,ijk}$ $r_{1,ijk} = \begin{Bmatrix} 1 \; if \; \Delta \; cost \\ 0 \; if \; \Delta \; effects \end{Bmatrix} \quad r_{2,ijk} = 1 - r_1$ With: $\begin{bmatrix} v_{0,0,k} \\ v_{0,1,k} \end{bmatrix} \sim BVN\,(0, \Omega_v)$ where $\Omega_v = \begin{bmatrix} \sigma^2_{v0,0} & \\ \sigma_{v0,01} & \sigma^2_{v0,1} \end{bmatrix}$ $\begin{bmatrix} u_{0,0,jk} \\ u_{0,1,jk} \end{bmatrix} \sim BVN\,(0, \Omega_u)$ where $\Omega_u = \begin{bmatrix} \sigma^2_{u0,0} & \\ \sigma_{u0,01} & \sigma^2_{u0,1} \end{bmatrix}$ $\begin{bmatrix} e_{0,0,ijk} \\ e_{0,1,ijk} \end{bmatrix} \sim BVN\,(0, \Omega_e)$ where $\Omega_e = \begin{bmatrix} \sigma^2_{e0,0} & \\ \sigma_{e0,01} & \sigma^2_{e0,1} \end{bmatrix}$
Models 1.c and 1.d Two-level variance components model with either: - cost-effectiveness data clustered in studies only (model 1.c) or - cost-effectiveness data clustered in countries only (model 1.d)		$y_{ij} \sim N(XB, \Omega)$ $y_{ij} = \beta_0 + u_{0j} + e_{0ij}$ With $u_{0j} \sim N(0, \sigma^2_{u0})$ $e_{0ij} \sim N(0, \sigma^2_{e0})$	$\begin{bmatrix} Y_{0,ij} \\ Y_{21ij} \end{bmatrix} \sim BVN(XB, \Omega)$ $y_{d,ij} = (\beta_{0d} + u_{0dj} + e_{0dij}) * r_{d,ij}$ $r_{1,ij} = \begin{Bmatrix} 1 \; if \; \Delta \; cost \\ 0 \; if \; \Delta \; effects \end{Bmatrix} \quad r_{2,ij} = 1 - r_1$ With: $\begin{bmatrix} u_{0,0,j} \\ u_{0,1,j} \end{bmatrix} \sim BVN\,(0, \Omega_u)$ where $\Omega_u = \begin{bmatrix} \sigma^2_{u0,0} & \\ \sigma_{u0,01} & \sigma^2_{u0,1} \end{bmatrix}$ $\begin{bmatrix} e_{0,0,ij} \\ e_{0,1,ij} \end{bmatrix} \sim BVN\,(0, \Omega_e)$ where $\Omega_e = \begin{bmatrix} \sigma^2_{e0,0} & \\ \sigma_{e0,01} & \sigma^2_{e0,1} \end{bmatrix}$
Model 1.e Empty ordinary least squares regression model which ignores complex data structures		$y_i \sim N(XB, \Omega)$ $y_i = \beta_0 + e_i$ With $e_i \sim N(0, \sigma^2_e)$	$\begin{bmatrix} Y_{0,i} \\ Y_{1,i} \end{bmatrix} \sim BVN(XB, \Omega)$ $y_{d,i} = (\beta_{0d} + e_{di}) * r_{d,i}$ $r_{1,i} = \begin{Bmatrix} 1 \; if \; \Delta \; cost \\ 0 \; if \; \Delta \; effects \end{Bmatrix} \quad r_{2,i} = 1 - r_1$ With: $\begin{bmatrix} e_{0,i} \\ e_{1,i} \end{bmatrix} \sim BVN\,(0, \Omega_e)$ where $\Omega_e = \begin{bmatrix} \sigma^2_{e0i} & \\ \sigma_{e01i} & \sigma^2_{e1i} \end{bmatrix}$

All models consist of an intercept term 'β_0', and an error term for each hierarchical level. The subscripts refer to these levels with 'i' representing level one, 'j' level two, and 'k' level three so that 'v_{0k}' is the error term for level three, 'u_{0jk}' the error term for level two and 'e_{0ijk}'the error term for level one respectively. As the cross-classified model, though conceptually a three-level model, is regarded as a two-level structure with countries and studies cross-classified at level two, the error terms are denoted with 'u_{0k}' and 'u_{0j}' respectively (Rasbash et al., 2009; Hox, 2010). For the univariate specification, it is assumed that INMB is normally distributed at each level of the model which is denoted with '$y_{ijk} \sim N(XB, \Omega)$'.

Decomposing the INMB statistic offers a number of advantages in this exercise (Nixon et al., 2005; Pinto et al., 2005; Manca et al., 2007; Grieve et al., 2007; Bachmann et al., 2007; Willan et al., 2008; Grieve et al., 2010). First, there is no need to run models at different threshold values λ, which is necessary to combine ΔC and ΔE to INMBs. Secondly, the correlation between the two stochastic components of the INMB statistic is explicitly modelled. Finally, once covariates are included, a bivariate model allows assessing the differential impact of covariates on each response variable whilst acknowledging that ΔC and ΔE are, themselves, correlated. The hierarchical structure assumed in these bivariate models is exactly identical to the hierarchy assumed in their univariate counterparts. However, a bivariate normal distribution is now assumed for the two response variables ΔC and ΔE. Furthermore, a response indicator 'r' is included which is 1 for ΔC and 0 for ΔE and a separate level for this response indicator has been fitted below the data-level. Finally, the bivariate model estimates one error variance for each response variable plus their respective covariance on each level. Again, further details on the multilevel methodology applied in this empirical exercise are also available from Chapter 3.

All datasets and the respective models to fit have now been specified for this first experiment. Hence, the next step is to actually implement and run these models within the software environment MLwiN using Markov Chain Monte Carlo (MCMC) estimation (Rasbash et al., 2009a; Browne, 2012). Though all models could also be implemented using iterative generalised least squares

(IGLS), it is a far more complex procedure, which is why it has been strongly advocated to use MCMC for cross-classified and especially for bivariate models in MLwiN (Rasbash et al., 2009 and personal communication with Professor W.J. Browne, CMM, Bristol). A detailed step by step guide on how to implement all models in MLwiN can be found in Appendix 3. The statistical software package MLwiN was chosen for this exercise as it is special MLM software with unique capacities to specify complex data structures as developed in this empirical exercise (Rasbash et al., 2009a). Though it is acknowledged that there are other software applications allowing to fit MLMs (e.g. HLM, STATA or R),the features offered by MLwiN proved particularly useful to deal with complex structures such as cross-classification, or multivariate MLMs, or even a combination thereof. In addition, the Centre for Multilevel Modelling in Bristol, which developed the software package MLwiN, is sponsored through the UK Economics and Social Science Research Council, which enables free access to an unrestricted version of MLwiN to all researchers based in a UK academic institution.

The following section reports on the results of this first exercise to determine the appropriate multilevel structure for secondary cost-effectiveness data from different geographic domains.

5.1.4. Results of exercise one

Table 5.6 provides the results of running the univariate models and Table 5.7 shows the corresponding results for the bivariate model specifications. When looking at the fixed part of each model (i.e. their intercepts) within the univariate framework first, one may conclude that results of the cross-classified and three-level hierarchical model (models 1.a. and 1.b) are very much in the same range (between £7191 and £7266). Only when applying the three-level hierarchical model to the reduced dataset, the intercept is slightly lower, producing an overall mean INMB of £6606 measured in 2010 £-Sterling. Assumptions about the country-level do not affect the two-level hierarchical model which clusters data in studies only (model 1.c.), which is why this model has only been run once with the full dataset and once with the reduced dataset. Results correspond very

strongly with models 1.a and 1.b (£7269 for the full dataset and £6707 for the reduced dataset without data from multinational trials). However, intercepts of the two-level model which clusters data in countries (model 1.d) and the OLS regression model (model 1.e) clearly differ from the results of the other models. As the intercept of an empty OLS regression model simply reduces to the raw mean of the response variable, results reported for model 1.e in Table 5.6 are identical to the raw mean INMB reported in Table 5.4. Interestingly, the two-level model, which clusters data in countries only, corresponds much stronger with this OLS regression model for the cross-classified and intermediate datasets whilst not assuming independence of data from multinational studies on country-level clearly affects the intercept of model 1.d. This observation already indicates the importance of appropriately reflecting complex data structures and making reasonable assumptions about dependencies in the data when analysing secondary cost-effectiveness data from published economic evaluation studies.

The intercepts of the bivariate model specifications reported in Table 5.7 are considered next. First, for the bivariate model, ΔC was linearly transformed by dividing each value of this response variable by 100. Though theoretically irrelevant, MLwiN may encounter convergence problems if, in a multivariate model, there is a large difference in error variances between the different response variables; as it is the case for the error variances of ΔC and ΔE (personal communication with Professor W.J. Browne and R. Pillinger, CMM, Bristol). When looking at the results, a reassuring observation is that re-combining the two response variables ΔC and ΔE to INMB's assuming a threshold value λ of £30,000 resembles the values reported for the intercepts of the univariate models in Table 5.6, though some variation is likely due to the random nature of the MCMC estimation process and the different model specifications (Browne, 2012). As for the univariate model specifications, models 2.a. and 2.b produce very similar intercepts. Likewise, model 2.c, which clusters data in studies only, shows similar results to the cross-classified and the three-level hierarchical model. However, when decomposing the INMB statistic, it suddenly becomes apparent that the similarity in the intercepts between model 1.d and 1.e in the univariate framework is likely to be a coincidence only as, in the bivariate framework, the corresponding values for ΔC and ΔE differ sharply. Only re-combining ΔC and ΔE to INMB' leads to the observed resemblance in results as in

the univariate models. This is a very compelling argument for decomposing the INMB statistic and indicates that some variability in measures of cost-effectiveness may simply '*disappear*' when combining ΔC and ΔE to INMBs.

Hence, a first look at the variance components models reported in Tables 5.6 and 5.7 already reveals that different assumptions about the hierarchical structure of a dataset may have strong implications on the fixed part of the model. However, if models which only differ with respect to their hierarchical structure do produce different results, one needs to choose one model specification which may fit the data best. For this purpose, one may compare the deviance of each model (Steele, 2008; Bartholomew et al., 2008; Rasbash et al, 2009; Hox, 2010). Generally, when estimating a multilevel model using IGLS, the deviance is the difference in the -2*log(likelihood) values for a fitted model and a saturated model (Steele, 2008; Bartholomew et al., 2008; Rasbash et al, 2009; Hox, 2010). When estimating a model using MCMC, the deviance information criterion (DIC) is the mean deviance at each iteration of the MCMC estimation procedure (Browne, 2012). This DIC also accounts for the number of parameters in the model so that the values reported in Tables 5.6 and 5.7 are directly comparable between competing model specifications (Browne; 2012). However, as Browne (2012) clarifies, the '*stochastic nature of the MCMC algorithm leads to some random variability in the DIC diagnostic depending on starting values and random number seeds*'. If differences in the DIC diagnostic are small, this should hence be confirmed with different seeds and/or starting values. (Browne, 2012)

Comparing the DIC diagnostic across different model specifications in Tables 5.6 and 5.7 clearly confirms the findings from the pilot study. The OLS regression model, which fails to capture any complex data structures, and the two-level hierarchical model, which clusters data in countries only, thereby ignoring that data is also clustered in economic evaluation studies, are clearly outperformed by the more elaborated model specifications which take into account the existence of a study-level. This fact holds both within the univariate and bivariate framework.

Table 5.6: Results for running univariate models in exercise one:

Dataset \ Model		Model 1.a Cross-classified model (data in studies and countries)	Model 1.b Three-level hierarchical model (data in studies and countries	Model 1.c. Two-level hierarchical model (data in studies)	Model 1.d. Two-level hierarchical model (data in countries)	Model 1.e. OLS regression model (no complex data structures)
		INMB	INMB	INMB	INMB	INMB
1.a. Cross-classified dataset $N_k = 23$ $N_j = 67$ $N_i = 2094$	Intercept (λ=£30,000)	7241	--	7269	10244	10207
	σ^2_{u0k} (Country)	247200	--	--	141402544	--
	σ^2_{u0j} (Study)	252676864	--	253045104	--	--
	σ^2_{e0} (Data)	290686048	--	290676352	557439040	646884928
	VPC - country	0.05%	--	--	20.23%	--
	VPC – study	46.48%	--	46.54%	--	--
	VPC - data	53.47%	--	53.46%	79.77%	100%
	DIC	46749	--	46749	48112	48424
Dataset		**INMB**	**INMB**	**INMB**	**INMB**	**INMB**
1.b. Intermediate dataset $N_k = 23+1$ $N_j = 67$ $N_i = 2094$	Intercept (λ=£30,000)	7266	--	**Same as above** (different assumptions regarding the data on country-level do not affect estimation results as the existence of a country-level has not been acknowledged in this model)	10389	**Same as above** (different assumptions regarding the data on country-level do not affect estimation results as the existence of a country-level has not been acknowledged in this model)
	σ^2_{u0k} (Country)	312401	--		149956464	
	σ^2_{u0j} (Study)	252744640	--		--	
	σ^2_{e0} (Data)	290719328	--		546511168	
	VPC - country	0.06%	--		21.53%	
	VPC – study	46.48%	--		--	
	VPC - data	53.46%	--		78.47%	
	DIC	46748	--		48111	
Dataset		**INMB**	**INMB**	**INMB**	**INMB**	**INMB**
1.c. Hierarchical dataset $N_k = 17+1$ $N_j = 67$ $N_i = 2094$	Intercept (λ=£30,000)	--	7191	**Same as above** (different assumptions regarding the data on country-level do not affect estimation results as the existence of a country-level has not been acknowledged in this model)	7725	**Same as above** (different assumptions regarding the data on country-level do not affect estimation results as the existence of a country-level has not been acknowledged in this model)
	σ^2_{u0k} (Country)	--	2482226		89529856	
	σ^2_{u0j} (Study)	--	251875168		--	
	σ^2_{e0} (Data)	--	290664096		555042432	
	VPC - country	--	0.46%		13.89%	
	VPC – study	--	46.21%		--	
	VPC - data	--	53.33%		86.11%	
	DIC	--	46749		48103	
Dataset		**INMB**	**INMB**	**INMB**	**INMB**	**INMB**
1.d. Reduced dataset $N_k = 17$ $N_j = 67$ $N_i = 1806$	Intercept (λ=£30,000)	--	6606	6707	6308	7885
	σ^2_{u0k} (Country)	--	4017363	--	62504128	--
	σ^2_{u0j} (Study)	--	233477872	236665968	--	--
	σ^2_{e0} (Data)	--	257023552	256972576	473493920	541458816
	VPC - country	--	0.81%	--	11.66%	--
	VPC – study	--	47.21%	47.94%	--	--
	VPC - data	--	51.97%	52.06%	88.34%	100%
	DIC	--	40097	40097	41200	41443

Table 5.7: Results for running bivariate models in exercise one:

Dataset \ Model		Model 2.a Cross-classified model (data in studies and countries)		Model 2.b Three-level hierarchical model (data in studies and countries		Model 2.c. Two-level hierarchical model (data in studies)		Model 2.d. Two-level hierarchical model (data in countries)		Model 2.e. OLS regression model (no complex data structures)	
		ΔC/100	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**
1.a. Cross-classified dataset $N_k = 23$ $N_j = 67$ $N_i = 2094$	Intercept (λ=£30,000)	62.36	0.448	--	--	65.18	0.456	53.47	0.494	88.72	0.636
	σ^2_{u0k} (Country)	212	0.004	--	--	--	--	2937	0.298	--	--
	σ^2_{u0j} (Study)	8161	0.419	--	--	8073	0.427	--	--	--	--
	σ^2_{e0} (Data)	10857	0.314	--	--	10867	0.313	18933	0.592	22717	0.826
	VPC - country	1.10%	0.54%	--	--	--	--	13.43%	33.48%	--	--
	VPC – study	42.44%	56.85%	--	--	42.62%	57.70%	--	--	--	--
	VPC - data	56.46%	42.61%	--	--	57.38%	42.30%	86.57%	66.52%	100%	100%
	DIC	28734		--		28734		31262		32134	
Dataset		**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**
1.b. Intermediate dataset $N_k = 23+1$ $N_j = 67$ $N_i = 2094$	Intercept (λ=£30,000)	62.06	0.438	--	--	**Same as above** (different assumptions regarding the data on country-level do not affect estimation results as the existence of a country-level has not been acknowledged in this model)		58.36	0.529	**Same as above** (different assumptions regarding the data on country-level do not affect estimation results as the existence of a country-level has not been acknowledged in this model)	
	σ^2_{u0k} (Country)	680	0.028	--	--			4863	0.388		
	σ^2_{u0j} (Study)	7302	0.386	--	--			--	--		
	σ^2_{e0} (Data)	10874	0.314	--	--			18776	0.587		
	VPC - country	3.61%	3.85%	--	--			20.57%	39.79%		
	VPC – study	38.73%	53.02%	--	--			--	--		
	VPC - data	57.67%	43.13%	--	--			79.43%	60.21%		
	DIC	28736		--				31265			
Dataset		**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**
1.c. Hierarchical dataset $N_k = 17+1$ $N_j = 67$ $N_i = 2094$	Intercept (λ=£30,000)	--	--	60.02	0.418	**Same as above** (different assumptions regarding the data on country-level do not affect estimation results as the existence of a country-level has not been acknowledged in this model)		62.10	0.465	**Same as above** (different assumptions regarding the data on country-level do not affect estimation results as the existence of a country-level has not been acknowledged in this model)	
	σ^2_{u0k} (Country)	--	--	2082	0.118			6270	0.365		
	σ^2_{u0j} (Study)	--	--	6984	0.362			--	--		
	σ^2_{e0} (Data)	--	--	10869	0.313			18818	0.603		
	VPC - country	--	--	10.44%	14.88%			24.99%	37.71%		
	VPC – study	--	--	35.03%	45.65%			--	--		
	VPC - data	--	--	54.52%	39.47%			75.01%	62.29%		
	DIC	--		28733				31267			
Dataset		**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**	**ΔC/100**	**ΔE**
1.d. Reduced dataset $N_k = 17$ $N_j = 67$ $N_i = 1806$	Intercept (λ=£30,000)	--	--	62.79	0.405	68.49	0.449	63.36	0.426	91.61	0.568
	σ^2_{u0k} (Country)	--	--	2585	0.145	--	--	7466	0.379	--	--
	σ^2_{u0j} (Study)	--	--	7338	0.318	8654	0.402	--	--	--	--
	σ^2_{e0} (Data)	--	--	11959	0.251	11954	0.251	20438	0.421	24912	0.648
	VPC - country	--	--	11.81%	20.31%	--	--	26.76%	47.38%	--	--
	VPC – study	--	--	33.53%	44.54%	41.99%	61.56%	--	--	--	--
	VPC - data	--	--	54.65%	35.15%	58.01%	38.44%	73.24%	52.63%	100%	100%
	DIC	--		24573		24573		26539		27464	

However, when comparing the cross-classified (model 1.a), the three-level hierarchical (model 1.b), and the two-level hierarchical model which clusters data in studies only (model 1.c), one recognises that their respective DIC values are almost identical. This also holds both for the univariate and bivariate model specifications. Furthermore, the minimal difference in DIC is likely to be a result of the random nature of the MCMC estimation process as very small changes in the DIC are likely to occur with each run of the model (Browne, 2012). As the DIC accounts for differences in the number of parameters estimated in each model (Browne, 2012), these model specifications may hence be regarded as equivalent in terms of their fit to the data (Note that for the reduced dataset the DIC is clearly lower in all model specifications. This is, however, not a result of improved model fit, but rather of dropping some observations from the dataset). As a result, including a study-level into the hierarchy modelled seems to be very important to appropriately reflect the structure of the data whilst adding a country-level does, at this point, not improve the model fit much further. Accordingly, and also in line with the pilot study results, models 1.d and 1.e may no longer be considered in this exercise.

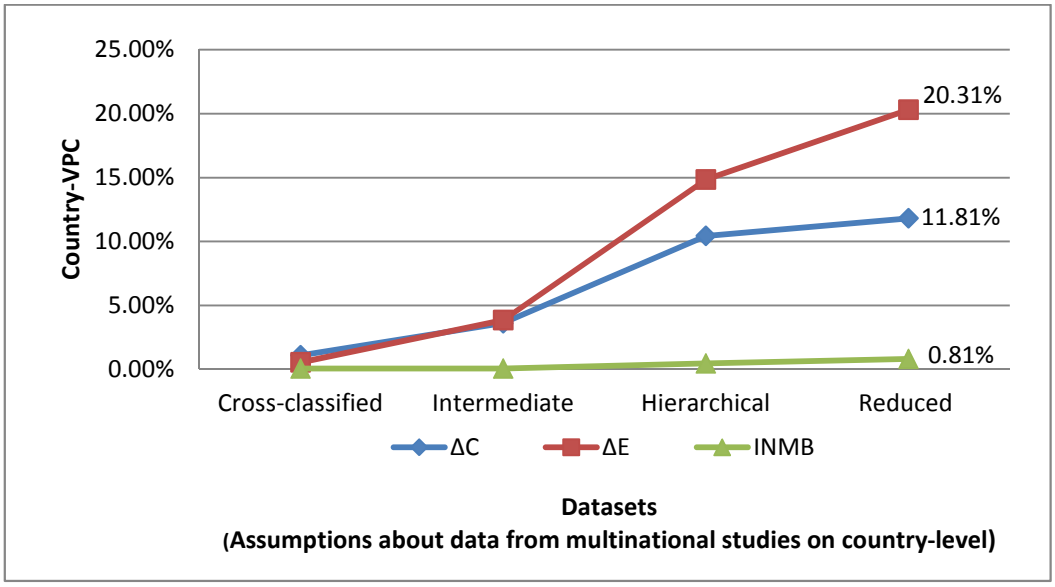
Though there is roughly equivalence in model fit between the two-level specification, which clusters data in studies only, and the more elaborated three-level hierarchical and cross-classified models, this does not mean that the more complex models will not fit the data better once covariates are considered. On the other hand, the two-level model does not permit the explicit assessment of country-level covariates, which is a key objective of this empirical exercise. For these reasons, it was decided to take a model specification which explicitly considers the country-level forward to the second part of this empirical exercise, which is concerned with covariate adjustment on data and study-level. However, this still leaves one choice open before considering covariates, which is the choice between the cross-classified model, which treats data from multinational studies as independent between countries (model 1.a), and the three-level hierarchical model, which assumes that data from multinational studies is not country-specific and thereby clusters this data in a separate group on country-level (model 1.b). As is was hypothesized that data from multinational studies draws a curtain over potentially existing country-level variability, it is indicated to have a closer look at the random part of the respective models.

Moving on to this random part of the MLMs presented in Tables 5.6 and 5.7 shows even more clearly the importance of making appropriate assumptions regarding dependencies within the data. When comparing the variance partitioning coefficient (VPC) between the cross-classified and the three-level hierarchical models, it can be observed that country-level variability is constantly increasing with decreasing influence of data from multinational studies on individual country-level parameters. In other words, it appears that data from multinational studies is much less affected by variability on country-level which '*drags*' country-means towards each other in the cross-classified model and thereby '*camouflages*' the country-level variability which actually does exist within the data from single-country-studies. Gradually removing this influence of data from multinational studies on individual country-level parameters increasingly uncovers this country-level variability in the rest of the data. This has also been visualised in Figure 5.3. It can be observed that the country-level VPC in Figure 5.3 constantly increases for all response variables in the univariate and bivariate framework starting from the cross-classified model, where all data from multinational studies was assigned to its respective target countries, up to the three-level hierarchical model applied to the reduced dataset, where all data from multinational studies was dropped. Interestingly, the effect of treating data from multinational studies as dependent on country-level is much stronger in the bivariate than in the univariate framework. This indicates that part of the country-level variability in international cost-effectiveness data may be disguised when combining ΔC and ΔE to INMBs, which shows ever more clearly the importance of decomposing the INMB statistic within the bivariate framework.

In conclusion, the assumption of independence of data from multinational studies on country-level may not be justified, which does not mean, however, that this data shall be dropped from the remainder of this empirical exercise. Rather, this first exercise indicates that the influence of this data on individual country-parameters may be removed by grouping it in a separate cluster on country-level, thereby acknowledging the fact that this data is not independent between countries. This is exactly what happens within the hierarchical dataset, which shows significantly increased country-level variability for both ΔC and ΔE in the bivariate model. To sum up, results from this first experiment show that the three-level hierarchical model applied to a dataset, which clusters all data from multinational studies in a separate group on country-level, may be carried

forward to subsequent analyses as the assumption of independence of multinational study data on country-level, which also introduces the cross-classification problem, appears to be questionable.

Figure 5.3: Country-level variability with respect to assumptions about (in-) dependence of data on country-level



After having made a choice regarding the appropriate model structure to take forward to the second part of this empirical exercise, it may be useful to take a step back to have a further look at the fixed part of this model before moving on to have a closer look at the impact of '*shrinkage*' within the multilevel framework in Section 5.1.5. Specifically, with information on the overall mean regression coefficient and the study residuals, which can be obtained from MLwiN (Rasbash et al., 2009), one can estimate the respective study means in the multilevel framework. Also, MLwiN provides the data required for estimating confidence intervals around these study means (Rasbash et al., 2009), so that it is possible to build forest plots for all the studies included in the dataset. These forest plots differ from the caterpillar plots which can be produced directly in MLwiN. Caterpillar plots report as '*zero*' on the vertical axis the overall regression mean, so that one may infer whether a study significantly departs from this overall regression mean (Rasbash et al., 2009). On the other hand, the forest plots presented in Figure 5.4 below, which were produced after estimating models in MLwiN and importing data to Ms Excel, show on the horizontal axis actual INMB,

ΔC or ΔE. Hence, one may infer, for instance, which studies have confidence intervals crossing zero, meaning that there is a chance that the response variable may also be negative. In addition, the forest plots provide a visual presentation of the data which makes it easy to locate outliers, and which studies report more or less robust results judged from the actual size of the confidence intervals. Finally, one can draw conclusions across the three forest plots due to the relationship of a studies mean INMB and the means of the components of the INMB statistic ΔC and ΔE.

From Figure 5.4.a. it gets apparent that, whilst the majority of studies report a mean INMB which is positive at a threshold value of £30.000, their respective confidence intervals almost always include zero. Few studies even report a negative mean INMB (updated to 2010 £-Sterling) which are highlighted in red in Figure 5.4.a (Hjalte et al., 1989; Glick et al., 1992; Drummond et al, 1993; Martens et al., 1994; Pharoah et al., 1996; Perreault et al., 1998; Hamilton et al., 1999; Morris & Godber, 1999; CDC Group, 2002; Nagata et al., 2005; Franco et al., 2007). One can also observe a few strong outliers reporting much higher INMBs than the rest of the studies in the dataset. Amongst these studies are Alonso et al. (2008), Grover et al. (1999), Grover et al. (2000), Grover et al. (2001), Grover et al. (2003) and a study from Spaans et al. (2003) which was also co-authored by S.A. Grover. The fact that the outliers in the dataset appear to be strongly related through '*common authorship*' is an observation which relates back to the genealogy study reported earlier in Chapter 4, and will also be further assessed in Section 5.2 of this empirical chapter when considering study-level covariates which resulted from the genealogy study.

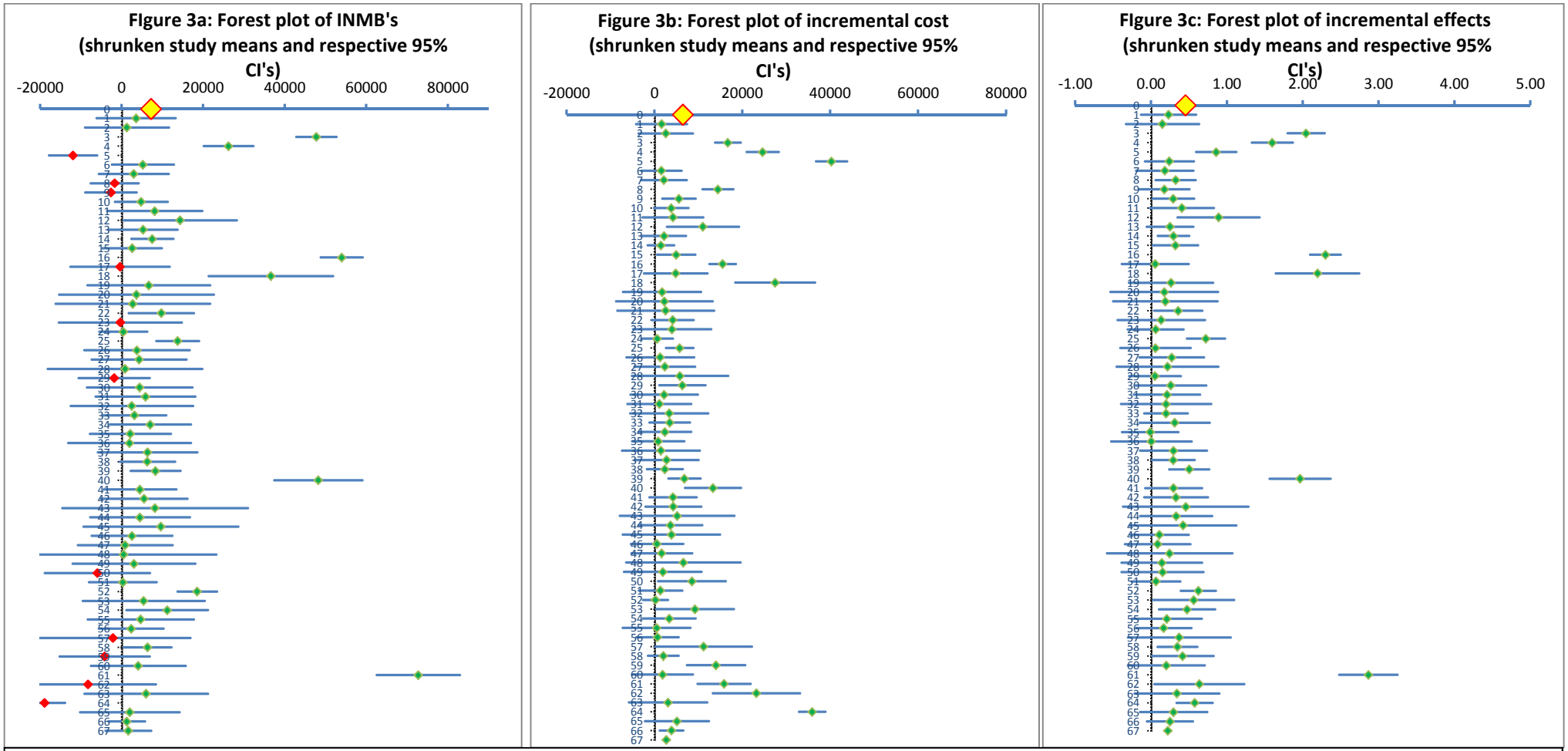
However, when looking across the three forest plots presented below, one can draw further conclusions in terms of these outliers. For instance, higher mean INMBs appears to be correlated to both higher ΔC and ΔE, so that part of the positive effect of elevated ΔE on INMBs may be offset by higher ΔC. The bivariate model specification, where the INMB statistic is decomposed into its stochastic components, may be useful to further assess the relationship between cost-effectiveness data and covariates whilst taking into account that ΔC and ΔE are, themselves, correlated. This is again a strong argument for decomposing the

INMB statistic in this exercise as it allows additional insights into causes of variability in international cost-effectiveness data.

Finally, the level-two residuals obtained from MLwiN and used to generate the forest plots below were subject to '*shrinkage*' when estimating the respective MLM (Rasbash et al., 2009). This means that not just the overall regression mean, but also the individual study means are affected depending on their respective number of data points and variances at each level of the data hierarchy. As a result, study means presented below are '*dragged*' towards the overall regression mean depending on the number of data points abstracted from those studies as well as the within and between study variability (Steele, 2008; Rasbash, 2009; Hox, 2010). A potential problem arises as this exercise does not deal with individual patient data, where shrinkage depends on the number of individuals included in a study, but rather with secondary cost-effectiveness data, where the number of data points per study depends on the rigour with which subgroup and sensitivity analyses were conducted and reported. Hence, whilst shrinkage is supposed to balance out the fact that studies in a dataset are of different size, this fact demands further attention when dealing with secondary cost-effectiveness data, where the number of data points abstracted from one study does not necessarily reflect the weight it should have relative to other studies in the dataset. The actual impact of shrinkage factors on the results of this exercise and the associated potential for bias when applying MLM techniques to secondary cost-effectiveness data from published economic evaluation studies is considered below in Section 5.1.5.

Figure 5.4: Forest Plots of INMBs, ΔCs, and ΔEs for all 67 studies included in the dataset

(Estimates from 2-level model with data clustered in studies – study means shrunken towards overal mean depending on within study variability and number of data points from that study)



Study IDs

01: Ashraf et al (1996)	02: Caro et al (1997)	03: Grover et al (1999)	04: Grover et al (2000)	05: Hamilton et al (1995)	06: Johanneson et al (1997)	07: Muls et al (1998)	08: Perreault et al (1998)	09: Pharoah et al (1996)	10: Szucs et al (1998)
11: Szucs et al (2000a)	12: van Hout et al (2001)	13: Jönsson et al (1996)	14: Jönsson et al (1999)	15: Ganz et al (2000)	16: Grover et al (2001)	17: Martens et al (1994)	18: Alonso et al (2008)	19: Annemans et al (2010)	20: Araujo et al (2007)
21: Lindgren et al (2010)	22: Grover et al (2008)	23: Franco et al (2007)	24: Greving et al (2011)	25: HPS Group (2009)	26: Khoury et al (2009)	27: Kongnakorn et al (2009)	28: Morris (1997)	29: Morris, Godber (1999)	30: Rosen (2010)
31: Scuffham et al (2005)	32: Scuffham et al (2006)	33: Tailor et al (2009)	34: Tonkin et al (2006)	35: Wagner et al (2009a)	36: Wagner et al (2009b)	37: Berger et al (1997)	38: Obermann et al (1997)	39: Davies et al (2006)	40: Spaans et al (2003)
41: Soini et al (2010)	42: Peura et al (2008)	43: Slejko et al (2010)	44: Nherera et al (2010)	45: Szucs et al (2000b)	46: Sigvant et al (2011)	47: Johannesson et al (1996)	48: Troche et al (1998)	49: Szucs et al (2004)	50: Nagata et al (2005)
51: Lindgren et al (2007)	52: HPS Group (2006)	53: Tsevat et al (2001)	54: Raikou et al (2007)	55: Ramsey et al (2008)	56: Scuffham et al (2004)	57: Hjialte et al (1989)	58: Caro et al (2003)	59: CDC Group (2002)	60: Chau et al (2001)
61: Grover et al (2003)	62: Glick et al (1992)	63: NICE (2008)	64:Drummond et al (1993)	65: Chan et al (2007)	66: Ward et al (2007)	67: Ara et al (2009)	◇: overall regression mean (not identical to raw mean estimates)		

*INMBs as reported in Figure 2.a may be calculated by multiplying ΔE from Figure 2c with a threshold value λ of 30000 and subtracting ΔC reported in Figure 2.b. Small differences are the result of shrinkage and the random nature of the MCMC estimation procedure.

5.1.5. Shrinkage as a potential source of bias in the analysis of secondary cost-effectiveness data

Before moving on to the assessment of covariates, it is important to consider the impact of shrinkage in the MLM framework. As mentioned in Chapter 3, assuming exchangeability, which is a necessary assumption to fit MLMs, is supposed to '*mediate*' between the two extreme viewpoints of either identical or independent study parameters in the sense that it allows for some sort of pooling, but without ignoring the fact that the data stems from different studies indeed (Spiegelhalter et al., 2000; Spiegelhalter et al., 2004; Gelman et al., 2004; Rasbash et al., 2009; Hox, 2010). As a result, shrinkage factors '*drag*' group means towards the overall regression mean depending on their respective number of lower-level units as well as the within and between group variability (Steele, 2008; Rasbash et al., 2009). This makes perfect sense, for instance, when the lowest level consists of patients in randomised controlled trials. The extent of shrinkage towards the overall regression mean would then depend on the number of patients in each trial as well as the within and between trial variability (Spiegelhalter et al., 2000).

However, this empirical exercise deals with secondary economic evaluation data abstracted from existing studies published on the cost-effectiveness of statins in the primary and secondary prevention of CVD. Hence, the data in this exercise does not represent individual patients which would instantly justify empirical Bayes shrinkage estimation. Rather, data points reflect estimates of cost-effectiveness of statins as reported by the authors of the respective studies included in this exercise. The number of data points within each study is hence depending on the extent to which authors performed and reported on subgroup and sensitivity analysis in their papers. This means that differences in the way data and models were utilised to perform subgroup and / or sensitivity analyses, or even just the way results have been reported in the respective papers, may already account for differences on group sizes, which then affects study means in the MLM through shrinkage estimates. As a result, there is an argument around the appropriateness of shrinkage estimation if the underlying data stems from existing economic evaluation studies and does not represent individual patients. Before moving on to covariate adjustment, this section therefore looks deeper

into this issue by discussing the underlying concept, show the actual impact of shrinkage estimates within the variance components models reported above, and discusses alternative strategies which were considered to '*counterweigh*' the effect of shrinkage within this secondary data integration exercise.

5.1.5.1. Empirical Bayes shrinkage estimation within the multilevel modelling framework

To quantify the impact of shrinkage on the results presented above, one needs to make explicit the connection between the MLM, where residuals are estimated for each hierarchical level and shrinkage estimates '*mediate*' between the two extreme assumptions of either identical or independent study parameters, and the respective non-hierarchical OLS model, which does not account for complex data structures. Empirical Bayes shrinkage estimates '*drag*' study means towards the overall regression mean (e.g. Steele, 2008; Rasbash et al., 2009). This section shows that the OLS regression model, which does not take into account that data is clustered in studies, is simply a special case of the MLM where shrinkage estimates are assumed to be zero (e.g. Willan et al., 2005; Steele, 2008; Rasbash et al., 2009). Further, in an empty OLS regression model, the intercept is simply the raw mean of the response variable (e.g. Maddala, 2001). In conclusion, we may quantify the effect of shrinkage by comparing study means as obtained from the variance components model presented above with the individual study means calculated from the raw data.

As detailed in the multilevel methods Chapter 3, to be able to make comparisons between studies in the multilevel framework, the study-level residual 'u_j' needs to be estimated (Steele, 2008; Rasbash et al., 2009). An estimate of 'u_j' may be derived by calculating the '*mean raw residual*', that is: (Steele, 2008; Rasbash et al., 2009; CMM-workshops/variance components).

$$\bar{r}_j = \bar{y}_j - \hat{\beta}_0 \qquad \text{(1, repeated from Chapter 3)}$$

Where' $\bar{y}_j'$ is the mean of the response variable in study 'j', and $'\hat{\beta}_0'$ is an estimator of the overall mean of the response variable. This raw residual is then multiplied by the shrinkage factor 'S' (Spiegelhalter et al., 2004; Willan et al., 2005; Steele, 2008; Rasbash et al., 2009):

$$\hat{u}_{0j} = S\bar{r}_j \quad \text{where} \quad S = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + (\frac{\hat{\sigma}_e^2}{n_j})} \qquad \text{(2, repeated from Chapter 3)}$$

In (2), $'n_j'$ is the sample size in study 'j', hence the number of data points reported in that study (Steele, 2008; Rasbash et al., 2009). $'\hat{\sigma}_e^2'$ and $'\hat{\sigma}_u^2'$ are estimates of the variances of the within-study and between-study error terms respectively (Steele, 2008; Rasbash et al., 2009). The respective level one residual is given as (CMM-workshops/variance components):

$$\hat{e}_{0ij} = y_{ij} - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1ij}\right) - \hat{u}_{0j} \qquad (3)$$

Now it can be illustrated why the assumption of exchangeability of study parameters mediates between the two extremes of either identical or independent parameters obtained from different cost-effectiveness studies. If the between study variance $'\sigma_u^2'$ is assumed to be zero, then this is equal to say that all variation in the reported cost-effectiveness measures stems from '*within-study'* variability, hence, the mean study effects are '*identical*' between studies. This would mean that there are no differences between studies and all cost-effectiveness estimates may be safely pooled together. In terms of equations (2) and (3), if 'S' is zero in (2), then $'\hat{u}_{0j}'$ will be zero too, and (3) reduces to:

$$\hat{e}_{0i} = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} \qquad (4)$$

This is simply the residual as defined for the OLS regression model (e.g. Maddala, 2001) If, on the other hand, $'\sigma_u^2 \to \infty'$, then the study effects are regarded to be independent, meaning that data from different sources may not be pooled together. Hence, assuming exchangeability allows to model that the '*reality*' might be somewhere in between those extreme viewpoints. To sum up,

$'\sigma_u^2 \to 0'$ is the special case where the shrinkage factor $'S = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \left(\frac{\hat{\sigma}_e^2}{n_j}\right)}'$ tends towards zero and the study effects are completely shrunken towards the overall mean of the response variable $'\beta_0'$. A pooled OLS regression model may be equivalent in this case. If $'\sigma_u^2'$ is high, then the shrinkage factor $'S = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + (\frac{\hat{\sigma}_e^2}{n_j})}'$ tends towards unity, meaning that the between country variance is high and shrinkage of study effects towards the overall mean $'\beta_0'$ is small (Steele, 2008)

However, the extent of shrinkage does not only depend on the amount of between study variability $'\sigma_u^2'$. The higher the number of data points provided by one single study (n_j), the more information is provided by that particular study to the overall model, and the less will the study mean be shrunken towards the overall mean (Steele, 2008; Rasbash et al., 2009; Hox, 2010). On the other hand, if a study provides only few estimates of cost-effectiveness, then shrinkage is high as this study '*borrows*' a lot of information from all other studies available (Steele, 2008; Rasbash et al., 2009; Hox, 2010).

It has now been shown that by assuming the shrinkage factor 'S' to be zero, a MLM, where data is grouped in economic evaluation studies, reduces to an OLS regression equation. Hence, to quantify the effect of shrinkage within the variance components models reported above, one may simply compare the study means of the variance components model with the intercepts estimated by an OLS regression model being fit to a) the pooled data and b) the data from each study separately to model both the assumptions of a) identical and b) independent parameters. However, we may further simplify this procedure. In an OLS regression equation, the intercept is given by (e.g. Maddala, 2001):

$$\hat{\beta}_{0i} = \overline{y}_i - \hat{\beta}_{1i} * \overline{x}_i \qquad (5)$$

Where $'\overline{y}'$ is the overall mean of the response variable and $'\hat{\beta}_{1i}'$ represents a vector of explanatory variables (e.g. Maddala, 2001). However, in the variance components models above covariates have not yet been considered so that the

equivalent OLS regression would be an empty model without any slope parameters. Hence, (5) reduces to:

$$\hat{\beta}_{0i} = \bar{y}_i \tag{6}$$

which means that, in an empty OLS regression model, the intercept is identical to the raw mean of the response variable. In conclusion, we may show the effect of shrinkage in the variance components model reported above by simply calculating the study means in the MLM as the group departure from the overall intercept ($\beta_0 + u_j$) and comparing the result to the raw study means calculated as (e.g. Maddala, 2001):

$$\bar{y}_{ij} = \frac{1}{n_j} * \sum_{i=1}^{nj} y_{ij} \tag{7}$$

to assume independent parameters or the raw mean of the pooled data for identical study parameters calculated as:

$$\bar{y}_i = \frac{1}{n} * \sum_{i=1}^{n} y_i \tag{8}$$

Note that, alternatively, one may simply insert values for the within and between group variances '$\hat{\sigma}_e^2$' and '$\hat{\sigma}_u^2$ ', which can be obtained from MLwiNs equation window after running the respective model as well as the study group sizes 'n_j' in equation (2) to calculate the shrinkage factor for each study in the MLM and then divide the level two residuals '$\hat{u}_{0j}$' which are also obtainable from MLwiN by 'S' to get the mean raw residual '$\bar{r}_j$' (Rasbash et al., 2009) When adding this mean raw residual to the overall regression mean 'β_0', one receives approximately the raw means calculated with equation (7) (Rasbash et al., 2009)
The following section reports both study means estimated by the variance components model reported above., the raw means calculated with equations (8) and (9), and the respective shrinkage factors calculated with equation (2). Results are then compared in a graph displayed in Figure 5.5. Finally, the impact of shrinkage in this empirical exercise and resulting concerns when applying MLM techniques to secondary cost-effectiveness data are discussed before moving on to part two of this empirical work, which is concerned with covariate adjustment in the MLM framework.

5.1.5.2. Quantifying the impact of empirical Bayes shrinkage estimates on study means

Table 5.8. shows raw study means, shrunken study means and shrinkage factors for INMB, ΔC and ΔE and the results from Table 5.8 are also visualised in Figure 5.5 below. Note that the shrinkage factor (S) is not the ratio of the raw study means and shrunken study means as it is applied to shrink the '*mean raw residuals*' ($\bar{r}_j$) to obtain study-level residuals '$\mathrm{u_j}$' which are then added to the overall regression mean 'β_0' to obtain shrunken study mean values (Steele, 2008; Rasbash et al., 2009, CMM workshops / variance components). For most studies in the dataset, the shrinkage factor, which is bound between 0 and 1, is above 0.9. Furthermore, for ΔE shrinkage factors are even closer to unity, indicating that study means for ΔE are even less shrunken towards the overall regression mean (Steele, 2008). The reason for this is given in equation (2) above. Shrinkage factors are closer to zero (i.e. so that shrinkage is stronger), if '$\mathrm{n_j}$' is small or '$\hat{\sigma}_e^2$' is large compared to '$\hat{\sigma}_u^2$' (Steele, 2008). In other words, if there are only few data points in one group (so that we have little information about that group), or if within group variability is high compared to between group variability (indicating low dependency of data within groups), then study means are pulled more strongly towards the overall mean (Steele, 2008; Rasbash et al., 2009; CMM workshops / variance components) Finally, as shrinkage factors are applied to the mean raw residual and not directly to study means, shrinkage has a stronger impact on outliers in the dataset than it has for studies close to the overall regression mean (Steele, 2008; Rasbash et al., 2009; CMM workshops / variance components)

Between group variability is generally high within the given dataset, indicating high dependency of the data within studies and therefore providing a strong justification for the use of MLM. Accordingly, shrinkage factors are generally high, meaning that shrinkage towards the overall regression mean is, at most, moderate. In addition, as ΔE shows even higher between group variability than INMB or ΔC (which is getting obvious when comparing the respective variance partitioning coefficients reported in Table 5.8), shrinkage factors are even closer to unity, meaning that study means are even less shrunken towards the overall regression mean. Nevertheless, though high dependency within groups generally

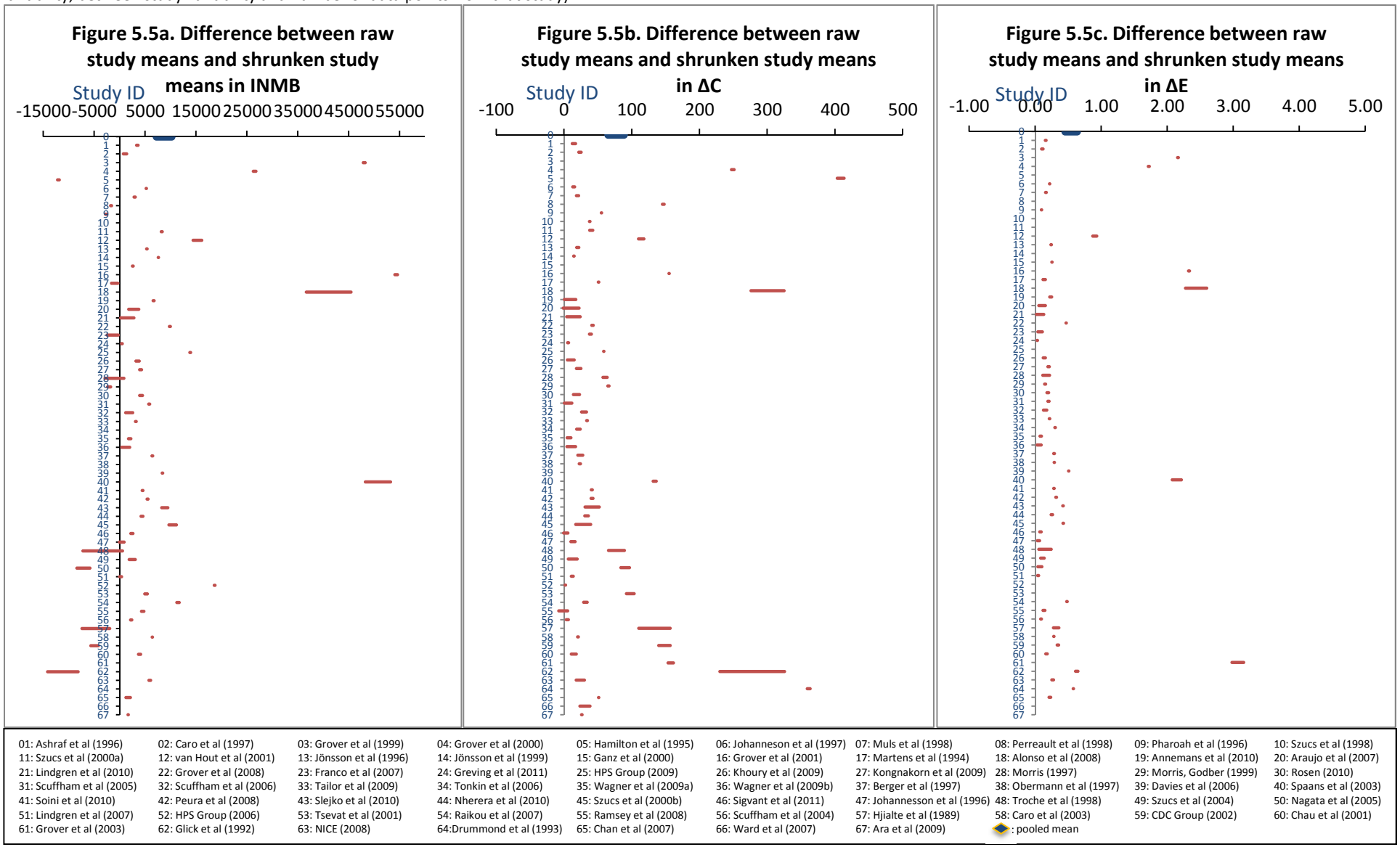
leads to higher shrinkage factors, which decreases the impact of shrinkage on study means, Figure 5.5 shows that for few studies shrinkage does have quite some impact within the MLM framework. The length of the horizontal line for each study in Figure 5.5 indicates the difference between the raw and the shrunken study means. It was mentioned that shrinkage has a stronger effect on outlying studies, and this is getting visible when comparing, for example, studies 18 and 19 (Alonso et al., 2008; Annemans et al., 2010) in Figure 5.5a. As 'n_j' is identical in both studies (n_j = 4), their respective shrinkage factors are also identical with 0.78. However, as the mean raw residual is much higher for the outlying study by Alonso et al. (2008), shrinkage has a much stronger impact on its respective study mean. In conclusion, the impact of shrinkage on study means in this exercise depends not just on the respective number of data points abstracted from each study, but also on the within and between group variability in the data and the location of each study mean relative to the overall regression mean.

Table 5.8: Raw and shrunken study means plus shrinkage factors

		INMB			ΔC			ΔE		
raw mean		10210			88.71			0.64		
β_0		7233			65.18			0.46		
$\hat{\sigma}_u^2$ (vpc)		253045104 (46.54%)			8073 (42.62%)			0.43 (58.11%)		
$\hat{\sigma}_e^2$ (vpc)		290676352 (53.46%)			10867 (57.38%)			0.31 (41.89%)		
Study	Nj	Raw mean	Shrunken mean	Shrinkage factor	Raw mean	Shrunken mean	Shrinkage factor	Raw mean	Shrunken mean	Shrinkage factor
1	13	3344	3626	0.92	12.49	16.97	0.91	0.15	0.17	0.95
2	11	750	1391	0.91	21.9	25.36	0.89	0.10	0.12	0.94
3	144	48242	47869	0.99	168.14	168.08	0.99	2.17	2.16	0.99
4	52	26763	26284	0.98	251.3	247.41	0.97	1.73	1.71	0.99
5	60	-12201	-11856	0.98	413.31	403.61	0.98	0.97	0.97	0.99
6	25	5216	5269	0.96	13.12	15.71	0.95	0.22	0.22	0.97
7	18	2837	3079	0.94	18.63	21.49	0.93	0.16	0.17	0.96
8	60	-1748	-1605	0.98	147.88	145.34	0.98	0.44	0.44	0.99
9	46	-2782	-2557	0.98	55.63	55.28	0.97	0.09	0.10	0.98
10	42	4836	4862	0.97	37.79	38.47	0.97	0.29	0.29	0.98
11	8	8356	8192	0.87	37.94	42.61	0.86	0.41	0.41	0.92
12	5	16143	14435	0.81	118.16	110.26	0.79	0.93	0.87	0.87
13	19	5320	5404	0.94	18.8	21.90	0.93	0.24	0.25	0.96
14	109	7639	7605	0.99	14.25	14.99	0.99	0.30	0.30	0.99
15	29	2537	2677	0.96	50.25	50.41	0.96	0.25	0.26	0.98
16	112	54616	54087	0.99	155.08	155.37	0.99	2.34	2.32	0.99
17	7	-1580	-344	0.86	51.54	50.65	0.84	0.12	0.15	0.91
18	4	45443	36677	0.78	324.81	276.24	0.75	2.60	2.28	0.85
19	4	6601	6751	0.78	0.74	17.30	0.75	0.22	0.25	0.85
20	2	1773	3774	0.64	-1.23	22.40	0.60	0.06	0.15	0.73
21	2	207	2808	0.64	3.93	24.02	0.60	0.02	0.13	0.73
22	22	9960	9790	0.95	41.25	43.27	0.94	0.47	0.47	0.97
23	4	-2470	-276	0.78	37.71	40.30	0.75	0.04	0.11	0.85
24	61	399	492	0.98	5.58	6.69	0.98	0.03	0.04	0.99
25	100	13937	13814	0.99	58.5	58.88	0.99	0.66	0.66	0.99
26	6	3199	3836	0.84	5.01	15.17	0.82	0.12	0.15	0.89
27	8	3960	4324	0.87	18.6	24.88	0.86	0.19	0.21	0.92
28	2	-2885	823	0.64	63.8	57.46	0.60	0.12	0.22	0.73
29	17	-2361	-1753	0.94	66.61	64.75	0.93	0.14	0.16	0.96
30	6	3961	4471	0.84	13.94	22.81	0.82	0.18	0.20	0.89
31	7	5718	5927	0.86	0.82	11.46	0.84	0.19	0.21	0.91
32	4	1220	2566	0.78	26.05	33.37	0.75	0.13	0.18	0.85
33	24	3101	3273	0.95	33.49	34.79	0.95	0.22	0.22	0.97
34	12	7103	7088	0.91	18.96	23.99	0.90	0.30	0.31	0.94
35	12	1737	2214	0.91	4.84	10.10	0.90	0.07	0.09	0.94
36	4	439	1973	0.78	4.46	16.98	0.75	0.03	0.09	0.85
37	7	6343	6481	0.86	20.57	27.92	0.84	0.28	0.29	0.91
38	36	6411	6401	0.97	22.54	24.16	0.96	0.29	0.29	0.98
39	52	8447	8375	0.98	68.41	68.41	0.97	0.51	0.51	0.99
40	10	53221	48311	0.90	131.69	136.26	0.88	2.21	2.08	0.93
41	16	4422	4562	0.93	40.45	41.99	0.92	0.28	0.29	0.96
42	10	5399	5579	0.90	40.21	42.99	0.88	0.31	0.32	0.93
43	1	9484	8302	0.47	31.15	52.05	0.43	0.42	0.43	0.58
44	7	4196	4585	0.86	30.94	35.83	0.84	0.24	0.26	0.91
45	2	11142	9706	0.64	17.57	39.46	0.60	0.43	0.42	0.73
46	12	2202	2593	0.91	-0.49	5.65	0.90	0.07	0.09	0.94
47	8	-17	900	0.87	10.4	16.36	0.86	0.03	0.07	0.92
48	1	-7224	557	0.47	89.34	65.71	0.43	0.06	0.24	0.58
49	4	1890	3085	0.78	6.6	19.52	0.75	0.09	0.14	0.85
50	6	-8381	-5848	0.84	96.56	84.09	0.82	0.04	0.10	0.89
51	20	3	369	0.95	11.05	13.65	0.94	0.04	0.05	0.96
52	168	18730	18608	0.99	1.72	2.53	0.99	0.63	0.63	1.00
53	4	4947	5467	0.78	103.52	92.40	0.75	0.51	0.51	0.85
54	12	11700	11260	0.91	29.1	34.19	0.90	0.49	0.48	0.94
55	6	4297	4781	0.84	-7.52	5.55	0.82	0.12	0.15	0.89
56	22	2178	2406	0.95	3.26	6.50	0.94	0.08	0.09	0.97
57	2	-7375	-1950	0.64	156.85	110.41	0.60	0.28	0.36	0.73
58	62	6432	6406	0.98	20.33	21.32	0.98	0.28	0.28	0.99
59	9	-5655	-4171	0.89	156.66	139.94	0.87	0.33	0.36	0.92
60	8	3729	4132	0.87	11.05	18.28	0.86	0.16	0.18	0.92
61	12	79387	72846	0.91	153.63	161.60	0.90	3.16	2.98	0.94
62	3	-14167	-8149	0.72	325.66	230.16	0.69	0.61	0.65	0.80
63	4	5746	6094	0.78	18.29	30.06	0.75	0.25	0.28	0.85
64	128	-19034	-18834	0.99	363.39	359.08	0.99	0.58	0.58	0.99
65	7	1230	2095	0.86	51.21	51.45	0.84	0.21	0.24	0.91
66	322	1352	1337	1.00	23.8	38.47	1.00	0.17	0.17	1.00
7	72	1659	1722	0.98	26.11	26.67	0.98	0.14	0.15	0.99

Figure 5.5: Differences between raw study means and shrunken study means

(Graph plots the mean INMB, ΔC and ΔE for each study as obtained from 1) the raw data and 2) from a two-level MLM so that means are shrunken towards overal mean depending on within study variability, between study variability and number of data points from that study)



5.1.5.3. Discussing the '*appropriateness*' of shrinkage within this secondary data integration exercise

Two questions follow from what has been reported above: first, may shrinkage lead to some sort of bias in the case of integrating secondary cost-effectiveness data where the number of data points per study is generally not indicative for the respective size of the underlying study sample. In other words, as shrinkage, or the lack thereof, may only be a consequence of differences in the way subgroup or sensitivity analyses were performed and reported in the respective studies, it may not be justified to shrink studies based on the respective number of data points abstracted per study. Secondly, if the answer to the first question is yes, is there anything we can do about it in the MLM framework. As mentioned, empirical Bayes shrinkage estimation follows directly the assumption of exchangeability, and any attempt to '*counterweigh*' shrinkage would impact on the very foundations of multilevel modelling (e.g. Spiegelhalter et al., 2000; Spiegelhalter et al., 2004; Steele, 2008). To this extend, this issue is nothing less than a fundamental critique on the use of MLM for the integration of secondary cost-effectiveness data, and the answer to this question may be much more a philosophical than a technical one.

To answer the first question, consider again the case of the two studies with the ID's 18 and 19 (Alonso et al., 2008; Annemans et al., 2010) in Figure 5.5a. As mentioned, both studies contribute four data points each to the overall dataset, so that their respective group sizes are relatively small compared to other studies. We do not judge a priori whether the data points provided by these studies are poor guesses of the cost-effectiveness of statins in the primary and secondary prevention of CVD. Accordingly, the mean raw residual ($\bar{r}_j$) calculated from the respective data points may be '*wrong*'. This idea holds for each group in the dataset, as we cannot judge a priori which study may provide the '*best guess*' regarding the cost-effectiveness of statins (though we will come back to this issue when adding covariates to the model). Now, when combining data from one group with data from all other groups, this will shrink residuals towards the overall average, so that level-two residuals will be less sensitive to outliers in the group (Steele, 2008). Figure 5.5a shows that, though group sizes and respective shrinkage factors are identical for both studies, the mean of the outlying study

18 is shrunken much stronger towards the overall average. Of course, when group sizes partly depend on nothing else than differences in the way authors reported cost-effectiveness estimates, one may argue that this process will bias results towards those studies which provide a large number of data points. However, the example of studies 18 and 19 makes clear that group size is only one of three factors determining the actual impact of shrinkage on study means in this exercise. The other two factors are 1) dependency (i.e. the relation between within and between group variability) and 2) location (i.e. the relative distance of each group mean from the overall regression mean) (Steele, 2008; Rasbash et al., 2009). Hence, to judge the appropriateness of shrinkage and to develop a method which may counterweigh potential bias introduced by '*artificial*' study sizes in the case of secondary data integration, one has to discuss each of the three factors involved in the process. This discussion starts with dependency and location before it comes back to the issue of group sizes.

In terms of '*dependency*' shrinkage appears to be justified just as much in the case of secondary data integration as it is for the integration of say individual patient data from RCTs. The idea that data from one study may be more similar to each other than it is to data from other studies makes intuitive sense in the case of secondary cost-effectiveness data. Furthermore, reflecting dependencies within the data is the whole purpose of MLM (e.g. Spiegelhalter et al., 2000; Spiegelhalter et al., 2004; Willan et al., 2005, Manca et al., 2005), and the fact that between group variability is generally high within the variance components models reported above provides a strong justification for the use of MLM. On the contrary, if one used a method to integrate secondary cost-effectiveness data which does not take into account dependencies within that data (as it is the case for example with the OLS model applied in the pilot study and in Chapter 3.4), one risks making plainly wrong inferences (e.g. Steele, 2008). This has been demonstrated quite impressively with the negative relationship between INMB and TCL obtained when applying the OLS model to the pilot dataset. Within each study, there is always a positive relationship between patients' cholesterol level and the cost-effectiveness of statins, which also accords expectations. However, as some studies assess patient groups which are sicker than others, the cost-effectiveness of statins may be lower even in the presence of higher cholesterol levels. Hence, the inability to discriminate between within and between group

effects leads to an overall negative relationship between INMB and TCL though this is clearly incorrect. Furthermore, the pilot study showed that precession was severely overestimated for higher-level variables like GDP per capita in the case of the OLS framework. For these reasons, acknowledging dependencies in the data through the process of shrinkage appears to be strongly justified within this secondary data integration exercise.

A similar conclusion has been reached with respect to '*location*' of study means relative to each other. If we treat each study in the dataset as a random sample of cost-effectiveness estimates from a wider '*population*' of cost-effectiveness estimates for a particular health technology (e.g. Spiegelhalter et al., 2000; Spiegelhalter et al. 2004), then it makes intuitive sense to drag outlying studies towards the overall study mean. In this sense, we simply combine the data from the outlying group with the information provided by all the other groups to bring study-level residuals closer to the overall average. This makes study-level residuals less sensitive to outlying elements. As mentioned, we do not judge a priori whether the information provided by one study is actually a '*poor guess'* of the cost-effectiveness of statins (though we may account for this later when including covariates to the equation). Rather we reduce the impact of outlying elements of the group by dragging them closer towards the overall average. In other words, if there was no shrinkage when integrating secondary cost-effectiveness data, then the impact of outliers on the overall regression mean would be much stronger. This can be observed when comparing the variance components models 1.a, 1.b. or 1.c above with the respective OLS regression model 1.e. Within the OLS regression model, full weight is given to outliers which report mean INMBs way above the other studies in the dataset, thereby pulling the overall mean INMB clearly above the regression mean within the multilevel framework. Hence, also with respect to '*location*', shrinkage appears to be justified for the integration of secondary cost-effectiveness data.

What is left is the impact of '*group size'*, and it is arguable whether the 'n_j' for each study appropriately reflect the size, and thereby the '*weight*' each study should carry in this exercise. On the other hand, it is certainly no straightforward task to determine an appropriate weight per study when some studies in this

exercise rely on actual patient data from RCTs, and others rely on combining data from different sources within a DAM. Even assigning the same weight to each study on these grounds may appear questionable. In other words, any attempt to re-weigh studies to '*counterweigh*' the effect of group sizes within the process of shrinkage (assuming there is a way to account for this in the model which is also discussed below) may be similar to trading one source of potential bias with another one.

However, let us consider once more the thought that INMBs in this exercise constitute a random sample from a hypothetical space of cost-effectiveness estimates (e.g. Spiegelhalter et al., 2000; Spiegelhalter et al., 2004). A priori, we do not judge whether one estimate may be more likely than another one (e.g. Spiegelhalter et al., 2004). If this assumption holds, however, then we have to accept that studies which provide more data points based on subgroup and sensitivity analyses also provide a larger '*slice*' of the space of potential cost-effectiveness estimates for that particular health technology. Even though these data points do not reflect individual patients, we ought to accept that, a priori, they should not be treated anyhow differently. From this perspective, shrinkage makes sense even when dealing with secondary cost-effectiveness data. On top of that, reporting subgroup and sensitivity analyses is regarded as a crucial factor when judging the quality of a particular economic evaluation study. In this respect, shrinkage simply assigns a higher weight to those studies which better comply with quality checklists in this matter (e.g. Drummond et al., 1996, Ofman et al., 2003; Evers et al., 2005). Furthermore, we may control for differences between studies through covariate adjustment (e.g. Drummond et al., 2009; Manca et al., 2010). For instance, one may rate study quality using an appropriate checklist and then control for this factor through the inclusion of a respective covariate on study-level. Within this empirical exercise, this is considered using the validated QHES instrument (Ofman et al., 2003). Furthermore, a number of study-level covariates are included to account for differences between studies. As a result, after adjusting for the appropriate covariates, data from different studies should become, in theory, exchangeable (Gelman et al., 2004; Manca et al., 2007; Drummond et al., 2009).

Before discussing potential strategies to counterweigh shrinkage, one further aspect of the problem needs to be considered which was not yet specifically addressed. The problem of shrinkage in the integration secondary data from health economic evaluations is very intuitive on study-level as it was detailed above. However, what does this mean for country-level parameters? Clearly, shrinkage on country-level follows the same logic in the sense that country-group size, location of the country-mean relative to other country parameters, and the relation of within and between country variability determines the amount of shrinkage per country (Steele, 2008, Rasbash et al., 2009). However, does the same critique apply with respect to country-group sizes? Unlike the number of data points abstracted from one particular study, which clearly depends on the rigour with which analysts exploited the data to perform subgroup and sensitivity analyses, country-group sizes additionally depend on the number of studies conducted in a specific country. In other words, countries for which more economic evaluation data from different studies are available receive a higher weight through shrinkage in the multilevel model. This is in complete accord with the logic behind empirical Bayes shrinkage estimation (e.g. Spiegelhalter et al., 2000; Spiegelhalter et al., 2004) as it gives higher weight to geographic domains for which more data is available and, conversely, drags countries for which the evidence base is poor closer to the overall mean. From this perspective, one may argue whether shrinkage is not perfectly justified in secondary data integration on country-level, irrespective of the nature of the underlying economic evaluation study (i.e. IPD analysis or DAM). In addition to that, data from different studies becomes, in theory, exchangeable once the researcher controlled for the appropriate set of covariates (Gelman et al., 2004; Manca et al., 2007; Drummond et al., 2009). Hence, it may be controlled for such differences in Section 5.2 of this empirical exercise, which is concerned with covariate adjustment on data and study-level, as this may also feed through to variability observed between geographic domains. The ultimate aim is to disclose the actual amount of variability between countries, which is then further assessed in Section 5.3 of this empirical exercise.

5.1.5.4. Strategies to address the issue of group sizes within secondary data integration

Though the above section provides a number of reasons to believe that shrinkage may also be appropriate in the case of integrating secondary data from published economic evaluation studies, it is acknowledged that the issue of group sizes on study-level may demand further attention. Hence, if one does not follow the argument as outlined above, it is indicated to develop a strategy to at least try to address the issue of shrinkage on study-level within the method of analysis chosen for this empirical exercise. Otherwise, as mentioned above, this issue may turn into a fundamental critique on the use of MLM for the integration of secondary cost-effectiveness data. For this reason, several strategies were considered to address the issue of study group sizes within the current exercise.

For example, as a relatively straightforward approach it was considered to bootstrap data from each study individually and then to draw a random sample for each study of identical size. This would have led to identical 'n_j' for each study in the dataset. However, it would have definitely changed the relationship between '$\widehat{\sigma}_e^2$' and '$\widehat{\sigma}_u^2$' so that shrinkage factors would not only be adjusted with respect to group size, but also with respect to the balance of within and between group variability. In other words, the potential bias introduced by artificial group sizes would have been traded against a potential bias with respect to changing the degree of dependency within the data. On top of that, this would have meant to bootstrap not just data for the response variable for each study individually, but also covariates once these are added to the model, which would have amounted to considerable efforts to implement this strategy.

As an alternative, it was considered to use MLwiNs weighting facility to adjust for differences in the 'n_j' for studies included in this empirical exercise (CMM, 2011). The idea is that information from particular studies is '*oversampled*' in the current dataset as it stems from studies where subgroup and sensitivity analyses were carried out more extensively (CMM, 2011). Hence, one may argue that out of the space of cost-effectiveness estimates, some did have a greater '*probability*' of being selected into the sample than others. Without weighting,

however, the model assumes that each data point did have the same chance of being selected into the sample (CMM, 2011). The probability of being selected into the sample may depend on certain variables referred to as 'Z'. Furthermore, a set of covariates is considered within the multilevel framework in the second part of this exercise, which is denoted with 'X'. According to (CMM, 2011), as long as there is no interaction between any variable out of Z on which the probability of being selected into the sample depends, and any variable out of X which enters the MLM as a covariate, results will be unbiased. Conversely, however, if there is a relationship between a covariate out of X and anything which may affect the probability of being selected into the sample out of Z, then the results may be biased (CMM, 2011).

The problem of weighting in this secondary data integration exercise is a very complex one, and, in line with what has been outlined above, many issues arise for instance in terms of i) is this the appropriate method to address the issue of study group sizes within secondary data integration, ii) if so, when to assign those weights, iii) how to assign those weights to data points of studies and, on a practical note, iii) how to implement the respective model in a particular software environment (MLwiN). Related to the latter issue, CMM (2011) state that '*at present the weights facility is not available when using MCMC estimation (since the method of implementing weights for MCMC would be radically different from the method for likelihood or quasi-likelihood estimation and require further methodological work and programming which has not yet been undertaken)*'. However, the models developed in this empirical exercise could, in line with what has been strongly advocated by members of the Centre for Multilevel Modelling in Bristol (personal communication with Professor W. Browne and R. Phillinger, CMM), only be implemented with success when using MCMC estimation procedures. This fact precludes the use of the weighting facility in MLwiN at this point in time so that it is strongly recommended to further look into this issue once an appropriate software environment exists which allows running complex multilevel models using MCMC whilst also applying weights to studies. This matter is therefore assigned to potential areas for further research as it cannot be fully addressed within this thesis.

5.1.5.5. Conclusion

In conclusion, this assessment showed that the impact of shrinkage is, at most, moderate within this empirical exercise and the problem arises predominantly on study-level. On the contrary, it is arguable whether the logic behind shrinkage, which gives higher gravity to geographic domains for which more data is available, is not justified between countries included in this empirical exercise. Differences in the nature of cost-effectiveness data may be addressed through covariate adjustment, so that data should become, in theory, exchangeable after controlling for the appropriate set of covariates. In addition to that, the moderate impact of shrinkage on study parameters within this empirical exercise is a result of high dependency of the data within studies, meaning that between study variability is high compared to within study variability, which leads to shrinkage factors being relatively close to unity. Furthermore, the relative location of study means with respect to the rest of the data seems to have a stronger impact on shrinkage than the actual number of data points provided by each study. Both '*dependency*' and '*location*', as factors influencing the impact of shrinkage, are perfectly justified for the integration of secondary cost-effectiveness data. For the third factor, '*group size'*, it is recommended to further look into methodologies for assigning appropriate weights to studies in the dataset and then to implement this information within the respective multilevel models, given that a software may soon be available which allows assigning weights when using MCMC estimation procedures.

5.1.6. Summary and conclusions for the first part of this empirical exercise

The first section of this empirical chapter was concerned with determining the appropriate MLM structure for the integration of published economic evaluation data from international studies on the cost-effectiveness of statins in the primary and secondary prevention of CVD. This assessment was not just concerned with testing which MLM structure previously developed works well on the data collected, but also whether assumptions made are justified for the data.

Starting off with the models previously developed in the MLM methods Chapter 3 and tested in the pilot study, this section showed that appropriate assumptions regarding (in-) dependencies of the data are crucial for making correct inferences when analysing secondary cost-effectiveness data. The pilot study already showed that, due to ignoring that data within studies is not independent, the OLS regression model and the two-level hierarchical model, which clusters data in countries only, were clearly outperformed by those MLM structures which did explicitly account for a study-level. This finding was clearly confirmed by the analysis reported in this section. However, it was also observed in the pilot study that the cross-classified model does not outperform the two-level hierarchical model which groups data in studies and that the country-level does not show noteworthy variability. As a potential reason for this finding, it was hypothesized that the assumption of independence between countries may not be adequate for those studies in the dataset which are '*multinational*' in nature and thereby introduce the issue of cross-classification. If data from multinational studies is less affected by variability on country-level, this potentially '*lays a curtain*' over the overall country-level variability present in the rest of the data from non-multinational studies and therefore casts into doubt whether the assumption of independence between countries actually holds for data of those studies.

For this reason, a three-level hierarchical model, which groups data in studies and studies in countries, was run both on a reduced dataset, where data from multinational studies was dropped, as well as the full dataset where this multinational study data was grouped in a separate cluster on country-level. To confirm that the lack of country-level variability in the cross-classified model

results from the data and not technical issues with model specification and implementation, a cross-classified model was also run on an '*intermediate*' dataset, where some multinational study data was assigned to its respective target countries, and the remaining data clustered in a separate country group, thereby introducing the cross-classification problem on '*a lower scale'*.

This analysis has clearly shown that the country-level variability observed in the model crucially depends on assumptions regarding independence of data from multinational studies on country-level. Whilst country-level variability was negligible for the cross-classified model, it increased dramatically within the bivariate three-level hierarchical framework, both for the full as well as the reduced dataset. In addition, running the cross-classified model for the intermediate dataset resulted in country-level variability somewhere in between the fully cross-classified model and the three-level hierarchical structure. These results clearly confirm that country-level variability may be underestimated in multinational studies, which constitutes an important finding from this research in its own right. Consequently, it was decided to take forward the three-level hierarchical model to the second part of this empirical exercise, as only this model makes appropriate assumptions regarding (in-) dependencies in the dataset, permits the simultaneous assessment of covariates on data, study, and country-level, and allows using the full dataset through grouping multinational study data in a separate group on country-level.

The analysis as detailed above also demonstrates the benefits of decomposing the INMB statistic into its components ΔC and ΔE within the bivariate framework. First, one does not need to re-run models for different threshold values. Secondly, the correlation between the two stochastic components of the INMB statistic is explicitly modelled. Finally, once covariates are included, a bivariate model allows assessing the differential impact of covariates on each response variable whilst acknowledging that ΔC and ΔE are, themselves, correlated. When comparing the univariate and bivariate versions of the tree-level hierarchical model, it got apparent that part of the variability in international cost-effectiveness data '*disappears'* when combining ΔC and ΔE to the INMB statistic. This very interesting finding is subject to further analysis in Section 5.3 of this empirical chapter, which is concerned with country-level variability, or the lack thereof, in both the univariate and bivariate MLM framework.

Finally, this section assessed, in depth, whether Empirical Bayes shrinkage estimation may be regarded as appropriate in a model which attempts to integrate secondary data from published economic evaluation studies, where the weight of a particular study does not depend on individual patients considered, but rather on the extent to which subgroup and sensitivity analyses have been reported. It could be shown that, due to high between group variability in the data, shrinkage factors are generally very high, which means that shrinkage is, at most, moderate. More importantly, however, this section argued that the impact of shrinkage on study means in this exercise depends not just on the respective number of data points abstracted from each study, but also on the within and between group variability in the data (i.e. dependencies) and the location of each study mean relative to the overall regression mean. With respect to '*dependency*' and '*location*', it was argued that shrinkage is perfectly justified within this secondary data integration exercise. For the third factor, '*group size'*, a distinction was made between country groups and study groups. With respect to countries, it was argued that the logic behind shrinkage, which gives higher gravity to geographic domains for which more data is available, may also be justified for countries included in this empirical exercise, even if the underlying data stems from published economic evaluation studies. With respect to studies, there may be arguments both in favour as well as against the appropriateness of shrinkage and it is recommended to further look into methods for assigning appropriate weights to studies in the dataset and then to implement this information in the respective MLM; provided that a software may soon be available which allows assigning weights when using MCMC estimation.

The discussion section (Chapter 6) elaborates further on the findings of this particular section. However, important for the next section is the fact that the three-level hierarchical model, which clusters data from multinational studies in a separate group on country-level, is regarded as the appropriate MLM to take forward for further analyses. This model is used to analyse covariates on data and study-level, which were drawn from a long list of potential variability factors as reported in the literature (Sculpher et al., 2004; Goeree et al., 2007). This assessment of data and study-level covariates is the focus in the next Section 5.2 of this empirical exercise.

5.2. Objective two: assessing variability factors on data and study-level

The previous section was concerned with determining the appropriate MLM structure which best describes complex data structures as present in a set of international cost-effectiveness data abstracted from multiple economic evaluation studies and applicable to multiple geographic domains. Within that section, a number of alternative model architectures were compared, ranging from an OLS regression model, which ignores the existence of complex data structures, up to a cross-classified MLM, which groups international cost-effectiveness data both in the studies it was abstracted from and the countries it refers to. Models which account for the existence of a study-level clearly outperformed the OLS regression model and the two-level hierarchical model which pools data in countries only. Furthermore, it turned out that the independency assumption of data from multinational studies on country-level disguises country-level variability, and that therefore a three-level hierarchical model, which pools data from multinational studies in a separate group on country-level, better fits the data. It was therefore concluded to carry forward this three-level hierarchical model to the second part of this empirical exercise, where the purpose is to assess variability factors which may account for part of the variation in international cost-effectiveness data.

Covariates on data and study-level are systematically assessed which were drawn from a long list of variability factors as obtained from the literature (Sculpher et al., 2004; Goeree et al., 2007) and abstracted from the studies included in the systematic literature review (as reported in Chapter 4). This may not just help gaining an insight into the most important variability factors on data and study-level, but also potentially disclose further variability on country-level (Hox, 2010), which is the focus of assessment in Section 5.3. of this empirical chapter. The following Section 5.2.1 outlines the plan of analysis for this part of the empirical exercise. Subsequently, methods and data to assess factors potentially causing variability in international cost-effectiveness data are introduced in Sections 5.2.2 and 5.2.3 Results of this assessment are presented in Section 5.2.4., before moving on to the third part of this empirical exercise, where the aim is to assess

country-level variability – or the lack thereof – in measures of cost-effectiveness elicited from international economic evaluation studies.

5.2.1. Plan of analysis for part two of this empirical exercise

The key objective of this exercise is to control for variability on data and study-level and thereby to potentially disclose further country-level variability (e.g. Hox, 2010). If successful, this may support the hypothesis that differences in cost-effectiveness results are, at least in part, due to differences between geographic domains, which would justify the three-level structure and also allow testing covariates on country-level. Within a MLM, anything introduced on a lower level might also impact on higher levels, but not vice versa (Hox, 2010). Hence, by introducing covariates on data and study-level, further country-level variability may be disclosed which may then be systematically assessed in Section 5.3 of this empirical exercise.

In other words, if differences between the cost-effectiveness of a health technology between studies are not just due to differences between those studies, but also due to differences between the geographic locations these studies were originally conducted for, then explicitly modelling a country-level better fits the data than applying a framework where data is clustered in studies only. The underlying assumption is that of (partial) exchangeability not just between studies but also between countries represented in the dataset (Spiegelhalter et al., 2000 & 2004; Drummond et al., 2009). In part one of this empirical exercise, this issue was addressed before including covariates to the model and equivalence of the three-level hierarchical model and its two-level counterpart, which ignores the existence of a country-level, was observed in terms of their respective DIC statistic. In this section, covariates are added to the model on data and study-level, and the explicit recognition of a country-level in the three-level hierarchical model allows the assessment of changes in country-level variability through the inclusion of lower-level covariates.

This is also in accord with what is recommended in the MLM literature. A bottom up approach is usually advocated, starting off with the simplest possible model and then including parameters which are tested for significance after they have been added (Hox, 2010). The simplest model is a variance components model which does not include any explanatory variables (Steele, 2008; Rasbash et al., 2009). This model specification was the focus of the previous Section 5.1. In this section, however, covariates are added to the model whilst assuming slopes of covariates to be fixed, resulting in a random intercepts specification (Steele, 2008, Rasbash et al., 2009; Hox, 2010). This is achieved by proceeding from the lower level to the higher level (Hox, 2010). Hence, data-level covariates are tested first, and only once the model is fully specified on data-level, study-level covariates may be added. Once a full random intercepts model with data and study-level covariates has been specified, the next step is to test covariates on country-level, given that sufficient variability in measures of cost-effectiveness exists between geographic domains. This, however, is the focus of the subsequent Section 5.3 of this empirical exercise.

Having outlined an overall analysis strategy for this exercise, it is necessary to operationalize this analysis strategy for the current dataset on the cost-effectiveness of statins in the primary and secondary prevention of CVD. As a first step, one has to determine which covariates in the dataset belong to which level, how covariates within levels may best be arranged in subgroups of 'similar tenor', and which sequence should be applied when analysing subgroups of covariates. This helps braking down the hugely complex task of determining the '*appropriate set of covariates*' (Drummond et al., 2009) out of a large pool of candidate variables and is the focus in Section 5.2.1.1. Subsequently, an analysis strategy is determined for assessing covariates within subgroups in Section 5.2.1.2. This involves producing bivariate statistics, checking for correlations, but also to look into potential additional covariates which may be derived from the raw data, e.g. through the use of data reduction techniques. Finally, further aspects of the analysis strategy with respect to missing values are addressed in Section 5.2.1.3, before moving on to Section 5.2.2 which introduces the full dataset and reports on descriptive statistics of covariates.

5.2.1.1. Covariates, levels, and subgroups of covariates within levels

The overall strategy of building a full random intercepts model is to proceed from the lower level to the higher level (Hox, 2010). Hence, data-level covariates are tested first, and only once the model is fully specified on data-level, study-level covariates may be added. This means that, as a very first step, the variables available from the dataset must be assigned to their respective levels. Note that the variables considered in this section are only those which are directly obtainable from the studies included in this empirical exercise and which consequently relate to the data and study-level only. Country-level covariates (including wider socioeconomic factors and healthcare system characteristics), which may be obtainable from alternative data sources (e.g. WHO databases), are considered in Section 5.3 of this empirical chapter, which is entirely involved with the assessment of country-level variability. To learn more about data and study-level covariates, Chapter 4 gives further details on how covariates were defined for this empirical exercise.

As already mentioned in Chapter 4, the theoretically unlimited space of variability factors potentially relevant for this empirical exercise was initially confined to a long list of factors based upon work from Sculpher et al. (2004) and Goeree et al. (2007), who both systematically reviewed the literature involved with economic evaluation in health to compile a list of 77 unique factors potentially causing variability in economic evaluation data. Based upon this list of variability factors, a data abstraction form was devised, which initially comprised more than 200 variables. Subsequently, this data abstraction form was tested and modified in a pilot study by abstracting data from a subset of 16 papers included in this empirical exercise. Details on both the development of the data abstraction form and the pilot study are available from Chapters 3 and 4. After the pilot study was completed, it was decided to drop around 90 variables, either because data was not available from the studies or because variation was minimal or entirely missing within as well as between studies included in the pilot. Hence, the main data abstraction exercise started off with a form (implemented in MS Excel), which comprised around 100 variables, both on data and study-level (this form is obtainable in MS Word format in Appendix 4.13).

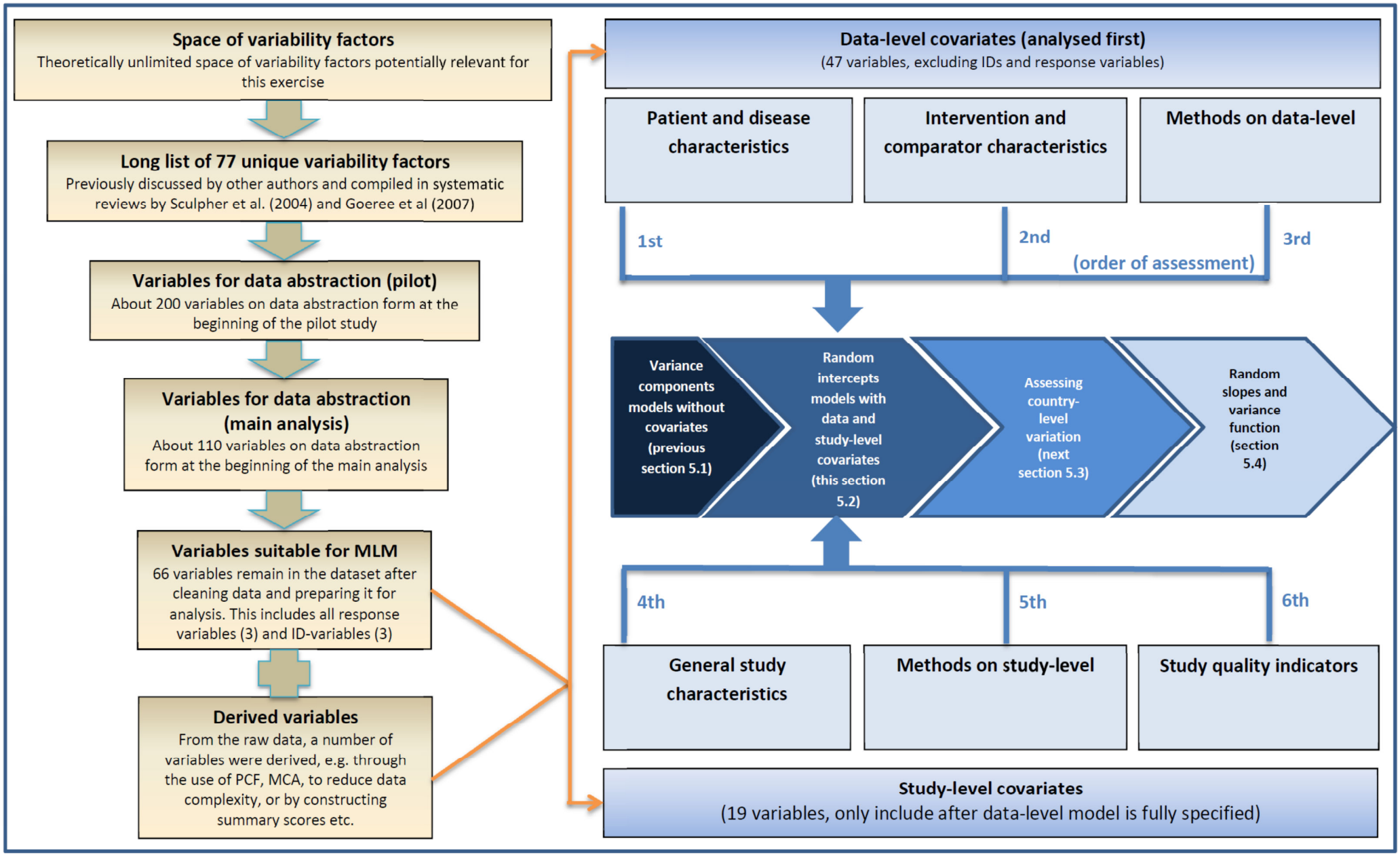
Following the systematic literature review, 67 studies were found includable in this empirical exercise, and data was abstracted from these studies. After cleaning and preparing the resulting data for MLM analysis, 66 variables were imported to STATA 12, which includes both the three response variables (INMB, ΔC and ΔE) as well as ID-variables for each level of the data hierarchy, but excludes any of the derived variables, e.g. from applying data reduction techniques (for instance principal components factor analysis). Descriptive statistics of these variables are presented below in Section 5.2.2.

During the course of data abstraction, it became apparent that some variables vary within studies, whilst others may only vary between studies. This observation is the only basis on which covariates are assigned to either the data or the study-level. For most variables, this decision is clear cut, for instance INMB values are usually reported for subgroups of patients with different cholesterol levels within a study, so that TCL varies within studies and is therefore a data-level covariate. On the other hand, timing is obviously a variable which never varies within, but most certainly between studies and should therefore be assigned to the study-level. However, for some variables this decision was not as clear cut. For instance, the economic perspective of the analysis was initially considered a study-level covariate, until some studies were found which varied their perspective (e.g. Wagner et al., 2009a). Hence, the economic perspective may no longer be regarded as a study-level covariate as there is variation within this categorical variable on data-level. Accordingly, the MLM software MLwiN, which is used to run the MLMs in this exercise, would automatically assign this variable to the data-level (Rasbash et al., 2009; Rasbash et al., 2009a). As a result, from 60 covariates (not considering response variables, id-variables and derived variables), 19 variables were assigned to the study-level, whilst the remaining 41 variables show variation within studies and were therefore assigned to the data-level. In this context it should also be noted that the inclusion of sensitivity analyses results in this empirical exercise was responsible for shifting a number of variables, initially considered as clear study-level covariates, to the data-level, as authors assessed the impact of varying the respective variables within their studies. This holds, for example, for the annual drug cost of the intervention, the discount rate, the economic perspective, the time horizon or the duration of treatment with statins in years.

Hence, in terms of the analysis plan, the 41 variables on data-level are assessed first before proceeding to include study-level covariates. However, though this already brings some structure into this exercise, one needs to devise an order with which variables are tested within each level. For this reason, subgroups of variables are defined and then ordered to further structure this exercise. As before when deriving variables for data abstraction, the studies by Sculpher et al. (2004) and Goeree et al. (2007) served as a starting point for this task. Sculpher et al. (2004) grouped variability factors into i) patient factors, ii) clinician factors, iii) healthcare system factors and iv) wider socioeconomic factors, whilst Georee et al. (2007), who basically confirmed and extended the list of variability factors reported by Sculpher et al. (2004), grouped such factors in i) patient characteristics, ii) disease characteristics, iii) provider characteristics, iv) healthcare system characteristics and v) methodological factors.

As country-level variation is not the focus of this part of the empirical exercise, variables relating to the healthcare system or wider socioeconomic factors are not considered at this point (but rather in Section 5.3 of this empirical chapter). Furthermore, variables relating to clinician factors or provider characteristics are not present in the dataset as studies did scarcely report data on these potential variability factors (the respective variables were dropped after completing the pilot study). Hence, most variables considered here encode patient characteristics, disease characteristics, or methodological factors. However, not all variables in the dataset fall into these categories, which is why further groups of variables were defined. These groups are intervention and comparator characteristics, general study characteristics and study quality indicators. This resulted in a system of six groups of covariates with three groups on data-level and study-level respectively, and this is also presented in the flow chart displayed in Figure 5.6 below.

Figure 5.6: Flow chart outlining the process of covariate assessment within this empirical exercise



Accordingly, Tables 5.9 and 5.10 contain groups of covariates on data and study-level, and also specify the anticipated relationship between each covariate and the respective response variable (INMB, ΔC or ΔE). The last two columns of Tables 5.9 and 5.10 report on existing studies which either support or contradict the anticipated relationship between covariates and response variables. For most studies which were used to inform prior expectations on the relationship between response variables and covariates, actual estimates of incremental cost, incremental effects and INMBs were compared within each study across subgroups of patients, different parameter ranges for sensitivity analyses or different scenarios considered within these studies. Most of these publications where decision analytic modelling studies, whilst three studies analysed individual patient data (Jönnson et al., 1999; Raiku et al., 2007; Tonkin et al.,2006). Only one study in Table 5.9 reports marginal effects from using a form of hierarchical modelling applied to secondary cost-effectiveness data from published studies (Franco et al., 2005).

As reported in Table 5.9, a large number of studies support the relationships anticipated between patient and disease characteristics and response variables (INMB, ΔC and ΔE). For instance, 12 studies are in support of a negative relationship between INMB and the age of patients, amongst those a high quality report by the National Institute for Health and Clinical Excellence, which states that better cost-effectiveness estimates associated with commencing treatment at younger age '*reflect the greater potential to prevent events, and thus the higher utility and cost benefits accrued from remaining event free health state'* (NICE, 2006). Only three studies where found which disagree with this view. Evidence is also very conclusive for the relationship between ΔC and ΔE and the age of patients, and also for other covariates encoding patient and disease characteristics. However, for all other groups of covariates reported in Tables 5.9 and 5.10, there is very little evidence available so that the empirical exercise is rather hypothesis generating than hypothesis testing for covariates other than patient and disease characteristics.

Table 5.9: Data-level covariates and their anticipated relationship to measures of cost-effectiveness

Variable name	Description	Nature of variable	Anticip. relationship			*Supporting evidence*	Contradicting evidence
			INMB	*ΔC*	*ΔE*		
Group 1a covariates: Patient and disease characteristics							
age_cat	*What was the age of the sub-population modelled*	*Ord. cat*	Neg	Neg	Neg	***INMB:** 7, 9-11, 14-16, 23, 24, 26, 31, 33 / **ΔC:**4, 6, 8-11, 14, 16, 23, 24, 26, 31, 33 / **ΔE:**4, 6, 9-11, 14-16, 23, 24, 26, 33*	***INMB:** 8, 13, 31 / **ΔC:** 15 / **ΔE:** 8, 13, 31*
gender_cat	*gender of the population (% of men in population)*	*Ord. cat.*	Pos	Neg	Pos	***INMB:** 5, 6 9-11, 13, 14, 20, 22, 24, 26, 27, 30, 33 / **ΔC:** 5, 6, 9-11, 14, 22, 26, 27, 30, 33 / **ΔE:** 5, 6, 8-11,13, 14, 20, 22, 26, 27, 30, 33*	***INMB:** 8 **ΔC:** 8, 13*
CVD_hist	*CHD related medical history (% of second. Prev. patients)*	*Unord. cat.*	Pos	Neg	Pos	***INMB:** 3, 7, 10, 24, 27, 32, 33 / **ΔC:** 3, 10, 27, 32 / **ΔE:** 3,10, 27, 32, 33*	***INMB:** 11 / **ΔC**: 11 / **ΔE:** 11*
Tcl	*Total cholesterol level baseline*	*Cont.*	Pos	Neg	Pos	***INMB & ΔE:** 6, 9, 10, 26 / **ΔC:** 6, 9, 10*	***ΔC**: 6, 26*
Hdl	*High density lipoprotein level at baseline*	*Cont.*	Neg	Pos	Neg	***INMB, ΔC & ΔE:** 14, 19*	---
Ldl	*Low density lipoprotein level*	*Cont.*	Pos	Neg	Pos	***INMB**:9, 26/ **ΔC**:9, 20, **ΔE:** 26*	***ΔC**: 26*
Hypert	*percentage of hypertensive people in the subsample*	*Cont.*	Pos	Neg	Pos	***INMB**:9, 14, 19, 26/ **ΔC:19 / ΔE: 9, 19,26***	
Sbp	*Mean systolic blood pressure at baseline*	*Cont.*	Pos	Neg	Pos	***INMB:** 6, 9, 14, 26/ **ΔC:19 / ΔE: 9, 19,26***	***INMB:** 19 **ΔC**: 6*
Diab	*Percentage of diabetic patients at baseline*	*Cont.*	Pos	Neg	Pos	***INMB:** 2, 6, 10, 11, 19, 21, 24, / **ΔC: 6, 10,** 11, 19, 31 / **ΔE:2, 10, 21***	***ΔC**: 21, / **ΔE: 6***
smokers	*Percentage of smokers at baseline*	*Cont.*	Pos	Neg	Pos	***INMB:** 2, 9, 14, 21, 26 / **ΔC: 9, 14, 28 / ΔE: 2, 9, 14, 21, 26***	***ΔC**: 21*
risk_cat	*Risk category of the subsample*	*Ord. cat.*	Pos	Neg	Pos	***INMB:** 7,17, 24*	---
			For primary prevent. only				
Group 1b covariates: Intervention and comparator							
intervention	*Brand name of the intervention drug?*	*Unord. cat.*	*?*	*?*	*No effect*	***INMB:** 25*	---
comparator	*Brand name of the comparator drug?*	*Unord. cat.*	*?*	*?*	*No effect*	***INMB:** 25*	---
act_comp	*Comparator 'doing nothing' or comparator 'other statin'*	*Unord. cat.*	Neg	Neg	Neg	---	---
			For 'other statin'				
tdd_int	*What was the total daily dose of the intervention*	*Ord. cat.*	Pos	Neg	Pos	***INMB:** 32 / **ΔC:** 1 /**ΔE:** 1*	***INMB:**1 **ΔE**: 32*
tdd_comp	*What was the total daily dose of the comparator*	*Ord. cat.*	Neg	Neg	Neg	---	---
Cost-int	*What are the annual drug cost of the intervention in 2010 £-Sterling*	*Cont.*	Neg	Pos	*n.a.*	***INMB:** 24*	---
Unitcost_int	*What was the unit cost of the intervention*	*Cont.*	Neg	Pos	*n.a.*	***INMB:** 24*	---
Cost_comp	*annual drug cost of the comparator in 2010 £-Sterling*	*Cont.*	Pos	Neg	*n.a.*	---	---
Unitcost_comp	*What was the unit cost of the comparator*	*Cont.*	Pos	Neg	*n.a.*	---	---
Incr_cost	*What was the incremental annual drug cost of the intervention*	*Cont.*	-	Pos	*n.a.*	---	---
Group 1c covariates: Methodological characteristics on data-level							
outc_measure	*How was health outcome reported in the study*	*Binary.*	*?*	*?*	*?*	---	---
elicitation	*If QALYS were used, what was the method of preference elicitation?*	*Unord. cat.*	*?*	*?*	*?*	---	---
population	*If QALYS were used, what do the utility values reflect (patient / population values)*	*Unord. cat.*	*?*	*?*	*?*	---	---
DRC	*Discount rate on costs*	*Cont.*	*?*	*?*	*n.a.*	---	---
DRB	*Discount rate on benefits*	*Cont.*	Neg	*n.a.*	Neg	---	---
duration	*Treatment duration modelled*	*Ord. cat.*	Pos	Pos	Pos	---	---
extrapol	*Any extrapolation beyond the latest follow up?*	*Binary*	*?*	*?*	*?*	---	---
horizon	*Time horizon?*	*Ord. cat.*	Pos	Pos	Pos	***INMB:** 24*	---
hor_eq_dur	*Does the time horizon equal the treatment duration?*	*Binary*	*?*	*?*	*?*	---	---
Persp_rep	*Study perspective as reported by the authors of the article*	*Unord. cat.*	Pos	Neg	*?*	***INMB:** 7*	---
			For 'societal' only				
Persp_cost_concl	*Study perspective on cost as concluded by the reviewer (health insurance perspective omitted)*	*Unord. cat.*	Pos	Neg	*n.a.*	***INMB:** 7*	---
			For 'societal' only				
data_class	*How was the datapoint classified*	*Unord. cat.*	*?*	*?*	*?*	--	---
basecase	*Was the data point result of a base case or sensitivity analysis?*	*Binary*	*?*	*?*	*?*	--	---
source_effects	*From which source (trial, meta-analysis) was effectiveness data taken from*	*Unord. cat.*	*?*	*?*	*?*	--	---
Barbieri_score_1	*How context specific is the CE estimate judged from the input parameters*	*Unord. cat.*	*?*	*?*	*?*	--	---
Barbieri_score_2	*How context specific is the CE estimate judged from the input parameters*	*Unord. cat.*	*?*	*?*	*?*	--	---

(1) Ara et al. (2009), (2) Ashraf et al (1996), (3) Caro et al (2003), (4) CDC-Group (2002), (5) Davies et al. (2006), (6) Drummond et al. (1993), (7) Franco et al. (2005), (8) Greving et al. (2011), (9) Grover et al (1999), (10) Grover et al (2000), (11) Grover et al. (2001), (12) Grover et al. (2003), (13) Grover et al. (2008), (14) Hamilton et al (1995), (15) HPS (2006), (16) HPS (2009), (17) Huse et al. (1998), (18) Johannesson et al (1997), (19) Jönsson et al (1999), (20) Lindgren et al (2007), (21) Muls et al. (1998), (22) Nagata-Kobayashi et al. (2005), (23) Nherera et al (2010), (24) NICE (2006), (25) NICE (2008b), (26) Perreault et al (1998), (27) Pharoah et al (1996), (28) Raikou et al. (2007), (29) Sigvant et al. (2011), (30) Soini et al (2010), (31) Tonkin et al. (2006), (32) van Hout et al. (2001), (33) Ward et al. (2007)

Table 5.10: Study-level covariates and their anticipated relationship to measures of cost-effectiveness

Variable name	Description	Nature of variable	Anticipated relationshio to			*Supporting evidence*	Contradicting evidence
			INMB	***ΔC***	***ΔE***		
Group 2a covariates: General Study characteristics							
language	*In which language was the paper written?*	*Binary*	*?*	*?*	*?*	---	---
paper_origin	*In which country was the paper written (if authors from several jurisdictions were involved, where is the lead author based?*	*Unord. cat.*	*?*	*?*	*?*	---	---
Timing	*What is the timing of the economic evaluation*	*Unord. cat.*	Pos	Neg	Pos	***INMB:** 1*	---
fund_inst	*primary source of funding (institution)*	*Unord. cat.*	Pos	Neg	Pos *For industry funding*	***INMB:** 1, 2*	---
fund_man	*If funding source was private, which manufacturer was involved?*	*Unord. cat.*	*?*	*?*	*?*	---	---
Author_group_long	*Variable which encodes relationships between published papers in terms of common authorship*	*Unord. cat.*	*?*	*?*	*?*	---	---
Author_group_short	*Variable which encodes relationships between published papers in terms of common authorship*	*Unord. cat.*	*?*	*?*	*?*	---	---
Group 2b covariates: Methodological characteristics on study-level							
gen_des	*What was the general study design?*	*Unord. cat.*	Neg	*?*	*?* *For secondary modelling*	***INMB:** 1*	---
prim_des	*If primary modelling, what was the specific study design?*	*Unord. cat.*	*?*	*?*	*?*	---	---
sec_des	*If secondary modelling, what was the specific study design*	*Unord. cat.*	*?*	*?*	*?*	---	---
effect_calc	*Method of effect calculation*	*Unord. cat.*	Neg	*?*	*?* *For modelling intermediate outcomes*	***INMB:** 1*	---
multinational	*Was the study multinational*	*Binary*	*?*	*?*	*?*	---	---
infl_adj	*Were cost estimates in the model adjusted for inflation?*	*Unord. cat.*	*?*	*?*	*n.a.*	---	---
adj_method	*If cost estimates were adjusted for inflation, what was the adjustment method*	*Unord. cat.*	*?*	*?*	*n.a.*	---	---
cur_conv	*Was currency converted*	*Binary*	*?*	*?*	*n.a.*	---	---
conv_method	*If currency was converted, what was the conversion method used by the authors?*	*Unord. cat.*	*?*	*?*	*n.a.*	---	---
scope	*What was the scope of assessment*	*Unord. cat.*	**Pos**	Neg	**Pos**	***INMB:** 3*	---
Group 2c covariates: Study Quality indicators							
qhes_cata	*What was the overall QHES category given a strict application of the QHES criteria*	*Ord. cat.*	*?*	*?*	*?*	---	---
qhes_catb	*What was the overall QHES category given a pragmatic application of the QHES criteria*	*Ord. cat.*	*?*	*?*	*?*	---	---
Qhes_conta	*What was the overall QHES score given a strict application of the QHES criteria?*	*Cont.*	*?*	*?*	*?*	---	---
Qhes_contb	*What was the overall QHES score given a practicable application of the QHES criteria?*	*Cont.*	*?*	*?*	*?*	---	---
(1) Franco et al (2005), (2) Miners et al (2005), (3) NICE (2006)							

Nevertheless, even without supporting literature, one can speculate about the relationship between covariates and response variables if there is a theory or logical argument which supports this prior expectation, which is why columns four to six in Tables 5.9 and 5.10 also report on anticipated relationships for those covariates for which there was no evidence available from the literature. Why a certain direction of change was anticipated for each covariate, however, will not be discussed here but rather in Chapters 5.2.4 and 6.3.1 in the light of the results from testing covariates within the multilevel models developed in this thesis and with reference to the supporting literature (if available).

Finally, it is important to define a sequence with which groups of covariates are considered in the multilevel analysis. As mentioned by Sculpher et al. (2004), '*In economic evaluation generally, arguably the most important source of variation in cost-effectiveness is between subgroups of patients defined in terms of demographic and clinical factors'*. Sculpher et al. (2004) further state that *'this important source of patient-level variation feeds through to centre or country variation in cost-effectiveness if these subgroups of patients are not evenly distributed between locations*. The amount of literature available to inform expected relationships between patient and disease characteristics and response variables as reported in Table 5.9 underpins the importance of this group of covariates. Therefore, patient and disease characteristics shall be considered first in this part of the empirical exercise. Doing so not just controls for differences between subgroups of patients on data-level, but also '*removes*' the potential impact on variation between studies and countries in the multilevel framework. Note that patient and disease characteristics only refer to differences between subgroups of patients as assessed within the studies included in this empirical exercise like age, gender, cholesterol level at baseline, CVD related medical history or smoking status. Patient and disease characteristics do not refer to differences in demographic or disease characteristics between countries. The latter will be the focus of Section 5.3 of this empirical study.

Following the assessment of patient and disease characteristics, variation caused by differences in the intervention and comparator are considered as this group of covariates is also suspected to be responsible for some variability in measures

of cost-effectiveness of statins in the primary and secondary prevention of CVD. There are a number of different statins on the market, which are considered in different dosages and assessed with respect to different comparators (either '*doing nothing*, '*other statin'* and /or '*different dosage*') within the studies included in this empirical exercise. Hence, one needs to control for this potential source of variation, as otherwise the effect of differences in the intervention or comparator may feed through to the variability in measures of cost-effectiveness observed on study or even country-level.

Once the model controls for the appropriate set of patient and disease as well as intervention and comparator characteristics, the next group of variability factors are methodological differences affecting measures of cost-effectiveness which may vary within studies included in the dataset. Hence, factors like the outcome measure (LYS or QALYs), the discount rate applied to costs or effects, the time horizon, the economic perspective, or whether the respective data point refers to base case or sensitivity analysis, may be assessed. This group of covariates is also concerned with variation with respect to the sources from which input data was drawn. This relates, for instance, to the source of effectiveness data in terms of the trial from which this data was obtained. However, another source of variation refers the '*geographic context specificity'* of input parameters.

Barbieri et al. (2005) grouped studies with respect to differences in the geographic origin of input parameters. Accordingly, there may be differing degrees of variability in measures of cost-effectiveness. This source of variation may even differ within the studies included in the dataset. To assess this potential source of variability, information was first collected from studies to record the geographic origin of the data used to populate the economic model. Secondly, covariates which group studies with respect to their '*geographic specificity'* of input parameters were derived from the raw data, which are subject to multilevel analysis in this section. The idea is analogous to what has been observed with respect to data from multinational studies in Section 5.1 of this empirical exercise, namely that the highest degree of variation on country-level may exist in cost-effectiveness data which is based on target country specific values for all input parameters. Chapter 4, which is concerned with the data abstraction for this empirical exercise, explains in detail how the ideas of

Barbieri et al. (2005) were applied to define a set of categorical variables which encode variation, or the lack thereof, with respect to geographic locations of input parameters. The raw data to derive such covariates has been omitted from further analysis.

Moving on to the study-level, it was decided that general study characteristics (such as timing, paper origin, or the respective funding source) are assessed before methodological aspects on study-level are considered. A reason for this sequence is that general study characteristics (such as paper origin or timing) may partly determine methods on study-level (for example as authors aim to follow a specific HTA guideline applicable to a certain geographic location at a certain point in time), so that it makes sense to assess general study characteristics first. Finally, study quality indicators, resulting from applying the QHES instrument (Ofman et al., 2003) to the studies included in this empirical exercise, are considered to conclude the assessment of data and study-level covariates in this section of the empirical analysis. How data from applying QHES was collected and combined to a single score is also described in Chapter 4.

5.2.1.2. Analysis strategy for covariate Assessment on data and study-level

In the previous subsection, the dataset was divided into subgroups of covariates, and a sequence was established with which groups of covariates are considered in this empirical exercise. This breaks down the complicated task of determining the appropriate set of covariates out of a large number of candidate variables. However, what is required to accomplish this task is a strategy to determine the appropriate set of covariates for inclusion in the final MLM, and this is the focus of this subsection.

First, univariate descriptive statistics are reported, i.e. means and standard deviations for continuous variables before and after imputation of missing values (the missing value strategy is also outlined further below), and proportions for

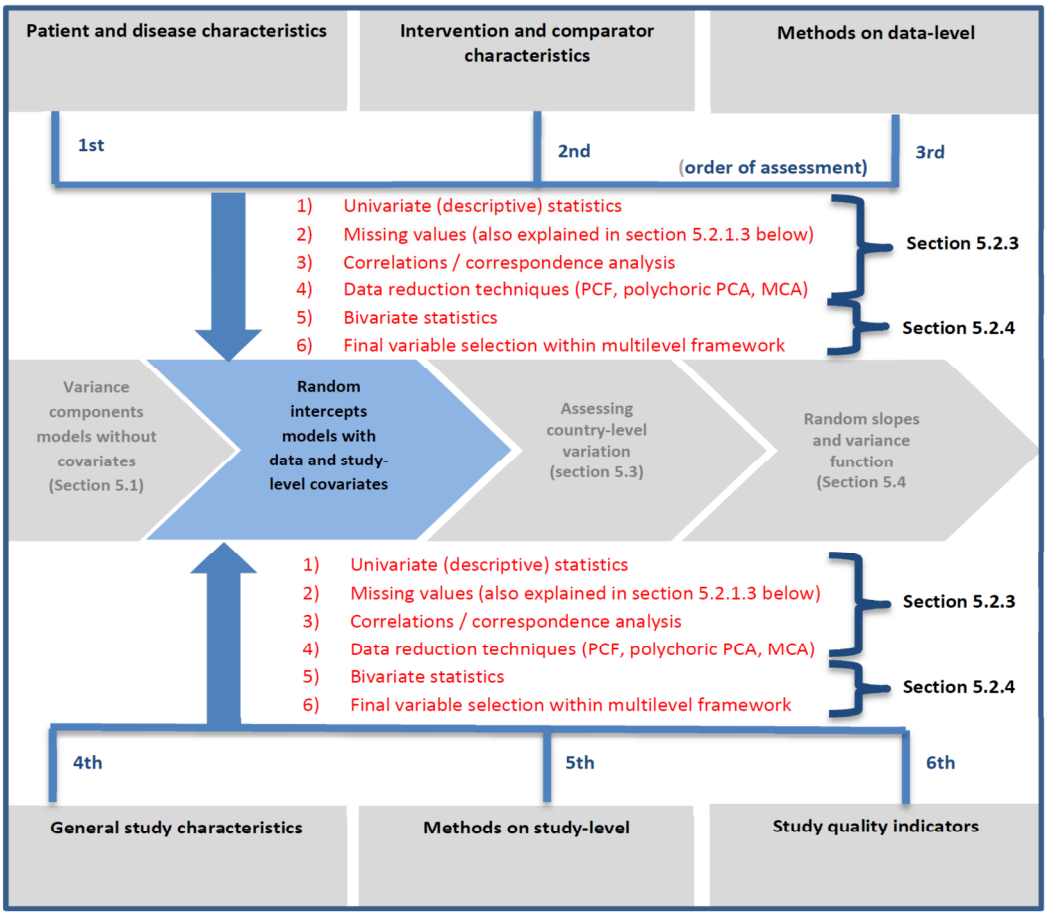
categorical variables respectively. At the very minimum, this establishes that variables considered in this exercise do, in fact, vary (Fielding & Pillinger, 2008). In addition, summarizing and describing the data may help hypothesising about potential patterns and relationships (Fielding & Pillinger, 2008). Apart from descriptive statistics, correlations between subsets of explanatory variables are assessed in depth and approaches to reduce redundancies in the dataset are applied (which is explained further below). Relationships between covariates and response variables are then analysed in more detail by individually entering covariates into the respective MLMs to test whether the anticipated relationship to the response variable (INMB, ΔC or ΔE) holds, and whether this relationship is statistically significant. Finally, a random intercepts model with covariates on data and study-level is developed to control for variability on data and study-level, thereby disclosing the maximum amount of country-level variation, which is then further assessed in Section 5.3. of this empirical exercise.

A systematic approach to covariate selection is imperative in this part of the empirical exercise. Though of tremendous informative value, it is not sufficient to choose variables on grounds of bivariate statistics only. The reason is that a covariate may not show the anticipated sign or may not reach statistical significance because of a confounding factor which has not been controlled for (e.g. Maddala, 2001). Conversely, covariates may individually reach significance, but not once tested in conjunction as they may be highly correlated (Maddala, 2001, Acock, 2010). To aid this choice and to avoid randomly testing subsets of covariates within the MLM framework, it is useful to hypothesize about potential relationships between explanatory variables, and to put such relationships to the test by assessing correlations between explanatory variables. An example in the current dataset are the patient and disease characteristics '*systolic blood pressure (SBP)*', '*hypertension status*' and '*smoking status*'. Hypertension status and SBP may be regarded as alternative measures of a similar physiological pattern, and, as it constitutes a risk factor for elevated blood pressure, smoking status is suspected to be correlated to both SBP and hypertension status. As all three variables are continuous (hypertension status and smoking status are measured as the percentage of patients affected), pairwise correlations may be produced as a first indication of statistical relationships between potential explanatory variables (Acock, 2010). Alternatively, polychoric correlations may

be produced for ordered categorical variables (Kolenikov & Angeles, 2004). For unordered categorical variables, dependencies may be tested using a chi^2 test of association, and the method of multiple correspondence analyses, which plots categories of variables in terms of their chi-squared distances on a low-dimensional space, may be used to further assess patterns in categorical data (Bartholomew et al., 2008; Le Roux & Rouanet, 2004)

Hence, subgroups of covariates defined in the previous subsection are analysed with respect to correlations in the data and based on these correlations, several strategies may be considered. If, for example, two variables are regarded as alternative measures of the same concept, then it may be indicated to simply choose one variable over the other for inclusion in the MLM based on significance, model fit and other issues such as the proportion of missing values imputed for each variable. However, if it is suspected that correlated variables are interrelated measures of a common underlying but unobserved construct, one may consider data reduction techniques. The choice of method thereby depends on the nature of the observed variables and potential candidate methods were already discussed in the context of the genealogy study reported in Chapter 4. One multivariate method often used to assess patterns in data is factor analysis (FA). As Rencher (2002) states, '*the goal of factor analysis is to reduce the redundancy among the variables by using a smaller number of factors.*' These factors are, unlike the observations, unobserved and therefore also referred to as '*latent variables'*. Of particular interest for this part of the empirical exercise is what Acook (2010) refers to as '*principal components factor analysis*' (PCF). PCF may be used if one has a set of items which all measure the same underlying concept (Acook, 2010) and it may be applied to obtain a '*factor score'* which could be used as a covariate instead of the set of items initially observed. The question for choosing amongst the set of covariates is whether the factor score itself, or any of the original items leads to better model fit when considered alternatively in the MLM.

Figure 5.7: Strategy of covariate assessment within subgroups of covariates



However, PCF is only valid for continuous data (Rencher, 2002; Acock, 2010;) and its application in the current exercise may therefore be rather limited. For this reason, a package developed by Kolenikov & Angeles (2004) was considered as an alternative, which implements '*polychoric principal components analysis'* into the software environment STATA. However, Kolenikov and Angeles (2004) state that this method is only valid for continuous as well as ordered categorical variables, which again leaves out the majority of variables relevant for this assessment. For this reason, multiple correspondence analysis (MCA) was considered, which is also referred to as the '*categorical equivalent to PCA'* (Le Roux && Rouanet, 2010). The aim of MCA is to visualize the raw data in a low-dimensional space (usually two dimensions) which then helps to identify patterns in this data. (Bartholomew et al, 2008). It does so by converting categories of variables into points on a plane (the biplot), and the researcher may then analyse the resulting cloud and sub-clouds of points in this geometric space (Le Roux, Rouanet, 2010). A summary score may then be derived which can be used in the MLM as an alternative to the initial set of correlated covariates.

To sum up, the choice of covariates follows a rigorous assessment of relationships between explanatory variables and response variables as well as intercorrelations between explanatory variables. Assessing correlations and patterns in the data as well as the application of data reduction techniques aims to avoid that multicollinearity between explanatory variables distorts the estimation process. This is the focus of Section 5.2.3 of this empirical exercise, which starts off with descriptive statistics before reporting on missing data and imputation thereof and concludes with a rigorous assessment of correlations between subsets of explanatory variables in the dataset. Subsequently, the results Chapter 5.2.4 starts off with reporting on bivariate statistics generated from testing covariates individually in the MLM framework, before a full random intercepts model is developed by applying the above detailed analysis strategy to each subgroup of covariates in the sequence outlined in the previous subsection. This results in a MLM with covariates on data and study-level, which may then be further developed through the inclusion of country-level covariates in Section 5.3, given that there is sufficient variability on country-level after adjusting for data and study-level covariates. Finally, random slopes and the variance function are assessed in a case study in Section 5.4 of this empirical chapter. Before moving on to study methods and introducing the data for this part of the empirical exercise, however, further aspects of the analysis strategy are explained with respect to missing values in the data.

5.2.1.3. Analysis strategy for missing observations

Missing data are, generally speaking, observations which we intended to make but, for some reason, haven't (Carpenter & Kenward, 2007). Within this empirical exercise, missing data arises from the failure of included studies and related sources to report on relevant aspects, for example, patient risk characteristics or methodological characteristics. Missing data occurs both on data and study-level, and it also occurs in categorical as well as continuous variables. There is not one single method, or gold standard, to address missing data issues, though multiple imputation of missing data has gained a lot of popularity in recent years (Carpenter & Kenward, 2007). The primary aim of

missing data imputation is to be able to make valid inferences even in the presence of missing data (Carpenter & Kenward, 2007; Carpenter & Goldstein). However, this validity depends upon the '*mechanism*' by which data is missing (Carpenter & Goldstein).

What is generally referred to as "*missingness mechanism*" goes back to Little & Rubin (1987). The missingness mechanism describes whether data is missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR) (Acock, 2005; Carpenter & Kenward, 2007). Data is missing completely at random if the missingness is completely unrelated to the outcome (Carpenter & Kenward, 2007). If data is MCAR, we may simply ignore data points with missing values, which leads to listwise deletion of the affected data (Carpenter & Kenward, 2007). However, the occurrence of MCAR is rare so that the assumption of data missing completely at random is restrictive. (Acook, 2005; Carpenter & Kenward, 2007; Acook, 2010; Carpenter & Goldstein). Less restrictive is the assumption of data missing at random (MAR) (Acook, 2005; Carpenter & Kenward, 2007). If we can explain the occurrence of missing data in one variable by another variable in the dataset, then the missing data may be MAR. (Acook, 2005; Carpenter & Kenward, 2007; Acook, 2010). In this empirical exercise, the MAR condition may be satisfied if, for example, missing values in the categorical age variable (age_cat) do not depend on the age of the respondents, after controlling for, say, gender. According to Acook (2005), the issue of MAR is not whether gender can predict age_cat, but whether gender is a mechanism to explain whether values in the age variable are missing. The variables which are used to explain whether data is missing or not (in this example '*gender*') may be included as auxiliary variables in a model when imputing missing values (Acook, 2010). A third mechanism of missingness is missing not at random (MNAR). In simple terms, if data are neither MCAR nor MAR, then they are MNAR or informatively missing (Carpenter & Kenward, 2007). Dealing with data MNAR is much more difficult as the observed data does not tell anything definite about the '*relationship between the chance of seeing a variable and its "unseen value" or to "describe how the distribution of the data differs among data points with missing values'* (Carpenter & Kenward, 2007).

The default strategy in many statistical packages, which is only valid under the restrictive MCAR assumption, is listwise deletion. In the case of MCAR it may provide unbiased estimates. However, the case of MCAR may be very rare, which casts into doubt the validity of this approach in many practical applications (Acock, 2005; Carpenter & Kenward, 2007). A further problem occurs if the sample size is small, which in the case of this empirical exercise, holds for the study and country-levels, so that listwise deletion may simply lead to an unacceptable loss of power. Apart from listwise deletion, traditional, or *'ad hoc'* approaches of dealing with missing values include pairwise deletion, mean substitution (overall mean, or mean of subgroups), regression based single imputation or creating a new category for missing values in categorical variables (Acook, 2005). A more elaborated approach, which is gaining a lot of popularity recently, is multiple imputation (MI) of missing values (Acook, 2005; Acock, 2010; Carpenter & Kenward, 2007; Carpenter and Goldstein).

Starting with MI, its idea, which goes back to Rubin (1987), is to treat missing values as random variables, and to impute several values for each missing value, resulting in a number of complete datasets (Rubin, 1987). The imputed values do not contain any unique information, so that the complete dataset restores the original variance-covariance matrix (Alemdar, 2009; Acook, 2010). The first step in MI is to model the variable with missing values as a response of the other variables in the dataset. Then, a (random) draw is being made from the normally distributed residuals, which results in the imputed values (Acock, 2005; Carpenter & Kenward, 2007, Acock, 2010). These values, as mentioned above, do not contain any unique information so that the completed dataset replicates the original variance-covariance matrix (Alemdar, 2009). However, as these values are imputed from a distribution of possible values, a single imputation is inappropriate. Therefore, several values are imputed for each missing value (Acock, 2005; Carpenter & Kenward, 2007, Acock, 2010) This results in a number of (k) datasets, typically somewhere between 5 and 10 (Acock, 2005). After having obtained 'k' imputed datasets, the next step involves analysing each dataset using the model of interest, in this case the hierarchical three-level model, which leads to 'k' regression outputs (Acock, 2005). The final step combines these 'k' regression outputs into one pooled regression-output Acock, 2005. The underlying assumption for MI is that data is, at least, MAR Acock, 2005; Carpenter & Kenward, 2007, Acock, 2010. In adding the variability in the

imputed values to the variability observed in the dataset, the regression results do not suffer from overestimated precision (Acock, 2005; Carpenter & Kenward, 2007, Acock, 2010). In other words, MI captures the uncertainty associated with data being missing (Acock, 2005; Carpenter & Kenward, 2007, Acock, 2010). As the imputed values do not contain any additional information than what is already contained in the dataset, the resulting inference is valid, i.e. unbiased (Acock, 2005; Carpenter & Kenward, 2007, Acock, 2010).

However, an important problem arises when considering the multilevel structure of the model of interest in this exercise. Carpenter & Goldstein highlight that an imputation model '*must have the right variance structure*'. Hence, '*if a dataset is multilevel, then the imputation model must be multilevel too*'. In a simulation study, Gibson and Oleynik (2003) compared different methods of treating missing values in a dataset with a two-level hierarchical data structure. Missing values (10% in one analysis and 40% in another) were generated in level two variables, and two datasets were produced, one with 30, and one with 160 higher-level units. Especially when missing values were high (40%) and the number of higher-level units was low (30), MI performed poorly compared to traditional '*ad hoc*' approaches. The authors concluded that the poor performance of MI in their study may have been due to the fact that the MI procedure was not suitable for hierarchical models when values are missing at level two.

For this reason, alternative options to reflect the multilevel nature of the dataset in the imputation model were considered in this thesis. For instance, Carpenter and Goldstein developed macros for multiple imputation in MLwiN, but unfortunately, it is only capable of handling missing data in level one variables which obviously limits its applicability to the current exercise. An alternative to Carpenter and Goldsteins macros is '*Realcom IMPUTE*', a freeware developed by the Centre for Multilevel Modelling in Bristol (Goldstein, 2009). The advantage of Realcom IMPUTE over the MI macros is that it deals more adequately with categorical and normal data and multilevel structures (Goldstein, 2009). The procedure has three stages. First, the model of interest is being set up in MLwiN where some of the variables have missing data. Secondly, REALCOM-IMPUTE is

run. Then, MLwiN uses the output from REALCOM-IMPUTE to produce the model estimates (Goldstein, 2009). However, Realcom IMPUTE is currently only capable of dealing with two-level hierarchical data, although the author mentions that, in some cases, it may be possible to substitute fixed for random effects. Nevertheless, the procedure may not yet be fully applicable to more complex data structures, as for example the three-level bivariate MLM in this exercise.

Finally, to further look into methods of implementing MI in the MLM framework, the author posted this problem at 'MULTILEVEL@JISCMAIL.AC.UK', an international mailing list for researchers involved with MLM. Though some suggestions were made to tackle this problem in alternative software environments (e.g. implementing the model of interest in MPlus, using STATA for both imputation and implementing the model of interest with GLAMM or STMIXED, or using AMELIA II instead of Realcom), the overwhelming feedback was that each solution would come with its own set of problems and that there is, at this point, probably no appropriate solution to this problem in the case of particular complex data structures.

In conclusion, MI may be, in general, the most sophisticated way to deal with missing data. However, problems arise when the multilevel data structure is not acknowledged appropriately in the imputation model. Whilst some solutions exist to implement multilevel MI, these methods may not yet be advanced enough to deal with more complicated data structures like the bivariate three-level model. This is an area of on-going research, and it may also constitute an interesting topic for future research beyond the scope of this project. However, for the current exercise, an alternative missing value strategy is required.

Hence, regression based single imputation was considered for continuous variables, whilst an extra '*missing category*' was introduced to address missing values in categorical data. In regression based single imputation, one predicts the missing value from a regression model where the dependent variable is the variable affected by missing values, and a vector of explanatory variables which may help predicting the value of the dependent variable (e.g. Briggs et al., 2002).

Hence, this strategy includes defining an imputation model and selecting a number of explanatory variables which one assumes to help predicting the missing values in the affected variable (Briggs et al., 2002). After this model has been fitted, one can use the results to impute the predicted mean from the regression equation given the observed values of the covariates for that data point (Briggs et al., 2002). This works well if the data is MAR (Briggs et al., 2002). In the case of this empirical exercise, one may, for instance, impute missing values on the annual drug cost of the intervention by fitting an OLS regression with the covariates '*timing', 'country', 'drug_name'* and '*drug_dose'*. After running the regression, one may then predict the missing value in '*drug_cost'* by applying the information on the coefficients applicable to the affected data points. As stated on missingvalues.org.uk, regression mean imputation *'can generate unbiased estimates of means, associations and regression coefficients in a much wider range of settings than simple mean imputation.'* Even though the variability of the imputations may be too small, this should be less of a concern compared to more ad hoc imputation techniques like simple mean imputation or imputing the mean for subgroups (missingvalues.org.uk).

To deal with missing values for categorical data, an extra category for missing values was created for each affected variable. This allows using the full dataset whilst showing how the missing cases differ from those where data has been reported. However, one needs to be aware of the fact that this approach potentially lumps together very dissimilar cases in one category. In other words, the "*impact of this strategy depends on how missing values are divided among the real categories, and how the probability of a value being missing depends on other variables*." (missingdata.org.uk). Bias in any direction may be the consequence.

In addition to the above, a binary indicator is used to show whether a value has been imputed in the variable of interest (e.g. Acook, 2005). If such an indicator variable is included in the model of interest, two scenarios are possible. If it is not significant, then imputation has increased the sample size without biasing the results (Morris et al., 2005). However, if the dummy is significant, then it allows estimating an effect for the non-missing values which is not affected by

the imputation of missing values (Morris et al., 2005). In other words, this method leads to the same regression estimates as listwise deletion as the binary indicator variable captures the departure of the imputed cases from the other cases in terms of the outcome variable (Acook, 2005). However, one needs to be aware that when there are several variables with missing values, and this missingness has a common pattern, then using several indicator variables may cause a multicollinearity problem (Acook, 2005).

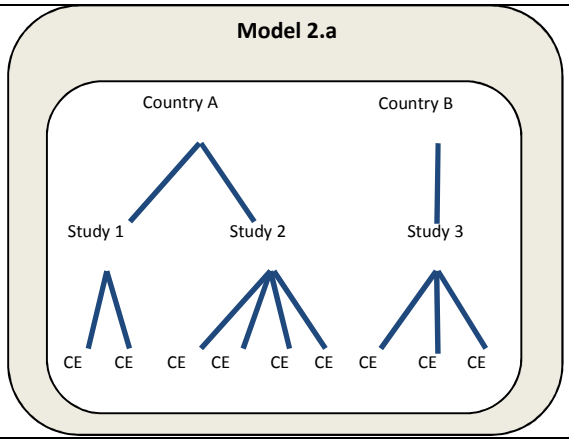
Imputing regression based values for continuous data, assigning missing values to an extra category for categorical variables, and including an indicator variable for missing values is not without controversy in the literature and certainly not the '*best one could do'* (e.g. Acock, 2005). Overestimation of precision and some potential for bias make it necessary to interpret results with caution. However this strategy might be the '*best we can currently do'*, as the use of MI in the presence of complex data structures is an area of on-going research. The next section briefly outlines the methods of analysis, which is basically to apply the three-level hierarchical model in a random intercepts specification to the data. Subsequently, the dataset is summarized in Section 5.2.3, where descriptive statistics before and after imputing missing values are reported. Results from bivariate statistics and the construction of a full random intercepts model with covariates on data and study-level are then presented in Section 5.2.4, before this part of the empirical exercise ends with a brief summary and conclusions in Section 5.2.5.

5.2.2. Methods of analysis for experiment two

This exercise applies the three-level hierarchical model with data grouped in studies and studies grouped in geographic domains, where data from multinational studies is clustered in a separate group on country-level, to a set of secondary cost-effectiveness data on statins in the primary and secondary prevention of CVD. The model is run in its univariate specification with INMB as the only response variable, as well as a bivariate model with ΔC and ΔE as a vector of response variables. Table 5.11 summarizes the models to determine the appropriate set of covariates on data and study-level.

In essence, the only difference between the models run in this part of the empirical exercise and the variance components models 1.b and 1.c in part one, is the inclusion of a vector of explanatory variables on each hierarchical level, denoted with '$\beta_1 x_{1\mathrm{ijk}}$' (data) and '$\beta_2 x_{2jk}$' (study) and '$\beta_3 x_{3k}$' (country) respectively. Note, however, that country-level covariates are only considered later in Section 5.3 of this empirical exercise. As before, the intercept term is denoted with 'β_0' and subscripts refer to each level of the data hierarchy with 'i' representing level one, 'j' level two, and 'k' level three so that '$\mathrm{v}_{0\mathrm{k}}$' is the error term for level three, '$\mathrm{u}_{0\mathrm{jk}}$' the error term for level two and '$\mathrm{e}_{0\mathrm{ijk}}$' the error term for level one respectively. For the univariate specification it is assumed that INMB is normally distributed at each level of the model whilst a bivariate normal distribution is assumed for the bivariate model. As in the bivariate variance components model before, the response indicator 'r' is 1 for ΔC and 0 for ΔE and a separate level for this response indicator is fitted below the data-level. Finally, the bivariate random intercepts model estimates one error variance for each response variable plus their respective covariance on each level. Again, further details on the multilevel methodology applied in this empirical exercise are available from Chapter 3.

Table 5.11: Multilevel models for exercise two

	Model of interest (Three-level hierarchical model)
Model summary	Three-level hierarchical random intercepts model with cost-effectiveness estimates being nested in economic evaluation studies and studies being nested in geographic domains. Data from multinational studies is being clustered in a separate group on country-level.
Unit diagram	
Univariate model specification	$y_{ijk} \sim N(XB, \Omega)$ $y_{ijk} = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + v_{0k} + u_{0jk} + e_{0ijk}$ With $v_{0k} \sim N(0, \sigma^2_{v0})$ $u_{0jk} \sim N(0, \sigma^2_{u0})$ $e_{0ijk} \sim N(0, \sigma^2_{e0})$
Bivariate model specification	$\begin{bmatrix} Y_{0,ijk} \\ Y_{1,ijk} \end{bmatrix} \sim BVN(XB, \Omega)$ $y_{d,ijk} = (\beta_{0d} + \beta_{1d} x_{1ijk} + \beta_{2d} x_{2jk} + \beta_{3d} x_{3k} + v_{0dk} + u_{0djk} + e_{0dijk}) * r_{d,ijk}$ $r_{1,ijk} = \begin{Bmatrix} 1 \; if \; \Delta \; cost \\ 0 \; if \; \Delta \; effects \end{Bmatrix} \quad r_{2,ijk} = 1 - r_1$ With: $\begin{bmatrix} v_{0,0,k} \\ v_{0,1,k} \end{bmatrix} \sim BVN\ (0, \Omega_v)$ where $\Omega_v = \begin{bmatrix} \sigma^2_{v0,0} & \\ \sigma_{v0,01} & \sigma^2_{v0,1} \end{bmatrix}$ $\begin{bmatrix} u_{0,0,jk} \\ u_{0,1,jk} \end{bmatrix} \sim BVN\ (0, \Omega_u)$ where $\Omega_u = \begin{bmatrix} \sigma^2_{u0,0} & \\ \sigma_{u0,01} & \sigma^2_{u0,1} \end{bmatrix}$ $\begin{bmatrix} e_{0,0,ijk} \\ e_{0,1,ijk} \end{bmatrix} \sim BVN\ (0, \Omega_e)$ where $\Omega_e = \begin{bmatrix} \sigma^2_{e0,0} & \\ \sigma_{e0,01} & \sigma^2_{e0,1} \end{bmatrix}$

As before, models were implemented in MLwiN using MCMC estimation procedures (Rasbash et al., 2009a; Browne, 2012). A detailed step by step guide on how to implement all models in MLwiN can be found in Appendix 3. As detailed in the analysis strategy above, covariate selection was supported by preliminary analyses, starting off from descriptive statistics for covariates, regression based imputation of missing values, as well as further analytical procedures to address collinearity between explanatory variables (pairwise correlations, polychoric correlations, correspondence analyses) and to reduce the complexity of the dataset (e.g. through principal components factor analyses or multiple correspondence analysis). Further details on these methods are obtainable from the genealogy study reported in Chapter 4. Preliminary analyses were performed in STATA 12 and are reported in the following section (5.2.3). A final dataset was then exported to MLwiN for bivariate statistics and to build a random intercepts model with covariates on data and study-level, and this is reported in Section 5.2.4 .

5.2.3. Data for experiment two and preliminary analyses

Details on response variables as well as data and study-level covariates are also available from Chapter 4 which focusses on the systematic literature review on the cost-effectiveness of statins in the primary and secondary prevention of CVD. In addition, further particulars and descriptive statistics for the response variables (INMB, ΔC, ΔE), are available from Section 5.1.2 of this empirical chapter. In brief, Chapter 4 reports on how the form for abstracting data from included studies was developed from a long list of potential variability factors drawn from the relevant economic evaluation literature and used to populate a dataset for this empirical exercise. The resulting dataset is used here for the analysis as outlined above.

67 studies were includable in this empirical exercise, providing 2094 estimates of incremental net monetary benefit where the authors also decomposed the INMB statistic, hence explicitly reporting data on ΔC and ΔE of the healthcare

intervention. Within the three-level hierarchical model, data is clustered in 17 countries which were the focus of the 61 single country studies included in this dataset. In addition, 288 data points from six multinational studies are clustered in a separate group on country-level, so that the final dataset consists of n_i=2094, n_j=67 and n_k=18. As some data is applicable to the UK as a whole, but other data only to England/Wales on the one hand or Scotland on the other, three distinct categories were introduced to the dataset to reflect these geographic entities. Table 5.2 in Section 5.1.2 of this empirical exercise provides an overview of the geographic locations represented in the data. Descriptive statistics for all covariates considered in this part of the empirical exercise are available from Appendix 5.

5.2.3.1. Descriptive statistics

1064 (50.81%) of all data points refer to secondary prevention, whilst 958 (45.75%) data points in the sample refer to patients who never experienced a CVD event. 72 data points stem from studies which did not discriminate between primary and secondary prevention of CVD. For those data points referring to primary prevention only, all risk categories are present, with most data referring to a 10 year CVD risk between 20% and 30%. This risk was calculated using the Framingham risk equation (Anderson et al., 1991) and individual risk factors collected within this study. Details are also reported in Chapter 4.

Whilst more than one third of all data points (38.16%) refer to male study populations, roughly one quarter refers to females (27.51%) or mixed populations (26.55%) respectively. In 7.78% of the cases there was no clear indication of gender. Descriptive statistics further reveal that mean TCL in the overall dataset was 6.68 (SD: 1.204), SBP for all patients was at a mean of 137.48 (SD: 13.35) and the hypertension status, diabetes status and smoking status was positive in 31.7%, 17.5% and 29.1% of all patients under assessment. Further patient and disease characteristics are available from Tables 5.1.1 and 5.1.2 of Appendix 5

Moving on to intervention and comparator characteristics, it turns out that simvastatin is by far the most prominent intervention under assessment, with 1080 (51.58) of all data points referring to this statin. In 87.58% of all cases, the intervention was compared to '*doing nothing*', whilst in 258 data points the comparator was a statin, either of a different kind as the intervention, or of the same kind but in different dosages. In two cases, the comparator was unclear. The mean annual drug cost of the intervention (converted to £-Sterling using Purchasing Power Parities (PPP) and updated to 2010 using country specific GDP deflators (Shemilt et al., 2008; OECD, 2010) are £521.59. The annual drug cost are highest for lovastatin at £932.14 (SD: 515.14), followed by pravastatin at £858.00 (SD: 236.67), atorvastatin at £503.89 (SD: 232.19), simvastatin at £477.70 (SD: 312.81), rosuvastatin at £337.28 (SD: 247.16) and fluvastatin at £293.08 (SD: 103.10) respectively. Finally, the unit cost of the intervention (across all statins under consideration) is around £0.05 per mg. Again, per statin, unit cost are highest for lovastatin at £0.075/mg (SD: 0.008), followed by pravastatin at £0.061/mg (SD: 0.021), simvastatin at £0.048/mg (SD: 0.033), atorvastatin and rosuvastatin at £0.044/mg each (SD: 0.041 and 0.021) and fluvastatin at £0.013 (SD: 0.012) respectively.

In the majority of cases (62.99%), life years saved was the outcome measure of choice compared to 37.01% of data points where QALYs were used to assess the incremental effect of statins. If QALYs were used, then utility weights usually reflect population preferences (447 vs. 215 of the 775 data points with QALYs). The duration of drug treatment in years peaked at 5 to 10 years (37.63%) and lifetime (38.87%), whilst almost all data points were affected by some sort of extrapolation beyond the latest follow up (92.74%). Accordingly, the time horizon was most frequently lifetime with 1333 data points (63.66%), and the second most common time horizon was between 10 and 15 years (15.75%). In the majority of cases (1328, 63.42%) drug treatment was assumed to last exactly as long as the time horizon modelled, and in 766 cases (36.58%) the duration of drug treatment in years was shorter than the time horizon under consideration. Few studies varied their economic perspective to show the impact of the perspective on cost-effectiveness results (e.g. Wagner et al., 2009a). Hence, the economic perspective is a variable on data-level. In most cases, authors reported to have used a health insurance (NHS) perspective (1369, 65.38%) and in 214

cases (10.22%), authors mentioned to have used a societal perspective. In 244 cases, no perspective was reported whatsoever. However, in this thesis, the economic perspective as reported by authors was compared to own judgement and in 635 cases there is disagreement between what studies report and what the author of this thesis concludes. Hence, the economic perspective, as judged in this thesis, is that of the provider in 20 cases (0.96%), that of the health insurance (NHS) in 1939 cases (92.60%) and a societal perspective was used in 135 cases (6.45%). Again, descriptive statistics of covariates are also reported in Appendix 5.1 of this empirical exercise.

About half (53.72%) of all data points refer to base case analysis, whilst 969 (46.28%) of data points refer to sensitivity analyses results. If a data point refers to sensitivity analyses, the most common variable under assessment was the treatment duration or time horizon (288, 13.75%), followed by the discount rate for cost or effects (257, 12.27%), and the cost of the intervention (annual drug cost), with 86 (4.11%) cases. When looking at sources of effectiveness data, the most prominent way to obtain this data was through a systematic review of the relevant effectiveness literature and/or a meta-analysis. The second most important source of effectiveness data was the 4S trial, a multinational study focussing on Scandinavian countries which looked into the effectiveness and cost-effectiveness of simvastatin (Scandinavian Simvastatin Survival Study Group, 1994). Other prominent sources of effectiveness data were the Heart Protection study (HPS collaborative group, 2002) with 280 (13.37%) data points and the Excel study (Bradford et al., 1990) with 120 (5.73) data points. In total, 19 different sources of effectiveness data were found in the relevant literature. Finally, looking at the geographic specificity of input parameters following the example of Barbieri et al (2005), one can see that by far the most common way to populate an economic model is to use country specific cost and resource use data, but effectiveness data and utility weights from other geographic domains (type CR, 1033 data points, 49.33%). Also quite common is the use of unit cost, resource use and utility weights from the target country, whilst effectiveness data is being transferred from other geographic domains (513, 24.50%). Only 56 data points (2.67%) were completely target specific, so that the economic model was fully populated with target country specific data.

Moving on to the study-level, the first group of covariates to consider are general study characteristics. 61 out of 67 studies were written in English language, and the most common geographic origins of publications included in this exercise were Canada and the UK, with 13 studies (19.50%) each, followed by the USA with 12 studies (17.91%). Most papers were somehow related through common authorship, and considering all relationships between papers published and included in this exercise, only 18 papers are not linked to each other, whilst 29 papers (43.28%) belong to one big group of common authorship. The most published authors in the field are Grover and Jonnsson, each involved in seven studies, Coupal (six studies) as well as Szucs and Zowal with five studies each. The timing of studies (which is not the year of publication) ranges from 1988 to 2009, with two peaks between 1995 to 1998 and 2005 to 2007. As mentioned, the majority of papers focussed on one single country (91.04%), whilst six papers (8.96%) were multinational in nature. Industry was involved in the funding of 39 publications (58.21%), whilst funding was unclear for 17 studies (25.37%). If industry funding was available, then the manufacturers most commonly involved were Pfizer with 13 studies and Merck with 12 studies in the dataset.

Methods on study-level show that most studies (61; 91.04%) relied on secondary modelling, whilst only six studies (8.96%) made direct use of individual patient data from RCTs. If secondary modelling, the most common model used was a Markov state transition model (41; 61.19%). Seven studies (10.45%) were based on decision trees, and other modelling approaches involved life tables, or discrete event simulation. An important question is how effectiveness was measured and modelled within a study, and there are two general approaches. Most studies (61.19%) modelled the reduction in risk to experience a CVD event in the future to estimate incremental effectiveness. 26 studies (38.81%), however, used the intermediate outcome of cholesterol reduction to approximate its impact on CVD risk, which then resulted in an estimation of life years or QALYs saved. The majority of 35 studies (52.24%) explicitly looked into the effect of statins on coronary heart disease (CHD) and cerebrovascular disease (CD), whilst 18 studies (26.87%) looked at CHD only. 11 studies (16.42%) looked at CHD, CD and peripheral arterial disease (PAD) simultaneously. Inflation adjustment of cost estimates to a common baseline year was explicitly reported in 18 studies (26.87%), whilst this was unclear in 35 cases (52.24%), If inflation

adjustment was applied, the healthcare component of the target countries consumer price index was most commonly used (10 studies, 14.93%). Finally, currency conversion was applied in 15 studies (22.39%), and the most common method of currency conversion was the use of real exchange rates (11, 16.42%).

Finally, an indication of study quality was provided by using the QHES instrument (Ofman et al., 2003). This quality checklist was applied to each of the 67 studies to obtain a summary score (between zero and 100), which may be used as an explanatory variable on study-level. Unfortunately, applying the QHES instrument was not entirely straightforward, for instance, as some dimensions on that checklist are comprised of several sub-categories, so that studies may tick part of that dimension, but not each aspect of it. For this reason, two approaches were considered when combining scores of individual QHES dimensions to one overall score of study quality. First, the instrument was applied in a strict sense, meaning that points were only given to a study if each sub-category in a particular dimension was ticked positive. Secondly, a pragmatic approach was applied, where the score for each dimension was divided by the number of subcategories in that dimension and points were allocated for each sub-category. Obviously, this resulted in higher scores for most studies in the dataset (more details on how the QHES instrument was operationalized are available in Chapter 4). Both scores are applied as continuous variables and the mean QHES score following strict criteria is 59.36 (SD: 16.33), whilst the '*pragmatic approach'* led to a mean score of 69.32 (SD: 13.89). However, scores were also converted to categorical variables to better reflect the lack of precision when using tools like the QHES instrument, and results indicate that a strict application of the QHES better discriminates between studies in the dataset.

5.2.3.2. Missing values

As mentioned in Section 5.1.3., there were no missing observations in any of the dependent variables (INMB, ΔC, ΔE). However, descriptive statistics show a multivariate pattern of missingness (Briggs et al., 2003) with missing values in both continuous and categorical variables on data and study-level. Table 5.12 summarizes missing values both on data and study-level.

Table 5.12: Missing values

Variable	Nature of variable	Level	Group	Obser-vations	Missing	% missing	Correlation between missingness indicator and response variable		
							INMB	ΔC	ΔE
TCL	Cont.	Data	Pat/Dis	1193	901	43.03	-0.1106***	0.3286***	-0.2849***
HDL	Cont.	Data	Pat/Dis	1147	947	45.22	-0.1326***	-0.3366***	-0.3098***
LDL	Cont.	Data	Pat/Dis	926	1168	55.78	-0.3316***	-0.0015	-0.3103***
Hypert	Cont.	Data	Pat/Dis	826	1268	60.55	-0.3231***	-0.1960***	-0.4099***
SBP	Cont.	Data	Pat/Dis	1140	954	45.56	-0.1341***	-0.3390***	-0.3125***
Smokers	Cont.	Data	Pat/Dis	1141	953	45.51	-0.1364***	-0.3395***	-0.3149***
Diab	Cont.	Data	Pat/Dis	1163	931	44.46	-0.1287***	-0.3365***	-0.3062***
Age_cat	Cat.	Data	Pat/Dis	2020	74	3.53	0.0178	-0.0631**	-0.0183
Gender_cat	Cat.	Data	Pat/Dis	1931	163	7.78	-0.0103	-0.0602***	-0.0429**
Risk_cat	Cat.	Data	Pat/Dis	2042	52	2.48	-0.0592***	-0.0229	-0.0679***
Intervention	Cat.	Data	Int/comp.	1746	348	16.62	-0.1468***	-0.1486***	-0.2191***
Cost_int	Cont.	Data	Int/comp.	1957	137	6.54	-0.0093	-0.101***	-0.0645**
Unitcost_int	Cont.	Data	Int/comp.	1738	356	17.00	-0.1495***	-0.1526***	-0.2239***
Tdd_int	Cat.	Data	Int/comp.	1738	356	17.00	-0.1495***	-0.1526***	-0.2239***
Comparator	Cat.	Data	Int/comp.	2092	2	0.10	-0.0076	-0.0140	-0.0148
Cost_comp	Cont.	Data	Int/comp.	2092	2	0.10	-0.0076	-0.0140	-0.0148
Unitcost_comp	Cont.	Data	Int/comp.	2083	11	0.53	-0.0175	-0.0830*	-0.0373*
Tdd_comp	Cat.	Data	Int/comp.	2083	11	0.53	-0.0175	-0.0830*	-0.0373*
Elicitation	Cat.	Data	Meth, DL	2077	17	0.81	-0.0238	-0.0224	-0.0346
Elicit_short	Cat.	Data	Meth, DL	2077	17	0.81	-0.0238	-0.0224	-0.0346
Population	Cat.	Data	Meth, DL	2008	86	4.11	-0.0311	-0.0604*	-0.0625**
Duration	Cat.	Data	Meth, DL	1974	120	5.73	-0.1667***	0.3141***	0.0181
Duration_short	Cat.	Data	Meth, DL	1974	120	5.73	-0.1667***	0.3141***	0.0181
Fund_inst	Cat.	Study	Gen, SL	50	17	25.37	0.0465	-0.0266	0.0289
Fund_man	Cat.	Study	Gen, SL	52	15	22.39	-0.0458	-0.0774	-0.0706
Sec_des	Cat.	Study	Meth, SL	54	13	19.40	-0.0971	0.0464	0.1019
Infl_adj	Cat.	Study	Meth, SL	32	35	52.24	0.0334	0.1606	0.0937
Adj_method	Cat.	Study	Meth, SL	32	35	52.24	0.0334	0.1606	0.0937
Conv_method	Cat.	Study	Meth, SL	63	4	5.97	-0.0149	0.1606	0.0937
scope	Cat.	Study	Meth, SL	64	3	4.48	-0.1104	0.0066	-0.0917

Most continuous variables are heavily affected by missing values. For instance, patient and disease characteristics show missing values between 43% (total cholesterol level) up to 60.55% (hypertension status in % of patients). In contrast, missing values for continuous intervention and comparator characteristics are much lower, with 6.5% for the annual drug cost of the intervention up to 17% for the unit cost of the intervention. Missing values in categorical variables on data-level are generally modest, between 0.1% (comparator) and 7.78% (gender), though two categorical variables are affected more strongly by missing values, namely 'intervention' and the total daily dose thereof with 17% missing values each. Moving to categorical variables on study-level, two variables show missing values of less than 10% (currency conversion method, 5.97% and scope, 4.48%). However, three variables show between 19.4% and 25,37% (secondary design, funding manufacturer, and funding institution, and two variables show even more than 50% missing values (inflation adjustment and adjustment method). It

is therefore questionable whether, from a missing values perspective, these two variables should enter multilevel analysis.

To gain further insights into the mechanism by which values are missing, a binary variable was created for each affected covariate (irrespective of whether continuous or categorical in nature) with 1 if a value is missing and 0 otherwise. Next, correlations between the response variables and the missing values indicators were generated. For this matter, point biserial correlations were chosen, a special case of Pearson correlations in which one variable (the missningness indicator) is dichotomous, and the other variable (the response variable) is continuous in nature (Cox, 1974). Results are presented in Table 5.12 above. The highly significant correlations between continuous patient and disease characteristics and the response variables do not speak in favour of any ad hoc imputation method (like simple mean imputation) as there is a high potential for bias. However, if auxiliary variables for the imputation model are chosen well, bias may be minimised when using a regression based imputation approach (missingvalues.org.uk). Note that bias is also addressed through the use of binary indicator variables in the MLM framework. If this indicator variable is not significant in the model of interest, then imputation increased the sample size without biasing the results. However, if significant, then regression estimates are equal to those obtained through listwise deletion as the binary indicator variable captures the departure of the imputed cases from the other cases in terms of the outcome variable (Acock, 2005; Morris et al., 2005).

To define a regression based imputation model for continuous variables, the dataset was first screened for covariates potentially related to missingness in affected variables. Subsequently, a logistic regression model was run with the binary missingness indicator of the affected variable as a response and each candidate for the imputation model as an explanatory variable (Acock 2010). As a result, the set of explanatory variables for the actual imputation model for each affected continuous variable was defined as those covariates significant in the logistic regression (Acock, 2010). The output from logistic regressions to determine variables for the imputation model are reported in Appendix 5.2. Note that logit models may experience convergence problems in the case of very few

missing values (Acook, 2010). For this reason, imputations for the annual drug cost of the comparator as well as the unit cost of the comparator were based on the same set of variables as have been chosen for the imputation models of the intervention cost and unit cost of the intervention respectively (though the variable '*intervention*' was obviously replaced by the variable '*comparator*'). Descriptive statistics of continuous variables before and after imputation of missing values can be found in Table 5.13 below.

Table 5.13: Descriptive statistics of continuous variables before and after regression based imputation of missing values

Variable	Obser-vations	Nr. of obser-vations imputed	Mean before impu-tation	SD before impu-tation	Mean after impu-tation	SD after impu-tation	Difference in means	H_0 diff ≠ 0 Pr(|T| > |t|)
tcl	1193	901	6.676	1.204	6.631	1.005	0.045	0.2533
hdl	1147	947	1.168	0.1023	1.177	0.0924	-0.009**	0.0165
ldl	926	1168	4.509	1.036	4.639	0.9659	-0.13***	0.0009
hypert	826	1268	0.317	0.381	0.341	0.275	-0.024*	0.057
sbp	1140	954	137.47	13.35	137.92	10.05	-0.45	0.2815
smokers	1141	953	0.291	0.3348	0.298	0.25	-0.007	0.5002
diab	1163	931	0.178	0.3491	0.195	0.2662	-0.017	0.1144
cost_int	1957	137	521.59	335.1	528.84	326.33	-7.25	0.4855
unitcost_int	1738	356	0.0504	0.0321	0.0459	0.0309	0.0045***	0.0001
cost_comp	2092	2	26.06	115.37	26.09	115.32	-0.03	0.9913
Unitcost_omp	2083	11	0.004	0.018	0.005	0.0285	-0.001	0.0299
* Difference in means significant at the 1% level , ** Difference in means significant at the 5% level , *** Difference in means significant at the 10% level								

As can be seen from Table 5.13, means of continuous variables before and after imputation are very much in the same range and only in three cases (HDL, LDL and unitcost_int) is the difference significant above the 95% confidence level. However, a binary variable is included in the MLM to capture the departure of the imputed from the non-imputed cases to make sure that imputation does not bias regression results (Acock, 2005; Morris et al., 2005). Further, looking at the standard deviations before and after imputation shows the main weakness of regression based imputation compared to MI. The standard deviations after imputation are generally lower than before imputation. This is because of the increased sample size and the fact regression based imputation generally leads to underestimated variability in the data (missingvalues.org.uk). Nevertheless, imputation of missing values allows using the full dataset with 2094 estimates of cost-effectiveness clustered in 67 studies and 18 geographic domains. The following Subsection 5.2.3.3 focuses on correlations in the data, before bivariate statistics are reported in Section 5.2.4.

5.2.3.3. Correlations between potential explanatory variables

Correlations are assessed between subgroups of potential explanatory variables. First, patient and diseases characteristics are considered. As all variables in this group are either continuous or ordered categorical, polychoric correlations are assessed using a STATA package developed by Kolenikov & Angeles (2004). Results are reported in Table 5.14 below. Starting off with measures of cholesterol, high collinearity was expected between TCL and LDL as TCL is comprised of LDL, HDL and triglycerides (e.g. Friedewald et al., 1972) and as LDL is the dominating factor for TCL. Hence, if LDL changes in one direction, so may TCL in most cases. As Table 5.14 shows, this expectation was met by the data as almost perfect collinearity was observed between both cholesterol measures. This observation, plus the previously reported fact that there are much less values missing for TCL, speaks in favour of dropping LDL from further analysis. Moving on to HDL, its impact on TCL may be offset by simultaneous changes in LDL or triglycerides in any direction, which is why neither a strong correlation was anticipated between HDL and TCL nor between HDL and LDL. This expectation was also met by the data.

Table 5.14: Polychoric correlations between patient and disease characteristics

	tcl	hdl	ldl	hypert	sbp	diab	smokers	age_cat	gender_cat	risk_cat	CVD_hist
tcl	1.00										
hdl	0.14	1.00									
ldl	0.96	0.12	1.00								
hypert	-0.12	0.00	-0.17	1.00							
sbp	0.03	0.18	0.00	0.84	1.00						
diab	-0.05	0.14	-0.06	-0.12	0.03	1.00					
smokers	-0.02	-0.15	-0.02	0.84	0.70	-0.20	1.00				
age_cat	-0.26	-0.12	-0.29	0.12	0.09	0.06	0.00	1.00			
gender_cat	-0.38	-0.06	-0.39	0.14	0.07	0.05	0.02	0.27	1.00		
risk_cat	-0.36	-0.17	-0.41	0.45	0.36	-0.12	0.29	0.43	0.59	1.00	
CVD_hist	-0.31	-0.20	-0.34	0.34	0.25	-0.25	0.19	0.27	0.56	0.86	1.00

The next group of variables for which strong correlations were anticipated are SBP, hypertension status and smoking status. Again, this expectation was confirmed by the data. However, unlike TCL and LDL, which may be used as

alternative measures of cholesterol, one may not simply drop say SBP in favour of hypertension or smoking status as all variables, though principally concerned with the circulation system, measure different aspects of it. On the other hand, if simultaneously used in a MLM, one almost certainly runs into problems of multicollinearity (e.g.Maddala, 2001). For this reason, principal components factor analysis (PCF) was applied to reduce the complexity of the dataset (Acock, 2010). When applied to the data, '*PCF identifies components that are composites of the measured variables*' (Acock, 2010). PCF may therefore be used when developing a scale '*where one dimension is identified to represent the core of a set of items*' (Acock 2010). This scale may then be used as an alternative covariate in the MLM, and referred to as circulation related CHD risk. Table 5.15 below summarizes the results from running a PCF on SBP, hypertension and smoking.

Table 5.15: Principal components factor analysis on SBP, hypertension and smoking

Factor	Eigenvalue	Difference	Proportion	Cummulative
Factor1	2.485	2.121	0.828	0.828
Factor 2	0.364	0.214	0.121	0.949
Factor 3	0.150	--	0.050	1
LR test: independent vs. saturated chi2(3) – 4171.76 Prob>chi2 = -0.0000				

In PCF, the sum of Eigenvalues equals the number of items, three in this case (Bartholomew et al., 2008; Acock, 2010). Hence, from the fact that Factor 1 has an Eigenvalue of almost 2.5 (82.5% of the sum of Eigenvalues), one can conclude that all three items fall along one dimension. The corresponding factor score, which is standardised with a mean of zero and a standard deviation of 1, was estimated in STATA 12. This score may be tested as an alternative to the components entering the PCF (SBP, hypertension and smoking).

Further patient and disease characteristics which are highly correlated are '*risk_cat*', an ordered categorical variable encoding CVD risk of patient subgroups, and '*CVD_hist*', an ordered categorical variable encoding the CVD related medical history of patients. This is not surprising as the variable '*risk_cat*' is comprised of five risk categories for primary prevention (from very low to extreme), plus an extra category for secondary prevention as the Framingham

risk equation, which was used to estimate CVD related risk, is not valid if patients already experienced a CVD event. As a result, the secondary prevention category of *'risk_cat'*, which comprises almost 50% of all data points, perfectly coincides with the secondary prevention category of the variable '*CVD_hist*'. Hence, one may only consider these variables alternatively and choose the one which best improves the fit of the model. Finally, correlations were also observed between individual patient risk characteristics and '*risk_cat*'. This is again not surprising as individual patient risk factors, such as *'TCL', 'LDL', 'gender', 'SBP', 'smoking status', 'diabetes'*, etc. were used as factors in the Framingham risk equation which was applied to derive the categorical risk variable '*risk_cat*'. For this reason it is questionable whether issues of multicollinearity permit simultaneous use of such variables in a multilevel model.

Moving on to characteristics of the intervention and comparator, high correlations were expected between the type, the unit cost and the total daily dose of either intervention or comparator on the one hand and their respective annual drug cost on the other. The reason is that '*unit cost', 'total daily dose'* and '*type*' may be regarded as determinants of '*annual drug cost'* of either the intervention or the comparator. Therefore, one may either use annual drug cost individually, or a combination of its determinants simultaneously in a MLM though the latter choice becomes more problematic as there is also high correlation between the total daily dose of the intervention and its respective unit cost, which again may cause a multicollinearity problem. Note that intervention and comparator have been omitted from Table 5.16 below as they are neither continuous nor ordered categorical variables, so that neither Pearson nor polychoric correlations apply. However, to confirm expectations, both variables were transformed into a set of binaries and point biserial correlations were computed to assess the association between brand and annual drug cost. In addition to the above, the two variables '*comparator*' and '*active_comparator'* are almost identical, as the latter is simply a reduced binary version of the former. Both variables coincide in 1834 (87.58%) cases where the comparator is '*doing nothing'*. Hence, simultaneous use in a multilevel model is not indicated.

Table 5.16: Polychoric correlations between patient and disease characteristics

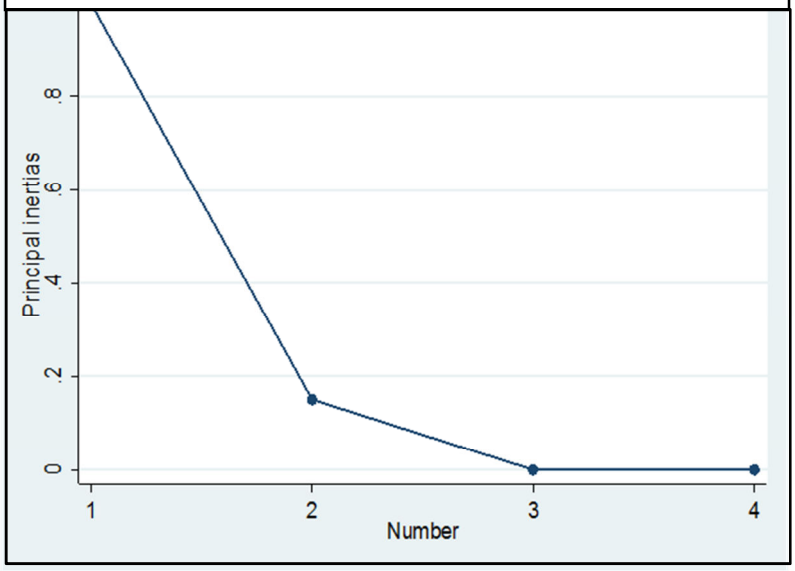
	Tdd_ int	Unitcost _int	Cost_ int	Tdd_ comp	Unitcost_ comp	Cost_ comp	Incr_ cost
Tdd_int	1						
Unitcost_int	-0.705	1					
Cost_int	-0.021	0.669	1				
Tdd_comp	0.407	-0.420	-0.159	1			
Unitcost_comp	0.449	-0.097	0.172	0.408	1		
Cost_comp	0.361	-0.046	0.284	0.416	0.847	1	
Incr_cost	-0.364	0.458	0.267	-0.390	-0.078	-0.055	1

Moving on to methodological characteristics on data-level, one can see that, in contrast to patient and disease or intervention and comparator characteristics, this set of variables is almost exclusively categorical in nature, with discount rates on cost and benefits being the only continuous exceptions. For this reason, another strategy was considered for assessing associations between categories of variables as Pearson correlations, polychoric correlations or point biserial correlations do not apply. The same method was used which was previously utilised for the genealogy study reported in Chapter 4. The two continuous variables DRC and DRB were first transformed into ordered categoricals and multiple correspondence analysis was applied to assess which categories of variables correspond strongly with each other (Bartholomew et al., 2008; Le Roux & Rouanet, 2010). A key advantage of MCA is that it provides a graphical presentation of correspondences between categories of variables on a plane with corresponding categories being in closer proximity on that plane. However, a related disadvantage is that such planes quickly become inextricable due to the number of categories under assessment (van Kerm, 1998). In other words, the principle '*garbage in garbage out'* certainly applies to MCA just as much as it does other quantitative methods.

For this reason, the method is used with caution and methodological variables on data-level are first further divided into groups of conceptually related variables, and these smaller groups of variables are then separately assessed using MCA if appropriate. Methodological variables on data-level are categorized in subgroups relating to a) outcome-measurement, b) time horizon and time preference c) perspective d) data classification and e) geographic source of input data. MCA

was deemed appropriate in three cases. First, '*outc_measure', 'elicitation'* and '*population*' were tested as they all relate to the way effectiveness is defined and elicited in a study. Secondly, '*DRC', 'DRB', 'duration' 'extrapol' 'horizon'* and '*hor_eq_dur*' are tested as variables concerned with time horizon and time preference. Finally, whilst including variables on data-level, different combinations of general study characteristics and methods on data and study-level are tested in an explorative exercise to find out whether there are correspondences between key characteristics of international methods guidelines for economic evaluation in health. If correspondences between categories of variables are revealed, a summary score is produced to reduce the complexity of the dataset analogously to what has been reported above when using PCA for measures of circulation related risk. This summary score may then be used as an alternative construct in the MLM.

Figure 5.8: Screeplot for MCA on outc_measure, elicitation and population

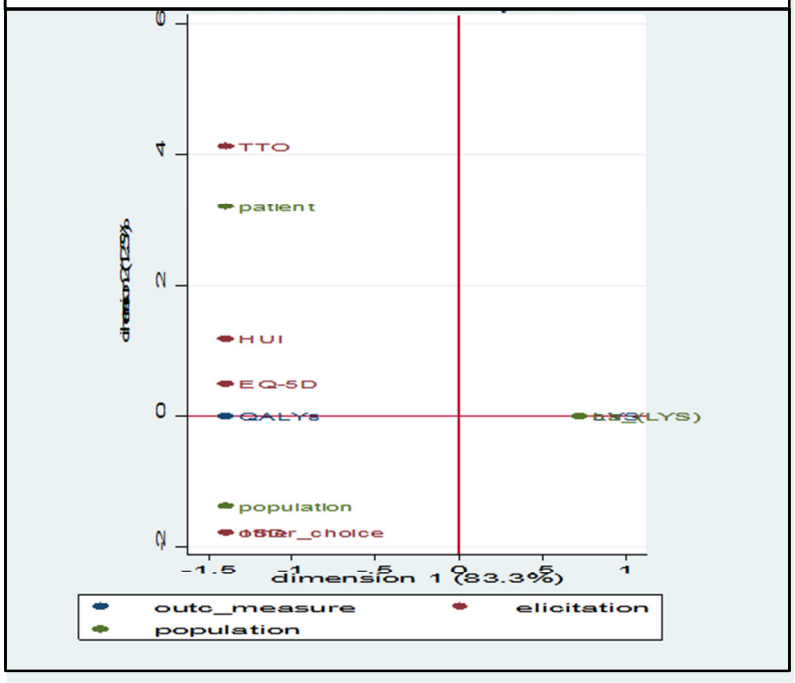


Results for running an MCA on '*outc_measure', 'elicitation'* and '*population*' are presented below. The total inertia is a measure of scatter, and the principal inertia tells how much of this scatter is captured by each dimension in the model (Bartholomew et al., 2008). This has been visualised in a screeplot above (Figure 5.8) where the principal inertia is plotted for each dimension in the model. In this case, the first two dimensions account for 95.82% of the total inertia. This indicates a very good fit of the model and the '*elbow*' in the screeplot, after

which there is only small decrease in inertia, indicates that plotting results on a two dimensional space may be appropriate (Bartholomew et al., 2008).

Next, the results of the MCA are visualised on a biplot presented in Figure 5.9., which clearly shows the correspondences between the categories of the three variables. One can see that the horizontal axis, which accounts for 83.3% of the total inertia, captures the two alternative measures of effectiveness 'QALYS' and 'life Years Saved. The vertical axis, on the other hand, captures different methods of preference elicitation and it also shows which of these methods were rather applied to patient populations, and which methods were rather used to elicit information from a general population sample. For instance, time trade off questions were usually asked to patients, whilst the health utility index or the EQ-5D instrument may have been applied to patients or members of the general population. Other choice based methods and the 15-D instrument clearly correspond with the category *'population'*, meaning that these methods were applied to assess utility weights from a general population sample.

Figure 5.9: Biplot for MCA on outc_measure, elicitation and population



The third step in this analysis is to estimate a correspondence score analogously to what has been done using PCF for components of circulation related CVD risk

above. This correspondence score may then be used as an alternative covariate capturing the concept of outcome measure within the MLM framework.

Table 5.17: MCA on 'outc_measure', 'elicitation' and 'population'

Number of obs: 1998 / Total inertia: 1.2008 / Number of axes: 2

	Principal Inertia	Percent	Cumulative
Dimension 1	1	83.28%	83.28%
Dimension 2	0.1506	12.54%	95.82%
Dimension 3	$1.11e^{-31}$	0%	95.82%
Dimension 4	$6.93e^{-33}$	0%	95.82%

	Overall			Dimension 1			Dimension 2		
Categories	**Mass**	**Quality**	**%inertia**	**Coord**	**sqcorr**	**contrib**	**Coord**	**sqcorr**	**contrib**
Outc_measure									
LYS	0.220	1.000	0.094	0.717	1.000	0.113	0.000	0.000	0.000
QALYs	0.113	1.000	0.183	-1.394	1.000	0.220	0.000	0.000	0.000
Elicitation									
n.a. (LYS)	0.220	1.000	0.094	0.717	1.000	0.113	0.000	0.000	0.000
TTO	0.019	0.841	0.083	-1.394	0.363	0.036	4.116	0.477	0.317
EQ 5D	0.039	0.994	0.065	-1.394	0.976	0.076	0.488	0.018	0.009
HUI	0.001	0.969	0.002	-1.394	0.876	0.002	1.168	0.093	0.001
15D	0.001	0.938	0.002	-1.394	0.753	0.002	-1.780	0.185	0.003
Other choice	0.054	0.938	0.115	-1.394	0.753	0.104	-1.780	0.185	0.170
Population									
n.a. (LYS)	0.220	1.000	0.094	0.717	1.000	0.113	0.000	0.000	0.000
patient	0.034	0.872	0.114	-1.394	0.487	0.066	3.195	0.385	0.349
population	0.079	0.959	0.153	-1.394	0.853	0.154	-1.382	0.124	0.151

MCA was also performed for a subset of variables relating to time horizon and time preference. The variables '*horizon', 'extrapol*' and '*hor_eq_dur*' led to a good overall fit of the model with 81.42% of the total inertia falling on the first two dimensions. Results of the respective MCA are presented in Appendix 5.3. Again, a correspondence score was calculated for use as a covariate in the MLM. Finally, the same procedure was repeated with different subsets of methodological characteristics on data-level as well as general study characteristics and methodological characteristics on study-level to assess whether key characteristics of international methods guidelines show correspondences on the biplot. However, further strong correspondences were not revealed.

5.2.4. Results

In Chapter 3 MLMs were developed for the analysis of secondary cost-effectiveness data from different geographic domains. This chapter also reported on a pilot study to test these models and Chapter 4 focussed on a systematic literature review and data abstraction exercise on the cost-effectiveness of statins for the primary and secondary prevention of CVD which provided the data for this empirical exercise. This chapter started off with determining the appropriate MLM structure to analyse the data obtained from the systematic literature review. Then, an analysis strategy was outlined to assess covariates within the MLM framework. The space of available covariates was divided in manageable groups of variables; descriptive statistics and missing values analysis were reported, as well as correlations between subsets of explanatory variables. The results from analysing covariates in the MLMframework assuming fixed slopes are reported next. First, bivariate statistics from entering each covariate individually in the multilevel framework are reported and discussed. Secondly, a model which aims to unravel the maximum amount of country-level variability is constructed by including the set of covariates which best controls for variability on data and study-level. This model is then carried forward to the analysis of country-level variability in Section 5.3 of this empirical chapter.

5.2.4.1. Bivariate statistics

Tables 5.18 to 5.23 below contain results of individually testing covariates in the univariate and bivariate versions of the three-level hierarchical model which groups cost-effectiveness data in the studies it was drawn from and the countries it refers to whilst accommodating a separate cluster on country-level for multinational study data. Data-level covariates are assessed first, starting off with patient and disease characteristics as the arguably most critical source of variability in cost-effectiveness data (Sculpher et al., 2004), followed by intervention and comparator characteristics and methods on data-level. Subsequently, general study characteristics, methods on study-level and study quality indicators are tested individually in the models of interest. A binary

missingness indicator is used in conjunction with variables where data was initially missing (Acock, 2005; Morris et al., 2005). If not significant, this indicates that missing data imputation increased the sample for analysis without biasing the results. If significant, then the missingness indicator captures the departure of the imputed cases from the non-imputed cases and thereby leads to the same regression coefficients one would have obtained with stepwise deletion of the missing observations (Morris et al., 2005). Models are implemented in MLwiN using MCMC estimation procedures (Rasbash et al., 2009a; Browne, 2012). Hence, the DIC diagnostic is used to compare the overall fit between models. The DIC diagnostic already accounts for differences in the number of parameters to estimate, so that its value is directly comparable between a saturated model and its respective comparator (Browne, 2012). For bivariate statistics, this comparator is obviously the variance components model without covariates. Hence, if the DIC decreases when including covariates into the models discussed in the first section of this empirical exercise, then this indicates an improved fit of the model. However, as Browne (2012) clarifies, the '*stochastic nature of the MCMC algorithm leads to some random variability in the DIC diagnostic depending on starting values and random number seeds*'. If differences in the DIC diagnostic are small, models were hence re-run with different seeds and/or starting values.

Starting with patient and disease characteristics (Table 5.18 below), almost all covariates show the expected sign and are significant on the 5% or even 1% level. This holds for all continuous variables (HDL, LDL, SBP, Hypertension, Smokers, BP_PCF and Diabetes) with the exception of TCL, for which the coefficient for ΔC in the bivariate model does show the anticipated sign but only borders at the 10% level of significance. Moving on to categorical variables, the coefficient for 46 to 55 year old patients is not significant in the univariate model. This only means that INMBs in this age cohort do not significantly differ from the omitted category, which represents patients below the age of 45. For older age groups, and especially in the bivariate model, where '*age_cat'* led to the largest drop in the DIC diagnostic compared to all other patient and disease characteristics, the coefficients are highly significant. The coefficients observed also accord expectations in a sense that better cost-effectiveness estimates associated with commencing treatment at younger age '*reflect the greater potential to prevent*

events, and thus the higher utility and cost benefits accrued from remaining event free health state' (NICE, 2006). Furthermore, the coefficient for patients above the age of 75, which indicates improved cost-effectiveness compared to patients aged 65 to 75, may be a result of more secondary prevention patients in this age cohort, for whom statin treatment is assumed to be more cost-effective. Also in accord with expectations are the coefficients for gender, as statins show lower incremental cost, higher incremental effects and hence improved cost-effectiveness measured in INMBs in males as compared to females.

Table 5.18: Bivariate statistics of patient and disease characteristics

		Univariate model		Bivariate model		
Explanatory variable (SE)	**Raw Mean (SD) / Proportion (%)**	**INMB** (2010 £ Sterling)	**DIC** (Benchmark : 46749 (%-change))	**ΔC/100** (2010 £ Sterling)	**ΔE**	**DIC** (Benchmark : 28735 (%-change))
TCL	6.676 (1.204)	4012 (640)***	46709 (-0.09%)	-6.62 (4.05)	0.103 (0.022)***	28699 (-0.13%)
HDL	1.168 (0.102)	-47062 (10302)***	46724 (-0.05%)	210.76 (59.10)***	-0.868 (0.353)**	28709 (-0.09%)
LDL	4.509 (1.036)	5200 (667)***	46687 (-0.13%)	-8.563 (4.224)**	0.136 (0.022)***	28677 (-0.20%)
SBP	137.48 (13.348)	640 (41.35)***	46509 (-0.51%)	-4.024 (0.252)***	0.008 (0.001)***	28352 (-1.33%)
Hypertension	31.70% (38.13%)	23718 (1661)***	46537 (-0.45%)	-92.29 (10.51)***	0.48 (0.06)***	28501 (-0.81%)
Smokers	29.10% (33.48%)	18681 (1859)***	46641 (-0.23%)	-25.60 (11.67)**	0.536 (0.061)***	28627 (-0.38%)
BP_PCF	0 (1.00)	6016 (424)***	46545 (-0.44%)	-25.49 (2.68)***	0.115 (0.014)***	28501 (-0.81%)
Diabetes	17.81% (34.91%)	13494 (1834)***	46692 (-0.12%)	-34.36 (11.34)***	0.341 (0.061)***	28679 (-0.19%)
Age_cat						
<45	322 (15.38%)	Omitted	46621 (-0.27%)	Omitted	Omitted	27705 (-3.58%)
46-55	439 (20.96%)	161 (1333)		-93.70 (7.363)***	-0.304 (0.039)***	
56-65	862 (41.17%)	-4819 (1271)***		-136.24 (7.06)***	-0.609 (0.037)***	
66-75	299 (14.28%)	-13295 (1364)***		-173.89 (7.51)***	-1.023 (0.040)***	
>75	98 (4.68%)	-7301 (2309)***		-120.61 (12.67)***	-0.641 (0.067)***	
Unclear	74 (3.53%)	10812 (8776)		-110.88 (43.67)**	-0.051 (0.271)	
Gender						
Female	576 (27.51%)	Omitted	46635 (-0.24%)	Omitted	Omitted	28619 (-0.40%)
Male	799 (38.16%)	10184 (955)***		-33.51 (5.88)***	0.226 (0.031)***	
Mixed sample	719 (34.34%)	5097 (3908)		-77.85 (23.58)***	-0.071 (0.148)	
Risk_cat						
<10%	193 (9.22%)	Omitted	46590 (-0.34%)	Omitted	Omitted	28103 (-2.20%)
10%-20%	367 (17.53%)	18319 (1637)***		-167.41 (9.06)***	0.053 (0.055)	
20%-30%	278 (13.28%)	17431 (1767)***		-216.62 (9.85)***	-0.139 (0.060)**	
30%-40%	140 (6.69%)	14547 (2055)***		-236.75 (11.44)***	-0.306 (0.069)***	
40%-50%	106 (5.06%)	23145 (2297)***		-241.10 (12.80)***	-0.032 (0.078)	
Secondary prevention	958 45.75%)	17452 (2057)***		-180.06 (11.66)***	0.000 (0.070)	
Unclear	52 (2.48%)	13698 (4975)***		-176.34 (26.28)***	-0.115 (0.178)	
CVD_history						
No	1064 (50.81%)	Omitted	46750 (0.00%)	Omitted	Omitted	28736 (-0.00%)
Yes	958 (45.75%)	1506 (1516)		3.33 (9.20)	0.081 (0.051)	
Mixed sample	72 (3.44%)	10325 (7418)		22.83 (42.62)	0.409 (0.270)	
Missingness indicators						
Miss_tcl	901 (43.03%)	-1134 (4271)	--	-24.16 (24.77)	0.471 (0.120)***	--
Miss_hdl	947 (45.22%)	-2719 (4296)	--	-21.89 (23.57)	-0.101 (0.168)	--
Miss_LDL	1168 (55.78%)	-8064 (3776)**	--	34.53 (23.41)	-0.109 (0.146)	--
Miss_sbp	954 (45.56%)	8188 (2583)***	--	-24.66 (22.65)	-0.091 (0.166)	--
Miss_hypert	1268 (60.55%)	-2553 (3978)	--	-29.90 (20.33)	-0.139 (0.145)	--
Miss_smokers	953 (45.51%)	-2269 (4154)	--	-31.10 (22.06)	-0.129 (0.163)	--
Miss_bp_pcf	1269 (60.60%)	-3795 (3942)	--	-22.53 (20.73)	-0.150 (0.147)	--
Miss_diabetes	931 (44.46%)	312 (4285)	--	-31.91 (22.52)	-0.053 (0.164)	--
Miss_age_cat	74 (3.53%)	-2603 (2381)	--	-3.76 (11.56)	-0.096 (0.061)	--
Miss_gend_cat	163 (7.78%)	-533 (1495)	--	-4.249 (9.32)	-0.033 (0.050)	--
Miss_risk_cat	52 (2.48%)	-558 (2383)	--	-3.41 (13.04)	0.419 (0.137)***	--

* significant at the 10%-level
* significant at the 5%-level
* significant at the 1%-level

More problematic, however, is the categorical variable '*risk_cat'*. Though highly significant both for INMBs in the univariate model and ΔC in the bivariate model, '*risk_cat'* does not show the pattern one would anticipate where cost-effectiveness improves continuously with increasing ten year CVD risk. Though it may be possible that cost-effectiveness is higher for the highest primary prevention group compared to secondary prevention (e.g. Huse et al., 1998), we should at least expect a continuous improvement of INMBs with increased CVD risk in primary prevention patients. This, however, is not the case, which casts into doubt the validity of the findings. What adds to the problem is that coefficients for ΔC in the bivariate model are mostly not significant. Obviously, one would expect a strong relationship between '*risk_cat'* and measures of cost-effectiveness for both ΔC, as statin treatment may prevent future healthcare cost, and ΔE, as higher risk groups may have more to benefit from statin treatment. Potential reasons for the problems associated with this explanatory variable may relate to the way it was constructed from a number of different variables abstracted from papers, which were used to estimate CVD risk using the Framingham risk equation (Anderson et al., 1991). If data to populate this risk equation was not available, a number of alternative means were exploited to obtain estimates of 10 year CVD risk (details are available from Chapter 4). Resulting risk estimates were then grouped for primary prevention in categories of 10% increments, and to capture the whole range of data points in the dataset, a category for secondary prevention was also introduced as the Framingham risk equation is not valid in patients who already experienced a CVD event (Anderson et al., 1991). The complexity of this process and the degree to which raw data was manipulated to obtain this categorical variable may have introduced additional noise which explains the problems with the coefficients as detailed above. In addition, it may be problematic to enter this variable in the final MLM in conjunction with individual risk factors such as TCL, HDL, or SBP, as '*risk_cat'* is a summary construct of these individual risk factors and therefore shows a considerable degree of collinearity (as observed in the previous Section 5.2.3.3). Therefore, multicollinearity problems may be the consequence, which may indicate dropping this variable from further analysis even though it leads to a considerable change in the DIC diagnostic in the bivariate model.

Finally, as can be seen from Table 5.18 above, the history of a CVD event is not a significant explanatory variable in the model. As mentioned above, the reason may be that very high risk primary prevention groups potentially benefit as much, if not more, than some secondary prevention patients (e.g. Huse et al., 1998), which then leads to a failure in observing significant results for this variable. However, simultaneously controlling for individual risk factors may address this problem and thereby reveal the actual relationship between CVD history and the cost-effectiveness of statins in the primary and secondary prevention of CVD.

Moving on to intervention and comparator characteristics (Table 5.19 below), highly significant coefficients are observed which also accord expectations for most continuous variables. Precisely, as annual intervention cost increase, so should ΔC, which consequently leads to decreasing INMBs. Annual drug cost may have an impact on treatment compliance and hence incremental effects as well, but this was not observed in the bivariate model. Likewise the continuous variable incremental drug cost, which is nothing but the annual drug cost of the intervention minus the annual drug cost of the comparator, shows a highly significant positive relationship with ΔC, and an inverse relationship with INMB which is also highly significant. However, the annual drug cost of the comparator is not in accord with prior expectations as ΔC of the intervention should decrease with increasing cost of the comparator, leading to better cost-effectiveness results. This was not confirmed by the data as an inverse relationship was observed in the univariate model. Finally, unit cost of the intervention and the comparator show the same sign as their annual drug cost counterparts, which again confirms expectations for the intervention but disagrees with expectations for the comparator. A reason may be some degree of correlation between intervention cost and comparator cost, meaning that changes in the cost of the comparator may be offset by changes in the intervention, so that its impact on incremental drug cost and, ultimately, INMBs remains ambiguous. This hypothesis may be tested when simultaneously including drug cost of the intervention and the comparator in one model in part two of this section.

Table 5.19: Bivariate statistics of intervention and comparator characteristics

		Univariate model		Bivariate model		
Explanatory Variable	**Raw Mean (SD) / Proportion (%)**	**INMB** (2010 £ Sterling)	**DIC** (Benchmark: 46749 (%-change))	**ΔC/100** (2010 £ Sterling)	**ΔE**	**DIC** (Benchmark: 28735 (%-change))
Cost_int	528.84 (326.32)	-9.289 (2.173)***	46733 (-0.03%)	0.112 (0.013)***	0.000 (0.000)	28658 (-0.27%)
Unitcost_int	0.046 (0.031)	-71638 (29179)**	46741 (-0.02%)	1172 (173)***	1.484 (0.957)	28696 (-0.14%)
Cost_comp	26.09 (115.31)	-10.59 (4.93)**	46747 (-0.00%)	-0.074 (0.029)**	-0.001 (0.000)	28726 (-0.03%)
Unitcost_comp	0.005 (0.029)	-67854 (44928)	46748 (-0.00%)	-655 (252)***	-4.43 (1.48)***	28722 (-0.05%)
Incr_cost	502.75 (314.47)	-7.331 (2.201)***	46739 (-0.02%)	0.134 (0.013)***	0.000 (0.000)	28643 (-0.32%)
Intervention						
Simvastatin	1080 (51.58%)	Omitted		Omitted	Omitted	
Fluvastatin	41 (1.96%)	-3508 (5660)		-30.16 (32.91)	0.179 (0.189)	
Atorvastatin	184 (8.79%)	-1508 (3073)		4.53 (17.34)	0.014 (0,104)	
Pravastatin	256 (12.23%)	-7058 (4078)*	46753 (0.01%)	-10.11 (20.20)	-0.203 (0.119)*	28729 (-0.02%)
Lovastatin	125 (5.97%)	-12292 (7360)*		75.26 (30.38)**	-0.349 (0.170)**	
Rosuvastatin	60 (2.87%)	-1607 (3177)		26.68 (18.62)	0.073 (0.104)	
unclear	348 (16.62%)	-7010 (9370)		23.52 (32.95)	0.030 (0.171)	
Tdd_intervention						
Up to 10 mg	65 (3.10%)	Omitted		Omitted	Omitted	
Up to 20mg	259 (12.37%)	-4876 (4370)		44.07 (25.96)*	-0.072 (0.148)	
Up to 30mg	554 (26.46%)	14940 (6466)**	46746 (-0.01%)	36.73 (35.04)	0.652 (0.255)**	28727 (-0.03%)
Up to 40mg	654 (31.23%)	-3194 (812)***		8.67 (25.58)	-0.054 (0.152)	
>60mg	206 (9.84%)	-7511 (4573)		32.48 (26.71)	-0.107 (0.158)	
unclear	356 (17.00%)	-6251 (8589)		-2.49 (46.58)	-0.310 (0.130)**	
Comparator						
Simvastatin	153 (7.31%)	Omitted		Omitted	Omitted	
Fluvastatin	3 (0.14%)	4010 (13100)		-76.51 (77.64)	-0.227 (0.431)	
Atorvastatin	44 (2.10%)	764 (9685)		-85.10 (56.96)	-0.359 (0.355)	
Pravastatin	19 (0.91%)	1735 (5148)	46751 (0.00%)	-27.36 (31.18)	-0.052 (0.169)	28727 (-0.03%)
Lovastatin	24 (1.15%)	75.09 (6117)		-104.7 (37.41)***	-0.427 (0.211)**	
Rosuvastatin	15 (0.72%)	2278 (7682)		-63.46 (46.15)	-0.136 (0.270)	
Doing nothing	1834 (87.58%)	7178 (4249)*		-23.63 (25.25)	0.091 (0.152)	
Unclear	2 (0.10%)	-614 (15598)		-25.86 (92.74)	-0.128 (0.398)	
Act_comparator						
no (doing nothing)	1834 (87.58%)	Omitted	46746 (-0.01%)	Omitted	Omitted	28729 (-0.02%)
yes (statin)	260 (12.42%)	-6545 (2770)**		-34.00 (16.27)**	-0.322 (0.097)***	
Tdd_comparator						
0mg	1834 (87.58%)	Omitted		Omitted	Omitted	
Up to 10mg	44 (2.10%)	-6110 (9228)		-70.81 (49.93)	-0.517 (0.332)	
Up to 20mg	26 (1.24%)	-3239 (4652)	46747 (0.00%)	-74.05 (26.81)***	-0.347 (0.153)**	28726 (-0.03%)
Up to 30mg	24 (1.15%)	-8588 (12580)		-89.10 (63.46)	-0.753 (0.444)*	
Up to 40mg	155 (7.40%)	-8062 (3342)**		-4.58 (19.39)	-0.269 (0.109)**	
unclear	11 (0.53%)	-5023 (12940)		-75.78 (75.84)	-0.400 (0.506)	
Missingness indicators						
Miss_cost_int	137 (6.54%)	2808 (3919)	--	-16.54 (22.26)	0.042 (0.128)	--
Miss_ucost_int	356 (17.00%)	-7388 (9008)	--	-26.76 (36.79)	-0.372 (0.346)	--
Miss_cost_comp	2 (0.10%)	-3502 (15396)	--	-6.97 (90.36)	-0.064 (0.509)	--
Miss_ucost_comp	11 (0.53%)	16038 (18449)	--	141 (104)	1.001 (0.682)	--
Miss_incr_cost	139 (6.64%)	2155 (3820)	--	-31.19 (21.17)	-0.041 (0.142)	--
Miss_intervention	348 (16.62%)	-4378 (3434)	--	-53.31 (20.40)***	-0.305 (0.118)***	--
Miss_tdd_int	356 (17.00%)	-1823 (3677)	--	-73.74 (20.96)***	-0.310 (0.130)**	--
Miss_comparator	2 (0.10%)	-2956 (12087)	--	-7.36 (73.64)	-0.128 (0.398)	--
Miss_tdd_comp	11 (0.53%)	1009 (5818)	--	-8.42 (25.47)	-0.082 (0.137)	--

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level

Moving on to categorical variables reported in Table 5.19 above, significant coefficients become rare. The fact that both the variable '*intervention*' and '*comparator*' show only few coefficients which reach statistical significance may speak in favour of the view taken by NICE, essentially saying that '*a statin is a statin is a statin*', implying there is no effectiveness evidence which would justify

to favour one statin over another when commencing treatment (Nice, 2008b). Non-significant coefficients for the ordered categorical variable encoding the total daily dose of the intervention may be explained by the offsetting effect of the total daily dose of the comparator technology, meaning that higher doses of the intervention may have been compared to higher doses of the comparator, so that measures of cost-effectiveness may remain in the same range. This hypothesis is supported by a positive polichoric correlation of above 0.4. Only the binary variable '*active comparator*' which indicates whether a statin has been compared to another statin or '*doing nothing*', shows significant relationships, which accords expectations. Precisely, comparing the intervention to another statin leads to lower ΔE as the comparator may lead to additional (quality adjusted) life years. This decrement in ΔE is likely to offset any decrement in ΔC, which may stem from the additional annual drug cost of the comparator. As a result, cost-effectiveness may be lower when comparing a statin to another statin, as opposed to comparing a statin to '*doing nothing'*.

The third group of variables to consider are methodological characteristics on data-level, i.e. study methods which may vary within individual studies, which are reported in Table 5.20 below. Starting with the annual discount rate for ΔC and ΔE, we observe highly significant negative relationships in the univariate model, as well as a positive relationship for the discount rate on costs and a negative relationship for the discount rate on effects in the bivariate model which both are highly significant too. For the discount rate on effects, the observed relationship is in accord with what one would expect as increasing the discount rate decreases the present value of future (quality adjusted) life years saved, and hence lowers estimates of INMBs. However, for incremental costs things are more difficult. Increasing the discount rate decreases the net present value of future drug cost for both the intervention and the comparator (if the comparator is not '*doing nothing'*). In addition, the net present value of future treatment cost without intervention decreases with increasing the discount rate on costs, which in sum may lead to increasing incremental cost of the intervention and this, in turn, may lead to a negative relationship between DRC and INMB.

Table 5.20: Bivariate statistics of methodological characteristics on data-level

		Univariate model		Bivariate model		
Explanatory variables	**Raw Mean (SD) / Proportion (%)**	**INMB** (2010 £ Sterling)	**DIC** (Benchmark: 46749 (%-change))	**ΔC/100** (2010 £ Sterling)	**ΔE**	**DIC** (Benchmark: 28735 (%-change))
DRC	0.039 (0.017)	-135022 (32104)***	46734 (-0.03%)	752.91 (187.16)***	--	28722 (-0.05%)
DRB	0.030 (0.018)	-215253 (33485)***	46708 (-0.09%)	--	-5915 (1036)***	28707 (-0.10%)
MCA_outc	0 (1)	624.83 (1213)	46751 (0.00%)	--	0.082 (0.042)*	28732 (-0.01%)
MCA_horizon	0 (1)	-1433 (1089)	46749 (0.00%)	-5.216 (6.325)	-0.053 (0.040)	28736 (0.00%)
Outc_measure						
LYS	1319 (62.99%)	Omitted	46750 (0.00%)	--	Omitted	28735 (0.00%)
QALYs	775 (37.01%)	-1879 (1667)			-0.080 (0.052)	
Elicitation						
n.a. (LYS)	1319 (62.99%)	Omitted	46751 (0.00%)	Omitted	Omitted	28738 (0.01%)
TTO	112 (5.35%)	-6617 (6320)		--	-0.193 (0.210)	
EQ-5D	313 (14.95%)	1745 (2795)		--	0.010 (0.088)	
HUI	6 (0.29%)	-3786 (12963)		--	-0.203 (0.431)	
15D	5 (0.24%)	-2203 (17657)		--	-0.073 (0.559)	
Other choice	322 (15.38%	-8966 (18374)		--	-0.047 (0.506)	
unclear	17 (0.81%)	-2193 (9456)		--	-0.045 (0.310)	
Elicitation_short						
n.a. (LYS)	1319 (62.99%)	Omitted	46751 (0.00%)	Omitted	Omitted	28738 (0.01%)
TTO	112 (5.35%)	-6388 (6436)		--	-0.188 (0.204)	
EQ-5D	313 (14.95%)	1734 (2788)		--	0.008 (0.089)	
Other choice.	333 (15.90%)	-3556 (9340)		--	-0.083 (0.316)	
unclear	17 (0.81%)	-2300 (9534)		--	-0.053 (0.317)	
Population						
n.a. (LYS)	1319 (62.99%)	Omitted	46750 (0.00%)	Omitted	Omitted	28737 (0.01%)
patient	215 (10.27%)	-4886 (4914)		--	-0.133 (0.155)	
population	474 (22.64%)	1930 (2860)		--	0.006 (0.091)	
unclear	86 (4.11%)	-2843 (8841)		--	-0.029 (0.288)	
Duration						
< 5 years	66 (3.15%)	Omitted	46733 (-0.03%)	Omitted	Omitted	28711 (-0.08%)
5 to <10 years	788 (37.63%)	1306 (5654)		25.09 (30.02)	0.097 (0.195)	
10 to <15 years	175 (8.36%)	2371 (5473)		11.97 (29.09)	0.097 (0.190)	
15 to <20 years	44 (2.10%)	3082 (6068)		23.36 (33.22)	0.145 (0.210)	
20 to <25 years	87 (4.15%)	5503 (5954)		-0.987 (32.50)	0.140 (0.209)	
> 25 years	814 (38.87%)	8753 (5535)		38.77 (29.11)	0.383 (0.194)**	
(lifet.)	120 (5.73%)	-10290 (12917)		215.97 (63.16)***	0.026 (0.503)	
Unclear						
Duration_short						
< 10 years	854 (4078%)	Omitted	46731 (-0.04%)	Omitted	Omitted	28715 (-0.07%)
10 to 20 years	219 (10.46%)	1532 (2301)		-3.31 (13.74)	0.040 (0.076)	
> 20 years	901 (43.03%	6974 (1916)***		10.47 (10.98)	0.251 (0.063)***	
unclear	120 (5.73%)	-13122 (11565)		221.93 (56.07)***	0.134 (0.449)	
Extrapol						
no	152 (7.26%)	Omitted	46750 (0.00%)	Omitted	Omitted	28736 (0.00%)
yes	1942 (92.74%)	3294 (5492)		30.29 (32.71)	0.163 (0.213)	
Horizon						
< 5 years	17 (0.81%)	Omitted	46741 (-0.02%)	Omitted	Omitted	28724 (-0.04%)
5 to <10 years	191 (9.12%)	888 (10654)		26.83 (59.53)	0.089 (0.404)	
10 to <15 years	330 (15.76%)	2710 (10544)		18.22 (59.51)	0.122 (0.400)	
15 to <20 years	132 (6.30%)	1804 (10620)		22.97 (59.92)	0.101 (0.403)	
20 to <25 years	91 (4.35%)	4993 (10732)		3.369 (60.88)	0.138 (0.408)	
> 25 years	1333 (63.6%)	8336 (10445)		43.83 (58.63)	0.373 (0.399)	
(lifet.)						
horizon_short						
< 10 years	208 (9.93%)	Omitted	46740 (-0.02%)	Omitted	Omitted	28727 (-0.03%)
10 to 20 years	462 (22.06%)	1424 (2570)		-6.02 (15.35)	0.020 (0.086)	
> 20 years	1424 (68%)	6600 (2326)***		8.62 (14.00)	0.222 (0.078)***	
hor_eq_dur						
no	1328 (63.42%)	Omitted	46742 (-0.01%)	Omitted	Omitted	28726 (-0.03%)
yes	766 (36.58%)	-5832 (2905)**		-1.39 (17.20)	-0.244 (0.104)**	
Persp_rep						
Health insurance	1369 (65.38%)	Omitted	46750 (0.00%)	Omitted	--	28736 (0.00%)
Societal	214 (10.22%)	-8945 (5019)*		40.54 (28.28)		
Not reported	511 (24.40%)	-5650 (5619)		32.28 (30.60)		
Persp_C_concl						
Health insurance	1939 (92.60%)	Omitted	46681 (-0.15%)	Omitted	--	28527 (-0.72%)
Societal	135 (6.45%)	13149 (4396)***		-147.19 (26.16)***		
Provider	20 (0.96%)	38984 (5074)***		-394.43 (29.26)***		

data_class						
base case	1125 (53.72%)	Omitted	46732 (-0.04%)	Omitted	Omitted	28685 (-0.17%)
efficacy	12 (0.57%)	-1636 (6017)		19.87 (36.26)	0.25 (0.199)	
baseline risk	49 (2.34%)	-6308 (3167)**		69.27 (19.18)***	0.025 (0.103)	
cost (not int.)	64 (3.06%)	-109.16 (2718)		4.18 (16.37)	0.017 (0.091)	
int. cost	86 (4.11%)	-632.80 (2198)		18.44 (13.43)	0.048 (0.071)	
QALYs	72 (3.44%)	-769.88 (3642)		1.40 (22.00)	-0.003 (0.124)	
Dur./hor.	288 (13.75%)	-6765 (1762)***		12.69 (10.68)	-0.176 (0.058)***	
Discount rate	257 (12.27%)	955.55 (1408)		37.69 (8.55)***	0.162 (0.046)***	
Other SA	141 (6.73%	865.04 (2214)		-2.46 (13.07)	0.028 (0.072)	
Basecase						
Yes	1125 (53.72%)	Omitted	46749 (0.00%)	Omitted	Omitted	28720 (-0.05%)
No	969 (46.28%)	-1337 (1080)		24.49 (6.64)***	0.045 (0.037)	
source_effects						
lit./meta	652 (31.14%)	Omitted	46750 (0.00%)	Omitted	Omitted	28736 (0.00%)
PLACI/II	38 (1.81%)	3497 (10297)		-7838 (5817)	-0.087 (0.421)	
CARE	88 (4.20%)	4750 (7842)		-44.62 (44.39)	0.067 (0.319)	
WOSCOPS	81 (3.87%)	-1070 (8090)		-31.44 (45.01)	-0.068 (0.320)	
4S	509 (24.31%)	24051 (6689)***		1.501 (38.35)	0.789 (0.261)***	
4S/ WOSCOPS	46 (2.2%)	-3978 (15536)		-9.63 (84.73)	-0.057 (0.605)	
EXCEL	120 (5.73%)	-10021 (11858)		152.38 (59.15)***	-0.034 (0.417)	
LIPID	23 (1.10%)	10950 (9692)		-55.55 (54.48)	0.230 (0.382)	
CARDS	28 (1.34%)	5147 (9387)		-111.79 (50.42)**	-0.229 (0.367)	
HPS	280 (13.37%)	12542 (10017)		-78.44 (52.70)	0.231 (0.379)	
TNT	42 (2.01%)	1104 (10450)		-84.50 (56.08)	-0.269 (0.385)	
LIPS	33 (1.58%)	2583 (10242)		-94-47 (56.40)*	-0.215 (0.427)	
IDEAL	24 (1.15%)	-1390 (12260)		-95.43 (67.59)	-0.449 (0.484)	
STELLAR	62 (2.96%)	6622 (11866)		-71.14 (66.68)	-0.090 (0.472)	
Brown et al	22 (1.05%)	7369 (16660)		-93.82 (84.76)	-0.302 (0.644)	
other	46 (2.20%)	8482 (6959)		-33.21 (38.16)	0.223 (0.272)	
4_S +						
No	1585 (75.69%)	Omitted	46748 (0.00%)	--	Omitted	28734 (0.00%)
yes	509 (24.31%)	21041 (5254)***			0.699 (0.172)***	
Barbieri_score_1						
Type C	186 (8.88%)	Omitted	46746 (-0.01%)	Omitted	Omitted	28731 (-0.01%)
Type CR	1033 (49.33%)	6964 (3738)*		-36.60 (22.53)	0.049 (0.131)	
Type CU	113 (5.40%)	1367 (6563)		-15.71 (39.21)	0.015 (0.227)	
Type CRE	193 (9.22%)	1452 (7308)		-82.61 (43.53)*	-0.235 (0.304)	
Type CRU	513 (24.50%)	515.29 (6631)		-59.34 (37.60)	-0.153 (0.262)	
Type CREU	56 (2.67%)	-1343 (7690)		-93.00 (46.20)**	-0.396 (0.314)	
Barbieri_score_2						
Type 1	186 (8.88%)	Omitted	46746 (-0.01%)	Omitted	Omitted	28730 (-0.02%)
Type 2	1146 (54.73%)	5818 (3321)*		-31.74 (19.50)	0.049 (0.114)	
Type 3	706 (33.72%)	23.25 (5144)		-64.11 (31.81)**	-0.185 (0.221)	
Type 4	56 (2.67%)	-3778 (5809)		-75.14 (35.96)**	-0.343 (0.237)	
Missingness indicators						
Miss_mca_outc	98 (4.68%)	-1269 (7184)	--	2.172 (38.57)	-0.001 (0.276)	--
Miss_elicitation	17 (0.81%)	-1755 (4226)	--	--	0.074 (0.132)	--
Miss_population	86 (4.11%)	-637.60 (2136)	--	--	-0.031 (0.068)	--
Miss_duration	120 (5.73%)	-1224 (1696)	--	22.27 (10.34)**	0.031 (0.055)	--

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level
\+ No variation on data-level. Will be treated as study-level covariate within further analysis.

Further continuous variables in this subgroup are MCA scores for variables concerned with outcome measurement ('*MCA_outcome'*) and time horizon ('*MCA_horizon'*). Neither leads to statistically significant results, as can be seen from Table 5.20 above. Likewise, the items used to generate these MCA scores fail to produce any significant results. However, very few methodological factors on data-level do have highly significant coefficients, one of them being the perspective on cost as judged by the author of this thesis. Interestingly, this is in sharp contrast to the results obtained when testing the perspective on cost as reported by the authors of studies included in this exercise, as results of this regression do not reach statistical significance. When looking at the perspective

as judged within this thesis, one observes that both a societal and a provider perspective are associated with lower incremental cost and improved cost-effectiveness as compared to the health insurance (NHS) perspective, which was left to the omitted category. For the provider perspective this may be due, for instance, to not including cost-items which may be considered from a health insurance but not from a provider perspective, and for the societal perspective this may be due to the inclusion of additional cost savings for the society due to CVD prevention with statins, such as avoiding future work loss.

As also reported in Table 5.20, whilst most types of sensitivity analyses did not significantly differ from base case results when testing the variable '*data_class'*, the binary variable basecase (yes/no) was highly significant for incremental cost in the bivariate model. Further, three covariates were tested in this group of variables which are concerned with data sources for populating the economic model. First, estimates of treatment effect were obtained from different randomised trials. However, the variable '*source effects'* shows that only very few coefficients are significant, indicating that trial results, even for different statins with different comparators, are in the same range. Nevertheless, results obtained using data from the 4S study, which is also the trial which most cost-effectiveness estimates in this empirical exercise are based upon, differ sharply from the rest of the dataset. A strong positive and highly significant relationship to ΔE in the bivariate model was observed and consequently, a positive and highly significant relationship with INMBs in the univariate model too. The fact that there was no relationship to ΔC also makes sense as different studies, though based on the same effectiveness data, may have used different costing methodologies. To control for this potential source of variability in effectiveness data, which also leads to a systematic difference in INMBs, a binary variable was created which is 1 if the estimate was obtained using 4S data, and zero otherwise. Note, however, that this binary variable does not show any variation within economic evaluation studies which is why it is treated as a study-level covariate in subsequent analyses. Obviously, it would be interesting to look into reasons why 4S is associated with better effectiveness and INMB values in the current analysis.

Finally, the geographic origin of input parameters was systematically assessed by creating a variable which captures the degree to which input values represent the target location. Precisely, the variable '*barbieri_score_1'* comprises six categories where the category 'C' refers to cost-effectiveness data with only unit-cost estimates being target location specific. Likewise, the category 'CU' refers to cost-effectiveness estimates based on target specific unit cost data and utility weights and so on, up to 'CUER', where all main input parameters were based on target location specific data. A shorter variant of this variable was also created ('*barbieri*_score_2'), which categorises data in four groups from '1' (only one input parameter target specific) to '4' (all input parameters target specific). Details on the way this variable was created can also be obtained from Chapter 4. As reported in Table 5.20 above, however, neither variant of this variable showed any conclusive results if tested individually in the multilevel framework. Maybe this changes when testing the variable in conjunction with other covariates.

Moving on to the study-level, bivariate statistics were first produced for general study characteristics, as reported in Table 5.21 below. First of all, timing was assessed as a continuous variable. A negative relationship was anticipated with ΔC, whilst the relationship with ΔE was assumed to be positive. The reason is that we may expect both statins becoming cheaper and more effective over time, consequently leading to improved cost-effectiveness. However, the data only confirms this expectation for ΔC in the bivariate model, whilst coefficients for ΔE and INMB are not significant. Moving on to the funding institution, we observe that studies mainly funded through industry show higher INMBs and this relationship is significant at the 5%-level. However, a similar relationship is not observed for ΔC or ΔE in the bivariate model.

Next, covariates encoding existing relationships between studies through common authorship are tested. These variables resulted from looking into the genealogy of economic evaluation studies in Chapter 4 and details on the way these covariates were created can be found there. Results reported in Table 5.21 show that only one group of papers with common authorship clearly depart from the rest of the dataset in terms of both ΔC and ΔE in the bivariate model as well as INMBs in the univariate model. This group of studies was co-authored by SA Grover and this finding confirms the interpretation of the forest plots presented

earlier in Section 5.1.4 of this chapter, which already indicated that papers co-authored by this researcher were clearly out of the range of other studies included in this empirical exercise. Apart from common authorship, another combining factor between those studies is the use of the same DAM, namely the '*CHD life expectancy model*'. An interesting topic for the discussion section is to look into potential reasons why this model may lead to much higher estimates of cost-effectiveness compared to other studies in this empirical exercise. To control for this potential source of variability, a binary variable was created which is '1' if a data point refers to a study included in this group and '0' otherwise.

Table 5.21: Bivariate statistics of general study characteristics

		Univariate model		Bivariate model		
Explanatory Variables	**Raw Mean (SD) / Proportion (%)**	**INMB** (2010 £ Sterling)	**DIC** (Benchmark: 46749 (%-change))	**ΔC/100** (2010 £ Sterling)	**ΔE**	**DIC** (Benchmark: 28735 (%-change))
Timing	2000 (5.40)	213.26 (378.04)	46749 (0.00%)	-8.09 (1.97)***	-0.018 (0.016)	28733 (-0.01%)
Language						
English	61 (91.04%)	Omitted	46749 (0.00%)	Omitted	Omitted	28735 (0.00%)
German	6 (8.96%)	-645.17 (7653)		-36.06 (50.19)	-0.124 (0.375)	
Funding_institution						
RC/gov/uni/other	11 (16.42%)	Omitted	46750 (0.00%)	Omitted	Omitted	28736 (0.00%)
Industry	39 (58.21%)	10803 (5430)**		-48.92 (32.18)	0.168 (0.212)	
unclear	17 (25.37%)	12184 (6343)*		-32.82 (40.93)	0.315 (0.269)	
Funding_manuf.						
No manufacturer	11 (16.42%)	Omitted	46750 (0.00%)	Omitted	Omitted	28737 (0.01%)
BMS	7 (10.45%)	5683 (7782)		-57.98 (45.18)	0.037 (0.318)	
MERCK	12 (17.91%)	16389 (6407)**		11.01 (40.15)	0.597 (0.291)**	
Pfizer	13 (19.40%)	14798 (6566)**		-71.02 (36.63)*	0.190 (0.284)	
Other	9 (13.43%)	5903 (7364)		-80.57 (41.90)*	-0.134 (0.294)	
unclear	15 (22.39%)	8807 (6546)		-35.84 (41.29)	0.214 (0.288)	
Author_group_long						
No relationships	18 (26.87%)	Omitted	46748 (0.00%)	Omitted	Omitted	28736 (0.00%)
Group 01	4 (5.97%)	-6581 (7736)		65.04 (45.72)	-0.004 (0.315)	
Group 02	5 (7.46%)	-1107 (8100)		-36.38 (49.14)	-0.099 (0.310)	
Group 03	8 (11.94%)	27522 (6222)***		141.69 (42.38)***	1.438 (0.297)***	
Group 04	4 (5.97%)	-6840 (8704)		94.59 (52.11)*	0.088 (0.309)	
Group 05	7 (10.45%)	-2107 (6949)		-11.39 (44.09)	-0.110 (0.287)	
Group 06	5 (7.46%)	1411 (7881)		-11.25 (46.36)	-0.006 (0.272)	
Group 07	3 (4.48%)	2051 (9142)		-16.51 (55.70)	-0.021 (0.351)	
Group 08	3 (4.48%)	3119 (9606)		-43.97 (55.79)	-0.033 (0.352)	
Group 09	3 (4.48%)	-2189 (8791)		-15.44 (48.64)	-0.115 (0.303)	
Group 10	3 (4.48%)	-1415 (9306)		-57.38 (55.99)	-0.244 (0.344)	
Group 11	2 (2.99%)	11601 (10848)		-33.69 (57.38)	0.289 (0.391)	
Group 12	2 (2.99%)	-6715 (11940)		0.98 (70.59)	-0.178 (0.437)	
Author_group_short						
No relationships	18 (26.87%)	Omitted	46748 (0.00%)	Omitted	Omitted	28735 (0.00%)
Group 01	29 (43.28%)	-2249 (4416)		6.86 (28.43)	-0.046 (0.167)	
Group 02	8 (11.94%)	27044 (6060)***		160.60 (41.37)***	1.522 (0.256)***	
Group 03	3 (4.48%)	1936 (8603)		-25.89 (58.21)	-0.060 (0.334)	
Group 04	2 (2.99%)	-929.24 (9788)		-10.87 (59.90)	-0.191 (0.370)	
Group 05	3 (4.48%)	-1292 (9286)		-61.32 (55.45)	-0.281 (0.327)	
Group 06	2 (2.99%)	10576 (9728)		-47.50 (58.82)	0.171 (0.361)	
Group 07	2 (2.99%)	-6883 (11661)		10.23 (66.79)	-0.130 (0.394)	
Author_Grover						
No	59 (88.06%)	Omitted	46747 (0.00%)	Omitted	Omitted	28735 (0.00%)
yes	8 (11.94%)	28410 (5196)***		147.15 (34.16)***	1.510 (0.223)***	
Missingness indicators						
Miss_fund_inst	17 (25.37%)	1064 (1185)	--	-9.215 (7.289)	0.006 (0.039)	--
Miss_fund_man	15 (22.39%)	1681 (4748)	--	-1.338 (10.36)	0.055 (0.056)	--

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level

Moving on to methodological characteristics on study-level for which bivariate statistics are reported in Table 5.22, we observe that studies capturing treatment effectiveness by modelling the impact of a change in cholesterol levels on (quality adjusted) life years show higher ΔC and ΔE than studies directly considering the change in CVD risk without making explicit the link from intermediate to final outcomes. This relationship is highly significant for both components of the INMB statistic in the bivariate model and, in addition to that, the impact on ΔE seems to outweigh the impact on ΔC, so that INMBs are also higher in studies which explicitly model from intermediate to final outcomes; though this relationship only borders at the 5%-level of significance.

Table 5.22: Bivariate statistics of methods on study-level

		Univariate model		Bivariate model		
	Raw Mean (SD) / Proportion (%)	**INMB** (2010 £ Sterling)	**DIC (**Benchmark: 46749 (%-change))	**ΔC/100** (2010 £ Sterling)	**ΔE**	**DIC (**Benchmark: 28735 (%-change))
General design						
Primary modelling	6 (8.96%)	Omitted	46749 (0.00%)	Omitted	Omitted	28736 (0.00%)
Secondary modelling	61 (91.04%)	2527 (7080)		51.30 (40.02)	0.244 (0.265)	
Secondary design						
N.a. (primary m.)	6 (8.96%)	Omitted	46750 (0.00%)	Omitted	Omitted	28737 (0.01%)
Markov model	41 (61.19%)	2672 (7033)		33.86 (46.77)	0.226 (0.287)	
Decision tree	7 (10.45%)	-286.35 (9344)		126.10 (48.48)***	0.470 (0.357)	
Other	13 (19.40%)	3399 (8230)		54.54 (46.77)	0.253 (0.333)	
Effect_calc						
CHD risk reduction	41 (61.19%)	Omitted	46749 (0.00%)	Omitted	Omitted	28735 (0.00%)
Cholest. reduction	26 (38.81%)	8188 (4368)*		84.48 (25.69)***	0.505 (0.194)***	
multinational						
no	60 (89.55%)	Omitted	46749 (0.00%)	Omitted	Omitted	28735 (0.00%)
yes	7 (10.45%)	4819 (6896)		-31.86 (54.79)	0.036 (0.403)	
Infl_adj						
n.a.	12 (17.91%)	Omitted	46750 (0.00%)	Omitted	Omitted	28737 (0.01%)
no	2 (2.99%)	-6110 (12836)		-20.49 (76.33)	-0.226 (0.507)	
yes	18 (26.87%)	-8759 (6458)		9.22 (40.18)	-0.275 (0.290)	
unclear	35 (52.24%)	-1834 (5777)		43.41 (32.53)	0.211 (0.236)	
Adj_method						
n.a.	12 (17.91%)	Omitted	46750 (0.00%)	Omitted	Omitted	28737 (0.01%)
simple CPI	8 (11.94%)	-12040 (7767)		15.64 (46.62)	-0.413 (0.324)	
healthcare CPI	10 (14.93%)	-5143 (7569)		-8.61 (47.75)	-0.180 (0.328)	
not though indicated	2 (2.99%)	-4734 (12604)		-32.73 (72.63)	-0.318 (0.478)	
unclear	35 (52.24%)	-1534 (5801)		35.73 (33.57)	0.192 (0.233)	
Cur_conv						
No	52 (77.61%)	Omitted	46750 (0.00%)	Omitted	Omitted	28737 (0.01%)
yes	15 (22.39%)	4797 (3878)		0.297 (23.13)	0.150 (0.151)	
Conv_method						
n.a.	52 (77.61%)	Omitted	46751 (0.00%)	Omitted	Omitted	28735 (0.00%)
Exchange rates	11 (16.42%)	6878 (4386)		-0.232 (24.73)	0.227 (0.158)	
unclear	4 (5.97%)	-2424 (7582)		17.80 (45.42)	0.009 (0.286)	
Scope						
CAD	18 (26.87%)	Omitted	46749 (0.00%)	Omitted	Omitted	28727 (-0.03%)
CAD and CD	35 (52.24%)	2610 (1781)		40.45 (10.70)***	-0.030 (0.059)	
CAD, CD and PAD	11 (16.42%)	1057 (5628)		-70.04 (28.55)**	-0.144 (0.226)	
unclear	3 (4.48%)	-8031 (11584)		-6.243 (67.14)	-0.182 (0.460)	
Missingness indicators						
Miss_infl_adj	35 (52.24%)	658.6 (887.1)	--	-2.825 (5.383)	0.012 (0.029)	--
Miss_adj_method	35 (52.24%)	661.8 (887.8)	--	-2.849 (5.383)	0.013 (0.222)	--
Miss_conv_meth.	4 (5.97%)	-3968 (3244)	--	-28.40 (19.82)	-0.227 (0.108)**	--
Miss_scope	3 (4.48%)	7132 (5290)	--	40.62 (32.32)	0.371 (0.176)**	--

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level

Apart from that, the assessment of methods on study-level did not lead to noteworthy results, as only one further categorical variable (scope) showed significant coefficients for the cost component of the INMB statistic. One may expect that studies which do not only look into the impact of statins on coronary heart disease (CHD), but also consider conditions as stroke or peripheral arterial disease, tend to have lower ΔC than studies which are confined to CHD only. This may be due to future cost avoided for a broader range of conditions due to prevention with statins. However, analysis shows a highly significant positive coefficient for studies which include stroke next to CHD but a negative coefficient for studies including CHD, stroke and peripheral arterial disease so that results are not entirely conclusive.

Finally, the data abstraction exercise reported in Chapter 4 was accompanied by applying the QHES instrument to the studies included in this empirical exercise (Ofman et al., 2003). This resulted in a score (bound between zero and 100) for each study in the dataset which supposedly gives an indication of study quality. Though a number of quality checklists exist and the application of such tools is not without controversy, this may be regarded as an attempt to control for some part of the variability in measures of cost-effectiveness due to differences in the methodological rigour with which economic evaluation studies were conducted and reported. Further details on why the QHES instrument was chosen out of a number of potential checklists, how it was applied to the studies in this empirical exercise, which problems were encountered throughout the process, and how these problems were addressed within this thesis are reported in Chapter 4.

For bivariate statistics, the information obtained from QHES was implemented in different ways, resulting in two continuous and two categorical variants of this variable. The reason is that the authors who developed QHES defined dimensions of study quality of which some are comprised of several subcategories (Ofman et al., 2003). This induces the problem that studies may score in some but not all subcategories in one dimension. Depending on how rigorous one is in the application of this instrument, this may lead to different scores for the same study. Therefore, two continuous variables were derived (again, further details are available from Chapter 4), and to better reflect the

immense uncertainty attached to the data obtained through QHES, the continuous data was also transformed to categorical variables with 5 categories with increments of 20 between zero and 100. Descriptive statistics show that a strict application of the QHES criteria (i.e. not assigning a score for a particular dimension if a study does not completely fulfil the criteria) leads to a lower mean but higher standard deviation for continuous data, and also a higher spread across categories for categorical data.

Bivariate statistics reported in Table 5.23 show negative relationships between continuous QHES variables and both components of the INMB statistic, whilst this relationship is significant at the 1% level for ΔC and at the 5%-level for ΔE. This result indicates that a higher study quality (as measured by the QHES instrument) is associated with more conservative estimates of incremental effects, and also lower estimates of incremental costs compared to studies which achieved a lower QHES score. However, significance was not achieved for coefficients in the univariate model. Testing the categorical versions of the QHES instrument confirms the previous findings, though it is interesting that results do not longer reach statistical significance for ΔE in QHES_cat_a, which resulted from a strict application of the QHES criteria. This observation indicates that results may be interpreted with caution, as they obviously are very sensitive to small variations in the way QHES has been operationalized in this empirical exercise.

Table 5.23: Bivariate statistics of study quality indicators

		Univariate model		Bivariate model		
	Raw Mean (SD) / Proportion (%)	**INMB** (2010 £ Sterling)	**DIC** (Benchmark: 46749 (%-change))	**ΔC/100** (2010 £ Sterling)	**ΔE**	**DIC** (Benchmark: 28735 (%-change))
QHES cont_a	66.23 (17.97)	-167.77 (131.06)	46749 (0.00%)	-2.244 (0.77)***	-0.013 (0.006)**	28736 (0.00%)
QHES cont_b	75.34 (15.36)	-193.81 (154.60)	46749 (0.00%)	-2.545 (0.903)***	-0.015 (0.007)**	28736 (0.00%)
QHES cat_a						
Up to 40 pts	7 (10.45%)	Omitted	46749 (0.00%)	Omitted	Omitted	28734 (0.00%)
41 to 60 pts	31 (46.27%)	7039 (7488)		-98.03 (38.03)***	-0.013 (0.293)	
61 to 80 pts	22 (32.84%)	4112 (7823)		-140.89 (41.25)***	-0.229 (0.321)	
81 to 100 pts	7 (10.45%)	-321 (9271)		164.97 (49.22)***	-0.481 (0.391)	
QHS cat_b						
Up to 60 pts	20 (29.85%)	Omitted	46749 (0.00%)	Omitted	Omitted	28735 (0.00%)
61 to 80 pts	40 (59.70%	-9528 (4565)**		-71.10 (24.70)***	-0.521 (0.172)***	
81 to 100 pts	7 (10.45%)	-12150 (7109)*		-110.83 (39.46)***	-0.762 (0.276)***	

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level

5.2.4.2. Random intercepts models with multiple covariates on data and study-level

In this section, a full random intercepts model is developed with covariates on data and study-level. The primary objective is to control for variability factors on both levels which potentially feed through to the country-level so that further country-level variability may be disclosed. This country-level variability is the focus of assessment in Section 5.3 of this empirical exercise. In this section, it is assumed that slopes of explanatory variables are fixed (this assumption is being relaxed in Section 5.4 through the inclusion of random slopes). As detailed in the analysis strategy for this section, the model is being build up from the lower level to the higher level, and covariates are added to the model if their coefficients are significant and accord expectations and the overall fit of the model improves.

The order with which variables are tested has also been laid out in the analysis strategy. Patient and disease characteristics are tested first as arguably the most critical source of variability in measures of cost-effectiveness (Sculpher et al., 2004), followed by intervention and comparator characteristics, methods on data-level, general study characteristics, methods on study-level and study quality indicators. Results from gradually building up the full random intercepts model can be obtained from Appendix 5.4. The results from running the final model fully specified with covariates on data and study-level are obtainable from Table 5.24 below. Each covariate is tested individually, and its impact on overall fit and other coefficients with their respective p-values is observed. This bottom up approach ensures that the most appropriate set of covariates is chosen out of the pool of candidates available in this dataset. Missingness indicators are included in the model; however, they are only reported if they reach statistical significance. After choosing the most appropriate set of covariates, the impact on variability observed on each hierarchical level throughout the development of the models of interest is analysed and discussed.

Combinations of patient and disease characteristics were tested first in the univariate and bivariate versions of the three-level hierarchical model. Results are detailed in Appendix 5.4.1. Total cholesterol (which was not significant for ΔC

when performing bivariate statistics), SBP and the percentage of patients diagnosed with diabetes in the study sample all turned out to be highly significant for INMB, ΔC and ΔE, whilst HDL was significant at the 10% level for ΔC and ΔE in the bivariate model and at the 5% level for INMB in the univariate model. SBP turned out to be the variable which best controls for variability in cost-effectiveness data due to circulation related CVD risk. As a result, hypertension status, smoking status and the factor score obtained from running a PCF on SBP, hypertension and smoking were all dropped from the model due to strong collinearity with SBP. As before when running bivariate statistics, coefficients for the categorical variable encoding the age of the study sample turned out to be highly significant and show the previously observed relationship with measures of cost-effectiveness. Likewise, the gender variable accords expectations and shows highly significant coefficients. Interestingly however, CVD-history, which previously failed to show statistically significant coefficients, now turns out to be highly significant both for INMBs in the univariate model and ΔE in the bivariate model. As hypothesized before, this may be due to the fact that controlling for individual patient risk factors unravels the actual relationship between measures of cost-effectiveness for statin treatment and CVD history, which should be positive for patients which previously experienced a CVD event.

Adding intervention and comparator characteristics to the model further improves its fit (Appendix 5.4.2). The annual drug cost of the intervention is highly significant and accords expectations for INMB in the univariate model and ΔC in the bivariate model. However, the annual drug cost of the comparator was dropped from the univariate model as improvement in fit was better when including the binary variable '*active comparator*'. The reason may be that both variables are highly correlated as the annual drug cost of the comparator is always zero when the comparator is '*doing nothing*' (which is the case in 1834 (87.58%) of data points). Hence, this binary variable was highly significant for ΔE in the bivariate model and significant at the 10% level in the univariate framework. In both cases, the coefficient shows the anticipated negative sign as one would expect both ΔE and INMBs to decrease if the intervention is compared to another statin rather than '*doing nothing*'. Instead of annual drug cost, a combination of the variables '*unit cost*', '*total daily dose*' and the specific

type of intervention or comparator was also tested in the model but results indicate a much better fit of the model as described above.

The third group of variables to add to the model are methodological characteristics on data-level (Appendix 5.4.3). It turned out that only few variables show significant coefficients, but these variables improved the fit of the model quite dramatically. First of all, the discount rates on cost and effects were tested. Whilst both turned out to be significant at first, the discount rate on cost was dropped subsequently after including the categorical variable encoding the economic perspective taken on costs. A potential explanation for the correlation between both variables may be that different studies complied with different methods guidelines, which have their idiosyncratic views on both discount rate and perspective. Nevertheless, the discount rate on effects remained in the model as it was highly significant both for ΔE in the bivariate model and INMBs in the univariate framework.

An interesting relationship was observed between the two explanatory variables '*horizon*' and '*duration_eq_horizon'*, which encodes whether treatment duration lasted as long as the time horizon of the economic model (or shorter). A time horizon of 20 years or more leads to higher INMBs and also higher ΔE in the bivariate model. Although one may think that longer treatment duration may also lead to higher ΔC, this was not observed in the bivariate model. In addition, INMBs and incremental effects turned out much more beneficial if the treatment with statins did not last as long as the models time horizon, meaning that statin treatment appears to be more beneficial when stopped before the end of the model lifetime. Moving on to the variable '*base case'*, which encodes whether a data point refers to sensitivity analyses or not, it turns out that sensitivity analyses results are associated with both higher ΔC and higher ΔE in the bivariate model. This effect was not observed in the univariate model, most likely because the positive effect on ΔC is simply offset by the positive effect on ΔE.

Finally, the variable 'barbieri_score' was tested in the model, and this led to significant coefficients for ΔC in the bivariate model. Precisely, continuously

decreasing ΔC is observed with increasing context specificity of the underlying input data. A related question is why this variable did not show significant coefficients for ΔE or INMBs. A reason may be that transferring effectiveness data is much more common than transferring economic data, so that there is simply no systematic variation in this relationship on the effectiveness side of the INMB statistic. This may then feed through to INMBs, so that the univariate model also fails to observe a relationship between context specificity of input data and measures of cost-effectiveness. Before moving on to the study-level, note that the source of effectiveness data also belongs to the group of methodological characteristics on data-level. Whilst this categorical variable did only show very little improvement in model fit, it also showed, however, that measures of ΔE and INMBs are much more beneficial if the study was based on data from the 4S trial. Hence, a binary variable was generated and tested when running bivariate statistics (more details are obtainable from the previous Section 5.2.4.1). Unlike the categorical variable with categories for each source of effectiveness data, which also differs within economic evaluation studies, this binary variable does not show any variation on data-level, and is therefore considered as a methodological characteristic on study-level further below.

Moving on to this study-level (Appendix 5.4.4), it becomes more difficult finding significant coefficients. Timing, for instance, was expected to be an important explanatory both for ΔC and ΔE- and therefore also for INMBs. It was, however, not statistically significant. In theory, one may expect that statins become both less costly and more effective over time, however this was not observed in the multilevel models. With respect to incremental cost and INMBs, the explanation for this failure to observe significant coefficients may be relatively simple. The model already contains variables which encode the annual drug cost of the intervention and comparator technologies. If statin treatment becomes less costly over time, for instance because of generics which enter the market after product patents run out, this should be captured by those variables. Consequently, removing those variables from the model results in significant coefficients for timing, which accords prior expectations as these coefficients are negative for incremental cost and positive for INMBs. However, as the model fit is much better with the variables encoding annual drug cost, it was decided to keep those variables in the model and rather to drop timing instead.

Unfortunately, other variables which encode general study characteristics failed to show significant coefficients, for instance language (as papers included in the systematic review and data abstraction exercise may have been written in English or German). A further number of variables showed only significant coefficients in the univariate model, whilst coefficients for ΔC and ΔE were not significant. For instance, there was no significant difference between studies which were primarily funded by the government, research councils or Universities, and those primarily funded by the industry in the bivariate model, whilst INMBs were significantly higher (at the 5% level) for industry funding in the univariate model. However, this significant relationship disappeared when including a binary encoding whether a study uses the CHD life expectancy model by Grover et al (1998) and this observation is discussed in more detail with respect to the relevant literature (e.g. Miners et al., 2005) in Chapter 6.

Assessing general study characteristics was not completely unsuccessful in terms of controlling for variability in measures of cost-effectiveness, and the reason for this is directly related to the genealogy study reported earlier in Chapter 4. It was hypothesized that studies which are '*genotypically*' related, e.g. through common authorship, may also have '*phenotypic*' characteristics in common, which may also lead to correlations in measures of cost-effectiveness. As a result, variables were created which encode links between studies included in this exercise due to common authorship, and analysis showed that one particular group of papers, all published by a group of authors around S.A. Grover, are strong outliers both in terms of ΔC and ΔE in the bivariate model and INMBs in the univariate framework. This finding confirms what was already indicated by the forest plots presented in Section 5.1.4. As a result, including a binary variable which is '1' if the paper refers to this group of authors and '0' otherwise leads to much higher estimates of INMBs and ΔE, and also elevated estimates of ΔC. Apart from common authorship, another similarity between the affected papers is the use of the identical DAM to assess cost-effectiveness which was already referred to above as the '*CHD life expectancy model*'. What may have caused this strong effect on measures of cost-effectiveness within the affected studies is also subject of further discussion in Chapter 6.

Table 5.24: Random intercepts model fully specified on data and study-level

	Univariate model	Bivariate model	
Explanatory variables	**INMB** (2010 £ Sterling)	**ΔC/100** (2010 £ Sterling)	**ΔE**
Fixed part:			
N (countries)	18	18	18
N (studies)	67	67	67
N (data)	2094	2094	2094
Intercept (λ=£30.000)	-50105	716	-1.450
TCL (SE)	4254 (562)***	-32.66 (3.13)***	0.040 (0.018)**
HDL (SE)	-31211 (9424)***	189.25 (53.17)***	-0.523 (0.303)*
SBP (SE)	703 (35.82)***	-3.37 (0.19)***	0.012 (0.001)***
Diabetes (SE)	15017 (1684)***	-19.44 (8.23)**	0.415 (0.053)***
Age_cat (SE)			
<45	Omitted	Omitted	Omitted
46-55	311.4 (1111)	-95.31 (6.05)***	-0.308 (0.036)***
56-65	-5934 (1091)***	-133.5 (5.95)***	-0.639 (0.035)***
66-75	-15211 (1157)***	-168.4 (6.27)***	-1.066 (0.037)***
>75	-6663 (1956)***	-130.3 (10.58)***	-0.651 (0.063)***
Unclear	-4437 (7421)	-116.1 (43.39)***	-0.535 (0.230)**
Gender (SE)			
Female	Omitted	Omitted	Omitted
Male	10329 (841)***	-32.76 (4.52)***	0.231 (0.027)***
Mixed sample	11082 (3493)***	-29.01 (22.94)	0.253 (0.108)**
CVD_history (SE)			
No	Omitted	--	Omitted
Yes	7985 (1632)***		0.228 (0.049)***
Mixed sample	8520 (6103)		0.334 (0.189)*
Cost_intervention	-9.08 (1.76)***	0.129 (0.010)***	--
Cost_comparator	--	-0.140 (0.023)***	--
Active_comparator			
No (doing nothing)	Omitted	--	Omitted
Yes (statin)	-3845 (2211)*		-0.274 (0.072)***
DRB	-133550 (29321)***	--	-3.477 (0.889)***
Persp_cost_concl.			
Health insurance (NHS)	Omitted	Omitted	--
Provider	33002 (4416)***	-399.8 (22.88)***	
Societal	15212 (3498)***	-173.9 (19.82)***	
Horizon			
< 20 years	Omitted	--	Omitted
>20 years (lifetime)	3812 (1349)***		0.159 (0.042)***
Duration=horizon			
yes	Omitted	--	Omitted
No (treatment duration< horizon)	13091 (2437)***		0.377 (0.075)***
Base case			
Yes	--	Omitted	Omitted
No		9.31 (4.95)*	0.066 (0.030)**
Barbieri_score_2			
Type 1		Omitted	
Type 2	--	-36.56 (15.29)**	--
Type 3		-64.08 (29.76)**	
Type 4		-95.75 (32.03)***	
Author_Grover			
No	Omitted	Omitted	Omitted
Yes	29138 (5849)***	162.6 (44.79)***	1.412 (0.194)***
4S			
No	Omitted	--	Omitted
yes	14295 (5294)***		0.495 (0.159)***
Scope			
CHD	Omitted	Omitted	--
CHD and stroke	5535 (1566)***	-36.63 (8.08)***	
CHD, stroke and PAD	8428 (4822)*	-33.78 (31.93)	
unclear	5984 (9227)	-104.5 (64.53)	
Random part:			
σ^2_{u0j} **(Country)**	444825	3515	0.056
σ^2_{u0k} **(Study)**	153583808	7797	0.136
σ^2_{e0} **(Data)**	188413984	5523	0.196
VPC - Country	0.13%	20.88%	14.43%
VPC – Study	44.85%	46.31%	35.05%
VPC - data	55.02%	32.81%	50.52%
DIC	45840	26412	
(benchmark)	(45844)	(26423)	

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level

Moving on to methods on study-level, only two variables showed significant coefficients, namely '*scope*' and '*4S*' (Table 5.24). The variable '*scope*' is a categorical which distinguishes between studies which 1) only consider the impact of statin treatment on CHD risk, 2) those which also consider the effect of prevention with statins on stroke and, 3) those studies which consider the effect of statins on CHD risk, stroke and peripheral arterial disease (PAD). Obviously, one would expect a positive relationship with ΔE as a broader scope leads to consideration of a broader range of beneficial effects from statin prevention. Likewise, ΔC may be lower, as a broader range of future healthcare cost may be avoided through statin treatment. However, significant coefficients which accord expectations were only observed for ΔC and INMBs, whilst coefficients were not significant for ΔE.

Finally, the variable '4S', which is '1' if a study based its results on evidence from the 4S trial, and zero otherwise, was included as a methodological characteristic on study-level. This variable arose out of analysing the source of effectiveness data, and results indicated that only the 4S study may result in outliers in terms of incremental effects and INMBs. The results from running bivariate statistics were confirmed when including the binary in the random intercepts model, which led to an improvement of fit and strong changes both in study-level and country-level variability.

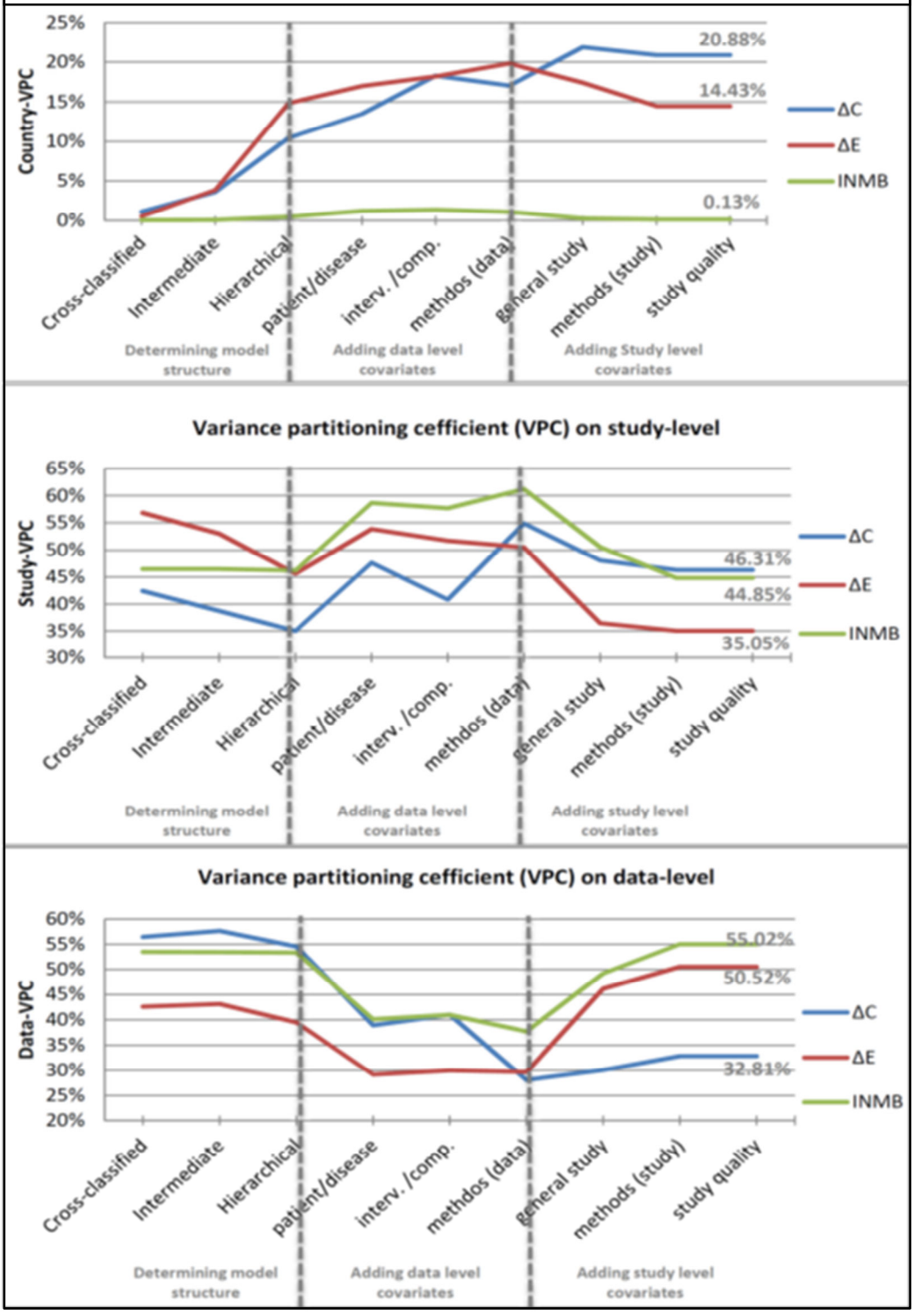
Whilst results from gradually adding covariates to the model are obtainable from Appendix 5.4, Table 5.24 above reports on the results of running the fully specified random intercepts model with covariates on data and study-level. One group of variables is missing entirely from this model, i.e. indicators of study quality. Unfortunately, results indicate that variables encoding information from applying the QHES instrument to the studies included in the dataset, are not related to measures of cost-effectiveness. Potential reasons are that either there is no relationship between study quality and measures of cost-effectiveness, or that the QHES instrument simply failed to capture study quality in an appropriate manner, so that it was not possible to disclose any relationship between methodological rigour of an economic evaluation study and cost-effectiveness

results. The author believes that this latter explanation may be more likely, and elaborates further on this issue in the discussion of this empirical exercise.

Before moving on to the next section, which is concerned with the assessment of country-level variability, it is indicated to make explicit once more how this whole exercise relates to the issue of variation in measures of cost-effectiveness between countries. For this matter, one may have a closer look at the random part of the models reported in Appendix 5.4. If lower-level variability factors feed through to higher levels, the actual amount of country-level variability may only be unravelled by including the appropriate set of covariates both on data and study-level. The multilevel framework offers an excellent platform to investigate this hypothesis as it allows higher-level variability to change in any direction with the inclusion of lower-level covariates (Hox, 2010). For instance, if Sculpher et al. (2004) are right with their statement that the impact of patient characteristics on measures of cost-effectiveness '*feeds through to differences in cost-effectiveness observed on higher levels*', than we ought to observe changes in error terms relating to each hierarchical level of the model by including such lower-level covariates. Obviously, the same holds for study-level covariates, as controlling for study characteristics may also impact on the variability in measures of cost-effectiveness observed between countries.

As the model was gradually built up from the lower to the higher level, it is possible to map the change in variability on each hierarchical level throughout the course of this empirical exercise. If the suspicion holds in the sense that variability on lower levels disguises variability on country-level, one should observe the country-level gradually becoming an ever more important source of variability in measures of cost-effectiveness. Figure 5.10 captures the change in the variance partitioning coefficient (VPC) on each hierarchical level throughout the course of this empirical exercise. During the first part of this study, which was concerned with determining the appropriate MLM structure, variability on data-level remained relatively constant, whilst the VPC increased on country-level and consequently decreased on study-level, especially for ΔC and ΔE in the bivariate model. The reason is that multinational study data disguised country-level variability in the cross-classified and intermediate models. Grouping data from multinational studies in a separate group on country-level disclosed further variability between countries, whilst study-level variability decreased respectively.

Figure 5.10. VPC on data, study and country-level throughout this empirical exercise



Secondly, testing data-level covariates in the model obviously had a very strong impact on level-one variability. It led to a sharp drop in VPC on data-level for both variability in INMBs in the univariate model as well as variability in ΔC and ΔE in the bivariate model. However, variability on study-level and country-level

increased simultaneously, and this not just in relative terms as measured with the VPC, but also in absolute terms, which is getting obvious when comparing the figures displayed in Tables 5.4.1 to 5.4.4 in Appendix 5.4. This is very compelling evidence in favour of the statement that the effect of lower-level variables feeds through to variability in measures of cost-effectiveness between studies and, ultimately, countries. It also shows one of the merits of the multilevel framework, as the impact of lower-level covariates on each subsequent hierarchical level is being made explicit. Unfortunately, however, variability on country-level remains very low for INMBs within the univariate framework, with just over 1% of the overall variability falling on that particular level. This is one focus of attention in the next Section 5.3 of this empirical exercise, which is concerned with country-level variability – or the lack thereof – in the MLM framework.

Finally, adding study-level covariates also helped controlling for some part of the overall variability, particular on study-level. This also led to a change in country-level variability, which increased further for ΔC but slightly decreased for ΔE and INMBs. It is important to mention, though, that the change in VPC observed on data-level due to the inclusion of study-level covariates is only a result of the overall variability decreasing. As mentioned before, changes on lower-levels may impact higher-levels, but not vice versa. Hence, as overall variability decreased, the error variances on data-level remained constant, which led their proportion of overall variability measured by the VPC to increase.

5.2.5. Summary and conclusions for part two of this empirical exercise

This section was predominantly concerned with testing covariates in the MLM framework on data and study-level and determining a model which best controls for variability on these levels, hence disclosing the maximum amount of country-level variability in measures of cost-effectiveness for statins. Covariates on data and study-level were systematically assessed which were drawn from a long list of variability factors as obtained from the literature (Sculpher et al., 2004;

Goeree et al., 2007) and abstracted from the studies included in the systematic literature review reported in Chapter 4.

Following descriptive statistics and the analysis and regression based imputation of missing values, a detailed assessment of correlations between potential explanatory variables, as well as the application of data reduction techniques, such as principal components factor analysis (Rencher, 2002; Acock, 2010) or multiple correspondence analysis (Le Roux & Rouanet, 2010), was carried out in this exercise. Thereafter, covariates were analysed in the multilevel framework assuming fixed slopes within a random intercepts specification (Steele, 2008; Rasbash et al., 2009, Hox, 2010) First, bivariate statistics from entering each covariate individually in the MLM were reported and discussed. Secondly, a model which aims to unravel the maximum amount of country-level variability was constructed by including the set of covariates which best controls for variability on data and study-level. This model, which was reported above in Section 5.2.4.2, is carried forward to the analysis of country-level variability in the subsequent Section 5.3 of this empirical chapter.

Some important conclusions can already be drawn from the analysis reported above. For instance, adding covariates to the model on data and study-level successfully disclosed further variability on country-level, which is a necessary condition for assessing country-level covariates in the subsequent section. Hence, the model was successful in showing that lower-level variability factors feed through to higher levels so that the actual amount of country-level variability may only be unravelled by including the appropriate set of covariates both on data and study-level. This demonstrates impressively the importance of both reflecting appropriately complex data structures and controlling for variability on lower levels, even if the main focus is on higher-level (i.e. country-level) variability. MLM is therefore regarded as an excellent analytic approach for this assessment as it allows for both complex data structures and covariate adjustment on each hierarchical level (Steele, 2008; Rasbash et al., 2009; Hox, 2010). The analysis could show that country-level variability was constantly increasing with the inclusion of lower-level covariates for the bivariate model, which may allow the assessment of covariates on country-level in the bivariate

framework. However, a different conclusion has been reached for the univariate framework, where country-level variability remained negligible throughout the course of this exercise. For this reason, the subsequent Section 5.3 focusses on both, covariate adjustment on country-level in the bivariate framework, and assessing reasons for a lack of country-level variability in the univariate MLM.

Next to determining a random intercepts model to be carried forward to Section 5.3 of this empirical exercise, some interesting findings were made along the way, which are also discussed in far more detail in Chapter 6. For instance, as Sculpher et al. (2004) hypothesized, this exercise could show the tremendous importance of variation in patient and disease characteristics, which feeds through to variation in measures of cost-effectiveness between studies and countries in this empirical exercise. Further, testing intervention and comparator characteristics in the model showed that NICE's view on statins is essentially confirmed, which says that '*for the purposes of initiating therapy, there were no data on clinical events to suggest the superiority of any one statin over all the others in reducing cardiovascular events*' (NICE, 2006). In other words, testing categorical variables which discriminate between different statins in the dataset, did not lead to significant results for incremental effects in the bivariate model.

Controlling for study-level variability factors proved far more difficult, especially with respect to methodological characteristics, even though the study-level appears to be an overriding source of variability in measures of cost-effectiveness. The non-significance of the timing variable may be explained with the inclusion of '*annual drug-cost'* as an explanatory, as this obviously reflects changes in price levels over time. However, only few variables on study-level were significant, amongst them a binary which captures whether effectiveness data was obtained from the 4S study, as well as a binary which captures whether a study uses the '*CHD life expectancy model'* by Grover et al. (1998). 4S data relates to higher effectiveness estimates which also feeds through to INMBs in the univariate model, whilst papers by Grover et al. using the CHD life expectancy model show higher levels in all outcome variables. These are both interesting findings, and the discussion section elaborates in more detail on potential explanations.

The results with respect to the QHES instrument are somewhat dissapointing. Considerable effort went into operationalizing this instrument and to apply it to all 67 studies included in this dataset. Though bivariate statistics showed statistically significant negative relationships to both components of the INMB statistic, which indicates that higher study quality is related to more conservative estimates of incremental effects but also lower incremental cost, this result was very sensitive to small variations in the way QHES was operationalized - so that results may be interpreted with caution. This issue is discussed further in Chapter 6. Another point for discussion relates to the funding source of papers included in this exercise, which indicates higher INMBs for industry funded studies, whilst coefficients were not significant in the bivariate model. Finally, a methodological aspect was identified which may be referred to future research. The use of multiple imputation methods for missing data imputation in multilevel analysis is an area of on-going research. Accordingly, methods which are appropriate for complex data structures as modelled within this empirical exercise are not yet developed. This is the reason why regression based imputation was used for the purposes of imputing missing data within this section. Hence, future research may look into multilevel multiple imputation methods for multivariate multilevel models, cross-classified models, and other complex data structures.

Before discussing research findings in more detail in Chapter 6, the following Section 5.3 focusses on the analysis of country-level variability, or the lack thereof, using the random intercepts model with data and study-level covariates as developed above. Subsequently, a case study shows the relevance of modelling random slopes and variance as a function of explanatory variables for the transferability problem of economic evaluation data in the final Section 5.4 of this empirical Chapter.

5.3. Objective three: assessing country-level variability within the multilevel model framework

In Chapter 5.1, we learned that data from multinational studies is likely to '*disguise*' country-level variability existing in international cost-effectiveness data from single-country studies as country-level variability significantly increased when grouping data from multinational studies in a separate cluster on country-level. This assumption follows the idea that data from multinational studies is not independent on country-level and therefore underestimates geographic variability in measures of cost-effectiveness. Assigning this data to their respective geographic domains '*infects*' individual country parameters and drags them towards each other, with the result that variability on country-level is severely underestimated in the cross-classified MLM. For this reason, a three-level hierarchical model was used to assess covariates on data and study-level in Section 5.2 of this empirical exercise, and this model assumes that multinational study data is not independent between geographic domains.

Assessing data-level and study-level covariates in section 5.2 led to a number of important findings. For instance, Sculpher et al. (2004) stated that lower-level variability factors, such as patient and disease characteristics, may feed through to higher levels, thereby affecting the variability observed between studies and countries reflected in the dataset. The MLM framework offers an excellent opportunity to make this relationship explicit, as it allows for changes in variation in any direction on higher levels induced by covariates on lower levels of the data hierarchy (e.g. Hox, 2010). Hence, analysis could show, for instance, that the inclusion of patient and disease characteristics in the bivariate model successfully controls for some variability on data-level, whilst variability between studies and countries actually increases. This is also a compelling argument for the use of MLM methodology within this empirical exercise, as, in theory, data should become exchangeable on country-level once we controlled for the appropriate set of covariates on each level of the data hierarchy (Drummond et al., 2009).

Furthermore, a number of variability factors on data and study-level were discovered which are all associated with variation in measures of cost-effectiveness for statins in the primary and secondary prevention of CVD. These findings are subject for further discussion in Chapter 6. However, two additional findings are of utmost relevance for this particular section, which is concerned with the assessment of variability in measures of cost-effectiveness between geographic domains.

First of all, determining the appropriate MLM structure in Section 5.1 and including data and study-level covariates in Section 5.2 significantly increased variability on country-level for both ΔC and ΔE in the bivariate MLM. The analysis started off in Section 5.1 with a VPC on country-level of around 1% for ΔC and negligible 0.5% for ΔE respectively. Though, at the end of Section 5.2., country-level variability is still not a dominant source of variability in measures of cost-effectiveness, a VPC of around 20% for ΔC and 15% for ΔE indicates that this exercise was successful in disclosing variability between geographic domains. However, it is also acknowledged that data and study-level variability remain the dominant source of variation in international cost-effectiveness data for statins in the primary and secondary prevention of CVD, so that one may conclude that the '*appropriate set of covariates*' which Drummond et al. (2009) refer to, has not yet been found on data and study-level.

Nevertheless, some country-level covariates may be tested within the bivariate framework to investigate whether there are characteristics of countries which explain part of the variability observed in cost-effectiveness data between geographic domains. This is the focus of the second part of this section. Before assessing country-level covariates in the bivariate framework, however, there is another issue which has to be addressed first. Whilst country-level variability increased for both response variables in the bivariate framework, variation between countries remains negligible for INMBs in the univariate model, even after controlling for variability factors on data and study-level. Therefore, the first part of this section focusses on this issue, which is in sharp contrast to the results observed in the bivariate model specification.

The remainder of this section is organised as follows: Section 5.3.1 outlines the plan of analysis for this part of the empirical exercise. Potential reasons for the failure to observe an increase in country-level variability in the univariate framework are considered, and experiments are designed to test hypothesized explanations for this observation. Secondly, the strategy for testing country-level covariates in the bivariate model is outlined. Subsequently, Section 5.3.2 considers the data required for part 3 of this empirical exercise. Country-level covariates are introduced and descriptive statistics are reported. A separate methods sections has been omitted from this part of the empirical chapter, as the models run are identical to those models developed in Section 5.2 of this empirical exercise. For this reason, Section 5.3.3 proceeds directly to the results of this analysis, starting off with the univariate model which, thus far, failed to detect considerable country-level variation, and proceeding to the analysis of country-level variability factors, which, at this point, may only be indicated within the bivariate framework. However, if further variability on country-level can be disclosed in the univariate model, then the inclusion of country-level covariates may even be an option in this framework.

5.3.1. Plan of analysis for part three of this empirical exercise

The aim of this part of the analysis is to assess country-level variability – or the lack thereof – in the MLM framework. The two specific objectives which follow from this aim are a) to test potential explanations for the failure of the univariate model to detect country-level variation and b) to assess country-level covariates, at least within the bivariate framework; but potentially also in the univariate framework, provided that country-level variation increases after addressing the first objective in this section.

5.3.1.1. Assessing the lack of country-level variation in the univariate model

Let us consider the lack of country-level variation in the univariate framework first. To generate potential explanations for this observation, the best way is to concentrate on anything which differs between the response variables in the

univariate model, which did not detect considerable country-level variation, and the bivariate framework, where the VPC on country-level achieved around 20% for ΔC and 15% for ΔE respectively. Obviously, the bivariate model decomposes the response variable INMB into its stochastic components ΔC and ΔE. Hence, what follows from this fact is also what should be considered in this analysis.

First, in order to combine ΔC and ΔE to INMBs, one needs to determine a threshold value λ. This threshold values was set constant at £30,000 in Sections 5.1 and 5.2. Hence, INMBs were calculated so far as follows:

$$INMB_{statins} = \Delta E * \lambda - \Delta C \quad (1)$$

$\lambda =$ £ 30,000

It may be the case that the value of the threshold λ impacts on variability in measures of INMB. For instance, if one sets the threshold value λ at zero, this means that INMBs reduce to ΔC and, consequently, the variability in measures of cost-effectiveness resembles the variability observed for ΔC in the bivariate model. Conversely, if one constantly increases λ, the impact of ΔC diminishes as λ tends towards infinity. Hence, variability observed resembles the variability in ΔE only. Whilst this shows that the threshold value λ needs to be considered when looking into the variability in measures of INMB, there is no a priori reason to believe that this would impact more strongly on the country-level than it does on variability within or between studies in the dataset.

Nevertheless, it is relatively straightforward to design an experiment which tests the impact of the threshold value on the variability observed on each level. We may simply re-run the variance components model developed in Section 5.1 with INMBs calculated at different values for the WTP threshold λ. In other words, a sensitivity analysis of variability observed on each hierarchical level is performed with respect to λ. As a result, we can map variability observed on each level as a function of λ, and if there is a significant effect of the WTP threshold on country-level variability, the random intercepts model developed in Section 5.2 could be re-run at this particular threshold value. This may provide a univariate model specification which discloses country-level variability through the inclusion of

data and study-level covariates at a WTP threshold which minimises the obscuring effect of λ on variability in measures of cost-effectiveness between countries. Results of this experiment are reported in Section 5.3.3.

There may be a second reason for dramatically reduced country-level variability in INMBs observed in the univariate model. The variability in one component of the INMB statistic may be, at least in part, offset by variability in the other component of INMBs. Sections 5.1 and 5.2 already provided some insights into the variability in measures of cost-effectiveness which may support this hypothesis. The forest plots presented in Section 5.1.4 showed that some studies, which are outliers in terms of INMB, did not just show dramatically increased measures of ΔE, but also elevated measures of ΔC. In addition, bivariate statistics produced in Section 5.2.4.1 and also the random intercepts models developed in Section 5.2.4.2 showed that studies produced by S.A. Grover et al. using the 'CHD life expectancy model' are characterised by elevated measures of INMB, ΔE, and ΔC. In other words, elevated mean INMB observed in these studies is most likely the result of ΔE being estimated above the average of other studies included in this exercise. However, as the CHD life expectancy model also led to higher estimates of ΔC, the overall departure from the average INMB may be lower than it is for ΔE, due to an offsetting effect of elevated ΔC in the same model. In conclusion, a common pattern of variability in both components of the INMB statistic may reduce overall variability in incremental net monetary benefits.

To test this hypothesis, country-level residuals are obtained from MLwiN after running the three-level hierarchical variance-components model. Residuals are then used to build forest plots analogously to the study-level forest plots of Section 5.1.4. If ΔC are positively correlated with ΔE, one should observe a similar pattern in both forest plots (i.e. higher incremental cost are associated with higher incremental effects and vice versa). As an increase in one component of the INMB statistic would then be, at least in part, offset by an increase in the other component, this should lead to lower variability observed in the forest plot for INMBs. However, interpretation of a visual presentation of the data is not unambiguous. For this reason, country-level means for ΔC and ΔE are transferred

back to STATA 12 to produce Pearson correlations for the two variables. To test correlations in the raw data, the same is done with raw country means for the components of the INMB statistic. If the suspicion articulated above holds, Pearson correlations should be high, positive, and statistically significant.

This test may lead to an explanation for drastically lower country-level variability in the univariate model as compared to what has been observed for incremental cost and incremental effects in the bivariate framework. However, if this offsetting effect between ΔC and ΔE causes lower variability in INMBs, there is not much that could be done within this exercise to increase country-level variation in the univariate framework. Rather, this finding should be interpreted as a compelling argument in favour of decomposing the INMB statistic in the bivariate model.

5.3.1.2. Testing country-level covariates in the bivariate model

Moving on to this bivariate framework, the plan of analysis is to include covariates on country-level to test whether they may, in part, explain variability in measures of cost-effectiveness between geographic domains. Though variability, whilst constantly increasing throughout the course of this empirical exercise, may still not be regarded as particularly high on country-level (country VPC of 20% for ΔC and roughly 15% for ΔE respectively), this exercise makes an attempt to look into potential causes of this variability on level three of the data hierarchy. As with data and study-level covariates, the relevant literature is used first to learn about potential causes of variability in measures of cost-effectiveness between countries. As before, the two publications by Sculpher et al. (2004) and Goeree et al. (2007), who reviewed the cost-effectiveness literature to compile a list of factors potentially causing variability in measures of cost-effectiveness, provided a number of candidates for this assessment. Unfortunately, however, operationalizing such candidate variables and obtaining data for assessment appeared to be more difficult than it was experienced for data and study-level covariates. For instance, country-level variability factors are usually not explicitly modelled in economic evaluation studies and therefore not

reported in the respective studies either. Hence, alternative data sources are required which are reliable and provide estimates which are comparable across the range of countries included in the dataset. Therefore, data bases of international organisations (WHO, 2012; OECD, 2012; World Bank, 2012) are screened for indicators which may serve as a proxy for variability factors mentioned by Sculpher et al. (2004) or Goeree et al. (2007). This resulted in a considerable number of potential covariates to test in the multilevel framework. If data for any of the indicators is missing for a country included in this exercise, national sources (i.e. data provided by national departments of health, statistical bureaus etc.) are used to fill these gaps (most problematic in this sense was the special administrative region of Hong Kong. It was the geographic area for which most missing data problems occurred as it is not represented in the WHO data repository). If data is still missing after searching for alternative sources, the strategy is either to drop the affected variable, or to listwise delete affected data points for the particular analysis. Listwise deletion has been chosen as regression based imputation is not deemed appropriate for a dataset with only a few countries included, and as other ad hoc imputation techniques have a potential for bias. Again, this strategy was only applicable to eight data points relating to Hong Kong so that, if data for this geographic domain was missing, this resulted in a dataset of 2086 data points from 66 studies and applicable to 17 geographic domains, including one cluster on country-level for multinational study data.

This data from multinational studies is responsible for another issue which has to be addressed before proceeding with the analysis. Not dropping this data increases the sample size on data and study-level, and grouping it in a separate cluster on country-level assures that this data does not disguise existing variability between countries. However, the resulting group on country-level represents a number of geographic domains, so that the question is how to assign values for country-level variables to this particular group. It was decided to run all analyses twice, once with the full dataset as used so far, and once with a reduced dataset, where all data from multinational studies has been dropped. For the full dataset, country-specific data is collected for the geographic domains represented in the *'multinational'* cluster, and a mean value is calculated for the affected data points. Though this lumps together a number of potentially dissimilar cases, it allows keeping this data in the analysis which improves

estimation on data and study-level, and it also assures that country-level covariates can be tested without the distorting effect of multinational data on country-level for the remaining geographic domains. Bias is not likely to occur in any direction as a mean value is calculated from actual, not imputed data, though this obviously means that a) results for the group '*multinational*' are essentially meaningless and b) there is likely to be an impact on the precision of the estimation procedure. Therefore, all analyses are also conducted using the reduced dataset from Section 5.1, where all data from multinational studies has been dropped. This procedure results in a dataset with 1806 data points clustered in 61 studies and 17 countries.

As mentioned before, the models run in this section are identical to those developed in Section 5.2 of this empirical exercise, with the exception of country-level covariates now being explicitly considered in the model. For this reason, methods of analysis have been omitted from this section and the interested reader is referred to Section 5.2.2. of the empirical chapter. The next section introduces the country-level data required for the assessment as outlined above, before results are presented and discussed in Section 5.3.3.

5.3.2. Data

Assessing reasons for a lack of country-level variability in INMBs observed in the univariate MLM framework builds up from data used before in Sections 5.1 and 5.2. Hence, the data introduced here relates to country-level covariates only, which are the focus of the second part of the analysis reported here. Table 5.25 below summarizes the country-level data obtained from different sources. The first column refers to the variable name, whilst the second column provides a definition for the respective covariate, including the reference year this data refers to. The third column explains how variables relate to the list of variability factors as reported by Goeree et al. (2007). Though not all variability factors potentially relevant on country-level have been acknowledged in this exercise, Table 5.25 shows that covariates selected relate to a considerable number of variability factors previously discussed in the health economic literature. Columns four to six describe the nature of country-level covariates, provide summary statistics and report on respective data-sources. The anticipated

relationship between country-level covariates and measures of cost effectiveness is reported in columns seven to nine. The last two columns list studies which either support or contradict the relationship anticipated between response variables and country-level covariates.

On the broadest level, GDP per capita and the percentage of GDP devoted to healthcare are considered as covariates predominantly encoding the resources available for healthcare in a particular country. The variables encoding private, governmental or social security expenditure as a percentage of total health expenditure or governmental expenditure on health respectively may be interpreted as indicators of the organisational structure and the type of healthcare system. Precisely, systems within which general government expenditure on health is high, may be those classified as NHS like systems, whereas social health insurance type systems show higher social security expenditure on health as a percentage of general government expenditure on health. If private expenditure on health as a percentage of total expenditure on health is high, this may be indicative of a predominantly private health insurance based system. User fees, co-payments and deductibles are important incentives for patients, which were previously mentioned as potentially variability factors (Sculpher et al., 2004; Goeree et al., 2007). Consequently, out-of-pocket expenditure (OOP) as a percentage of private expenditure on health is assessed to address this potential source of country-level variability.

As statin prevention may be part of a national action plan to address CVD, the existence of such an action plan is encoded in the categorical 'CVD_policy'. The next four variables in Table 5.25 ('GPs', 'nurses', 'pharmacists', 'beds') relate to variability in available resources (staff, facilities, equipment), and input mix (personnel / equipment). The remaining variables in Table 5.25 relate either to patient characteristics or disease characteristics respectively. In particular, 'age', 'life expectancy at birth' and 'urban' relate to variability due to differences in demographics, mortality and population density, whilst '*CVD_death'* may be regarded as a measure of mortality due to the disease under consideration. A problem with the latter variable is the fact that the WHO, which provided data on this covariate, only gives a combined measure of cardiovascular and diabetes death per 100,000, so that it is not clear how much of these death`s relate to which disease. Results may therefore be interpreted with caution.

Table 5.25: Country-level covariates and their anticipated relationship to measures of cost-effectiveness

Name of covariate	Definition	Related variability factors from Goeree et al. (2007)	Nature of variable	Mean (SD) / Proportion (%)	Data sources	Anticipated relationship to			*Supporting evidence*	Contradicting evidence
						INMB	***ΔC***	***ΔE***		
GDP	GDP per capita, PPP (current international $), 2009	Available resources	Cont.	35169 (8969)	World Bank	Pos	Pos	Pos	7, 19	---
THE_GDP	Total expenditure on health as a percentage of gross domestic product, 2009	Available resources	Cont.	9.96 (2.38)	WHO / World Bank	Pos	Pos	?	6, 24	---
GOV_EXP_THE	General government expend. on health as a percentage of total expend. on health, 2009	Type of insurance coverage	Cont.	68.18 (11.44)	WHO / World Bank	Pos	Neg	Pos	17	---
PRIV_EXP_THE	Private expend. on health as a percentage of total expenditure on health, 2009	Type of insurance coverage	Cont.	29.67 (12.85)	WHO / World Bank	Neg	Pos	?	6	---
SOCSEC_GGE	Social security expend. on health as a % of general gov. expend. on health, 2009	Type of insurance coverage	Cont.	39.47 (37.91)	WHO / World Bank	?	?	?	---	---
OOP_PRIV_EXP	Out-of-pocket expenditure as a percentage of private expenditure on health, 2009	User fees, co-payments, deductibles, incentives for patients	Cont.	65.49 (18.00)	WHO / World Bank	Pos	Neg	?	6	---
CVD_POLICY	Existence of operational policy / strategy / action plan for cardiovascular disease? 2010	Access to programmes and services, conventions, norms, guidelines	Unord. cat.	No: 4 (22.22%) Yes: 12 (66.67%) Unclear: 2 (11.11%)	WHO	Pos	Neg	Pos	---	---
GPs	Physicians density (per 10 000 population) 2005-2009	Available resources, input mix	Cont.	29.13 (7.49)	WHO / Hong Kong, Centre for Health Protection, DH	Pos	Neg	Pos	---	---
Nurses	Nursing and midwifery personnel density (per 10 000 population) 2005-2009	Available resources, input mix	Cont.	82.03 (45.27)	WHO / Hong Kong, Centre for Health Protection, DH	Pos	Neg	Pos	---	---
Pharmacists	Density of pharmaceutical personnel (per 10 000 population) 2005-2009	Available resources, input mix	Cont.	7.74 (3.24)	WHO / Hong Kong, Centre for Health Protection, DH	Pos	Neg	Pos	---	---
Beds	Hospital beds per (10 000 population) 2008 – 2010	Available resources, input mix	Cont.	51.37 (28.88)	WHO / Hong Kong, Centre for Health Protection, DH	Pos	Pos	Pos	---	---
Age	Population median age (years), 2009	Demographics	Cont.	39.87 (3.50)	WHO / Hong Kong, Centre for Health Protection, DH	Neg	Neg	Neg	*2, 3, 4, 5, 8, 9, 10, 12, 14, 15, 20, 21, 22, 25, 26*	***INMB:*** *2, 3, 5, 11, 25 /* ***ΔC:*** *14 /* ***ΔE:*** *5, 11, 25*
						Relationship may reverse for age >65				
Urban	Population living in urban areas (%), 2009	Population density	Cont.	80.88 (10.39)	World Bank	?	?	?	---	---
Life Expectancy	Life expectancy at birth (in years), 2009	Demographics	Cont.	80.06 (2.80)	WHO / Hong Kong, Centre for Health Protection, DH	Pos	Pos	Pos	17	---
CVD_death	Cardiovascular and diabetes death per 100000, 2008	Mortality due to disease	Cont.	152.92 (59.84)	WHO	Pos	Neg	Pos	---	---
BMI_25	BMI ≥ 25, crude estimate, 2008	Disease interact., comorb., epidemiol.	Cont.	53.35 (10.37)	WHO	Pos	Neg	Pos	13	---
BMI_30	BMI ≥ 30, crude estimate, 2008	Disease interact., comorb., epidemiol.	Cont.	22.50 (6.25)	WHO	Pos	Neg	Pos	13	---
MEAN_BMI	Mean BMI, crude estimate, 2008	Disease interact., comorb., epidemiol.	Cont.	26.52 (1.22)	WHO	Pos	Neg	Pos	13	---
TCL_6.2	Raised TCL ≥ 6.2 mmol/L, crude estimate, 2008	Epidemiology	Cont.	19.81 (4.15)	WHO	Pos	Neg	Pos	*3, 8, 9, 22*	***ΔC****: 3, 22*
MEAN_TCL	Mean TCL (crude estimate), 2008	Epidemiology	Cont.	5.26 (0.21)	WHO	Pos	Neg	Pos	*3, 8, 9, 22*	***ΔC****: 3, 22*
SBP_140	Raised blood pressure (SBP ≥ 140 or DBP ≥ 90 or on medication, crude estimate, 2008	Disease interact., comorb., epidemiol.	Cont.	42.18 (4.89)	WHO	Pos	Neg	Pos	*3, 8, 12, 16, 22*	***INMB:*** *16* ***ΔC****: 3*
MEAN_SBP	Mean Systolic blood pressure, crude estimate, 2008	Disease interact., comorb., epidemiol.	Cont.	128.48 (3.42)	WHO	Pos	Neg	Pos	*3, 8, 12, 16, 22*	***INMB:*** *16* ***ΔC****: 3*
GLUCOSE_7	Raised blood glucose ≥ 7.0 mmol/L, 2008	Disease interact., comorb., epidemiol.	Cont.	9.58 (1.50)	WHO	Pos	Neg	Pos	*1, 3, 9, 10, 16, 18, 21, 23, 25*	***ΔC****: 3, 18, 23*
MEAN_GLUCOSE	Mean fasting glucose in mmol/L, crude estimate, 2008	Disease interact., comorb., epidemiol.	Cont.	4.49 (0.17)	WHO	Pos	Neg	Pos		

(1) Ashraf et al (1996), (2) CDC-Group (2002), (3) Drummond et al. (1993), (4) Franco et al. (2005), (5) Greving et al. (2011), (6) Grieve et al. (2005), (7) Grieve et al. (2007), (8) Grover et al (1999), (9) Grover et al (2000), (10) Grover et al. (2001), (11) Grover et al. (2008), (12) Hamilton et al (1995), (13) Hippisley-Cox et al. (2008), (14) HPS (2006), (15) HPS (2009), (16) Jönsson et al (1999), (17) Manca et al. 2007), (18) Muls et al. (1998), (19) Taghreed et al. (2003), (20) Nherera et al (2010), (21) NICE (2006), (22) Perreault et al (1998), (23) Soini et al (2010), (24) Thompson et al. (2006), (25) Tonkin et al. (2006), (36) Ward et al. (2007)

The remaining ten variables are measures of CVD risk factors on population level relating to obesity, total cholesterol level, blood pressure and blood glucose. An important question is why CVD risk factors may enter the MLM on country-level, as they were already included as variability factors on data-level. The reason is that patient and disease characteristics on data-level refer to differences in study populations, whilst the idea with population risk factors is that the cost-effectiveness of statins on country-level may partly depend on the existence of CVD risk factors within the target population of a country. The MLM framework offers an excellent opportunity to consider both, differences in study populations on data-level, and differences in potential target populations on country-level.

5.3.3. Results

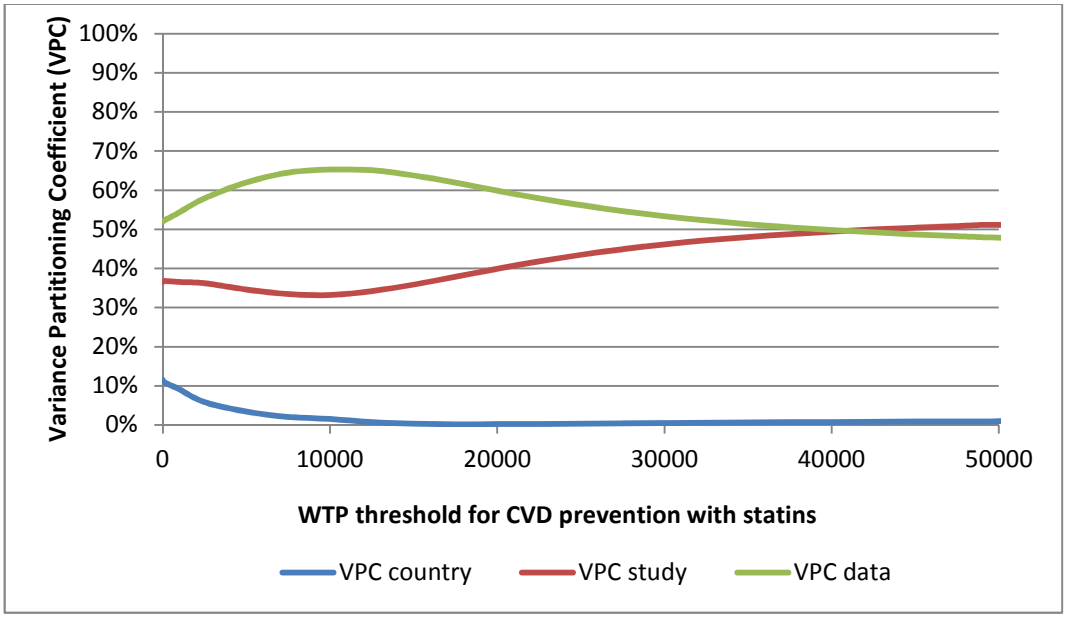
The first part of this section focuses on the assessment of variability in the univariate framework, within which it was not possible to disclose a substantial amount of variability in INMBs between geographic domains. The second part refers to the assessment of country-level variability factors in the bivariate MLM.

5.3.3.1. Reasons for a lack of country-level variability in the univariate model

As detailed in Section 5.3.1.1., two potential causes of a lack of country-level variability are being further assessed in this section, both originating from differences between the univariate model, where country-level variability is negligibly low, and its bivariate counterpart, where country-level variability was observed in both components of the INMB statistic. Results from performing a sensitivity analysis on the threshold value λ, which was set constant at £ 30,000 in Sections 5.1 and 5.2 of this empirical chapter, are reported first, before the focus is on the suspicion of a common pattern in measures of ΔC and ΔE on country-level. Figure 5.11 below summarizes the results of running the univariate variance components model as developed in Section 5.1 of this empirical exercise with different values for the WTP threshold λ. Analysis starts with a

threshold value of zero. This reduces the INMB statistic to ΔC only, and results in terms of variability on each level resemble those observed for incremental cost in the bivariate framework.

Figure 5.11: Variability in INMBs in the univariate model as a function of the threshold value λ



Gradually increasing λ reduces the country-VPC from just above 10% to below 0.5%, which is in complete accord with results of the univariate model reported in previous sections. Simultaneously, the VPC on data-level increases, whilst the study-level VPC remains relatively constant up to a WTP threshold of roughly £10,000. Beyond this level, however, the study-level VPC increases gradually as the data-level VPC decreases until both measures first cross and then level out beyond a threshold of around £40,000 per unit of health gain.

Figure 5.11 above leads to several important conclusions. First of all, the univariate model run with incremental cost only (which is equivalent to assuming λ = 0), resembles the findings for ΔC in the bivariate framework. This validates the results of the bivariate MLM. Secondly, increasing the threshold value reduces country-level variation, and at a WTP-threshold of £30,000, findings are also in accord with the results reported for the univariate model in previous

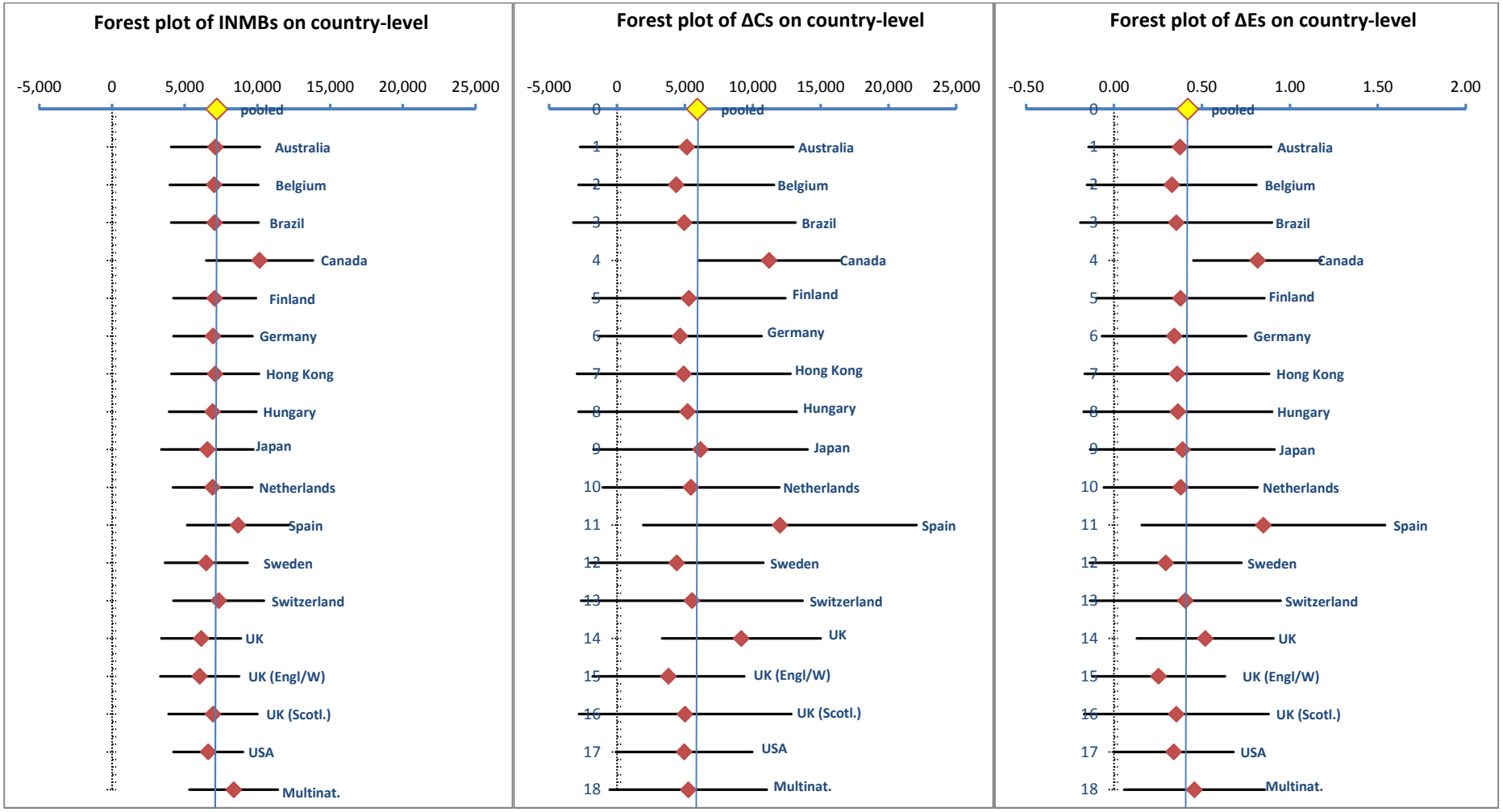
sections. However, we cannot confirm that country-level variation increases drastically at any value for λ, and even increasing WTP beyond £500,000, would only resemble the results obtained for ΔE in the bivariate framework (which again indicates that results are valid). As mentioned before, there is no reason why changing the threshold value should have a particular impact on the country-level beyond that observed at its extreme values, i.e. λ=0 and λ→∞, resembling the variation in ΔC and ΔE respectively. In conclusion, we may rule out the threshold value as a potential reason for the failure to observe significant country-level variation in the univariate model.

However, the sensitivity analysis showed that variability does increase at λ=0 and λ→∞, and this is in accord with the suspicion that there is a common pattern in ΔC and ΔE on country-level; so that variability in one component of the INMB statistic is partly offset by variability in the other component of INMBs. This would explain why country-level variability reduces with higher values for λ (as variability in ΔC is increasingly offset by corresponding variability in ΔE), and only increases again once λ is so high that variability in INMBs only reflects that of ΔE. This suspicion is supported by some of the previous findings, especially regarding studies using the CHD life-expectancy model, which report both higher incremental cost and higher incremental effects for statins in the primary and secondary prevention of CVD.

Hence, the next task is to further assess the idea that variability in one component of the INMB statistic is partly offset by variability in the other component of INMBs. This is done by simultaneously looking into country-level measures of ΔC, ΔE and INMBs respectively. For this reason, consider the forest plots in Figure 5.12 below.

Figure 5.12: Forest plots of incremental costs, incremental effects and INMBs on country-level

(Forest plots generated after running a three-level hierarchical random intercepts model and exporting country-level residuals from MLwiN to MS Excel)



There appears to be a common pattern in ΔC and ΔE, i.e. if an estimate of incremental effects for a particular country is above the pooled mean, so is the corresponding measure of incremental cost for the same country. This results in variability in ΔE being partly offset by corresponding variability in ΔC in the INMB statistic, and this is confirmed by the forest plot on the left showing that variability in INMBs is much lower compared to that observed on country-level in ΔCs and ΔEs. To rule out that this finding is an artefact from running MLMs providing country-level residuals which are subject to shrinkage, the same analysis was repeated with raw country means for ΔCs, ΔEs and INMBs. The forest plots (omitted here) appeared to be more '*noisy*' than that displayed in Figure 5.12 above as country means were not shrunken towards the pooled mean. Though this somewhat complicated the interpretation of results, a similar pattern appeared to be present in the data. However, to confirm this result, which is based on the interpretation of a graphical display of the data and therefore not unambiguous, Pearson correlations were produced for country-level mean ΔCs and ΔEs both after running the three-level model and also for country-level means calculated from the raw data. If there is a common pattern in mean incremental cost and incremental effects on country-level, then Pearson correlations should be high, positive and statistically significant.

Generating Pearson correlations in STATA 12 confirmed the results from interpreting Figure 5.12 above. A highly significant and positive correlation coefficient of 0.968 was found for country-level ΔC and ΔE obtained from the three-level hierarchical model. In addition to that, correlation coefficients for the raw mean ΔC and ΔE for each country were just as much correlated, with a correlation coefficient reaching 0.913. This result is also significant at the 1% level and rules out that shrinkage has anything to do with the common pattern observed in country-level mean incremental cost and incremental effects.

Hence, we may conclude that the lack of country-level variation observed in INMBs in the univariate model stems from a common pattern in the country-level variability for Δ and ΔE, so that variability in one component of the INMB statistic is partly being offset by corresponding variability in the other component of the INMB statistic. The discussion in Chapter 6 hypothesizes about

potential causes for this common pattern in ΔC and ΔE on country-level. At this point, however, one may only conclude that the lack of variability observed in the univariate model on country-level is likely to be a result of combining ΔCs and ΔEs to INMBs, so that within the scope of this exercise, there are no further means to increase country-level variability in the univariate framework. One should interpret this finding as a compelling argument for decomposing the INMB statistic within the bivariate model. This bivariate model is the focus of the second part of this section, which is entirely concerned with covariates encoding potential variability factors on country-level.

5.3.3.2. Country-level covariates in the bivariate multilevel model

Table 5.26 below reports results from individually testing each country-level covariate in the bivariate random intercepts model with covariates already included on data and study-level. Results from running the same model on the reduced dataset where data from multinational studies was dropped are reported in Appendix 5.5. Given the relatively low VPC for both ΔC and ΔE observed at the end of Section 5.2, the a priori expectation of ascertaining significant coefficients on country-level was low. Nevertheless, results show a number of coefficients moderately significant at the 10% level or 5% level respectively. In addition, coefficients mostly show the anticipated sign and reduce the DIC statistic, indicating an improved fit of the multilevel model. Results for running the same model on the reduced dataset are almost identical. Though coefficients obviously differ in their magnitude, these differences are generally negligible, and both signs of coefficients as well as their statistical significance accords between both analyses.

Whilst GDP per capita as an indicator of a countries economic performance was not significant when tested individually in the random intercepts model with data and study-level covariates already specified, there was a small but significant positive relationship between the total health expenditure as a percentage of GDP and incremental effects in the bivariate model. There may be

several potential explanations for this observation. Statin utilisation may be higher in countries which devote a higher percentage of their GDP to healthcare, or the prevalence of CVD risk factors may be different in those countries. In fact, when looking at correlations between explanatory variables on country-level, one can observe positive and highly significant correlations between '*THE_GDP*' on the one hand and '*mean_bmi', 'mean_sbp', 'mean_tcl'* and 'mean_glucose' on the other. Hence, the positive and significant coefficient of '*THE_GDP*' may either be explained by the fact that countries with higher GDP per capita are also characterised by higher levels of CVD risk-factors, or the significant coefficient may simply be a statistical artefact because of existing collinearity between the above named variables.

Table 5.26: Individually testing country-level covariates in the three-level bivariate random intercepts model

Bivariate model				
	Raw Mean (SD) / Proportion (%)	**ΔC/100** (2010 £ Sterling)	**ΔE**	**DIC (**Benchmark: 26413 (%-change))
GDP	35169 (8969)	0.001 (0.001)	0.000 (0.000)	26413 (0.000%)
THE_GDP	9.96 (2.38)	0.109 (1.280)	0.020 (0.008)**	26410 (-0.011%)
GOV_EXP_THE	68.18 (11.44)	-0.215 (0.258)	-0.001 (0.002)	26415 (0.008%)
PRIV_EXP_THE	29.67 (12.85)	0.223 (0.253)	0.000 (0.002)	26414 (0.004%)
SOCSEC_GGE+	39.47 (37.91)	-0.096 (0.131)	0.001 (0.001)	26413 (0.000%)
OOP_PRIV_EXP+	65.49 (18.00)	0.031 (0.182)	-0.000 (0.001)	26415 (0.008%)
CVD_POLICY No Yes Unclear	 4 (22.22%) 12 (66.67%) 2 (11.11%)	 Omitted 26.72 (19.24) 19.91 (19.63)	 Omitted 0.108 (0.109) -0.094 (0.110)	26413 (0.000%)
GPs	29.13 (7.49)	-1.005 (0.981)	0.011 (0.005)**	26409 (-0.015%)
NURSES	82.03 (45.27)	0.041 (0.130)	0.001 (0.001)	26414 (0.004%)
PHARMACISTS	7.74 (3.24)	1.642 (1.832)	0.019 (0.010)*	26409 (-0.015%)
BEDS	51.37 (28.88)	-0.121 (0.256)	0.002 (0.002)	26410 (-0.011%)
AGE	39.87 (3.50)	-2.003 (1.828)	0.020 (0.001)*	26408 (-0.019%)
URBAN	80.88 (10.39)	0.268 (0.564)	-0.001 (0.003)	26415 (0.008%)
LIFE_EXPECTANCY	80.06 (2.80)	-3.059 (3.408)	0.004 (0.021)	26414 (0.004%)
CVD_DEATH	152.92 (59.84)	0.114 (0.153)	0.000 (0.001)	26414 0.004%)
BMI_25+	53.35 (10.37)	0.434 (0.644)	0.004 (0.004)	26414 0.004%)
BMI_30+	22.50 (6.25)	0.559 (0.965)	0.010 (0.006)*	26413 (0.000%)
MEAN_BMI+	26.52 (1.22)	2.353 (5.056)	0.060 (0.031)*	26412 (-0.004%)
TCL_6.2+	19.81 (4.15)	-0.562 (0.858)	0.004 (0.005)	26414 (0.004%)
MEAN_TCL+	5.26 (0.21)	-18.91 (21.18)	0.177 (0.130)	26412 (-0.004%)
SBP_140+	42.18 (4.89)	-0.232 (0.748)	0.001 (0.005)	26415 (0.008%)
MEAN_SBP+	128.48 (3.42)	-0.284 (1.017)	-0.001 (0.007)	26415 (0.008%)
GLUCOSE_7+	9.58 (1.50)	1.572 (2.061)	0.025 (0.013)*	26410 (-0.011%)
MEAN_GLUCOSE+	4.49 (0.17)	14.03 (23.92)	0.356 (0.147)**	26409 (-0.015%)

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level

\+ Eight data points referring to the special administrative region Hong Kong have been dropped due to country-level data missing for this geographic domain

Positive and significant coefficients were also observed for variables encoding the availability of resources (GPs and pharmacists). More GPS or pharmacists per 10,000 population may be a reason for better utilisation of statins as they may be made available to a higher number of eligible patients and as those patients may be better monitored and motivated to comply with statin therapy (for GPs, however, the coefficient was not significant when running the model on the reduced dataset). Finally, significant coefficients were observed for age, BMI and the percentage of population with a blood glucose level of above 7 mmol/L. These findings accord expectations as statins should be more effective in populations with higher CVD risk.

When testing country-level covariates individually in the random intercepts model, significant coefficients were only observed for incremental effects, whereas none of the covariates tested showed a significant relationship to incremental cost in the bivariate model. The same observation was made when running the model on the reduced dataset. A potential reason may be the presence of confounding factors, which potentially disguise relationships which may exist within the data. For this reason, country-level covariates were tested simultaneously in the three-level bivariate multilevel model. Results for the full dataset are obtainable from Table 5.27 below. Results for the reduced dataset are reported in Appendix 5.6.

Table 5.27: Bivariate random intercepts model fully specified on data, study and country-level

	ΔC/100 (2010 £ Sterling)	**ΔE**
N (countries)	18	18
N (studies)	67	67
N (data)	2094	2094
Intercept (λ=£30,000)	739	-1.500
TCL (SE)	-33.44 (3.13)***	0.040 (0.018)**
HDL (SE)	183.11 (53.91)***	-0.521 (0.301)*
SBP (SE)	-3.39 (0.19)***	0.013 (0.001)***
Diabetes (SE)	-19.35 (8.30)**	0.441 (0.053)***
Age_cat (SE)		
<45	Omitted	Omitted
46-55	-95.34 (6.02)***	-0.308 (0.036)***
56-65	-133.9 (5.91)***	-0.639 (0.035)***
66-75	-168.8 (6.27)***	-1.066 (0.037)***
>75	-131.6 (10.62)***	-0.656 (0.063)***
Unclear	-113.3 (45.68)**	-0.512 (0.224)**
Gender (SE)		
Female	Omitted	Omitted
Male	-33.13 (4.52)***	0.234 (0.027)***
Mixed sample	-31.11 (22.93)	0.234 (0.107)**
CVD_history (SE)		
No	--	Omitted
Yes		0.267 (0.054)***
Mixed sample		0.345 (0.188)*
Cost_intervention	0.127 (0.010)***	--
Cost_comparator	-0.139 (0.022)***	--
Active_comparator		
No (doing nothing)	--	Omitted
Yes (statin)		-0.274 (0.072)***
DRB	--	-4.127 (0.892)***
Persp_cost_concl.		
Health insurance (NHS)	Omitted	--
Provider	-401.19 (22.86)***	
Societal	-175.26 (19.70)***	
Horizon		
< 20 years	--	Omitted
>20 years (lifetime)		0.184 (0.045)***
Duration=horizon		
yes	--	Omitted
No (treatment duration< horizon)		0.394 (0.076)***
Base case		
Yes	Omitted	Omitted
No	9.42 (4.81)*	0.089 (0.030)***
Barbieri_score_2		
Type 1	Omitted	
Type 2	-37.11 (15.26)**	--
Type 3	-66.07 (30.58)**	
Type 4	-99.09 (32.85)***	
Author_Grover		
No	Omitted	Omitted
Yes	156.35 (43.68)***	1.443 (0.200)***
4S		
No	--	Omitted
yes		0.434 (0.153)***
Scope		
CHD	Omitted	
CHD and stroke	-38.71 (8.76)***	--
CHD, stroke and PAD	-33.40 (34.46)	
unclear	-104.78 (64.86)	
GDP_CAPITA	-0.002 (0.001)**	--
GOV_EXP_THE	-0.759 (0.407)*	--
THE_GDP	--	0.042 (0.010)***
AGE_POPULATION	--	0.050 (0.013)***
Random part:		
σ^2_{u0j} **(Country)**	3640	0.069
σ^2_{u0k} **(Study)**	7946	0.137
σ^2_{e0} **(Data)**	5517	0.194
VPC - Country	21.28%	17.25%
VPC – Study	46.46%	34.25%
VPC - data	32.26%	48.50%
DIC (benchmark)	26387 (26412)	

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level

As a result of testing different combinations of country-level covariates in the three-level MLM with data and study-level covariates already specified, GDP per capita and government expenditure on health as a percentage of total health expenditure turn out to be negatively related to ΔC in the bivariate model and significant at the 5% and 10% level respectively. The same holds for running the model on the reduced dataset, with even higher significance levels (1% and 5% respectively). Countries with high government spending on health may also be characterised by higher market regulations, which may result in differences in prices for healthcare which potentially explains lower incremental cost. Conversely, countries with higher GDP may have higher prices for healthcare, so that future health care cost avoided through statin prevention may result in lower ΔC in countries with higher GDP. Finally, both total health expenditure as a percentage of GDP and age turn out positive and highly significant for ΔE, which confirms findings from testing both variables individually in the model. Again, results from running the model on the reduced dataset are in accord with these findings, with coefficients even being significant on the 1% level.

5.3.4. Summary and conclusions for part three of this empirical exercise

This section was concerned with variability in measures of cost-effectiveness between the countries reflected in the dataset. The analysis consisted of two parts. The objective of part one was to analyse potential causes for a lack of country-level variability in the univariate multilevel framework with INMBs as single response variable. Part two, on the other hand, was concerned with covariates on country-level in the bivariate framework, within which considerably more country-level variability was disclosed before in Sections 5.1 and 5.2 of this empirical exercise.

Two potential reasons where considered for a lack of country-level variability in the univariate framework. First, the threshold value λ, which was set constant in previous analyses, was suspected to be responsible for lower country-level

variability in INMBs as compared to its stochastic components ΔC and ΔE. For this reason, a sensitivity analysis was performed within which the model was run at different threshold values. For extreme values, this analysis showed similar country-level variability as for ΔC and ΔE in the bivariate model. This makes intuitive sense as variability in INMBs stems from ΔC alone if λ equals zero, and ΔE alone if λ tends towards infinity. However, with respect to country-level variability, the threshold value may be excluded as a potential reason for a lack thereof in the univariate model.

Secondly, it was suspected that variability in one component of the INMB statistic may be partly offset by variability in the other component. Previous analyses in Sections 5.1 and 5.2 of the empirical chapter already indicated that this may be the case. For this reason, forest plots with country means and their respective confidence intervals were produced for each response variable both from the raw data and using country-residuals from running the three-level hierarchical model. Analysing these forest plots strengthened the suspicion that variability on country-level in one component of the INMB statistic may be partly offset by variability in the other component, ultimately leading to drastically reduced country-level variability in INMBs. To confirm these results, Pearson correlations were produced both for the raw data as well as the country means obtained from running the three-level MLM. Correlations were well above 0.9 and highly significant. It may hence be concluded that the lack of country-level variability in INMBs results from combining ΔC and ΔE, which have similar patterns of variability, so that variability in one component of the INMB statistic is partly being offset by variability of the other component. Potential reasons for this observation are considered in the discussion section in Chapter 6.

Finally, this section was also concerned with country-level covariates for the bivariate model, within which the proportion of country-level variability was considerably higher. To rule out that lumping country-level data from multinational studies in a separate group on country-level distorts results, this analysis was performed both with the full dataset, and a reduced dataset which drops data points from multinational studies. Results were almost identical for both datasets. Testing country-level covariates individually in the model did not

result in any significant covariates for ΔC, whilst government expenditure on health as a percentage of total health expenditure, the number of GPs and pharmacists per 10,000 inhabitants, the mean population age as well as measures of BMI and blood glucose showed significant relationships to ΔE. When simultaneously assessing covariates in the final random intercepts model with variability factors considered on each hierarchical level, GDP per capita and the government expenditure on health as a percentage of total health expenditure turned out to be significantly related to ΔC, whilst coefficients for total health expenditure as a percentage of GDP and the population age were highly significant for ΔE. These results are further discussed in Chapter 6.

Assessing country-level covariates in the three-level hierarchical random intercepts model concludes the systematic assessment of variability factors on all hierarchical levels of the dataset. Results are further discussed in Chapter 6, together with findings from the two previous Sections 5.1 and 5.2 of this empirical exercise. Before that, however, Section 5.4 of this empirical chapter shifts the focus towards a methodological feature of the MLM framework, which is hypothesized to relate at its core to the transferability problem of economic evaluation in health. A case study shows how random slopes may be added to the model, which then allows to explicitly model variation in measures of cost-effectiveness as a function of explanatory variables. This so called '*variance function'* is then constructed for a number of covariates in the model to show how this concept may be utilised to focus research efforts more specifically to those questions for which disagreement (i.e. variation) in international cost-effectiveness data is particularly high.

5.4. Objective four: random slopes and the variance function

The first section of this empirical exercise was concerned with determining the appropriate MLM structure for the analysis of secondary cost-effectiveness data of statins in the primary and secondary prevention of CVD. A three-level hierarchical model with measures of cost-effectiveness grouped in studies and studies grouped in countries was developed where data from multinational studies was clustered in a separate group on country-level. This model was carried forward to section two of this empirical exercise, where data-level and study-level covariates were tested in a random intercepts model to control for variability on both levels. Throughout the course of this exercise, further country-level variability was disclosed for incremental effects and incremental cost in the bivariate model, whilst variability on country-level remained negligible for INMBs in the univariate framework. Accordingly, in Section 5.3, potential reasons for a lack of country-level variation in the univariate model were assessed, and covariates on country-level were included in the bivariate model. The analysis carried out in Section 5.3 concluded the assessment of variability factors for measures of cost-effectiveness on all hierarchical levels of the dataset.

This final section of the empirical chapter shifts the focus towards a methodological feature of the MLM framework which may address the transferability problem of economic evaluation in health at its core. In a first step, random slopes are introduced to the models developed. Allowing slopes of covariates to vary within the MLM framework is something which other health economists already considered in their applications of MLM to economic evaluation data (Sculpher et al., 2004; Manca et al., 2005; Thompson et al., 2006; Bachmann et al., 2007). Doing so, however, also allows modelling the variation in the relationship between explanatory variables and the response variable as a function of explanatory variables and this is, to the knowledge of the author, something which has neither been considered in the area of health economics in general, nor in health economic evaluation in particular. The MLM literature refers to this concept as the '*variance function*' (e.g. Steele, 2008; Rasbash et al., 2009).

This final section of the empirical analysis shows how the variance function relates to the transferability problem. The idea is that for the range of values of explanatory variables where variance in international cost-effectiveness data is low, the transfer of existing evidence to the target location is rather indicated, compared to ranges of explanatory variables which show high variability in measures of cost-effectiveness. The following Subsection 5.4.1 starts off by outlining the plan of analysis for introducing random slopes to the model and modelling the variance function, before the methods of analysis are detailed in Section 5.4.2. The data used for this analysis has already been described in detail before, which is why results are reported directly after the methods of analysis in Section 5.4.3.

5.4.1. Plan of analysis for part four of this empirical exercise

It is important to clarify the rationale behind fitting random slopes to the multilevel models developed thus far. In the variance components model applied in Section 5.1 of this chapter, the effect of explanatory variables on measures of cost-effectiveness was not acknowledged. Therefore, the random intercepts model was introduced with multiple explanatory variables on data and study-level in Section 5.2 of this chapter. Country-level covariates were then added to the bivariate model in Section 5.3. However, thus far it was assumed that the effect of explanatory variables on measures of cost-effectiveness of statins in the primary and secondary prevention of CVD is the same across all higher-level units (i.e. studies and/or countries). In other words, the individual regression lines for each study (and country) in the dataset were assumed to be parallel to each other (Steele, 2008; Rasbash et al., 2009).

It is very likely though that the relationship between explanatory variables and measures of cost-effectiveness differs between different studies and/or countries in the dataset. For instance, it is not unreasonable to assume that, even if all studies report a positive relationship between total cholesterol and INMB, the slope of this relationship, i.e. the coefficient of the covariate, differs between studies. Likewise, the relationship between INMBs and say the annual

drug cost of the intervention may be different for different countries in the dataset. The MLM framework offers a unique opportunity to account for differences in the relationship between explanatory variables and response variables across higher-level units (Steele, 2008; Bartholomew et al., 2008; Rasbash et al., 2009; Hox, 2010).

When allowing the relationship between explanatory variables and the response variable to differ between higher-level units, regression lines are no longer parallel to each other as they are in the random intercepts model (Steele, 2008; Hox, 2010). As long as regression lines run parallel, higher-level variability is only reflected in differences between intercepts, whilst variability in slopes is, by definition, not existent (Steele, 2008; Rasbash et al., 2009; CMM workshops / random slopes). Once slopes are allowed to differ, the variability between group regression lines may be different at any value of the explanatory variable (Steele, 2008; Rasbash et al., 2009; CMM workshops / random slopes). Hence, this variability may now be expressed as a function of the explanatory variable (Steele, 2008; Rasbash et al., 2009; CMM workshops / random slopes). Further details are also available from the MLM Chapter 3 and from the methods of analysis section below. Modelling this variance function for a number of explanatory variables is the objective of this final section of the empirical chapter.

To demonstrate the potential value of the variance function for addressing the transferability problem in economic evaluation in health, the analysis in this section concentrates on patient and disease characteristics, which Sculpher et al. (2004) rightfully identified as a critical source of variability in measures of cost-effectiveness, potentially feeding through to variability between higher-level units. The analysis within this project thus far showed that patient and disease characteristics are, in fact, an important source of variability, and that the inclusion of respective covariates also changed the variation observed on higher levels. Additionally, relationships observed between patient and disease characteristics and response variables were always highly significant and in accord with prior expectations. This is in sharp contrast to study characteristics, where only few categorical variables were identified out of a large number of

candidates which successfully controlled for some variability on study and / or country-level – even though the study-level constitutes a major source of overall variability in the data. Likewise, country-level variability was observed to be low, even in the bivariate model, and only few variables were found to be related to the response variable.

Having decided to use data-level covariates which encode characteristics of the patient and the disease to demonstrate the concept of the variance function, one needs to determine on which level random slopes may be fitted. This could be, in theory, any level of the model. However, it was decided to model random slopes on the study-level for the following reasons. When considering transferability problems in practice, decision makers may have to choose out of a number of existing economic evaluation studies applicable to other geographic domains. This choice may be based on the '*degree of similarity'* between study characteristics and the target location, which accords the principles of analogical reasoning as outlined in Chapter 2. To assess the extent to which available studies meet the requirements of the target jurisdiction, decision makers hence compare attributes of the studies available with attributes of the target context, and decide which of the available international cost-effectiveness studies may be most appropriate to inform decisions in the target country. This process is also reflected in available transferability checklists, decision charts or indices (e.g. Heyland et al., 1996; Späth et al., 1999; Welte et al., 2004; Boulenger et al., 2005; Turner et al., 2009; Nixon et al; 2004 & 2009; Antonanzas et al., 2009). In short, the choice to make is a choice between existing studies, not geographic domains. Only once the decision maker has identified a number of candidate studies which meet the requirements of the target jurisdiction, country characteristics may be considered to the extent to which they are reflected in the economic evaluation in question. However, the analysis within the empirical exercise thus far clearly showed that country-level variability is low, even after controlling for a large number of potential variability factors on data and study-level. Hence, the geographic context within which the available evidence was originally produced turned out to be a far less important source of variability in international cost-effectiveness data than differences between economic evaluation studies, which this analysis proved to be the overriding source of the variability in measures of

cost-effectiveness. This, in conclusion, constitutes a compelling argument for fitting random slopes on study-level.

Before proceeding to the methods of analysis, it needs to be clarified once more that this chapter aims to demonstrate the value of fitting random slopes and modelling the variance function, it is not aimed to provide a systematic assessment of random slopes within the MLMs developed in previous sections. Doing so would require testing random slopes for all covariates on a '*parameter by parameter*' basis, to avoid building an '*overparameterized model which suffers from serious estimation problems*' (Hox 2011). Secondly, covariates which were not significant with fixed slopes may be significant in a random slope specification, so that previously excluded covariates would have to be tested again (Hox, 2011). Third, random slopes may be fitted to continuous and categorical data, and in the case of categorical variables, they may be fitted to each category separately (Rasbash et al., 2009). As a result, a systematic and full assessment of random slopes within this exercise would include considering an enormous amount of variables and categories of categorical variables with random slopes being tested on each level in the univariate model as well as each side of the bivariate model respectively, and this is clearly not feasible within the scope of this exercise.

As a result, the following steps are taken for a number of patient and disease characteristics (TCL, SBP and smoking) to demonstrate the potential value of the variance function for addressing the transferability problem of economic evaluation data:

1. running an OLS-regression model to plot an overall mean regression line without assuming complex data structures (this serves as baseline for analysis and helps demonstrating the relative merits of the multilevel model methodology).
2. running a random intercepts model which results in an overall mean regression line and, parallel to that, individual study regression lines for each study in the dataset (this model specification is equivalent to the models run in Sections 5.2 and 5.3 of this empirical exercise)

3. running a random slopes model which allows study-level regression lines to differ both with respect to their intercepts and slopes (this model specification relaxes the assumption of an identical relationship between explanatory variables and response variables for different studies)
4. Model the variance in the relationship between patient and disease characteristics and measures of cost-effectiveness between studies as a function of explanatory variables.

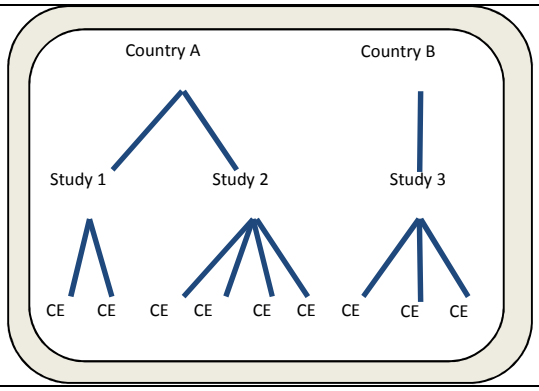
Models are run both with INMBs in the univariate framework as well as incremental cost and incremental effects in the bivariate framework using MCMC estimation in MLwiN (Rabsash et al., 2009a, Browne, 2012)

5.4.2. Methods of analysis

The random slopes model in this section of the empirical analysis is, as before, run both in a univariate specification, with INMB as the only response variable, as well as a bivariate model, with ΔC and ΔE as a vector of response variables. The model of interest is a three-level hierarchical model, which groups data in studies and studies in countries respectively. In addition, data from multinational studies is being clustered in a separate group on country-level. Table 5.28 below summarizes the models relevant for this section.

As before in the random intercepts model, we can interpret the parameter 'β_0' as the intercept and '$\beta_1 x_{1ij}$' as the slope of the pooled regression line. Also identical to the univariate random intercepts model, 'σ_{e0}^2' is the variance for the within-study error term 'e_{0ijk}' and 'σ_{v0}^2' is the variance of the between country error term 'v_{0k}' respectively. However, in contrast to the random intercepts model, the slope is no longer identical between the pooled regression line and the individual study regression lines.

Table 5.28: Multilevel models for exercise four

	Model of interest (Three-level hierarchical model)
Model summary	Three-level hierarchical random slopes model with cost-effectiveness estimates being nested in economic evaluation studies and studies being nested in geographic domains. Data from multinational studies is being clustered in a separate group on country-level. Random slopes were fitted on level two to allow the relationship between data-level covariates and response variables (INMB, ΔC, ΔE) to differ between different studies in the dataset.
Unit Diagram	
Univariate model specification	$y_{ijk} \sim N(XB, \Omega)$ $y_{ijk} = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + v_{0k} + u_{0jk} + u_{1jk} + e_{0ijk}$ With $[v_{0k}] \sim N(0, \Omega_v)$ where $\Omega_v = [\sigma^2_{v0}]$ $\begin{bmatrix} u_{0,jk} \\ u_{1,jk} \end{bmatrix} \sim N\ (0, \Omega_u)$ where $\Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$ $[e_{0ijk}] \sim N(0, \Omega_e)$ where $\Omega_e = [\sigma^2_{e0}]$
Bivariate model specification	$\begin{bmatrix} Y_{0,ijk} \\ Y_{1,ijk} \end{bmatrix} \sim BVN(XB, \Omega)$ $y_{d,ijk} = (\beta_{0d} + \beta_{1d} x_{1ijk} + \beta_{2d} x_{2jk} + \beta_{3d} x_{3k} + v_{0dk} + u_{0djk} + u_{1djk} + e_{0dijk}) * r_{d,ijk}$ $r_{1,ijk} = \begin{Bmatrix} 1\ if\ \Delta\ cost \\ 0\ if\ \Delta\ effects \end{Bmatrix} \quad r_{2,ijk} = 1 - r_1$ With: $\begin{bmatrix} v_{0,0,k} \\ v_{0,1,k} \end{bmatrix} \sim BVN\ (0, \Omega_v)$ where $\Omega_v = \begin{bmatrix} \sigma^2_{v0,0} & \\ \sigma_{v0,01} & \sigma^2_{v0,1} \end{bmatrix}$ $\begin{bmatrix} u_{0,0,jk} \\ u_{0,1,jk} \\ u_{1,0,jk} \\ u_{1,1,jk} \end{bmatrix} \sim BVN\ (0, \Omega_u)$ where $\Omega_u = \begin{bmatrix} \sigma^2_{u0,0} & & & \\ \sigma_{u00,01} & \sigma^2_{u0,1} & & \\ \sigma_{u01,00} & \sigma_{u01,10} & \sigma^2_{u1,0k} & \\ \sigma_{u01,01} & \sigma_{u01,11} & \sigma_{u11,01} & \sigma^2_{u1,1k} \end{bmatrix}$ $\begin{bmatrix} e_{0,0,ijk} \\ e_{0,1,ijk} \end{bmatrix} \sim BVN\ (0, \Omega_e)$ where $\Omega_e = \begin{bmatrix} \sigma^2_{e0,0} & \\ \sigma_{e0,01} & \sigma^2_{e0,1} \end{bmatrix}$

Hence, special attention needs to be placed on the interpretation of $'\Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$, in which '$\sigma^2_{u1}$' is the variance in slopes between studies, 'σ^2_{u0}' is the variance in intercepts between studies, and 'σ_{u01}' is the covariance between intercepts and slopes (Steele, 2008; Rasbash et al., 2009). This issue is also further addressed below. For the bivariate model, the variance and covariance

between the stochastic components of the INMB statistic need to be considered as well, resulting in a 4*4 variance-covariance matrix for 'Ω_u' on study-level (Bartholomew et al., 2008). For the univariate model it is assumed that INMB is normally distributed at each level of the model whilst a bivariate normal distribution is assumed for the bivariate model. As in the bivariate random intercepts model before, the response indicator 'r' is 1 for ΔC and 0 for ΔE and a separate level for this response indicator has been fitted below the data-level. Further details on the multilevel methodology applied in this empirical exercise are also available from Chapter 3.

As mentioned above, special attention needs to be placed on the variance in intercepts between studies 'σ_{u0}^2', the variance between their respective slopes 'σ_{u1}^2', as well as the intercept slope covariance, denoted with 'σ_{u01}'. In general, a random slope model allows not just intercepts, but also the slopes of individual regression lines to differ between studies included in the dataset (Steele, 2008; Rasbash et al., 2009). This means, if random slopes are assumed in the relationship between an explanatory variable and measures of cost-effectiveness, this allows study regression lines to vary both in their intercepts and slopes across the range of values of the explanatory variable (Steele, 2008; Rasbash et al., 2009). In other words, study-level variability becomes a function of the explanatory variable (Steele, 2008; Rasbash et al., 2009). This has very important implications: first of all, the covariance between slope and intercept variance tells us about the pattern of individual study regression lines with respect to changes in the explanatory variable (Rasbash et al., 2009; MLM workshops / random slopes). If this covariance is negative, then study regression lines are *'fanning in'* over the range of the explanatory variable (Steele, 2008; Rasbash et al., 2009; MLM workshops / random slopes). Conversely, if 'σ_{u01}' is positive, study predictions are fanning out (Steele, 2008; Rasbash et al., 2009; MLM workshops / random slopes). An intercept slope covariance of zero indicates no particular pattern in individual study regression lines (Steele, 2008; Rasbash et al., 2009; MLM workshops / random slopes).

Secondly, special attention needs to be placed on where the explanatory variable for which a random slope has been fitted is centred (Steele, 2008; Rasbash et al.,

2009). The matrix '$\Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{01} & {}^{2}_{u1} \end{bmatrix}$' defines the relationship between the error terms of the intercepts 'u_{0j}' and the error term of the slopes 'u_{1j}' (Steele, 2008; Rasbash et al., 2009). Depending on where the y-axis cuts the x-axis, one would obtain different measures of the between group error term 'u_{0j}', the variance of the intercepts between studies 'σ^2_{u0}' and the covariance between slopes and intercepts 'σ_{u01}', which is why all parameters must be interpreted simultaneously and in the light of where x=0 was placed. (Steele, 2008; Rasbash et al., 2009; MLM workshops / random slopes). To illustrate this issue, consider Figures 5.13 and 5.14 below. The black line in Figure 5.13 represents the overall mean regression line, whilst the green line represents the relationship between the explanatory variable x and the measures of cost-effectiveness in one particular study. Depending on the value of x, the variability between studies changes, which is shown by the red arrows between the overall mean regression line and the individual study prediction.

Figure 5.13: Study-level variability depends on value of explanatory variable in a random slopes model

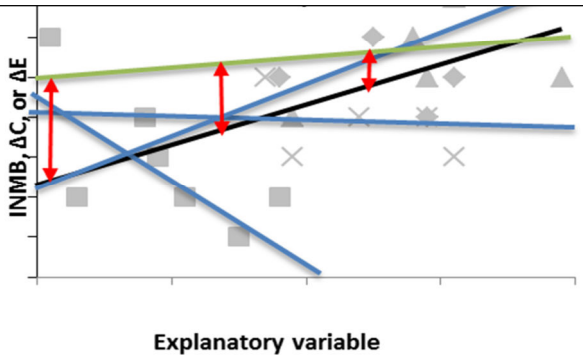
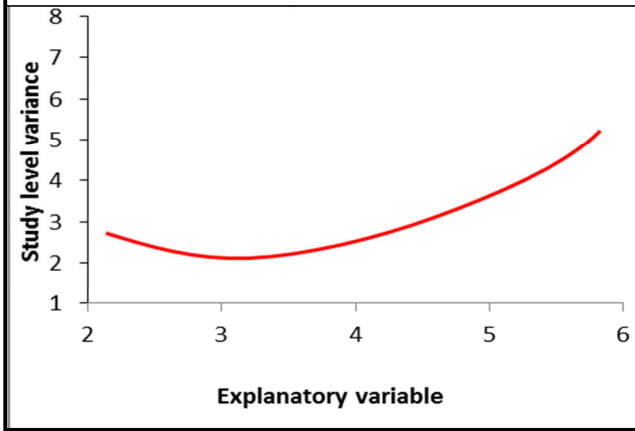


Figure 5.14: Resulting variance function for sample dataset



In relation to that, the variance partitioning coefficient, which tells us about the amount of variability attributable to each hierarchical level of the model, now depends on the value of the explanatory variable too. It can be shown that study-level variability is now a quadratic function of the explanatory variable (Steele, 2008), which is shown in equation (1)

$$Var(u_{0j} + u_{1j}x_{1ij}) = \sigma^2_{u0} + 2\sigma_{u01}x_{1ij} + \sigma^2_{u1}x^2_{1ij} \quad (1)$$

Accordingly, the VPC may now be expressed as:

$$VPC_{study} = \frac{\sigma^2_{u0}+2\sigma_{u01}x_{1ij}+\sigma^2_{u1}x^2_{1ij}}{\sigma^2_{u0}+2\sigma_{u01}x_{1ij}+\sigma^2_{u1}x^2_{1ij}+\sigma^2_{e0}} \quad (2)$$

Equation (1) is for obvious reasons also referred to as the '*variance function'* which is also displayed in Figure 5.14 above (Steele, 2008; Rasbash et al., 2009). This variance function is being modelled for patient and diseases characteristics TCL, SBP and smoking. Results are reported in Section 5.4.3 below. It is hypothesized that additional information generated through conducting new studies in the target country may be particularly valuable for (ranges of) explanatory variables where study-level variability is high. Conversely, there may be regions of the variance function for which variability in existing data is low, and a study produced for the target country would be more likely to produce results which are in the same range then existing studies; so that additional research may not be indicated and one may rather transfer evidence from existing studies to the target domain. Hence, the variance function may be used to target research resources more specifically to those questions for which study-level variability in measures of cost-effectiveness is particularly high.

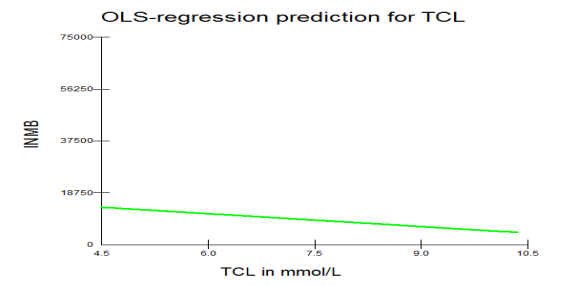
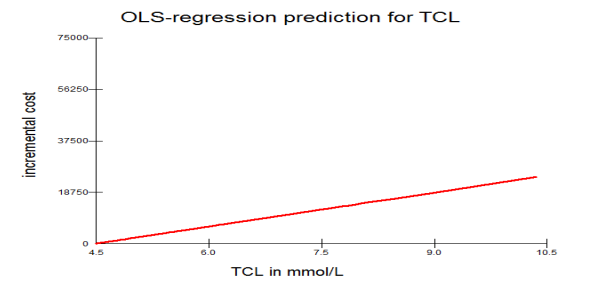
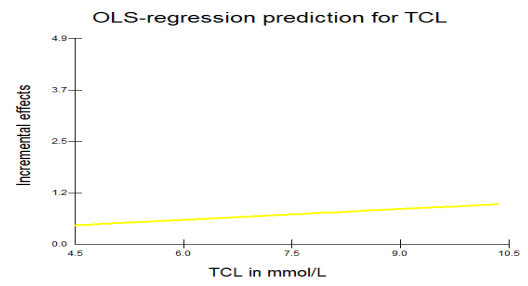
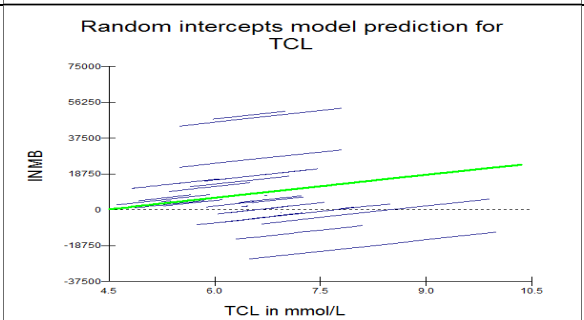
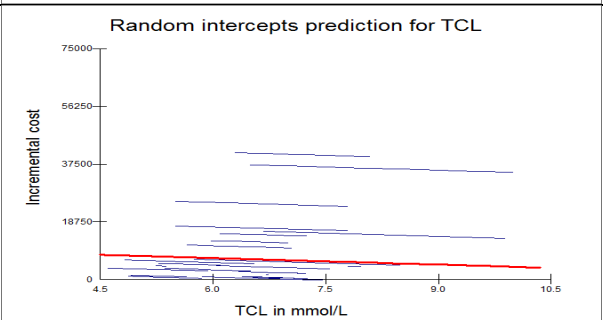
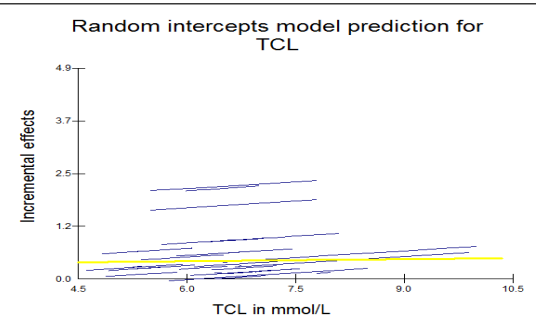
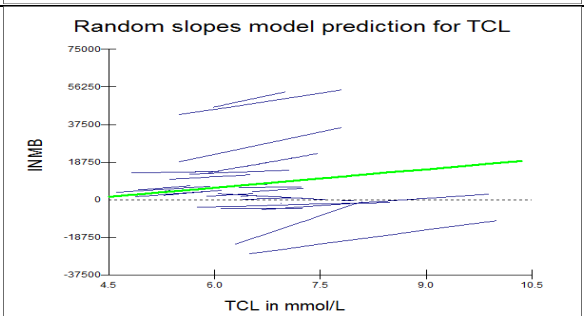
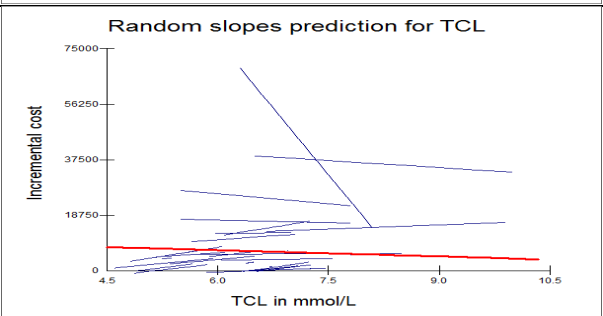
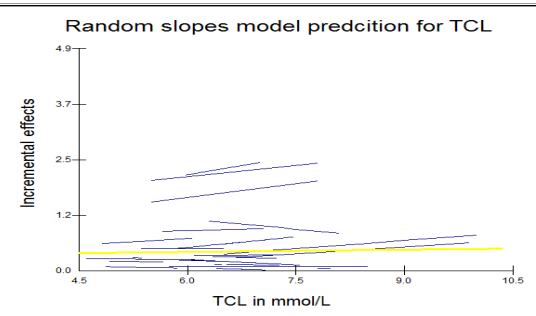
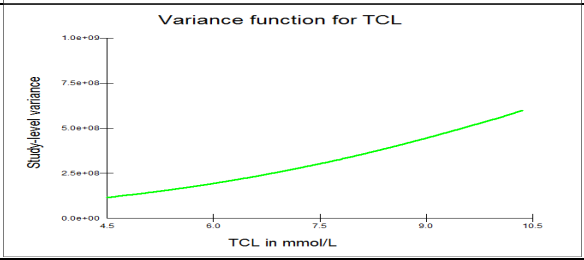
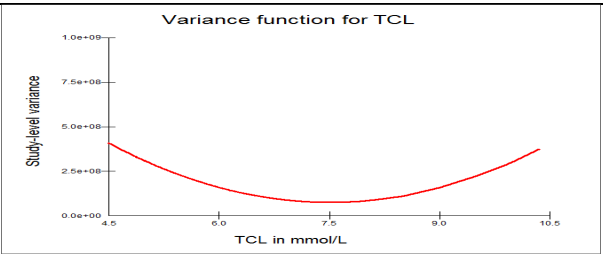
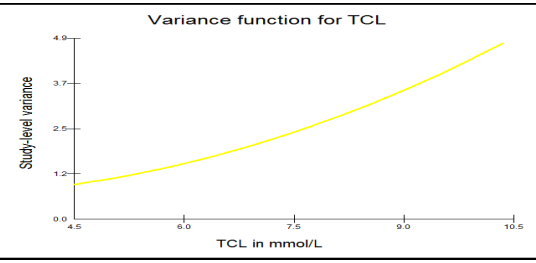
5.4.3. Results

Table 5.29 below summarizes the results from gradually building up a multilevel model with random slopes and modelling the study-level variance as a function of total cholesterol level both for INMBs in the univariate model and ΔC and ΔE in the bivariate framework. The first row of Table 5.29 shows the predictions

from running an OLS regression. Against prior expectations but in full accord with pilot study results reported in Chapter 3.4, a negative relationship is obtained between INMB and TCL and a positive relationship between ΔC and TCL. As elaborated on in Chapter 3.4, the result is likely to be explained by the failure of the OLS regression model to capture the difference between within and between study effects. Precisely, the relationship between INMB and TCL may be positive within each study in the dataset and this is what one would expect to observe when running a regression analysis on the respective dataset. However, some studies refer to patient populations which are characterised by a poorer health status, and this also includes an elevated cholesterol level. Though the positive relationship between cholesterol level and statin cost-effectiveness holds within these studies, their overall result in terms of INMB may be lower than that of studies with otherwise healthier patient populations. As the OLS regression model is unable to capture this between-study effect, it results in an overall negative relationship between TCL and INMB, and for the same reason, an overall positive relationship between TCL and ΔC respectively.

The random intercepts model, for which predictions are displayed in the second row of Table 5.29 below, takes into account within and between study effects and therefore shows the anticipated positive relationship between TCL and INMB as well as TCL and ΔE, and also the anticipated negative relationship between TCL and ΔC. The pooled regression lines are displayed in red for INMBs, green for ΔCs and yellow for ΔEs respectively. Around these pooled regression lines, we can see individual study predictions for each of the 67 studies included in the dataset. These predictions, however, are all parallel to the pooled regression line and therefore only differ by their respective intercepts, which also explains the term '*random intercepts model*' for this class of multilevel models (e.g. Steele, 2008). In other words, the relationship between TCL and the response variable (INMB, ΔC and ΔE) is assumed to be identical between all 67 studies, and this is also what has been assumed in Sections 5.2 and 5.3 of this empirical exercise.

Table 5.29: Gradually building up a multilevel model with random slopes and modelling study-level variance as a function of explanatory variables

	INMB (univariate model)	Incremental cost (bivariate model)	Incremental effects (bivariate model)
OLS-regression prediction			
Random intercepts model prediction			
Random slopes model prediction			
Variance function			

This assumption of zero variation in the relationship between TCL and the response variable between studies has been relaxed in the third row of Table 5.29. By allowing the slope of the data-level covariate TCL to differ between studies included in the dataset, we now obtain individual study predictions which are no longer parallel to each other (e.g. Steele, 2008). Some studies, for instance, show a steeper relationship between TCL and INMB than others, which results in a pattern of study-predictions which is '*fanning out*' for INMBs over the range of the explanatory variable. The same fanning out pattern is observed for the relationship between TCL and ΔE, whilst the pattern for the relationship between TCL and ΔC appears to be '*fanning in*' at first, but then fanning out again for higher values of TCL.

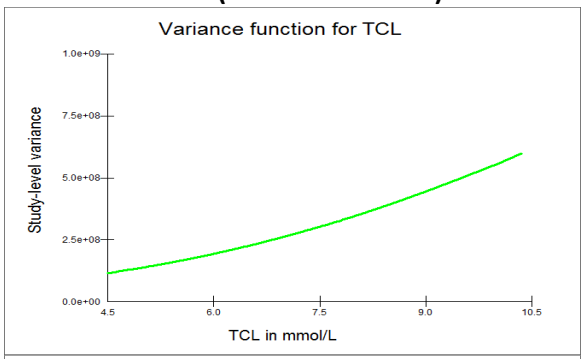
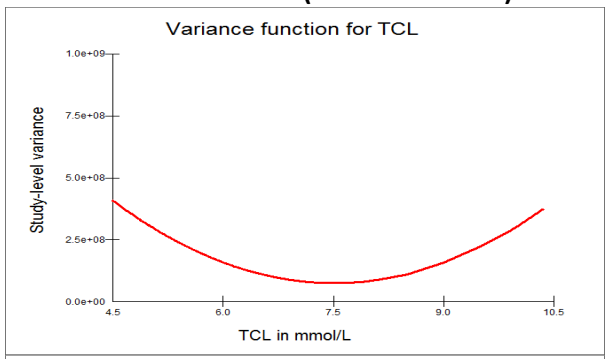
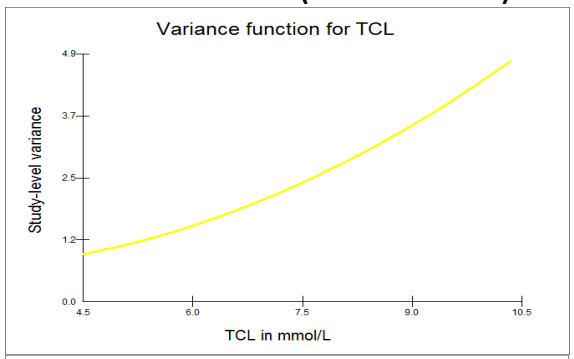
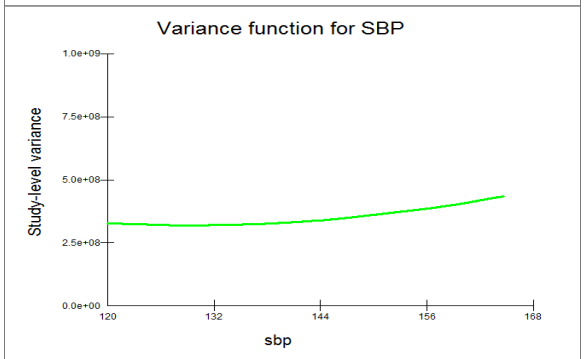
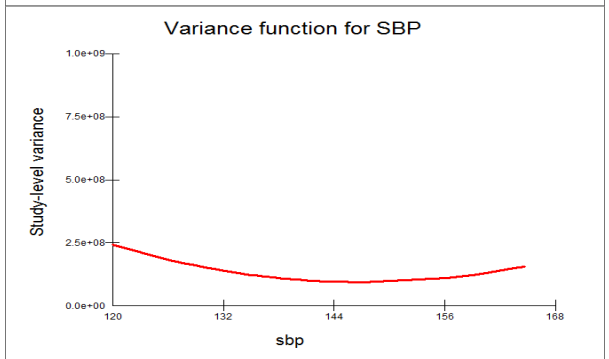
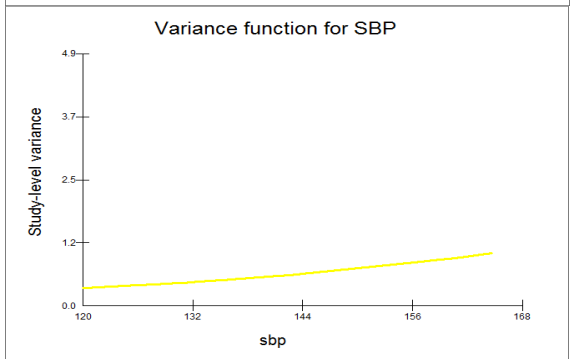
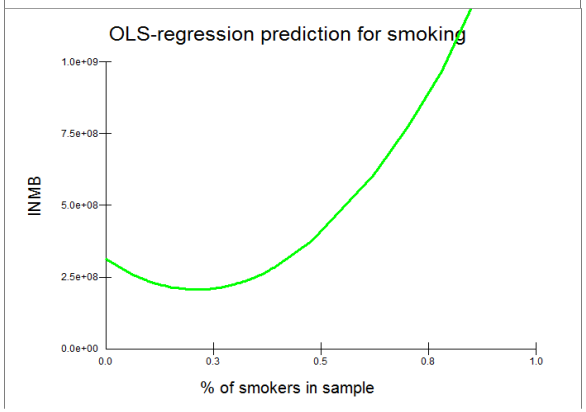
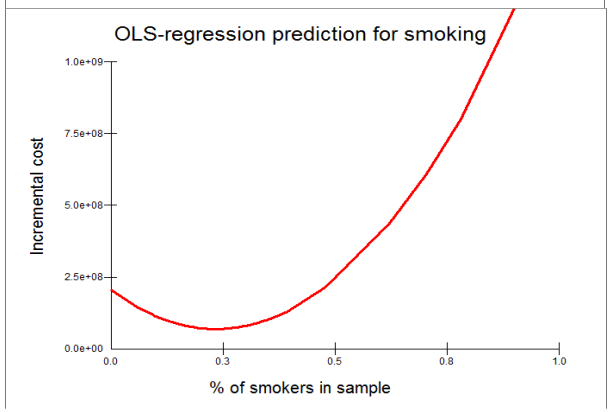
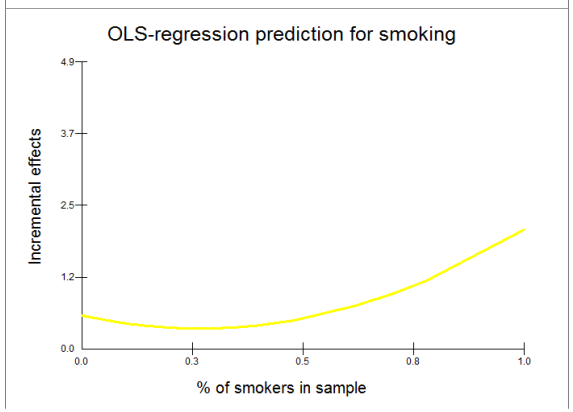
Random slopes have been fitted to economic evaluation datasets before, for instance by Sculpher et al. (2004), Manca et al. (2005), Thompson et al. (2006) or Bachmann et al. (2007). The covariate most commonly modelled with a random slope was the treatment effect across centres in multicentre studies, however, Thompson et al (2006) also assumed random slopes for the patient covariate '*incontinence status*' across centres in a multicentre observational study. The idea was analogous to what is being modelled in this exercise, namely to consider variation in patient characteristics between centres which may cause variability in cost-effectiveness data across centres participating in the study.

Most importantly for the purposes of this particular exercise, the variation between individual study predictions for the relationship between TCL and the response variable is no longer assumed to be zero, and this variation changes with the value of the explanatory variable. The study-level variance is now a quadratic function of TCL, and this has been plotted in the fourth row of Table 5.29 above. For INMBs and ΔEs, we observe constantly increasing study-level variation in the relationship between TCL and the response variable, and for ΔCs we can see a concave function where variation in the relationship between TCL and incremental cost decreases up to a TCL level of 7.5 mmol/L. Beyond that level of TCL, however, study-level variability is beginning to increase again.

Next, the same analysis was repeated for the continuous patient and disease characteristics systolic blood pressure (SBP) and the percentage of smokers in the sample (smoking). Respective variance functions (including the function for TCL which was already displayed in Table 5.29 above)) are shown in Table 5.30 below. To enhance comparability of results, functions for INMB and ΔC were all plotted on the same scale.

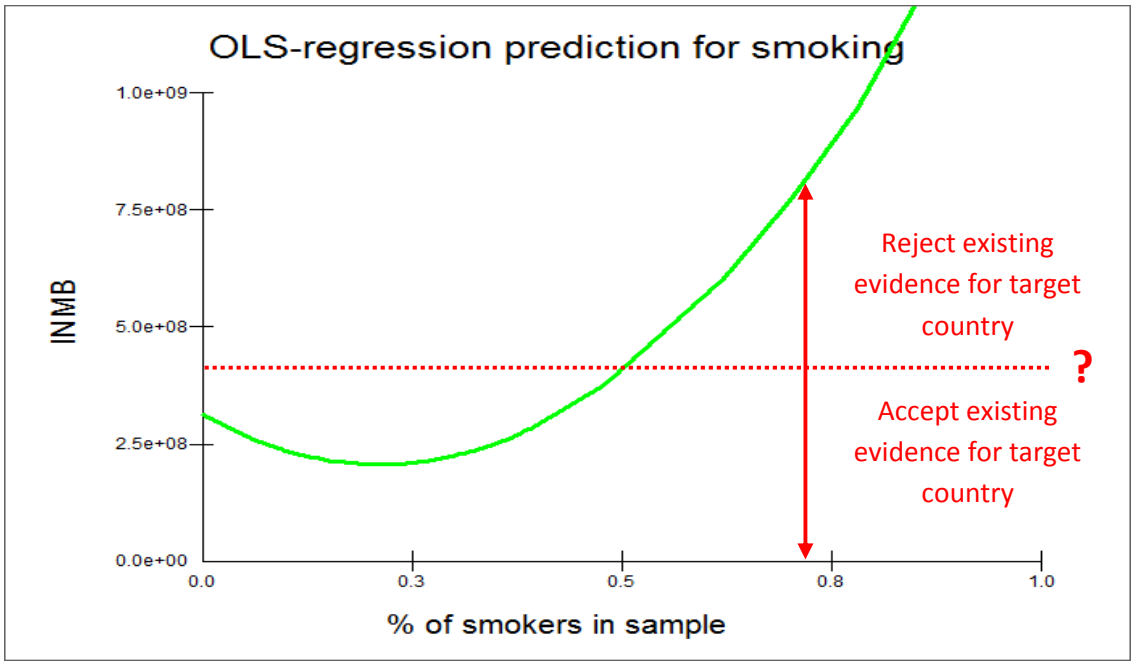
From Table 5.30 below we may draw several conclusions with respect to the transferability of measures of cost-effectiveness for the primary and secondary prevention of CVD. First of all, variability between studies is constantly increasing for the relationship between INMBs and TCL and ΔE's and TCL respectively, so that results may be less transferable the higher the total cholesterol level of the study population. The concave variance function for study-level variability in the relationship between TCL and ΔC indicates that results may be most transferable at a TCL level of around 7.5 mmol/L as variability between studies in measures of cost-effectiveness is lowest at this level of total cholesterol. Moving on to SBP, there are two important observations to make. First of all, variance functions are almost flat, indicating that variability in measures of cost-effectiveness between studies is pretty much the same over the whole range of values for SBP. However, as the scale used for plotting variance functions both for TCL and SBP are the same, we also observe that the variance function is at a much higher level for SBP. This demonstrates an important issue when using the variance function to make judgements regarding the transferability of results. One has to make a judgement regarding the '*tolerable level of variation*' to accept transferred evidence for the target domain. In other words, one has to decide on a '*threshold value*' for variability. Once the variance function exceeds this threshold, one may regard the existing evidence as non-transferable and therefore consider the conduct of a new study for the target country.

Table 5.30: Variance functions for TCL, SBP and smoking

	INMB (univariate model)	Incremental cost (bivariate model)	Incremental effects (bivariate model)
TCL			
SBP			
Smoking			

Finally, variance functions for smoking indicate that agreement between studies is highest at around 30% of smokers in the study sample. Beyond that, study-level variability increases drastically. However, the same problem as mentioned for systolic blood pressure applies, and this has been demonstrated in Figure 5.15 below. As already detailed for SBP, it is unclear where to place the horizontal line which represents the threshold for level-two variability. Several questions follow. First of all, is there a way to determine such a threshold value for study-level variability which may be helpful to guide the decision on whether or not to transfer existing evidence to the target country. Secondly, if this threshold value may not be determined, is the concept of the variance function still helpful to guide transferability decisions, for instance by comparing variance functions for several explanatory variables with respect to their relative location and shape and thereby determining research priorities for the target country. Finally, this exercise utilised data from published economic evaluation studies and integrates this data using MLM methodology. Hence, are there alternative application areas for the methodology demonstrated above, for instance using individual patient data from randomized controlled trials? These questions constitute important areas for future research and are addressed further in the discussion reported in Chapter 6.

Figure 5.15: Determining a threshold value for study-level variance



5.4.4. Summary and conclusions for part four of this empirical exercise

The primary aim of this section was to demonstrate a methodological feature of the MLM framework which allows modelling variation in international cost-effectiveness data directly as a function of explanatory variables. Starting off from an OLS regression equation, multilevel models were gradually built up both in the univariate and bivariate framework. Random slopes were then introduced for patient and disease characteristics (TCL, SBP and smoking), and slopes were allowed to vary randomly on study-level. This follows the thought that the decision makers choice with respect to the transferability problem is one between existing studies, and only once the decision maker has identified a number of candidate studies which meet the requirements of the target jurisdiction, country characteristics may be considered to the extent to which they are reflected in the economic evaluation in question. However, the analysis in this thesis has shown that the geographic context within which the available evidence was originally produced for is a far less important source of variability in international cost-effectiveness data than differences between economic evaluation studies. This, in conclusion, constitutes a compelling argument for fitting random slopes on study-level.

After random slopes were fitted to patient and disease characteristics, variation in cost-effectiveness data was plotted as a function of explanatory variables. It turned out, for instance, that variability between studies is constantly increasing for the relationship between INMBs and TCL and ΔE's and TCL respectively, so that results may be less transferable the higher the total cholesterol level of the study population. In addition, variance functions turned out to be almost flat for SBP, indicating that variability in measures of cost-effectiveness between studies is pretty much the same over the whole range of values for SBP. However, the overall level of variability in measures of cost-effectiveness with respect to SBP was much higher compared to other explanatory variables, so that it may be more indicated to focus research resources on this particular patient characteristic for the target domain. Finally, variance functions for smoking showed that study-level variability increases drastically with an increasing

percentage of smokers in the study sample, indicating that cost-effectiveness data for populations with a high percentage of smokers may not be transferable as evidence from existing studies is not conclusive.

Finally, a number of issues have been identified which ought to be addressed in future research and are therefore discussed in more detail in Chapter 6. For instance, is there a way to determine a '*threshold value'* for study-level variability which may be helpful to guide the decision on whether or not to transfer existing evidence to the target country. The discussion section addresses this question by drawing an analogy to the '*value of information'* concept (Claxton et al., 2000) Secondly, if there is no rational for determining a threshold value, could the concept of the variance function still be helpful to guide transferability decisions, for instance by comparing variance functions for several explanatory variables with respect to their relative location and shape, and thereby determining research priorities for the target country. Finally, are there additional application areas where modelling the variance function may be useful which go beyond the scope of the integration of secondary cost-effectiveness data from existing economic evaluation studies (e.g. applying this concept to individual patient data from international multicentre RCTs). These questions are addressed in more detail in the next Chapter 6, which entails the overall discussion of this thesis.

6. Discussion

This chapter provides a discussion of the results generated in this thesis. Findings across all chapters are summarised in Section 6.1, before strength and weaknesses of the work carried out are being discussed in Section 6.2. Section 6.3 then discusses findings in the context of existing research in the field. Policy implications are summarized in Section 6.4 and recommendations for further research are given in Section 6.5. Concluding remarks are provided in the final Section 6.6 of this chapter.

6.1. Summary of findings

The primary aim of this thesis was to address the transferability problem of economic evaluation in health by analysing what causes variability in measures of cost-effectiveness within and between studies, and ultimately, between geographic domains. The review of the economic evaluation literature in Chapter 2 confirmed that '*the methods that have been proposed to address the transferability issue have often been relatively ad hoc, with the obvious consequence that the methodological literature in this area has evolved somewhat nonlinearly over time*' (Manca, 2009). In addition, the available literature confirms that '*there is a lack of empirical studies which prevents stronger conclusions regarding which transferability factors are most important to consider and under which circumstances' (*Goeree et al., 2007). To address these issues, MLM was used to analyse variability factors for measures of cost-effectiveness.

In Chapter 2, the transferability problem was defined as an '*analogical inference*', within which a mapping of relevant attributes between a source domain, about which more is apparently known, and a, less studied, target domain is produced to infer whether the information of interest may also hold in the target setting

(Gentner & Markman, 1997; Forbus, 2001; French, 2002). This theoretical framework was then linked to the statistical concept of '*exchangeability*', which forms the conceptual basis of multilevel statistical modelling. MLM makes explicit the exchangeability assumption and allows for the assessment of variability factors of measures of cost-effectiveness within studies, between studies, and ultimately, between geographic domains through the assumption of conditional independence (Drummond et al., 2009). Chapter 2 concluded with a review of the use and applications of MLM within the area of economic evaluation in health, and showed that all applications of MLM in this area focus on the analysis of IPD from multicentre trials or observational studies, within which a strict two-level hierarchical data structure is commonly assumed.

In other words, using MLM as a mode for meta-regressing secondary cost-effectiveness data from published economic evaluation studies is a novelty in this area, which is why methods to integrate cost-effectiveness data from different studies and across different geographic domains were developed and tested in Chapter 3. Starting off from a simple OLS regression equation, complex data structures were gradually introduced, resulting in a number of strictly hierarchical as well as cross-classified models. Within these models, INMB as a single response variable was considered in a univariate framework, as well as the stochastic components of the INMB statistic ΔC and ΔE as a vector of response variables in a bivariate MLM. Chapter 3 concluded with a pilot study to test these models using a subset of cost-effectiveness data on statins for the primary and secondary prevention of CVD; as this is an extensively studied area which allows for the assumption of random parameters on study and country-level (Snijders, 2005). Results from the pilot study were promising, and in line with Gelman et al. (2004), who state that '*the valid concern is not about exchangeability, but encoding relevant knowledge as explanatory variables where possible.'* However, results also showed negligible country-level variation in the cross-classified model, and it was suspected that data from multinational studies may disguise actual country-level variability. This idea was supported by existing literature, which found lower levels of country-level variation in multinational study data (e.g. Barbieri et al., 2005), and also the invaluable feedback received from presenting pilot study results at various conferences and seminars.

After carrying out the pilot study, a systematic literature review was conducted to populate a dataset of cost-effectiveness estimates (ΔC, ΔE and INMBs), as well as additional data encoding potential variability factors. This was the focus of Chapter 4. As in the pilot study, statins for the primary and secondary prevention of CVD were the focus of assessment, and the dataset previously developed for the pilot was complemented with information from a large number of additional studies, as well as an extensive list of potential explanatory variables. In total, 67 studies were included in this empirical exercise, reporting 2094 cost-effectiveness estimates applicable to 23 geographic domains. Covariates were defined from a long list of potential variability factors, as previously reported in the literature (Sculpher et al., 2004; Goeree et al., 2007). Results of most studies referred to one geographic domain only, whilst six studies were multinational in nature, hence causing the strict hierarchical data structure with data clustered in studies and studies clustered in countries to break down.

Systematically reviewing this literature revealed that studies may be related to each other, for instance, through common authorship, the use of identical data sources, reuse of a previously published DAM, or simply a common source of funding. This, however, may violate the assumption of independence between studies, which is necessary to fit the MLMs developed in Chapter 3. If studies are related, it might therefore be more appropriate to pool them in one group on the study-level. Though this may reduce the number of level-two units, it may be more appropriate to fit MLMs to the data, especially when studies are very similar. Therefore, the aim of the final section of Chapter 4 was to look into the '*genealogy*' of economic evaluation studies on the cost-effectiveness of statins. Multiple correspondence analysis was used to ascertain whether studies are similar with respect to key characteristics, and once a '*phenotypic*' similarity was disclosed, a '*genotypic*' relationship between studies was sought. As a result, some relationships amongst studies were identified, however, the method did not (yet) prove accurate enough to justify the use of alternative MLM structures based upon its results. Instead, further explanatory variables were derived with the aim of encoding relationships between studies. This exploration into possible relationships between published evidence and potential similarities between studies' results may be followed up in future research.

Chapter 5 presented the main empirical analysis, and the first section of this chapter was predominantly concerned with determining the appropriate MLM structure for the integration of published cost-effectiveness data. Starting off with the models developed and tested in Chapter 3, Section 5.1 showed that appropriate assumptions regarding dependencies in the data are crucial for making correct inferences. The pilot study already showed the importance of explicitly acknowledging the existence of a study-level in the model, and this finding was clearly confirmed by the analysis reported in this section. However, the pilot study also showed negligible country-level variability in the cross-classified model, and as a potential reason, it was hypothesized that the assumption of independence between geographic domains may not be adequate for multinational study data. If data from multinational studies underestimates country-level variation, this potentially disguises country-level variability in the rest of the data. Therefore, a three-level hierarchical model, which groups data in studies and studies in countries, was run both on a reduced dataset without multinational studies, as well as on the full dataset, where data from multinational studies was grouped in a separate cluster on country-level. To confirm that the lack of country-level variability in the cross-classified model results from the data and not from technical issues with model specification and implementation, a cross-classified model was also run on an '*intermediate*' dataset, where some multinational study data was assigned to its respective target countries, and the rest of the data clustered in a separate group on country-level, thereby introducing the cross-classification problem on '*a lower scale'*.

As a result, whilst country-level variability was negligible for the cross-classified model, it increased dramatically within the bivariate three-level hierarchical framework, both for the full as well as for the reduced dataset. In addition, running the cross-classified model for the intermediate dataset resulted in country-level variability somewhere in between the fully cross-classified model and the three-level hierarchical structure. These results clearly confirm that country-level variability may be underestimated in multinational studies. Potential reasons could be identical trial protocols across centres and countries in multinational RCTs (e.g. Ramsey et al., 2005), or assuming transferability of input parameters between countries, which consequently results in lower

country-level variation in measures of cost-effectiveness (e.g. Barbieri et al., 2005). As a result, it was decided to take forward the three-level hierarchical model to the second part of this empirical exercise, as only this model makes appropriate assumptions regarding dependencies in the dataset, permits the simultaneous assessment of covariates on data, study, and country-level, and allows use of the full dataset by grouping multinational study data in a separate group on country-level.

Finally, Section 5.1 assessed, in depth, whether '*empirical Bayes shrinkage estimation*' may be regarded as appropriate in a model which integrates secondary data from published economic evaluation studies, where the weight of a particular study does not depend on the number of individual patients in the sample, but rather on the extent to which subgroup and sensitivity analyses have been reported. It could be shown that, due to high between-group variability, shrinkage factors are generally very high in this case study, which means that shrinkage is, at most, moderate. More importantly, however, this section argued that the impact of shrinkage on study means in this exercise depends not just on the number of data points abstracted from each study, but also on the within and between group variability in the data (i.e. dependencies) and the location of each study mean relative to the overall regression mean. This issue is further discussed in Section 6.2 of this chapter, which is concerned with the main strength and weaknesses of this empirical exercise.

Section 5.2 was concerned with testing covariates on data and study-level and determining a model which best controls for variability on these levels; thereby disclosing the maximum amount of country-level variability Covariates were drawn from a long list of possible variability factors suggested in the literature (Sculpher et al., 2004; Goeree et al., 2007). Following descriptive statistics, regression based imputation of missing values and a detailed assessment of correlations between potential explanatory variables were applied. Further, data reduction techniques were used to derive further covariates, including principal components factor analysis (Rencher, 2002; Acock, 2010) and multiple correspondence analysis (Le Roux & Rouanet, 2010). Thereafter, covariates were analysed in a random intercepts model (Steele, 2008; Rasbash et al., 2009, Hox,

2010) First, covariates were tested individually in the MLM. Secondly, a model with multiple covariates on data and study-level was constructed to reveal the maximum amount of country-level variability. This model was also carried forward to the analysis of country-level variability in Section 5.3 of the empirical chapter.

Adding covariates to the model on the data and study-levels successfully disclosed further variability on the country-level. Hence, the model could show that lower-level variability factors may feed through to higher levels, which demonstrates impressively the importance of both appropriately reflecting complex data structures and controlling for variability on lower levels, even if the main focus is on higher-level (i.e. country-level) variability (Sculpher et al., 2004). The analysis also showed that country-level variability was constantly increasing with the inclusion of lower-level covariates in the bivariate (ΔC, ΔE) model, allowing assessment of covariates on country-level. However, a different conclusion was reached for the univariate (INMB) framework, where country-level variability remained negligible throughout the course of this exercise. This finding was assessed in much more detail in Section 5.3 of the empirical chapter and is summarized further below.

In terms of covariates, the analysis carried out in Section 5.2 showed, for instance, the importance of patient and disease characteristics for variability in measures of cost-effectiveness, which also feeds through to the study and country-level. Further, testing intervention and comparator characteristics in the model showed that NICE's view on statins is essentially confirmed, which says that '*for the purposes of initiating therapy, there were no data on clinical events to suggest the superiority of any one statin over all the others in reducing cardiovascular events*' (NICE, 2006). On the other hand, controlling for study-level variability factors proved far more difficult, especially with respect to methodological characteristics; even though the study-level turned out to be a major source of variability in cost-effectiveness data. Only few variables on study-level were statistically significant, amongst them a binary which captures if effectiveness data was obtained from the 4S study, as well as a binary which captures whether the study uses the '*CVD life-expectancy model*' by Grover et al

(1998). Using 4S data significantly increased estimates of ΔE in the bivariate model and INMB in the univariate model, whilst papers by Grover et al., which all use the CVD life-expectancy model, showed higher levels in all outcome variables. The discussion in Section 6.2 elaborates in more detail why studies utilising 4S data or the CVD life-expectancy model provide results which are so different to other studies included in this empirical exercise.

Section 5.3 addressed variability in measures of cost-effectiveness between countries. The first objective was to analyse potential causes for a lack of country-level variability in the univariate MLM with INMB as single response variable. The second objective was to analyse covariates on country-level in the bivariate framework with ΔC and ΔE as response variables, within which considerably more country-level variability was disclosed in Sections 5.1 and 5.2.

Two potential reasons were considered for a lack of country-level variability in the univariate framework. First, the threshold value λ, which was set constant in previous analyses, was altered and models were run at different threshold values. For extreme values, this analysis showed similar country-level variability as for ΔC and ΔE in the bivariate model. This makes intuitive sense, as variability in INMBs stems from ΔC alone if λ equals zero, and ΔE alone if λ tends towards infinity. However, for the continuum between these extreme cases, variability on country-level remained negligibly low, so that we may exclude λ as a potential reason for a lack of country-level variability in the univariate model. Secondly, based on analyses carried out in Sections 5.1 and 5.2, it was suspected that variability in one component of the INMB statistic may be partly offset by variability in the other component. For this reason, forest plots with country means and their respective confidence intervals were produced for each response variable both from the raw data and using country-residuals from the three-level hierarchical model. In addition, Pearson correlations were calculated for the raw data as well as the country means obtained from the three-level MLM. Forest plots showed similar patterns in ΔC and ΔE and correlations were well above 0.9 and highly significant. It may hence be concluded that the lack of country-level variability in INMBs results from combining ΔC and ΔE, which have similar patterns of variability, so that variability in one component of the INMB

statistic is partly offset by variability in the other component. Potential reasons for this observation are considered further below.

Finally, this section was also concerned with country-level covariates for the bivariate model. To test whether *'lumping'* country-level data from multinational studies in a separate group on country-level distorts results, this analysis was performed both with the full dataset, and a reduced dataset in which data points from multinational studies were dropped. Results were almost identical for both datasets. Testing country-level covariates individually in the model did not result in any significant covariates for ΔC, whilst government expenditure on health as a percentage of total health expenditure, the number of GPs and pharmacists per 10,000 inhabitants, the mean population age, as well as measures of BMI and blood glucose showed significant relationships to ΔE. When covariates were simultaneously assessed in the final random intercepts model with variability factors considered on each hierarchical level, GDP per capita and the government expenditure on health as a percentage of total health expenditure was significantly related to incremental cost, whilst coefficients for total health expenditure as a percentage of GDP and mean population age were highly significant for ΔE. These results are also discussed below.

Section 5.4 shifted the focus towards a methodological feature of the MLM framework. The primary aim was to demonstrate how modelling variation in international cost-effectiveness data directly as a function of explanatory variables relates to the transferability problem. First, random slopes were modelled for patient and disease characteristics (TCL, SBP and smoking). Then, variation in cost-effectiveness data was plotted as a function of explanatory variables. It turned out, for instance, that variability between studies is constantly increasing for the relationship between INMBs and TCL and ΔE's and TCL respectively, so that results may be less transferable the higher the total cholesterol level of the study population. In addition, variance functions were almost flat for SBP, indicating that variability in measures of cost-effectiveness between studies is pretty much the same over the whole range of values for SBP. However, the overall level of variability in measures of cost-effectiveness with respect to SBP was much higher compared to other explanatory variables, so

that it may be more indicated to focus research resources for the target domain on this particular patient characteristic. Finally, variance functions for smoking showed that study-level variability increases drastically with an increasing percentage of smokers in the study sample, indicating that cost-effectiveness data for populations with a high percentage of smokers may not be transferable as evidence from existing studies shows high variability.

The following section discusses the strength and weaknesses of the empirical analysis. Thereafter, findings are discussed in the context of existing research in the field in Section 6.3, before summarizing policy implications in Section 6.4. Finally, Section 6.5 discusses potential areas for further research, before a conclusion is provided in Section 6.6.

6.2. Strength and weaknesses of the analysis

Probably the biggest strength of the empirical analysis also constitutes its main weakness, namely the use of secondary cost-effectiveness data from published economic evaluation studies. On the one hand, this ensured quick access to the data required to carry out this research. More importantly, however, secondary data from published studies may be more appropriate for analysing variability factors than the use of IPD from multinational trials. Trials usually implement strict protocols which may be identical across centres and countries (e.g. Ramsey et al., 2005). These protocols, though crucial to ensure the internal validity of trial results, artificially reduce the variability which may exist between centres and countries under real world conditions (Ramsey et al., 2005; Barbieri et al., 2005). Secondary data from published economic evaluation studies, which are more likely to have been designed to inform decisions under real world conditions, may better reflect this variability. Furthermore, using IPD from one trial may not allow for the assessment of factors causing variability in cost-effectiveness between studies, which has been identified as an overriding source of variability within this project.

However, several problems arise when using secondary data from published economic evaluation studies in a multilevel statistical analysis of factors causing variability in measures of cost-effectiveness. For instance, the assumption of random parameters on study and country level, which is necessary to fit multilevel models, may be violated for the current exercise. On study level, strict inclusion and exclusion criteria specified, for instance, that only studies which explicitly report values for components of the INMB statistic may be considered. However, as earlier studies may be more likely to report their results in terms of ICERs (without making explicit values for ΔE and ΔC), this may bias the dataset towards studies which were published more recently. This, in turn, may affect results as evidence suggests that not just study methods evolved, but statins also became cheaper and more effective over time. It may also affect the variability observed between studies, as more recent studies may build upon experiences (and results) of earlier studies, which may lead to a converging effect in terms of variability over time.

Methodological standards may be more widely adapted in some countries compared with others and therefore favour the inclusion of studies from particular countries into this exercise. This may bias the dataset towards particular jurisdictions, (e.g. Canada, UK). On the other hand, it was not possible to acquire data for countries of low or medium levels of economic attainment, with the exception of Hungary. As a result, the low country-level variability observed in the empirical analysis may partly be a result of the fact that only high income countries were represented in the data. Finally, issues with the assumption of randomness may not only relate to higher level units. Problems did arise, for instance, if a study reported cost-effectiveness as ICERS and data on either ΔC or ΔE was available, which allowed decomposing the ICER-statistic. If, however, the use of statins resulted in cost-savings, studies rightly omitted resulting negative ICERs, which meant that ΔC and ΔE could not be calculated and the respective data point had to be dropped. This resulted in a loss of data points providing evidence in strong support of statins, which in turn, may have biased response variables downwards in the current exercise.

Apart from issues with the assumption of randomness, the problem of '*empirical Bayes shrinkage estimation*' was identified as a major cause for concern within the MLM framework if data stems from published economic evaluation studies. Shrinkage may only be a consequence of differences in the way subgroup or sensitivity analyses were performed and reported, which is not indicative of the size of the underlying study sample. This problem was considered and extensively discussed in Section 5.1.5 of the empirical chapter. In brief, due to high between study variability, shrinkage factors are generally high in this exercise, which means that shrinkage is, at most, moderate. More importantly, however, the impact of shrinkage on study means depends not just on the number of data points per study, but also the within and between study variability in the data (i.e. dependencies) and the location of study means relative to the overall regression mean. With respect to '*dependency*', shrinkage makes sense as one would rather drag studies towards each other if between-study variability is low. With respect to location, shrinkage may also be justified as one would not want outlying studies to bias regression results. For '*group size'*, however, a distinction was made between country groups and study groups. With respect to countries, giving higher gravity to countries for which more evidence is available, may also be justified, even if the underlying data stems from published economic evaluation studies. With respect to studies, there may be arguments both in favour as well as against the appropriateness of shrinkage and it is recommended to further look into methodologies for assigning weights to studies and to implement this information within the MLM framework.

Despite these issues, the use of MLM is considered a major strength of this thesis. MLM makes explicit the exchangeability assumption, which '*mediates*' between assumptions of either identical or independent parameters (e.g. Spiegelhalter et al., 2000 & 2004), and allows integrating secondary cost-effectiveness data from different studies and different countries without ignoring that study and country residuals are not independent. Further, through the assumption of conditional independence, one may assess the impact of variability factors modelled as covariates on each level of the data hierarchy. As Gelman et al. (2004) put it *'in this way exchangeable models become almost universally applicable, because any information to distinguish different units*

should be encoded.' (Gelman et al., 2004 cited from Manca et al., 2007). As multilevel models treat country-parameters as randomly drawn from a common, prior distribution, we have to ask ourselves, however, about the potential consequences of violations to this assumption within the current exercise. On study-level and country-level, bias may occur if some studies did have a greater chance to be selected into the sample than others, and potential reasons for this to happen have been considered further above. However, unlike OLS regression, which treats β coefficient as if they were fixed constants, the MLM approach controls for systematic differences between higher level units if they are reflected in the dataset. And even if these differences are not reflected in the data collected (as it is, for instance, the case with the level of economic attainment on country-level) the MLM approach ensures at least that parameter estimates do not suffer from wrongly estimated precision (i.e. suggesting a relationship, when in fact, there is none).

In this respect, the analysis carried out in this thesis goes far beyond the work of other researchers in the field. In particular, Barbieri et al. (2005) shared a similar aim and the authors claim to have provided '*the most comprehensive analysis to date of the variation in the results of cost-effectiveness studies, of drugs, in Europe'*. Nevertheless, their approach was based on a rather descriptive analysis of relevant studies with the major limitation that all variability in cost-effectiveness data was assigned to the country-level. This may be misleading as the results of this project show that differences within and between studies account for most of the variability in measures of cost-effectiveness. Furthermore, Barbieri et al. (2005) were not able to control for a number of variability factors simultaneously. However, there may be complicated interactions between different variability factors, which require careful consideration and systematic assessment. MLM allows simultaneous assessment of variability factors within studies, between studies and between countries, and therefore provides a more appropriate methodological framework for the analysis of factors causing variability in cost-effectiveness data.

Further to that, MLMs were also specified in a bivariate framework, hence allowing for the simultaneous assessment of ΔC and ΔE as a vector of response

variables (Bartholomew et al., 2008; Nixon et al., 2005; Pinto et al., 2005; Manca et al., 2007; Grieve et al., 2007; Bachmann et al., 2007; Willan et al., 2008; Grieve et al., 2010). This has several advantages: First, the approach does not require a new regression to be estimated for each value of the cost-effectiveness threshold (Manca et al., 2007). Secondly, the correlation between the two stochastic components of the INMB statistic is explicitly modelled. Finally, the bivariate model allows assessing the differential impact of covariates on each response variable whilst acknowledging that ΔC and ΔE are, themselves, correlated.

Finally, the method applied in this thesis allows explicit modelling of variability in measures of cost-effectiveness as a function of explanatory variables (e.g. Rasbash et al., 2009). Allowing slopes of covariates to vary randomly in the MLM framework is something other health economists have implemented before within their applications of MLM to IPD from multinational trials or observational studies (Sculpher et al., 2004; Manca et al., 2005; Thompson et al., 2006; Bachmann et al., 2007). However, doing so also allows modelling variation in the response variable as a function of explanatory variables and this is, to the knowledge of the author, something which has neither been considered in the area of health economics in general, nor in health economic evaluation in particular. The MLM literature refers to this concept as the '*variance function*' (e.g. Steele, 2008; Rasbash et al., 2009) and Section 5.4 of the empirical chapter showed how the variance function relates to the transferability problem. A number of questions were phrased with respect to the variance function, and these are discussed further as potential areas for future research in Section 6.5 of this chapter.

Moving on from the methodological framework of MLM, a further strength of the analysis is that explanatory variables were systematically derived from a list of factors previously suggested in the literature as possible constraints on the transferability of cost-effectiveness evidence (Sculpher et al., 2004; Goeree et al., 2007). This is believed to be the first attempt of a systematic assessment of such variability factors. Other researchers have tested selected covariates on patient, centre or country-level within a two-level hierarchical model applied to IPD (e.g.

Grieve et al., 2005 & 2007; Thompson et al., 2006; Manca et al., 2007; Petrinco et al., 2009; Edbrooke et al., 2011), and only one study (Barbieri et al., 2005) aimed to analyse country-level variability in cost-effectiveness data. However, all of these studies limited their analysis to a small number of selected variables. The research carried out within this thesis, however, began with a long list of candidates ascertained from the literature (Sculpher et al., 2004; Goeree et al., 2007) to develop and test a data abstraction form to populate a dataset for the empirical analysis. This resulted in an extensive list of covariates on data, study and country-level, and a random intercepts model with multiple covariates on each hierarchical level was specified which best controls for variability in cost-effectiveness data for statins in the primary and secondary prevention of CVD.

Before moving on to a discussion of the findings generated in this project, it should be highlighted that the empirical analysis was based on data collected for one intervention area only. The intervention of statins for the primary and secondary prevention of CVD was chosen as it has been extensively researched in the past, meaning that a sufficient number of includable studies and countries was hypothesized to be present to justify the assumption of random parameters on study and country-level. This resulted in a large dataset with a 2094 data points referring to 67 studies and 23 geographic domains. Nevertheless, all results are intervention specific, and though it was not feasible to carry out another case study within this project, this may be regarded as a general weakness of this thesis. Furthermore, all countries represented in the dataset have high levels of economic attainment and even variation in health related indicators was generally low. The implication is that the quantitative results should be interpreted with caution. Section 6.4 below will provide examples of how the results of this thesis may be used in a policy environment and elaborate on the potential pitfalls in doing so. However, the generalisability of findings beyond statins in particular and pharmacological interventions more generally should be established in future analysis through replication in other intervention areas and a wider range of countries at different levels of economic attainment. The generalisability of the method may be limited to those intervention areas which received similar attention than statins in the primary and secondary prevention of CVD.

6.3. Discussion of findings

6.3.1. Variability in measures of cost-effectiveness on data, study and country-level

Probably one of the most important findings of both the pilot study and the empirical analysis is low country-level variation across all models and subsets of the data analysed. Or put differently, whilst country-level variation was present, its proportion compared to variability within and between studies was low for ΔC and ΔE in the bivariate model, and negligible for INMB in the univariate model. A number of potential reasons for this observation were considered throughout the analysis carried out within this project.

First of all, a negligible country VPC in the cross-classified MLMs relates to inappropriate assumptions about the independence of data from multinational studies on country-level. Due to much lower country-level variability in data from multinational studies, this data should not be treated as independent on country-level. Rather, pooling it in a separate group on country-level, thereby removing its disguising effect from other country-parameters, disclosed further country-level variability in the bivariate model. In 2005, Barbieri et al. reached a similar conclusion, stating that '*it can be seen that the extent of variability is lower for multicounty studies than single country studies'* and '*this may be because, in a multicounty study, the analysts give more active consideration to the harmonization of data and analyses*.'

Related to that, two of the six multinational studies are primary modelling studies, i.e. directly assessing the cost-effectiveness of statins using IPD from multinational RCTs. The remaining four studies are secondary modelling studies, employing DAM to estimate the cost-effectiveness of statins. With respect to primary modelling studies (Johannesson et al., 1996; Jonnsson et al, 1999), the lack of country-level variation may relate to a widely discussed trade-off between the internal and external validity of trial evidence. For instance, Baltussen et al. stated already in 1999 that '*policy-makers need cost-effectiveness information that is both internally and externally valid. The latter aspect is often*

ignored and refers to the relevance of the results of economic trials to the specific decision-making context of the policy-maker'. Increasing internal validity in a trial may, however, come at the cost of lower external validity for each target jurisdiction. For instance, stringent inclusion and exclusion criteria, which are identical across centres, increase the internal validity of results - but treatment populations may not reflect target populations in any of the trial countries (Marshal & Hux, 2009). Likewise, standardised trial protocols minimise the potential for bias. However, variability with respect to differences in clinical practice between countries is literally non-existent in the data. As a result, a multinational RCT '*doesn't have any value or existence except in the degree to which it captures the reality as observed in the original setting'* (Sleigh, 1997).

Further to that, two of the six multinational studies in question (Johannesson et al., 1996; Jonnsson et al., 1999) achieve '*context specificity'* of their results only through the use of country-specific unit cost data, whilst all other input parameters are identical across the jurisdictions included. Three studies (Grover et al., 2001; Szucs et al., 2004; Taylor et al., 2009) vary both unit cost and resource use data to achieve context-specificity, whilst one study (Lindgren et al., 2007) applies context-specific unit cost and utility values, whilst resource use data and effectiveness data was transferred from a different country. In conclusion, as Barbieri et al (2005) already emphasized, this practice may lead to differing degrees of country-level variability within the affected data. As a result, decision makers ought to be very critical about the applicability of multinational study data to their particular context, even if their country is explicitly considered in the study. A careful consideration of each aspect of the study which potentially impacts on the context-specificity of its results may be indicated.

Another reason for low country-level variation, especially in the univariate model, relates to the INMB statistic itself. This exercise showed that country-level variability is much higher in ΔC and ΔE than it is in the INMB statistic. Section 5.3 provides an explanation for this observation, as variability for both ΔC and ΔE share a similar pattern, which means that variability in one component of the INMB statistic is partly '*being offset'* by variability in the other component,

hence leading to reduced country-level variability in INMBs. The author is not aware of studies which hypothesized or empirically looked into this issue, which is why, at this point, we can only guess about potential causes for this '*similar pattern'* in the country-level variability of ΔC and ΔE. A potential reason may be a structural relationship between the two components of the INMB statistic. CVD prevention with statins may lead to an increased life expectancy, which, in turn, may expand the period within which statins are consumed. Hence, higher ΔE may naturally be associated with higher ΔC, which would explain a common pattern in their variability. As a result, variability in one component may partly be offset by variability in the other component of the INMB statistic. Future research should assess whether this common pattern in the components of the INMB statistic is specifically related to statins, or more generally present in pharmacological interventions, preventative interventions or even other intervention areas.

If the proportion of country-level variability is low in all models run within the empirical analysis, this also means that the proportion of variability in cost-effectiveness data due to within and between-study differences must be high. Hence, another key message from the empirical analysis is that, for statins in the primary and secondary prevention of CVD, differences within studies (e.g. relating to subgroup and sensitivity analyses) and differences between studies (e.g. methodological study characteristics), are far more important sources of variability in published cost-effectiveness data than differences between the countries reflected in that data. However, whilst the analysis in Section 5.2 of this empirical chapter was very successful in controlling for within-study variability, it was generally difficult to ascertain significant covariates on study-level.

6.3.2. Variability factors on data-level

Starting off with data-level covariates, patient and disease characteristics turned out to be key variability factors, feeding through to variability between studies and countries in the dataset as previously hypothesized by Sculpher et al. (2004).

MLM provides an excellent framework to make explicit the impact of patient and disease characteristics on study and country-level variability, as it assigns respective variance partitions to each hierarchical level modelled (e.g. Steele, 2008). Further, testing intervention and comparator characteristics in the model showed that this is also an important source of variability on data-level, especially with respect to annual drug cost. However, on the effectiveness side, NICE's view on statins was essentially confirmed, which says that '*for the purposes of initiating therapy, there were no data on clinical events to suggest the superiority of any one statin over all the others in reducing cardiovascular events*' (NICE, 2006).

Some data-level covariates may also be discussed in the light of the little existing empirical evidence on variability factors for measures of cost-effectiveness. Barbieri et al. (2005) explored issues surrounding variability of economic evaluations of pharmaceuticals in Western Europe and concluded that '*the extent of variation across countries in effectiveness, resource use or unit costs, allowed by the researcher's chosen methodology was the most important variability factor for measures of cost-effectiveness*.' In other words, the more studies utilize context-specific data, the higher the variability in measures of cost-effectiveness between countries. To confirm this hypothesis, categorical variables were developed and tested which classify data points with respect to their degree of '*context specificity*' of input parameters. Descriptive statistics showed that, within this exercise, by far the most common way to populate a model was to use country specific unit cost and resource use data, but effectiveness data and utility weights from other geographic domains (type CR: 1033 data points, 49.33%). Only 56 data points (2.67%) were generated with a model fully populated with target specific data. Within the final random intercepts model specified with covariates on data and study-level, coefficients of this categorical variable were significant at least at the 5%-level, negative and constantly decreasing for ΔC in the bivariate model. This indicates lower incremental cost with higher context specificity.

Hence, results of this thesis essentially confirm the conclusions of Barbieri et al (2005). However, the analysis within this thesis goes much further. The impact of

'*context Specificity*' was actually quantified after controlling for a number of potential confounders. This means that, after controlling for differences in study populations, intervention cost and methods on data-level, significant coefficients could be ascertained for context specificity. Secondly, by decomposing the INMB statistic, it could be shown that the impact of context specificity is strongest for ΔC, whilst coefficients were not significant for ΔE or INMBs. It may be the case that transferring effectiveness data is much more common than transferring economic data, so that there is simply no systematic country variation on the effectiveness side of the INMB statistic. This may then feed through to INMBs, so that the univariate model also fails to show significant relationships between context specificity and measures of cost-effectiveness. Finally, whilst the analysis of Barbieri et al. (2005) relied on a rather descriptive analysis of ICERS within multinational studies, the analysis within this thesis utilised data from both single country and multinational studies within a quantitative analysis accounting for the hierarchical structure of the underlying data, confounding effects of other explanatory variables in the model, and the fact that multinational study-data does not justify the assumption of independence on country-level, thereby producing much more reliable results.

6.3.3. Variability factors on study-level

Moving on to the study-level, the '*appropriate set of covariates'* proved to be a far more difficult to determine. Out of a large number of candidate variables, there were only few significant results: a binary capturing whether the study used the '*CHD life expectancy model'* by Grover et al. (1998); a binary capturing whether effectiveness data was elicited from the 4S study; and a categorical encoding the scope of assessment (i.e. whether studies looked into CHD only, CHD and stroke, or CHD, stroke and PAD). Whilst it makes intuitive sense that the scope of assessment alters cost-effectiveness results, it may be indicated to look into potential reasons why 4S and studies by Grover et al. resulted in significant coefficients.

With respect to studies published by Grover et al., features of the CVD life-expectancy model as well as other study characteristics were analysed to elicit potential explanations for elevated cost-effectiveness results. One reason may be that the CVD life-expectancy model combines a number of features which all turned out to be positively related to measures of cost-effectiveness when individually tested in the MLM. In particular, the CVD life-expectancy model is a Markov model and the authors estimated treatment effectiveness by modelling the effect of statin treatment on cholesterol reduction and multivariate regression was used to estimate the effect of cholesterol reduction on CVD related mortality (e.g. Grover et al., 1998). More importantly, however, when validating their model, Grover et al (1998) compare the difference in CHD and stroke related mortality between intervention and control arms predicted by the model with the mortality observed in a number of clinical trials. Whilst predicted values were almost all within the range of CIs generated from trial data, these CIs were large and almost all of the predicted values exceeded observed trial results. Hence, the CVD life-expectancy model may overestimate the impact of statin treatment on CVD related mortality, which may lead to higher ΔC, ΔE and INMBs compared to other studies.

Moving on to 4S data, existing meta-analyses on the effectiveness of statins in the primary and secondary prevention of CVD do not provide an indication of higher effectiveness estimates. (Ward et al., 2007). In other words, whilst 4S data favours treatment over control, estimates were well within the range of other statin trials (RR for 4S: 0.71, 95%CI: 0.59-0.85 versus all placebo controlled trials in Ward et al., 2007, RR: 0.84, 95%CI: 0.78-0.9). Hence, there must be other reasons for elevated ΔC observed in the bivariate model for 4S related data. One reason could relate to the prevention category, as 4S relates to secondary prevention for which statin effectiveness is well established. However, respective covariates have been tested extensively within this exercise so that this should have been picked up before. In addition, other data points within the analysis also refer to secondary prevention, whilst they elicited effectiveness data from other sources (e.g. CARE, LIPID or PLAC I & II). In conclusion, it was not possible to identify potential causes for elevated estimates of ΔE for 4S in the bivariate model. It may hence be indicated to further assess differences between

economic evaluation studies relying on 4S and those which elicited effectiveness data from other sources to find out what causes this significant difference in ΔE.

A number of covariates on study-level expected to be related to variability in measures of cost-effectiveness were not included in the final model. For instance, industry funding was positively related to INMBs at the 5% level, whilst coefficients were not significant for ΔC or ΔE in the bivariate model. A positive relationship between industry funding and the cost-effectiveness of health technologies has previously been discovered by Miners et al. (2005). The authors undertook a retrospective pairwise comparison of evidence submitted to NICE's technology appraisal programme by manufacturers of the relevant health technologies and by contracted university based assessment groups and conclude that *'the estimated incremental cost-effectiveness ratios submitted by manufacturers were on average significantly lower than those submitted by the assessment groups. These results show that an important role of NICE's appraisal committee, and of decision makers in general, is to determine which economic evaluations, or parts of evaluations, should be given more credence'.*

One variability factor on study-level, which has been previously identified by Barbieri et al (2005), was not significantly related to measures of cost-effectiveness within the empirical analysis of this project. Precisely, Barbieri et al (2005) showed that the general study design (i.e. trial based or model based) explains some of the variability in measures of cost-effectiveness across countries. This was not confirmed within the current analysis. Further to that, results were especially disappointing with respect to the QHES instrument. Considerable effort was undertaken to operationalize this instrument and apply this quality checklist to all 67 studies included in this dataset. Though bivariate statistics showed statistically significant negative relationships to both components of the INMB statistic, results were very sensitive to small variations in the way QHES was operationalized for this empirical exercise, so that results may be interpreted with highest caution.

6.3.4. Variability factors on country-level

As mentioned, country-level variability was negligible in the univariate framework, and though considerably higher in the bivariate model, the country VPC did not exceed 21% for ΔC and 15% for ΔE respectively. Nevertheless, a number of country-level covariates were tested in the bivariate model, and very small but statistically significant negative relationship between ΔC and GDP per capita as well as between ΔC and government spending as a percentage of total health expenditure were found. Likewise, a number of explanatory variables showed significant positive relationships with ΔE, amongst them total health expenditure as a percentage of GDP per capita, the number of GPs and pharmacists per 10,000 population, as well as population age, BMI and blood glucose levels. Though coefficients of country-covariates were mostly moderately significant and in accord with prior expectations, results should be interpreted with some caution. As mentioned, the country VPC was low, even for ΔC and ΔE in the bivariate model. This may not accord findings of Barbieri et al. (2005) who found considerable variation between countries. However, in their study, the authors did not assign variation in measures of cost-effectiveness to their respective levels (i.e. data, study, and country) and therefore, assign all variability observed to the country-level. This exercise applied a more sophisticated method using MLM, which allows partitioning the variation in the data, thereby making explicit how much variability refers to differences within studies, between studies and between geographic domains.

Barbieri et al. (2005) conclude that '*differences in cost-effectiveness results between countries are not systematic'*. Though only few significant country-coefficients were found in this thesis and country-level variability was generally low, a different conclusion has been reached within this thesis. The observation of small country-coefficients may be explained by the fact that countries included in this empirical exercise were quite similar with respect to the country-characteristics tested in the model. In particular, only developed countries were considered, showing mostly similar levels of economic attainment and variation in health related indicators is also generally low. In addition, unlike single level OLS regression models, MLMs ensure that standard errors for higher-level

covariates are not underestimated (Steele, 2008; Rasbash et al., 2009; Hox, 2010). Hence, if country-level covariates still turn out to be significant in the three-level hierarchical model, this provides a strong indication that there may be systematic differences between countries. In conclusion, whilst differences within and between studies constitute the overriding source of variability in measures of cost-effectiveness, this project did show that there may be small but statistically significant systematic differences in measures of cost-effectiveness between countries. Further research may replicate this analysis within other intervention areas which also allows the inclusion of a wider range of countries at different levels of economic attainment.

6.4. Key policy implications

A number of researchers concerned with the transferability of economic evaluation results believe that '*the economic question of whether an activity adds more to well-being than the alternative uses of the same resources in a particular community cannot be answered by reference to the costs and consequences of the same activity in a different community (Birch & Gafni, 2003).* Other researchers argue that economic evaluation is transferable if '*(a) potential users can assess their applicability to their setting and (b) they are applicable to that setting' (e.g. Späth et al., 1999; Boulenger et al., 2005)*

Nevertheless, as Birch and Gafni further state, *the validity of the method of valuation cannot be established independent of the setting in which it is to be used*' and even if the methods of valuation are valid in each setting, the authors pose the question of whether this implies that '*numbers produced by application of the methods are generalisable across individuals and settings*.' Birch & Gafni (2003) conclude that the '*generalisability of the validity of a method of valuation does not imply generalisability of the resulting valuations*' A similar opinion is being shared by Vale (2010).

However, one may argue that findings of this thesis mediate between these extreme viewpoints; just like the exchangeability assumption, which provides the theoretical foundation for MLM, mediates between the opposing assumptions of either identical or independent study and country parameters in a multilevel model. In line with Drummond et al (2009), this research builds up from the believe that the transfer of evidence from one country to another may hold if *'the analyst has identified the appropriate set of covariates for the exchangeability assumption to hold; and the characteristics of the country of interest are represented appropriately by countries in the dataset'* (Drummond et al., 2009).

The question is then how the results generated within the empirical exercise may be used and interpreted for policy purposes. Having quantified variability factors for measures of cost-effectiveness, one may use the bivariate random intercepts model with covariates on data, study and country level developed in chapter 5.3 to predict incremental cost and incremental effects for countries for which data is missing. For instance, for a male patient within the age cohort of 56 to 65, with a mean TCL of 6.5, HDL of 1.02, SBP of 133 and a history of CVD, predicting ΔC and ΔE for a country with a mean population age of 39, an annual GDP per capita of £25.000, 75% government expenditure on health as a percentage on total health expenditure and 11% of total health expenditure as percentage of GDP, results in 1.63 ΔE and £6,727 ΔC respectively. This result refers to either LYS or QALYS (as the outcome measure was not a significant variability factor for statins) and further assumes annual drug cost for statins of £520, a lifetime horizon and a health insurance perspective, and that statins are compared to 'doing nothing' at a 3.5% discount rate on incremental effects. Having predicted both components of the INMB statistic for the country of interest, we can now calculate the cost-effectiveness for statins for the above specified patient cohort in this country, resulting in an ICER of £4,128, or an INMB of £42,165 at a WTP threshold for a unit of health gain of £30,000. (Likewise, for a WTP threshold of say 15.000, the respective INMB would be £17.723) Hence, in the absence of location specific data, prediction results support the use of statins for the specified patient cohort in the country of interest.

However, these results should be interpreted with caution. Most importantly, immense uncertainty results from high variability in measures of cost-effectiveness observed in the data. This uncertainty may be quantified for the prediction results by applying statistical simulation methods which allow computation of the predicted values, their standard deviation as well as confidence intervals around predicted values (Taghreed et al., 2003). However, additional work would be required to allow performing such a simulation exercise within the bivariate MLM to obtain relevant parameters for predicting uncertainty. In this context, the concept of the variance function may be considered as in interesting alternative. The MLM framework allows modelling variability in measures of cost-effectiveness explicitly as a function of covariates, and chapter 5.4 introduced examples of plotting variability in INMB, ΔC and ΔE as a function of TCL, SBP, smoking status and diabetes status. It could be shown, for instance, that variability in measures of cost effectiveness increases with increasing blood pressure, and this could be utilised to focus research resources on further work for high blood pressure patients. The variance function will also be addressed as a potential area for future research below.

Nevertheless, variability in the existing data is vast, and this exercise showed the difficulties of finding the "appropriate set of covariates" for the exchangeability assumption to hold. This proved to be a major challenge in this empirical exercise, especially with respect to differences between studies, which account for a large proportion of the overall variability in international cost-effectiveness data. In addition, potential violations of the MLM assumptions, for instance with respect to random parameters, or the appropriateness of shrinkage estimation within secondary data analysis, may lead to biased coefficients, which may ultimately lead to biased prediction results for ΔC and ΔE if one was using the models generated within this exercise for prediction. The policy implication which follows with respect to using the models developed for prediction is that after careful consideration of factors potentially causing variability in cost-effectiveness data, one may use prediction results to inform decisions in countries for which this data is missing. However, because of the immense variability in the data, the problem of defining the 'appropriate set of covariates' and additional uncertainty from transferring data to locations for which it was not originally produced, decision makers ought to be very critical about the

information provided from prediction. Finally, the aim of this thesis was not to provide a tool for prediction, but – more fundamentally - to assess potential variability factors for cost-effectiveness data and to provide a method with which this could be done using secondary data from published economic evaluation studies.

When considering transferability problems in practice, decision makers may have to choose from a number of existing economic evaluation studies applicable to different geographic domains. This choice should be based on the '*degree of similarity*' between study characteristics and the target location, which accords the principles of analogical reasoning as outlined in Chapter 2. To assess the extent to which available studies meet the requirements of the target jurisdiction, decision makers ought to compare attributes of the studies available with attributes of the target country, and decide which of the available international cost-effectiveness studies may be most appropriate to inform decisions in the target country. In short, the choice within the existing data is, first of all, a choice between existing studies, not countries. Only once the decision maker has identified a number of candidate studies which meet the minimum requirements of the target jurisdiction, country characteristics may be considered to the extent to which they are reflected in the economic evaluation in question. However, the analysis within the empirical exercise of this thesis showed that the proportion of country-level variability is low, even after controlling for a large number of potential variability factors on data and study-level. Hence, the geographic context within which the available evidence was originally produced turned out to be a less important source of variability in international cost-effectiveness data than differences between economic evaluation studies. This constitutes a key take home of this thesis, though it needs to be emphasized that the generalizability of these results should be evaluated in future research.

Testing the generalisabilty of the results outside of statins could be done by replicating this analysis for different interventions to test the findings of this thesis, and a number of recommendations may simplify such a potential future exercise. For instance, a future study could, on the basis of this thesis, focus only on bivariate multilevel models for secondary data integration which cluster data in studies and countries, either with or without considering cross-classified data

structures. Secondly, populating a dataset may be simplified. For instance, experiences from this thesis may be used to more rapidly develop an appropriate data abstraction form for a similar exercise. However, this thesis also showed that such an exercise crucially depends on a certain degree of consistency across studies with respect to the reporting of economic evaluation data. Improving consistency would also improve the transparency and comparability of economic evaluation results in general, which is why it is strongly recommended to further develop and improve the acceptance of international standards for conducting and reporting economic evaluation studies in health.

6.5. Areas for future research

Probably one of the most interesting methodological areas for future research relates to the use of the variance function to express variability in measures of cost-effectiveness as a function of explanatory variables. In Section 5.4 of the empirical chapter, random slopes were fitted to the model, and variation in cost-effectiveness data was plotted as a function of explanatory variables. It turned out, for instance, that variability between studies is constantly increasing for the relationship between INMBs and TCL and ΔE's and TCL, so that results may be less transferable from the existing data to out of sample locations the higher the TCL of the target population. In addition, variance functions were almost flat for SBP, though the overall level of variability was much higher, suggesting to focus research resources on this patient characteristic in the target domain. In conclusion, for ranges of values of explanatory variables where variation in cost-effectiveness data is high, uncertainty increases when transferring cost-effectiveness results, which also increases the value of new evidence for the target country

A number of issues for future research have been identified with respect to the variance function. For instance, is there a way to determine a '*threshold value*' for variability in cost-effectiveness data which may guide decisions on whether or not to transfer existing evidence to the target country? To address this issue, it may be helpful to draw an analogy to the '*value of information*' concept.

Precisely, Claxton et al. (2000) state that '*the expected costs of uncertainty can also be interpreted as the expected value of perfect information (EVPI) since perfect information (an infinite sample) can eliminate the possibility of making the wrong decision. It is also the maximum a decision maker should be willing to pay for additional evidence to inform this decision in the future. If the EVPI exceeds the expected costs of additional research then it is potentially cost-effective to acquire more information by conducting additional research'.* In the context of transferability of economic evaluation data, one may interpret higher variability in measures of cost-effectiveness as higher uncertainty for the target setting. In other words, if variability in measures of cost-effectiveness in the existing evidence increases, so does the expected cost of uncertainty within the target country. Hence, one may transfer evidence for ranges of explanatory variables for which variability in measures of cost-effectiveness is low, and consider additional research for ranges of explanatory variables which show high variability in cost-effectiveness data. However, how to assign monetary values to different levels of variability, or how else to determine a threshold value for variability to decide upon the transfer of evidence to the target setting, are future research questions.

In addition to that, one may also consider the application of the variance function in different research contexts, for instance within economic evaluation using IPD collected alongside multinational multicentre studies. A number of researchers already considered two-level hierarchical models for the analysis of multicentre trial or observational data (Sculpher et al., 2004; Grieve et al., 2005; Willan et al., 2005; Manca et al., 2005; Nixon et al., 2005; Pinto et al, 2005; Thompson et al., 2006; Manca et al., 2007; Grieve et al., 2007; Bachmann et al., 2007; Coupe et al., 2007; Willan et al., 2008; Petrinco et al., 2009; Grieve et al., 2010; Edbrooke et al., 2011). Some of these applications also allowed for random slopes of explanatory variables (Sculpher et al., 2004; Manca et al., 2005; Thompson et al., 2006; Bachmann et al., 2007). However, none of the existing studies modelled the variance function to show how variability in measures of cost-effectiveness changes as a function of explanatory variables. This, however, could be very informative within the context of a multicentre study, as it makes explicit how variability changes between centres, or countries, with respect to explanatory variables. In conclusion, heterogeneity between patients, centres

and countries with respect to explanatory variables could be made explicit through the use of the variance function within the MLM framework, and it is suggested to address this issue in future research.

Another area for future research has been identified when systematically reviewing the economic evaluation literature on statins in the primary and secondary prevention of CVD. Studies within this area seemed to be strongly related through common authorship, identical data sources, re-use of a previously published DAM, etc. As such relationships may cast into question the independence assumption of data on study-level, a '*genealogy study*' was conducted which has been reported in Section 4.3. This exercise used MCA to explore '*phenotypic similarities*' between published economic evaluation studies, and subsequently, it was aimed to identify '*genotypic relationships*' which may explain the similarities across studies. The initial aim of this exercise was to confirm the appropriate MLM structure and to make sensible assumptions regarding dependencies of data on study-level. However, though the method was not (yet) accurate enough to base alternative MLM structures upon its findings, the method did show some potential so that future research should look into the use of MCA within the context of study-genealogy.

Two technical issues for further research were identified with respect to MLM. First, the issue of shrinkage has been discussed earlier within the context of integrating secondary data from published economic evaluation studies. Within this discussion, the use of MLwiN's weighting facility was mentioned. The idea is that information from particular studies could be regarded as being '*oversampled*' in the current dataset as it stems from studies where subgroup and sensitivity analyses were carried out more extensively. Hence, one may argue that out of the space of cost-effectiveness estimates, some estimates did have a greater '*probability*' of being selected into the sample than others. Without weighting, however, the model assumes that each data point did have the same chance of being selected into the sample (CMM, 2011). As MLwiNs weighting facility is currently not available within MCMC estimation, it is suggested to pick up on the issue of weighting to 'counteract' shrinkage when using MLM for secondary data integration in future research.

The second technical issue within the MLM framework to be addressed in future research relates to multilevel multiple imputation of missing data. It has been previously pointed out that an imputation model '*must have the right variance structure*'. Hence, '*if a dataset is multilevel, then the imputation model must be multilevel too*' (Carpenter & Goldstein). Therefore, a number of software tools were considered which may allow multilevel multiple imputation, for instance, MI macros for MLwiN (Carpenter & Goldstein) or Realcom IMPUTE (Goldstein, 2009). However, none of these tools were capable of dealing with more complicated data structures like the bivariate three-level model. This is an area of on-going research, and it would also constitute an interesting area for future research which goes beyond the scope of this project.

Finally, with respect to the empirical findings, the overriding priority for future research is, as mentioned above, to replicate this work within other intervention areas to test the robustness of its results. As previously highlighted, the empirical analysis was based on data collected for one intervention area only. Hence, results are intervention specific. Further to that, all countries represented in the dataset are developed countries, showing mostly similar levels of economic attainment and even variation in health related indicators was generally low. In conclusion, it is strongly recommended to replicate this research in other intervention areas to test its key findings. For instance, is there is the same '*disguising effect*' of multinational study data on variability in measures of cost-effectiveness between countries in other intervention areas? Secondly, how does variability in measures of cost-effectiveness spread across different levels of the data hierarchy. In particular, does the finding that variability in measures of cost-effectiveness is primarily due to differences within and between studies, not countries, also hold for other intervention areas? Third, is there a similar pattern in country-level variability in ΔC and ΔE which reduces country-level variation in INMBs in other disease areas, e.g. outside the area of disease prevention, or pharmacological interventions? If so, what may be potential reasons for such a similar pattern in variability within the components of the INMB statistic. Finally, what are the most important variability factors for measures of cost-effectiveness in different intervention areas, and to which level do they belong. All these questions should be addressed when replicating this research within other intervention areas in future research.

6.6. Concluding remarks

The aim of this thesis was to explore the transferability of economic evaluation results produced for one geographic area to another location of interest. Multilevel statistical models were developed for the integration of published international cost-effectiveness data to assess the impact of contextual effects on country-level; whilst controlling for baseline characteristics within, and across, a set of economic evaluation studies. Explanatory variables were derived from a list of factors suggested in the literature as possible constraints on the transferability of cost-effectiveness evidence. The approach was illustrated using published estimates of the cost-effectiveness of statins for the primary and secondary prevention of cardiovascular disease Results show that the proportion of variation at the country-level observed depends on the appropriate multilevel model structure and never exceeds 15% for incremental effects and 21% for incremental cost respectively. Key sources of variability are patient and disease characteristics, intervention cost and a number of methodological characteristics defined on the data-level. There were fewer significant covariates on the study and country-levels. The findings of the empirical work carried out within this thesis suggest that variability in cost-effectiveness data is primarily due to differences between studies, not countries. Further, comparing different models suggests that data from multinational studies severely underestimates country-level variability. Additional research is needed to test the robustness of these conclusions on other sets of cost-effectiveness data, to further explore the appropriate set of covariates, and to foster the development of multilevel statistical modelling for economic evaluation data in health.

References

ACOCK, A.C., 2010. *A gentle introduction to Stata.* 3 edn. Stata Press; Texas, USA.

ACOCK, A.C., 2005. Working with missing values. *Journal of Marriage and Family,* 67, pp. 1012-1028.

ALEMDAR, M., 2009. *A monte carlo study: the impact of missing data in cross-classification random effects models.* Department of Educational Policy Studies, Georgia State University, USA.

ALONSO, R., FERNANDEZ DE BOBADILLA, J., MENDEZ, I., LAZARO, P., MATA, N. & MATA, P., 2008. Cost-effectiveness of managing familial hypercholesterolemia using atorvastatin-based preventive therapy. *Rev. Esp. Cardiol.,* 61(4), pp. 382-393.

ANDERSON, K.M., ODELL, P.M., WILSON, P.W. & KANNEL, W.B., 1991. Cardiovascular disease risk profiles. *American Heart Jounal,* 121, pp. 293-298.

ANNEMANS, L., MARBAIX, S., WEBB, K., VAN GAAL, L. & SCHEEN, A., 2010. Cost effectiveness of atorvastatin in patients with type 2 diabetes mellitus A pharmacoeconomic analysis of the collaborative atorvastatin diabetes study in the Belgian population. *Clinical Drug Investigation,* 30(2), pp. 133-142.

ANTONANZAS, F., RODRIGUEZ-IBEAS, R., JUAREZ, C., HUTTER, F., LORENTE, R., PINILLOS, M., 2009. Transferability indices for health economic evaluations: methods and applications. *Health Economics,* 18, pp. 629-643.

ARA, R., PANDOR, A., STEVENS, J., REES, A. & RAFIA, R., 2009. Early high-dose lipid-lowering therapy to avoid cardiac events: a systematic review and economic evaluation. *Health Technology Assessment (Winchester, England),* 13(34), pp. 1-74.

ARAUJO, D.V., BAHIA, L., SOUZA, C.P.R. & PAVÃ£O, A.L.B., 2007. Cost-effectiveness and budget impact analysis of rosuvastatin and atorvastatin for LDL-cholesterol and cardiovascular events lowering within the SUS scenario. *International Journal of Atherosclerosis,* 2(3), pp. 189-194.

ASHRAF, T., HAY, J.W., PITT, B., WITTELS, E., CROUSE, J., DAVIDSON, M., FURBERG, C.D. & RADICAN, L., 1996. Cost-effectiveness of pravastatin in secondary prevention of coronary artery disease. *American Journal of Cardiology,* 78(4), pp. 409-414.

BACHMANN, M.O., FAIRALL, L., CLARK, A. & MUGFORD, M., 2007. Methods for analyzing cost effectiveness data from cluster randomized trials. *Cost Effectiveness and Resource Allocation,* 5(12).

BALTUSSEN, R., LEIDL, R. & AMENT, A., 1999. Real world designs in economic evaluation. Bridging the gap between clinical research and policy-making. *PharmacoEconomics,* 16(5), pp. 449-458.

BARBIERI, M., DRUMMOND, M., WILLKE, R., CHANCELLOR, J., JOLAIN, B. & TOWSE, A., 2005. Variability of cost-effectiveness estimates for pharmaceuticals in Western Europe: lessons for inferring generalisability. *Value in Health,* 8(1), pp. 10-23.

BARTHOLOMEW, D.J., STEELE, F., MOUSTAKI, I. & GALBRAITH, J., 2008. *Analysis of multivariate social sciences data (2nd edn)*. Chapman & Hall/CRC, Tailor & Francis Group; Boca Raton.

BENZERCI, J.P. et al., 1972. *L'Analyse des Donnees. Vol.1: Taxinomie. Vol.2: Analyse des correspondences (Data analysis. Vol.1: Taxinomy. Vol.2: Correspondence Analysis);* Paris, France.

BERGER, K., KLOSE, G., SZUCS, T.D., 1997. Economic aspects of drug therapy exemplified by pravastatin. A socioeconomic analysis of cholesterol synthase enzyme inhibition in coronary heart disease patients. *Medizinische Klinik,* 92(6), pp. 363-369.

BERNARDO, J.M. & SMITH, A.F.M., 1994. *Bayesian theory.* John Wiley & Sons, LTD; Chichester, UK

BICKEL, R., 2007. *Multilevel analysis for applied research - it's just regressio*n! The Guilford Press; New York, USA

BIRCH, S. & GAFNI, A., 2003. Economics and the evaluation of health care programmes: Generalisability of methods and implications for generalisability of results. *Health Policy,* 64(2), pp. 207-219.

BLAUG, M., 1980. *The Methodology of Economics – or how economists explain.* Cambridge University Press; New York, USA

BOULENGER, S., NIXON, J., DRUMMOND, M., ULMANN, P., RICE, S. & DE POUVOURVILLE, G., 2005. Can economic evaluations be made more transferable? *European Journal of Health Economics,* 6(4), pp. 334-336.

BRADFORD, R.H., SHEAR, C.L., CHREMOS, A.N., DUJOVNE, C., FRANKLIN, F.A., HESNEY, M., HIGGINS, J., LANGENDÖRFER, A., POOL, J.L., SCHNAOER, H. & STEPHENSON, W.P., 1990. Expanded clinical evaluation of lovastatin (EXCEL) study: design and patient characteristics of a double-blind, placebo-controlled study in patients with moderate hypercholesterolemia. *The American Journal of Cardiology,* 66(8), pp. B44-B55.

BRIGGS, A., CLARK, T., WOLSTENHOLME, J. & CLARKE, P., 2002. Missing.... presumed at random: cost-analysis of incomplete data. *Health Economics,* 12(5), pp. 377-392.

BRIGGS, A. & FENN, P., 1998. Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane. *Health Economics,* 7, pp. 723-740.

BROWNE, W.J., 2012. *MCMC estimation in MLwiN v.2.25.* Centre for Multilevel Modelling (CMM), University of Bristol; Bristol, UK:

BURGESS, J.F., CHRISTIANSEN, C.L., MICHALAK, S.E. & MORRIS, C.N., 2000. Medical profiling: improving standards and risk adjustments using hierarchical models. *Journal of Health Economics,* 19, pp. 291-309.

CAREY, K., 2000. A multilevel modelling approach to analysis of patient costs under managed care. *Health Economics,* 9(5), pp. 435-446.

CARO, J., KLITTICH, W., MCGUIRE, A., FORD, I., NORRIE, J., PETTITT, D., MCMURRAY, J. & SHEPHERD, J., 1997. The West of Scotland coronary prevention study: economic benefit analysis of primary prevention with pravastatin. *BMJ,* 315(7122), pp. 1577-1582.

CARO, J.J., HUYBRECHTS, K.F., KLITTICH, W.S., JACKSON, J.D. & MCGUIRE, A., 2003. Allocating funds for cardiovascular disease prevention in light of the NCEP ATP III guidelines. *American Journal of Managed Care,* 9(7), pp. 477-489.

CARPENTER, J.R. & GOLDSTEIN, H., *Multiple Imputation in MLwiN.* ESRC Research Methodology Programme Grant H 333250047.

CARPENTER, J.R. & KENWARD, M.G., 2008. *Missing data in randomised controlled trials - a practical guide.* National Institute for Health Research, Publication RM03/JH17/MK; Brimingham, UK

CDC DIABETES COST-EFFECTIVENESS Group, 2002. Cost-effectiveness of intensive glycemic control, intensified hypertension control, and serum cholesterol level reduction for type 2 diabetes. *Journal of the American Medical Association (JAMA),* 287(19), pp. 2542-2551.

CENTRE FOR MULTILEVEL MODELLING (CMM), 2011. *Weighting in MLwiN.* Centre for Multilevel Modelling, Bristol, UK

CENTRE FOR MULTILEVEL MODELLING (CMM), Introduction to multilevel modelling using MLwiN - workshop presentation / random intercepts models. Available at: http://www.bristol.ac.uk/cmm/software/support/workshops/materials/randomintweb.pdf [May, 06, 2012].

CENTRE FOR MULTILEVEL MODELLING (CMM), Introduction to multilevel modelling using MLwiN - workshop presentation / random slopes models. Available at: http://www.bristol.ac.uk/cmm/software/support/workshops/materials/randomslopeweb.pdf; [May, 06, 2012].

CENTRE FOR MULTILEVEL MODELLING (CMM), Introduction to multilevel modelling using MLwiN - workshop presentation / variance components models. Available at: http://www.bristol.ac.uk/cmm/software/support/workshops/materials/multilevel-m.html [May, 06, 2012].

CHAN, P.S., NALLAMOTHU, B.K., GURM, H.S., HAYWARD, R.A. & VIJAN, S., 2007. Incremental benefit and cost-effectiveness of high-dose statin therapy in high-risk patients with coronary artery disease. *Circulation,* 115(18), pp. 2398-2409.

CHAU, J., CHEUNG, B.M., MCGHEE, S.M., LAUDER, I.J., LAU, C.P. & KUMANA, C.R., 2001. Cost-effectiveness analysis of applying the Cholesterol and Recurrent Events (CARE) study protocol in Hong Kong. *Hong Kong Medical Journal,* 7(4), pp. 360-368.

CLAXTON, K., NEUMANN, P.J., ARAKI, S.S. & WEINSTEIN, M.C., 2000. *Bayesian value of information analysis: an application to a policy model of Alzheimer's disease*. University of York, Department of Economics and Related Studies; York, UK

CLAXTON, K. & POSNETT, J., 1996. An economic approach to clinical trial design and research priority-setting. *Health Economics,* 5(6), pp. 513-521.

COOPER, A., NHERERA, L., CALVERT, N., O'FLYNN, N., TURNBULL, N., ROBSON, J., CAMOSSO-STEFINOVIC, J., RULE, C., BROWNE, N., RITCHIE, G., STOKES, T., MANNAN, R., BRINDLE, P., GILL, P., GUJRAL, R., HOGG, M., MARSHALL, T., MINHAS, R., PAVITT, L., RECKLESS, J., RUTHERFORD, A., THOROGOOD, M. & WOOD, D., 2008. *Clinical guidelines and evidence review for lipid modification: cardiovascular risk assessment and the primary and secondary prevention of cardiovascular disease.* National Collaborating Centre for Primary Care and Royal College of General Practitioners; London, UK

COUPÉ, V.M., VEENHOF, C., VAN TULDER, M.W., DEKKER, J., BIJLSMA, J.W., & VAN DEN ENDE, C.H., 2007. The cost effectiveness of behavioral graded activity in patients with osteoarthritis of hip and/or knee. *Annals of the Rheumatic Diseases,* 66(2), pp. 215-221.

COX, N.R., 1974. Estimation of the correlation between a continuous and a discrete variable. *Biometrics,* 30, pp. 171-178.

DAVIES, A., HUTTON, J., O'DONNELL, J. & KINGSLAKE, S., 2006. Cost-effectiveness of rosuvastatin, atorvastatin, simvastatin, pravastatin and fluvastatin for the primary prevention of CHD in the UK. *British Journal of Cardiology,* 13(3), pp. 196-202.

DRUMMOND, M.F., MCGUIRE ,A. & FLETCHER, A., 1993. *Economic evaluation of drug therapy for hypercholesterolemia in the UK. CHE discussion paper No 104. University of York, UK*

DRUMMOND, M.F. & JEFFERSON, T.O., 1996. Guidelines for authors and peer reviewers of economic submissions to the BMJ. *British Medical Journal,* 313, pp. 275.

DRUMMOND, M.F., SCULPHER, M.J., TORRANCE, G.W., O'BRIEN, B.J. & STODDART, G.L., 2005a. *Methods for the economic evaluation of health care programmes.* 3rd edn. Oxford University Press.

DRUMMOND, M., BARBIERI, M., COOK, J., GLICK, H.A., LIS, J., MALIK, F., REED, S.D., RUTTEN, F., SCULPHER, M. & SEVERENS, J., 2009. Transferability of economic evaluations across jurisdictions: ISPOR good research practices task force report. *Value in Health,* 12(4), pp. 409-418.

DRUMMOND, M., MANCA, A. & SCULPHER, M., 2005. Increasing the generalisability of economic evaluations: recommendations for the design, analysis, and reporting of studies. *International Journal of Technology Assessment in Health Care,* 21(2), pp. 165-171.

DRUMMOND, M. & PANG, F., 2001. Transferability of economic evaluation results. In: *Economic Evaluation in Health Care: Merging Theory with Practice,* pp. 256-276. Oxford University Press, Oxford, UK

DRUMMOND, M.F., BLOOM, B.S., CARRIN, G., HILIMAN, A.L., HUTCHINGS, H.C., KNILL-JONES, R.P., DE POUVOURVILLE, G. & TORFS, K., 1992. Issues in the cross-national assessment of health technology. *International Journal of Technology Assessment in Health Care,* 8(4), pp. 671-682.

DUNCAN, C., JONES, K. & MOON, G., 1996. Health-related behaviour in context: a multilevel modelling approach. *Social Science and Medicine,* 42(6), pp. 817-830.

DUNCAN, C., JONES, K. & MOON, G., 1998. Context, composition and heterogeneity: using multilevel models in health research. *Social Science and Medicine,* 46(1), pp. 97-117.

EDBROOKE, D.L., MINELLI, C., MILLS, G.H., IAPICHINO, G., PEZZI, A., CORBELLA, D., JACOBS, P., LIPPERT, A., WIIS, J., PESENTI, A., PATRONITI, N., PIRRACCHIO, R., PAYEN, D., GURMAN, G., BAKKER, J., KESECIOGLU, J., HARGREAVES, C., COHEN, S.L., BARAS, M.,ARTIGAS, A. & SPRUNG, C.L., 2011. Implications of ICU triage decisions on patient mortality: a cost-effectiveness analysis. *Critical Care,* 15(1):R56

ESSERS, B.A.B., SEFERINE, S.C., TJAN-HEIJNEN, V.C.G., SEVERENS, J.L., NOVAK, A., POMPEN, M., ORON, U.H. & JOORE, M.A., 2010. Transferability of model-based economic evaluations: the case of trastuzumab for the adjuvant treatment of HER2-positive early breast cancer in the Netherlands. *Value in Health,* 13, pp. 375-380.

EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT (EUnetHTA) WP4 core model. Available: http://www.eunethta.eu/Public/About_EUnetHTA/Tools/HTA-Core-Model/ [June 02, 2012].

EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT (EUnetHTA), 2008. *HTA core model for medical and surgical interventions - Work Package 4 - The HTA Core Model.* v. 1.0. FinOHTA, Finnish Office for HTA, Finland

EVERS, S., GOOSSENS, M., DE VET, H., VAN TULDER, M. & AMENT, A., 2005. Criteria list for assessment of methodological quality of economic evaluations: concensus on health economic criteria. *International Journal of Technology Assessment in Health Care,* 21(2), pp. 240-245.

FIELDING, A. & PILLINIGER, R., 2008. Introduction to quantitative data analysis. *LEMMA - Learning Environment for Multilevel Methods and Applications.* Centre for Multilevel Modelling; Bristol, UK.

FORBUS, K.D., 2001. Exploring analogy in the large. In: D. GENTNER, K.J. HOLYOAK & B. KOKINOV, eds., *The Analogical Mind: Perspectives from Cognitive Science.* Massachusetts Institute of Technology (MIT), USA

FRANCO, O.H., DER KINDEREN, A.J., DE LAET, C., PEETERS, A. & BONNEUX, L., 2007. Primary prevention of cardiovascular disease: cost-effectiveness comparison. *International Journal of Technology Assessment in Health Care,* 23(1), pp. 71-79.

FRANCO, O.H., PEETERS, A., LOOMAN, C.W. & BONNEUX, L., 2005. Cost effectiveness of statins in coronary heart disease. *Journal of Epidemiology & Community Health,* 59(11), pp. 927-933.

FRENCH, R.M., 2002. The computational modeling of analogy-making. *Trends in Cognitive Sciences,* 6(5), pp. 200-205.

FRIEDEWALD, W.T., LEVY, R.I. & FREDRICKSON, D.S., 1972. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical Chemistry,* 18, pp. 449-502.

GANZ, D.A., KUNTZ, K.M., JACOBSON, G.A. & AVORN, J., 2000. Cost-effectiveness of 3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitor therapy in older patients with myocardial infarction. *Annals of Internal Medicine,* 132(10), pp. 780-787.

GELMAN, A., CARLIN, J.B., STERN, H.S. & RUBIN, D.B., 2004. *Bayesian data analysis.* Chapman & Hall/CRC; New York, USA.

GENTNER D., 1997. Structure mapping in analogy and similarity. *American Psychologist,* 52(1), pp. 45-56.

GIBSON, N.M. & OLEJNIK, S., 2003. Treatment of missing data at the second level of hierarchical linear models. *Educational and Psychological Measurement,* 63(2), pp. 204-238.

GLICK, H., HEYSE, J.F., THOMPSON, D., EPSTEIN, R.S., SMITH, M.E. & OSTER, G., 1992. A model for evaluating the cost-effectiveness of cholesterol-lowering treatment. *International Journal of Technology Assessment in Health Care,* 8(4), pp. 719-734.

GOEREE, R., HE, J., O'REILLY, D., TARRIDE, J., XIE, F., LIM, M. & BURKE, N., 2011. Transferability of health technology assessments and economic evaluations: a systematic review of approaches for assessment and application. *ClinicoEconomics and Outcomes Research,* 3, pp. 89-104.

GOEREE, R., BURKE, N., O'REILLY, D., MANCA, A., BLACKHOUSE, G. & TARRIDE, J., 2007. Transferability of economic evaluations: approaches and factors to consider when using results from one geographic area for another. *Current Medical Research and Opinion,* 23(4), pp. 671-682.

GOLDMAN, L., GOLDMAN, P.A., WILLIAMS, L.W. & WEINSTEIN, M.C., 1993. Cost-effectiveness considerations in the treatment of heterozygous familial hypercholesterolemia with medications. *American Journal of Cardiology,* 72(10), pp. 7579-79D.

GOLDMAN, L., WEINSTEIN, M.C., GOLDMAN, P.A. & WILLIAMS, L.W., 1991. Cost-effectiveness of HMG-CoA reductase inhibition for primary and secondary prevention of coronary heart disease. *JAMA,* 265(9), pp. 1145-1151.

GOLDSTEIN, H., 2009. *REALCOM-IMPUTE: multiple imputation using MLwiN.* Centre for Multilevel Modelling, University of Bristol, UK.

GOLDSTEIN, H., 1999. *Multilevel statistical models.* Institute of Education, Multilevel Models Project,, London, UK.

GOLDSTEIN, H. & SAMMONS, P., 1997. The influence of secondary and junior schools on sixteen year examination performance: a cross-classified multilevel analysis. *School Effectiveness and School Improvement,* 8(2), pp. 219-230.

GREENLAND, S., 2000. Principles of multilevel modelling. *International Journal of Epidemiology,* 29, pp. 158-167.

GREVING, J., VISSEREN, F., DE WIT, G. & ALGRA, A., 2011. Statin treatment for primary prevention of vascular disease: whom to treat? Cost-effectiveness analysis. *BMJ (Clinical Research Ed.),* 342, pp. d1672.

GRIEVE, R., NIXON, R. and THOMPSON, S.G., 2010. Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials. *Medical Decision Making,* 30(2), pp. 163-175.

GRIEVE, R., NIXON, R., THOMPSON, S.G. & CAIRNS, J., 2007. Multilevel models for estimating incremental net benefits in multinational studies. *Health Economics,* 16(8), pp. 815-826.

GRIEVE, R., NIXON, R., THOMPSON, S.G. & NORMAND, C., 2005. Using multilevel models for assessing the variability of multinational resource use and cost data. *Health Economics,* 14(2), pp. 185-196.

GROVER, S., COUPAL, L. & LOWENSTEYN, I., 2008. Preventing cardiovascular disease among Canadians: is the treatment of hypertension or dyslipidemia cost-effective? *Canadian Journal of Cardiology,* 24(12), pp. 891-898.

GROVER, S.A., COUPAL, L., PAQUET, S. & ZOWALL, H., 1999. Cost-effectiveness of 3-hydroxy-3-methylglutaryl-coenzyme A reductase inhibitors in the secondary prevention of cardiovascular disease: forecasting the incremental benefits of preventing coronary and cerebrovascular events. *Archives of Internal Medicine,* 159(6), pp. 593-600.

GROVER, S.A., COUPAL, L., ZOWALL, H., ALEXANDER, C.M., WEISS, T.W. & GOMES, D.R.J., 2001. How cost-effective is the treatment of dyslipidemia in patients with diabetes but without cardiovascular disease? *Diabetes care,* 24(1), pp. 45-50.

GROVER, S.A., COUPAL, L., ZOWALL, H. & DORAIS, M., 2000. Cost-effectiveness of treating hyperlipidemia in the presence of diabetes: who should be treated? *Circulation,* 102(7), pp. 722-727.

GROVER, S.A., HO, V., LAVOIE, F., COUPAL, L., ZOWALL, H. & PILOTE, L., 2003. The importance of indirect costs in primary cardiovascular disease prevention: Can we save lives and money with statins? *Archives of Internal Medicine,* 163(3), pp. 333-339.

GROVER, S.A., PAQUET, S., LEVINTON, C., COUPAL, L. & ZOWALL, H., 1998. Estimating the benefits of modifying risk factors of cardiovascular disease - A comparison of primary vs. secondary prevention. *Archives of Internal Medicine,* 158(6), pp. 655-662.

HAMILTON, V.H., RACICOT, F.E., ZOWALL, H., COUPAL, L. & GROVER, S.A., 1995. The cost-effectiveness of HMG-CoA reductase inhibitors to prevent coronary heart disease. Estimating the benefits of increasing HDL-C. *JAMA,* 273(13), pp. 1032-1038.

HAY, J.W., WITTELS, E.H. & GOTTO, A.M. JR, 1991. An economic evaluation of lovastatin for cholesterol lowering and coronary artery disease reduction. *American Journal of Cardiology,* 67(9), pp. 789-796.

HEART PROTECTION STUDY (HPS) COLLABORATIVE GROUP, 2002. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. *The Lancet,* 360(9326), pp. 7-22.

HEART PROTECTION STUDY (HPS) COLLABORATIVE GROUP, 2009. *Circulation Cardiovascular Quality & Outcomes,* 2(2), pp. 65-72.

HEART PROTECTION STUDY (HPS) COLLABORATIVE GROUP, MIHAYLOVA, B., BRIGGS, A., ARMITAGE, J., PARISH, S., GRAY, A. & COLLINS, R., 2006. Lifetime cost effectiveness of simvastatin in a range of risk groups and age groups derived from a randomised trial of 20,536 people. *BMJ,* 333(7579), pp. 1145.

HEYLAND, D.K., KERNERMAN, P., GAFNI, A. & COOK, D.J., 1996. Economic evaluations in the critical care literature: Do they help us improve the efficiency of our unit? *Critical Care Medicine,* 24(9), pp. 1591-1598.

HIPPISLEY-COX, J., COUPLAND, C., VINOGRADOVA, Y., ROBSON, J., MAY, M. & BRINDLE, P., 2007. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ,* (335), pp. 136.

HIPPISLEY-COX, J., COUPLAND, C., VINOGRADOVA, Y., ROBSON, J., MINHAS, R., SHEIKH, A. & BRINDLE, P., 2008. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ,* 336, pp. 1475.

HJALTE, K., LINDGREN, B., PERSSON, U. & OLSSON, A.G., 1989. Lipid lowering therapy: cost estimates in Sweden. In: Lewis B, Assman G, eds. *The social and economic context of coronary prevention*. *Current medical Literature: Proceedings of the international task force for prevention of coronary heart disease*.

HJELMGREN, J., BERGGREN, F. & ANDERSSON, F., 2001. Health economic guidelines - similarities, differences and some implications. *Value in Health,* 4(3), pp. 225-250.

HOCH, J.S., BRIGGS, A.H. & WILLAN, A.R., 2002. Something old, something new, something borrowed, something blue: A framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics,* 11(5), pp. 415-430.

HONG KONG SPECIAL ADMINISTRATIVE REGION GOVERNMENT - INFORMATION SERVICES / CENSUS AND STATISTICS DEPARTMENT, Hong Kong Census. Available at: http://www.gov.hk [January, 08, 2012].

HOX, J.J., 2010. *Multilevel Analysis - Techniques and Applications.* Routledge; New York, USA

HUSE, D.M., RUSSELL, M.W., MILLER, J.D., KRAEMER, D.F., D'AGOSTINO, R.B., ELLISON, R.C. & HARTZ, S.C., 1998. Cost-effectiveness of statins. *American Journal of Cardiology,* 82(11), pp. 1357-1363.

IMF (THE INTERNATIONAL MONETARY FUND), World Economic Outlook Database / frequently asked questions. Available at: http://www.imf.org/external/pubs/ft/weo/faq.htm#q3d [May, 26, 2010].

INTERNATIONAL SOCIETY FOR PHARMACOECONOMICS AND OUTCOMES RESEARCH (ISPOR), ISPOR Pharmacoeconomic Guidelines Around The World. Available at: http://www.ispor.org/PEguidelines/index.asp. [May, 06, 2011].

JACKMAN, S., 2008. *Bayesian Analysis for the Social Sciences.* John Wiley & Sons, LTD; Chichester, UK

JEFFERSON, T.O., MUGFORD, M., GRAY, A. & DEMICHELI, V., 1996. An exercise on the feasibility of carrying out secondary economic analyses. *Health Economics,* 5(2), pp. 155-165.

JEFFREY, R., 2002. *Subjective Probability - The Real Thing.* Cambridge University Press; Cambridge, UK

JOHANNESSON, M., BORGQUIST, L., JONSSON, B. & LINDHOLM, L.H., 1996. The cost effectiveness of lipid lowering in Swedish primary health care. The CELL Study Group. *Journal of Internal Medicine,* 240(1), pp. 23-29.

JOHANNESSON, M., JONSSON, B., KJEKSHUS, J., OLSSON, A.G., PEDERSEN, T.R. & WEDEL, H., 1997. Cost effectiveness of simvastatin treatment to lower cholesterol levels in patients with coronary heart disease. Scandinavian Simvastatin Survival Study Group. *New England Journal of Medicine,* 336(5), pp. 332-336.

JÖNSSON, B., COOK, J.R. & PEDERSEN, T.R., 1999. The cost-effectiveness of lipid lowering in patients with diabetes: Results from the 4S trial. *Diabetologia,* 42(11), pp. 1293-1301.

JONSSON, B., JOHANNESSON, M., KJEKSHUS, J., OLSSON, A.G., PEDERSEN, T.R. & WEDEL, H., 1996. Cost-effectiveness of cholesterol lowering. Results from the Scandinavian Simvastatin Survival Study (4S). *European Heart Journal,* 17(7), pp. 1001-1007.

JUTHE, A., 2005. Argument by Analogy. *Argumentation,* 19, pp. 1-27.

KHOURY, H., WAGNER, M., MERIKLE, E., JOHNSON, S.J. & ROBERTS, C., 2009. Cost-effectiveness of atorvastatin in the primary prevention of major cardiovascular events in patients with type 2 diabetes in Canada. *Canadian Journal of Diabetes,* 33(4), pp. 363-374.

KLIX, F., 2004. Problem solving: deduction, induction, and analogical reasoning. *International Encyclopedia of the Social and Behavioral Sciences,* pp. 12123-12130.

KNIES, S., AMENT, A.A., EVERS, S.M.A.A. & SEVERENS, J.L., 2009. The transferability of economic evaluations: testing the model of Welte. *Value in Health,* 12, pp. 730-738.

KOLENIKOV, S. & ANGELES, G., 2004. The Use of discrete data in PCA: theory, simulations and applications to socioeconomic indices. Working Paper of Measure / Evaluation project No. WP-04-85, Carolina Population Centre, UNC.

KONGNAKORN, T., WARD, A., ROBERTS, C.S., O'BRIEN, J.A., PROSKOROVSKY, I. & CARO, J.J., 2009. Economic evaluation of atorvastatin for prevention of recurrent stroke based on the SPARCL trial. *Value in Health,* 12(6), pp. 880-887.

LE ROUX, B. & ROUANET, H., 2010. *Multiple Correspondence Analysis.* SAGE publications, Inc.

LEE, C.C., HO, H.C., LEE, J.C.C., SU, Y.C., LEE, M.S., HUNG, S.K. & CHOU, P., 2010. Association between surgeon volume and hospitalisation costs for patients with oral cancer: a nationwide population base study in Taiwan. *Clinical Otolarynology,* 35(1), pp.46-52

LINDGREN, P., ERIKSSON, J., BUXTON, M., KAHAN, T., POULTER, N., DAHLOF, B., SEVER, P., WEDEL, H., JONSSON, B. & ANGLO-SCANDINAVIAN-CARDIAC OUTCOMES TRIAL INVESTIGATORS, 2010. The economic consequences of non-adherence to lipid-lowering therapy: results from the Anglo-Scandinavian-Cardiac Outcomes Trial. *International journal of Clinical Practice,* 64(9), pp. 1228-1234.

LINDGREN, P., GRAFF, J., OLSSON, A.G., PEDERSEN, T.J., JONSSON, B. & IDEAL TRIAL INVESTIGATORS, 2007. Cost-effectiveness of high-dose atorvastatin compared with regular dose simvastatin. *European Heart Journal,* 28(12), pp. 1448-1453.

LITTLE, J.R. & RUBIN, D., 1987. *Statistical Analysis with missing data.* 2 edn. Wiley; New York, USA.

MADDALA, G.S., 2001. *Introduction to Econometrics.* 3 edn. John Wiley & Son, LTD; Chichester, UIK.

MANCA, A., 2009. Economic data transferability for HTA: Are we there yet? *Value in Health,* 12(4), pp. 407.

MANCA, A., LAMBERT, P.C., SCULPHER, M. & RICE, N., 2007. Cost-effectiveness analysis using data from multinational trials: the use of bivariate hierarchical modeling. *Medical Decision Making,* 27(4), pp. 471-490.

MANCA, A., RICE, N., SCULPHER, M.J. & BRIGGS, A.H., 2005. Assessing generalisability by location in trial-based cost-effectiveness analysis: the use of multilevel models. *Health Economics,* 14(5), pp. 471-485.

MANCA, A., SCULPHER, M.J. & GOEREE, R., 2010. The analysis of multinational cost-effectiveness data for reimbursement decisions: A critical appraisal of recent methodological developments. *PharmacoEconomics,* 28(12), pp. 1079-1096.

MANCA, A. & WILLAN, A.R., 2006. 'Lost in translation': accounting for between-country differences in the analysis of multinational cost-effectiveness data. *PharmacoEconomics,* 24(11), pp. 1101-1119.

MARSHALL, D.A. & HUX, M., 2009. Design and analysis issues for economic analysis alongside clinical trials *Medical Care,* 47(7 (Suppl. 1)), pp. 14-20.

MARTENS, L.L. & GUIBERT, R., 1994. Cost-effectiveness analysis of lipid-modifying therapy in Canada: comparison of HMG-CoA reductase inhibitors in the primary prevention of coronary heart disease. *Clinical Therapeutics,* 16(6), pp. 1052-1062.

MASON, J.M. & MASON, A.R., 2006. The generalisability of pharmacoeconomic studies - issues and challenges ahead. *Pharmacoeconomics,* 24(10), pp. 937-945.

MINERS, A.H., GARAU, M., FIDAN, D. & FISCHER, A.J., 2005. Comparing estimates of cost-effectiveness submitted to the National Institute for Clinical Excellence (NICE) by different organisations: retrospective study *BMJ,* 330(65).

MISSINGVALUES.ORG.UK, ESRC research methods programme grant RES-189-25-0103. Available at: www.missingvalues.org.uk [June, 06, 2012].

MORELLE, M., REMONNAY, R., GIRAUD, P. & CARRÈRE, M.O., 2009. Analyzing multiple learning effects in health care using multilevel modeling: application to radiotherapy at an early stage of innovation. *International Journal of Technology Assessment in Health Care,* 25(2), pp. 232-239.

MORRIS, S., SUTTON, M. & GRAVELLE, H., 2005. Inequity and inequality in the use of health care in England: an empirical investigation. *Social Science & Medicine,* 60, pp. 1251-1266.

MORRIS, S., 1997. A comparison of economic modelling and clinical trials in the economic evaluation of cholesterol-modifying pharmacotherapy. *Health Economics,* 6(6), pp. 589-601.

MORRIS, S. & GODBER, E., 1999. Choice of cost-effectiveness measure in the economic evaluation of cholesterol-modifying pharmacotherapy. An illustrative example focusing on the primary prevention of coronary heart disease in Canada. *PharmacoEconomics,* 16(2), pp. 193-205.

MULS, E., VAN GANSE, E. & CLOSON, M.C., 1998. Cost-effectiveness of pravastatin in secondary prevention of coronary heart disease: comparison between Belgium and the United States of a projected risk model. *Atherosclerosis,* 137(Suppl), pp. S111-6.

MURRAY, C.J.L., EVANS, D.B., ACHARYA, A. & BALTUSSED, R.M.P.M., 2000. Development of WHO guidelines on generalized cost-effectiveness analysis. *Health Economics,* 9, pp. 235-251.

TAGHREED A., EVANS, D.B., MURRAY, J.L. 2003. Econometric estimation of country-specific hospital costs. *Cost-Effectiveness and Ressource Allocation,* 1 (3).

NAGATA-KOBAYASHI, S., SHIMBO, T., MATSUI, K. & FUKUI, T., 2005. Cost-effectiveness of pravastatin for primary prevention of coronary artery disease in Japan. *International journal of cardiology,* 104(2), pp. 213-223.

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELENCE (NICE), 2008a (reviewed 2010). *NICE clinical guideline 67 lipid modification, cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease.* London, UK

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE (NICE), 2008. *Guide to the Methods of Technology Appraisal.* London, UK

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE (NICE), 2006. *Technology Appraisal 94: Statins for the prevention of cardiovascular events.* London, UK.

NHERERA, L., CALVERT, N.W., DEMOTT, K., HUMPHRIES, S.E., NEIL, H.A., MINHAS, R. & THOROGOOD, M., 2010. Cost-effectiveness analysis of the use of a high-intensity statin compared to a low-intensity statin in the management of patients with familial hypercholesterolaemia. *Current Medical Research & Opinion,* 26(3), pp. 529-536.

NIXON, R.M. & THOMPSON S.G., 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics,* 14, pp. 1217-1229.

NIXON, J., RICE, S., DRUMMOND, M., BOULENGER, S., ULMANN, P. & DE POUVOURVILLE, G., 2009. Guidelines for completing the EURONHEED transferability information checklists. *European Journal of Health Economics,* 10(2), pp. 157-165.

NIXON, J., ULMANN, P., GLANVILLE, J., BOULENGER, S., DRUMMOND, M. & DE POUVOURVILLE, G., 2004. The European Network of Health Economic Evaluation Databases (EURONHEED) project. *European Journal of Health Economics,* 5(2), pp. 183-187.

OBERMANN, K., MATTIAS, G., SCHULENBURG, J.M. & MAUTNER, G.C., 1997. Economic analysis of secondary prevention of coronary heart disease with simvastatin (Zocor) in Germany. *Medizinische Klinik,* 92(11), pp. 686-694.

O'BRIEN, B.J., 1997. A tale of two (or more) cities: geographic transferability of pharmacoeconomic data. *American Journal of Managed Care,* 3(5 SUPPL. 1).

OFMAN, J., SULLIVAN, S., NEUMANN, P.J., CHIOU, C., HENNING, J.M., WADE, S.W. & HAY, J.W., 2003. Examining the value and quality of health economic analyses: implications of utilizing the QHES. *Journal of Managed Care Pharmacy,* 9(1), pp. 53-61.

OR, Z., WANG, J. & JAMISON, D., 2005. International differences in the impact of doctors on health: a multilevel analysis of OECD countries. *Journal of Health Economics,* 24, pp. 531-560.

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD), Main Economic Indicators. Available at: www.oecd.org/std/ppp [May, 26, 2010].

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD), OECD-iLibrary. Available at: http://www.oecd-ilibrary.org/statistics [January, 07, 2012].

PERREAULT, S., HAMILTON, V.H., LAVOIE, F. & GROVER, S., 1998. Treating hyperlipidemia for the primary prevention of coronary disease. Are higher dosages of lovastatin cost-effective? *Archives of Internal Medicine,* 158(4), pp. 375-381.

PETRINCO, M., PAGANO, E., DESIDERI, A., BIGI, R., GHIDINA, M., FERRANDO, A., CORTIGIANI, L., MERLETTI, F. & GREGORI, D., 2009. Information on center characteristics as costs' determinants in multicenter clinical trials: Is modeling center effect worth the effort? *Value in Health,* 12(2), pp. 325-330.

PEURA, P., MARTIKAINEN, J., SOINI, E., HALLINEN, T. & NISKANEN, L., 2008. Cost-effectiveness of statins in the prevention of coronary heart disease events in middle-aged Finnish men. *Current Medical Research & Opinion,* 24(6), pp. 1823-1832.

PHAROAH, P.D. & HOLLINGWORTH, W., 1996. Cost effectiveness of lowering cholesterol concentration with statins in patients with and without pre-existing coronary heart disease: life table method applied to health authority population. *BMJ,* 312(7044), pp. 1443-1448.

PICKIN, D.M., MCCABE, C.J., RAMSAY, L.E., PAYNE, N., HAQ, I.U., YEO, W.W. & JACKSON, P.R., 1999. Cost effectiveness of HMG-CoA reductase inhibitor (statin) treatment related to the risk of coronary heart disease and cost of drug treatment. *Heart,* 82(3), pp. 325-332.

PINTO, E.M., WILLAN, A.R. & O'BRIEN, B.J., 2005. Cost-effectiveness analysis for multinational clinical trials. *Statistics in Medicine,* 24, pp. 1965-1982.

RAIKOU, M., MCGUIRE, A., COLHOUN, H.M., BETTERIDGE, D.J., DURRINGTON, P.N., HITMAN, G.A., NEIL, H.A.W., LIVINGSTONE, S.J., CHARLTON-MENYS, V. & FULLER, J.H., 2007. Cost-effectiveness

of primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes: results from the Collaborative Atorvastatin Diabetes Study (CARDS). *Diabetologia,* 50(4), pp. 733-740.

RAMSEY, S., WILLKE, R., BRIGGS, A., BROWN, R., BUXTON, M., CHAWLY, A., COOK, J., GLICK, H., LILJAS, B., PETITTI, D. & REED, S., 2005. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA task force report. *Value in Health,* 8(5), pp. 521-533.

RAMSEY, S.D., CLARKE, L.D., ROBERTS, C.S., SULLIVAN, S.D., JOHNSON, S.J. & LIU, L.Z., 2008. An economic evaluation of atorvastatin for primary prevention of cardiovascular events in type 2 diabetes. *PharmacoEconomics,* 26(4), pp. 329-339.

RASBASH, J., 2008. Module 4 - Multilevel structures and classifications. In: CENTRE FOR MULTILEVEL MODELLING (CMM), *LEMMA - Learning Environment for Multilevel Methods and Applications.* Centre for Multilevel Modelling; Bristol, UK

RASBASH, J., CHARLTON, C., BROWNE, W.J., HEALY, M. & CAMERON, B., 2009a. *MLwiN Version 2.1*. Centre for Multilevel Modelling; Bristol, UK

RASBASH, J., STEELE, F., BROWNE, W.J. & GOLDSTEIN, H., 2009. *A User's Guide to MLwiN.* Centre for Multilevel Modelling; Bristol, UK

RENCHER, A.C., 2002. *Methods of Multivariate Analysis.* 2nd edn. John Wiley & Sons, Inc.

RICE, N. & JONES, A., 1997. Multilevel models and health economics. *Health Economics,* 6(6), pp. 561-575.

RIVIERE, M., WANG, S., LECLERC, C., FITZSIMON, C. & TRETIAK, R., 1997. Cost-effectiveness of simvastatin in the secondary prevention of coronary artery disease in Canada. *Canadian Medical Association Journal,* 156(7), pp. 991-997.

ROSEN, V.M., TAYLOR, D.C., PAREKH, H., PANDYA, A., THOMPSON, D., KUZNIK, A., WATERS, D.D., DRUMMOND, M. & WEINSTEIN, M.C., 2010. Cost effectiveness of intensive lipid-lowering treatment for patients with congestive heart failure and coronary heart disease in the US. *PharmacoEconomics,* 28(1), pp. 47-60.

RUBIN, D., 1987. *Multiple Imputation for Non-Response in Surveys.* Wiley; New York, USA

RUSSELL, M.W., HUSE, D.M., MILLER, J.D., KRAEMER, D.F. & HARTZ, S.C., 2001. Cost effectiveness of HMG-CoA reductase inhibition in Canada. *Canadian Journal of Clinical Pharmacology,* 8(1), pp. 9-16.

SCANDINAVIAN SIMVASTATIN SURVIVAL STUDY GROUP, 1994. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *The Lancet,* 19(344), pp. 1383-1389.

SCOTT, A. & SHIELL, A., 1998. *The use of multilevel analysis in health economics: An application to examining the effect of competition on general practitioners' behaviour.* KLUWER ACADEMIC PUBLISHERS; USA

SCUFFHAM, P.A. & CHAPLIN, S., 2005. A cost-effectiveness analysis of fluvastatin in patients with diabetes after successful percutaneous coronary intervention. *Clinical Therapeutics,* 27(9), pp. 1467-1477.

SCUFFHAM, P.A. & CHAPLIN, S., 2004. An economic evaluation of fluvastatin used for the prevention of cardiac events following successful first percutaneous coronary intervention in the UK. *PharmacoEconomics,* 22(8), pp. 525-535.

SCUFFHAM, P.A. & KOSA, J., 2006. The cost-effectiveness of fluvastatin in Hungary following successful percutaneous coronary intervention. *Cardiovascular Drugs & Therapy,* 20(4), pp. 309-317.

SCULPHER, M.J. & DRUMMOND, M.F., 2006. Analysis sans frontières: can we ever make economic evaluations generalisable across jurisdictions? *PharmacoEconomics,* 24(11), pp. 1087-1099.

SCULPHER, M.J., PANG, F.S., MANCA, A., DRUMMOND, M.F., GOLDER, S., URDAHL, H., DAVIES, L.M. & EASTWOOD, A., 2004. Generalisability in economic evaluation studies in healthcare: a review and case studies. *Health Technology Assessment,* 8(49), pp. iii-iv, 1-192.

SHEMILT, I., MUGFORD, M., BYFORD, S., DRUMMOND, M., EISENSTEIN, E., KNAPP, M., MALLENDER, J., MCDAID, D., VALE, L. and WALKER, D. ON BEHALF OF THE CAMPBELL & COCHRANE ECONOMICS METHODS GROUP, 2008. Chapter 15: Incorporating economics evidence. In: J.P.T. HIGGINS and S. GREEN, eds, *Cochrane Handbook for Systematic Reviews of Interventions.* John Wiley & Sons Ltd, Chichester, UK

SIGVANT, B., HENRIKSSON, M., LUNDIN, F. & WAHLBERG, E., 2011. Asymptomatic peripheral arterial disease: is pharmacological prevention of cardiovascular risk cost-effective? *European journal of cardiovascular prevention and rehabilitation : official journal of the European Society of Cardiology, Working Groups on Epidemiology & Prevention and Cardiac Rehabilitation and Exercise Physiology,* 18(2), pp. 254-261.

SLEIGH, J.W., 1997. Logical Limits of Randomized Controlled Trials. *Journal of Evaluation in Clinical Practice,* 3(2), pp. 145-148.

SLEJKO, J.F., PAGE, R.L.,2nd & SULLIVAN, P.W., 2010. Cost-effectiveness of statin therapy for vascular event prevention in adults with elevated C-reactive protein: implications of JUPITER. *Current Medical Research & Opinion,* 26(10), pp. 2485-2497.

SNIJDERS, T.A.B., 2005. Power and sample size in multilevel modelling. In: B.S. EVERITT and D.C. HOWELL, eds, *Encyclopedia of Statistics in Behavioural Science.* 3rd edn. Wiley, pp. 1570-1573; Chichester, UK

SOINI, E.J., DAVIES, G., MARTIKAINEN, J.A., HU, H.X., TUNCELI, K. & NISKANEN, L., 2010. Population-based health-economic evaluation of the secondary prevention of coronary heart disease in Finland. *Current Medical Research & Opinion,* 26(1), pp. 25-36.

SPAANS, J.N., COYLE, D., FODOR, G., NAIR, R., VAILLANCOURT, R., GROVER, S.A. & COUPAL, L., 2003. Application of the 1998 Canadian cholesterol guidelines to a military population: health benefits and cost effectiveness of improved cholesterol management. *Canadian Journal of Cardiology,* 19(7), pp. 790-796.

SPÄTH, H.-., CARRÈRE, M.-., FERVERS, B. & PHILIP, T., 1999. Analysis of the eligibility of published economic evaluations for transfer to a given health care system: Methodological approach and application to the French health care system. *Health Policy,* 49(3), pp. 161-177.

SPIEGELHALTER, D.J., ABRAMS, K.R. & MYLES, J.P., 2004. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* John Wiley & Sons, LTD; Chichester, UK

SPIEGELHALTER, D.J., MYLES, J.P., JONES, D.R. & ABRAMS, K.R., 2000. Bayesian methods in health technology assessment: A review. *Health Technology Assessment,* 4(38).

STEEL, P.D., 2008. *Across the Boundaries - Extrapolation in Biology and Social Science.* Oxford University Press; New York, USA.

STEELE, F., 2008. Introduction to Multilevel Modelling Concepts. *LEMMA - Learning Environment for Multilevel Methods and Applications.* Centre for Multilevel Modelling; Bristol, UK

STERNE, J.A.C. & HARBORD, R.M., 2004. Funnel Plots in Meta-Analysis. *The Stata Journal,* 4(2), pp. 127-141.

STEUTEN, L., VALLEJO-TORRES, L., YOUNG, T. & BUXTON, M., 2008. Transferability of economic evaluations of medical technologies: a new technology for orthopedic surgery. *Expert Review of Medical Devices,* 5(3), pp. 329-336.

STINNETT, A.A. & MULLAHY, J., 1998. Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making,* 18(2 SUPPL.).

SZUCS, T.D., BERTEL, O., DARIOLI, R., GUTZWILLER, F. & MORDASINI, R., 2000. Pharmacoeconomic evaluation of pravastatin in coronary secondary prevention in patients with myocardial infarct or unstable angina pectoris. An analysis based on the LIPID Study. *Praxis,* 89(18), pp. 745-752.

SZUCS, T.D., BERGER, K., MARZ, W. & SCHAFER, J.R., 2000. Cost effectiveness of pravastatin in secondary coronary prevention in patients with myocardial infarct or unstable angina in Germany. An analysis on the basis of the LIPID trial. *Herz,* 25(5), pp. 487-494.

SZUCS, T.D., GUGGENBERGER, G., BERGER, K., MAERZ, W. & SCHAEFER, J.R., 1998. Pharmacoeconomic evaluation of pravastatin in the secondary prevention of coronary heart disease in patients with average cholesterol levels. An analysis for Germany based on the CARE study [Pharmakoekonomische Bewertung von Pravastatin in der Sekundaerpraeivention der koronaren Herzkrankheit bei Patienten mit durchschnittlichen Cholesterinwerten - Eine Analyse für Deutschland auf der Grundlage der CARE-Studie]. *Herz,* 23, pp. 319-329.

SZUCS, T.D., KLOSE, G. & DUSING, R., 2004. Cost-effectiveness of atorvastatin for the prevention of coronary disease. An analysis of the ASCOT study. *Deutsche medizinische Wochenschrift,* 129(25-26), pp. 1420-1424.

TAMBOUR, M., ZETHRAEUS, N. & JOHANNESSON, M., 1998. A note on confidence intervals in cost-effectiveness analysis. *International Journal of Technology Assessment in Healthcare,* 14(3), pp. 467-471.

TAN-TORRES EDEJER, T., BALTUSSEN, R., HUTUBESSY, R., ACHARYA, A., EVANS, D.B. & MURRAY, C.J.L., 2003. *WHO Guide to Cost-Effectiveness Analysis.* Geneva, Switzerland.

TAYLOR, D.C.A., PANDYA, A., THOMPSON, D., CHU, P., GRAFF, J., SHEPHERD, J., WENGER, N., GRETEN, H., CARMENA, R., DRUMMOND, M. and WEINSTEIN, M.C., 2009. Cost-effectiveness of intensive atorvastatin therapy in secondary cardiovascular prevention in the United Kingdom, Spain, and Germany, based on the Treating to New Targets study. *European Journal of Health Economics,* 10(3), pp. 255-265.

THOMPSON, S.G., NIXON, R.M. & GRIEVE, R., 2006. Adressing the issues that arise in analysing multicentre cost data, with application to a multinational study. *Journal of Health Economics,* 25, pp. 1015-1028.

TONKIN, A.M., ECKERMANN, S., WHITE, H., FRIEDLANDER, D., GLASZIOU, P., MAGNUS, P., KIRBY, A., MULRAY, S., DENTON, M., SALLABERGER, M., HUNT, D., SIMES, J. & LIPID STUDY GROUP, 2006. Cost-effectiveness of cholesterol-lowering therapy with pravastatin in patients with previous acute coronary syndromes aged 65 to 74 years compared with younger patients: results from the LIPID study. *American Heart Journal,* 151(6), pp. 1305-1312.

TROCHE, C.J., TACKE, J., HINZPETER, B., DANNER, M. & LAUTERBACH, K.W., 1998. Cost-effectiveness of primary and secondary prevention in cardiovascular diseases. *European Heart Journal,* 19(Suppl C), pp. 59-65.

TSEVAT, J., KUNTZ, K.M., ORAV, E.J., WEINSTEIN, M.C., SACKS, F.M. & GOLDMAN, L., 2001. Cost-effectiveness of pravastatin therapy for survivors of myocardial infarction with average cholesterol levels. *American Heart Journal,* 141(5), pp. 727-734.

TURNER, S., CHASE, D.L., MILNE, R., COOK, A., HICKS, N.J., ROSTEN, C., PAYNE, L., COLES, S., BELL, E. FOR THE EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT (EUnetHTA), 2009. The health technology assessment adaptation toolkit: Description and use. *International Journal of Technology Assessment in Health Care,* (25 suppl. 2), pp. 37-41.

URDAHL, H., MANCA, A. & SCULPHER, M.J., 2006. Assessing generalisability in model-based economic evaluation studies: A structured review in osteoporosis. *PharmacoEconomics,* 24(12), pp. 1181-1197.

VALE, L., 2010. Health technology assessment and economic evaluation: arguments for a national approach. *Value in Health,* 13(6), pp. 859-861.

VAN HOUT, B.A. & SIMOONS, M.L., 2001. Cost-effectiveness of HMG coenzyme reductase inhibitors; whom to treat?. *European Heart Journal,* 22(9), pp. 751-761.

VAN KERM, P., 1998. Simple and multiple correspondence analysis in Stata. *Stata Technical Bulletin,* 42, pp. 32.

WAGNER, M., GOETGHEBEUR, M., MERIKLE, E., PANDYA, A., CHU, P. & TAYLOR, D.C., 2009b. Cost-effectiveness of intensive lipid lowering therapy with 80 mg of atorvastatin, versus 10 mg of atorvastatin, for secondary prevention of cardiovascular disease in Canada. *Canadian Journal of Clinical Pharmacology/Journal Canadien de Pharmacologie Clinique,* 16(2), pp. e331-45.

WAGNER, M., LINDGREN, P., MERIKLE, E., GOETGHEBEUR, M. & JONSSON, B., 2009a. Economic evaluation of high-dose (80 mg/day) atorvastatin treatment compared with standard-dose (20 mg/day to 40 mg/day) simvastatin treatment in Canada based on the Incremental Decrease in End-Points Through Aggressive Lipid-Lowering (IDEAL) trial. *Canadian Journal of Cardiology,* 25(11), pp. e362-e369.

WARD, S., LLOYD, JONES, M., PANDOR, A., HOLMES, M., ARA, R., RYAN, A. et al., 2007. *systematic review and economic evaluation of statins for the prevention of coronary events.* Health Technology Assessment, 11(14).

WELTE, R., FEENSTRA, T., JAGER, H. & LEIDL, R., 2004. A decision chart for assessing and improving the transferability of economic evaluation results between countries. *PharmacoEconomics,* 22(13), pp. 857-876.

WILLAN, A.R. & KOWGIER, M.E., 2008. Cost-effectiveness analysis of a multinational RCT with a binary measure of effectiveness and an interacting covariate. *Health Economics,* 17, pp. 777-791.

WILLAN, A.R., PINTO, E.M., O'BRIEN, B.J., KAUL, P., GOEREE, R., LYND, L. & ARMSTRONG, P.W., 2005. Country specific cost comparisons from multinational clinical trials using empirical Bayesian shrinkage estimation: The Canadian ASSENT-3 economic analysis. *Health Economics,* 14(4), pp. 327-338.

WILLKE, R.J., 2003. Tailor-made or off-the-rack? The problem of transferability of health economic data. *Expert Review of Pharmacoeconomics and Outcomes Research,* 3(1), pp. 1-4.

WOLFENSTETTER, S.B. & WENIG, C.M., 2010. Economic evaluation and transferability of physical Activity programmes in primary prevention: A systematic review. *International Journal of Environmental Research and Public Health,* 7, pp. 1622-1648.

WORLD BANK, World Development Indicators. Available at: http://data.worldbank.org/indicator [January, 06, 2012].

WORLD HEALTH ORGANIZATION (WHO), Global Health Observatory Data Repository. Available at: http://apps.who.int/ghodata/ [January, 06, 2012].

Appendices

Appendices for Chapter 2

Appendix 2.1: Search strategy for SCOPUS (Medline, Embase, Science Direct)

Initial search performed in SCOPUS on September 13th 2010 and updated on May, 23rd 2012. 166 hits exported from SCOPUS to RefWorks

1	ALL(multilevel OR hierarchical OR empirical bayes OR shrinkage OR random effects)	70696
2	TITLE-ABS-KEY-AUTH(economics OR economic evaluation OR cost-effectiveness OR cost effectiveness OR cost OR effectiveness or technology assessment)	40224
3	#1 and #2	166

Appendix 2.2: Search strategy for Web of Knowledge

Initial search performed in Web of Science on September 13th 2010 and updated on May, 23rd 2012. 246 hits exported from SCOPUS to Rewforks

1	TITLE(multilevel OR hierarchical OR empirical bayes OR shrinkage OR random effects)	246
2	TITLE(economics OR economic evaluation OR cost-effectiveness OR cost effectiveness OR cost OR effectiveness or technology assessment)	254945
3	#1 and #2	246

Appendix 2.3: Search strategy for HEED

Initial search performed in HEED on September 13th 2010 and updated on May, 23rd 2012. 246 hits exported from SCOPUS to Rewforks

1	ALL(multilevel OR hierarchical OR empirical bayes OR shrinkage OR random effects)	51

Appendices for Chapter 3

Appendix 3.1: Step by step guide for implementing univariate models in MLwiN using MCMC estimation

1. If not already included, create data-id variable which is a vector of consecutive numbers from 1 to 2094. (using data manipulation / generate)
2. sort data on 'data_ID' within 'Study ID' within 'Country ID' (using data manipulation / sort)
3. Create a constant vector taking the value 1 for every data point in the dataset. This will be needed to model the intercepts. (using data manipulation, generate vector, constant vector, copies (1) value (1)
4. Set up a two-level hierarchical model with data clustered in studies
5. Run the two-level model in IGLS to obtain starting values
6. Add random structure at level 3 (country)
7. Change to MCMC estimation
8. Click on model/MCMC/classifications and tick 'treat levels as cross-classified (for the cross-classified model only)
9. Go to C1096 and change country-level variance from 0 to 0.001
10. Click start

Appendix 3.2: Step by step guide for implementing bivariate models in MLwiN using MCMC estimation

1. If not already included, create data-id variable which is a vector of consecutive numbers from 1 to 2094. (using data manipulation / generate)
2. Divide incremental cost by 100. Though in theory irrelevant, a vast difference in the error variance between both response variables may distort the estimation process.
3. Sort data on 'data_ID' within 'country-ID' within 'study-ID' (using data manipulation / sort)
4. create a constant variable, taking the value 1 for every individual in the dataset. This will be needed to model the intercepts.
5. Press the 'responses' button and select 'incr_cost_div' and 'incr_effect' in equations window. This will set up the bivariate model

6. Click on ‘resp’ in the Equations window and specify a 4 level structure with level 4 being ‘Study ID’, level 3 being ‘Country ID’, level 2 being the sequence generated in Step 4, and level 1 being the ‘resp_indicator’ (which is already specified as level 1 since we are fitting a multivariate model)
7. Press Add Term, selected ‘cons’, and click ‘add separate coefficients’.
8. Click on each of the newly entered terms and tick all the boxes so that there will be a random effect at all three conceptual levels of the model (study, country, and data).
9. Click on ‘Add Term’ and add explanatory variables with separate coefficients for each response and centred around the grand mean of the variable.
10. Run the model in IGLS
11. If a warning appears about the variance-covariance matrix not being positive definite several times: this happens sometimes during IGLS estimation and the model may not converge, but sometimes if pressing Yes (perhaps several times) to continue estimation the model will eventually converge.
12. Some variances in the model were estimated zero. Zero variances are not acceptable starting values for MCMC. Go to the names window, column 1096, click on view data. This will display the variances and covariances calculated. If there are zero values, edit them to 0.001 as MCMC estimation requires a positive definite starting value of the variances.
13. Change to MCMC estimation
14. Click on Model, MCMC, classifications, tick the box ‘treat levels as cross-classified’ (for the cross-classified model only)
15. Click on model, hierarchy viewer to check model structure: Model should display 2094 units in 67 studies and 23 countries.
16. Click on start in the equations window, model will run 5000 iterations. However burning in and iterations may be changed when switching to MCMC estimation.

Appendices for Chapter 4

Appendix 4.1: Search strategy for OVID (Medline and British Nursing Index)

Search performed in OVID on April, 15th, 2011. Search strategy adapted from Ward et al (2007). 703 hits exported from OVID to Rewforks

1	statin$.tw.	17605
2	simvastatin.tw.	4709
3	pravastatin.tw.	2802
4	lovastatin.tw.	1232
5	fluvastatin.tw.	2706
6	atorvastatin.tw.	3620
7	rosuvastatin.tw.	1043
8	Hmg$.tw.	13658
9	Co-reductase inhibitor$.tw.	3
10	Hydroxymethylglutaryl-CoA Reductase Inhibitors/	15387
11	Anticholesteremic Agents/ or pravastatin/ or simvastatin/ or lovastatin.mp. (mp=title, original title, abstract, name of substance word, subject heading word, unique identifier)	19281
12	Lipid lowering.tw.	8035
13	Or/1-12	48550
14	Coronary disease/	120990
15	(coronary or heart or arter$).mp. (mp=title, original title, abstract, name of substance word, subject heading word, unique identifier)	1516737
16	Cerebrovascular disorders/	41308
17	Stroke.tw.	103629
18	Or/14/17	1585408
19	13 and 18	15788
20	Economics/	25995
21	Exp “Costs and Cost Analysis”/	38645
22	Cost allocation/	1893
23	Cost-benefit analysis/	50450
24	Cost control/	18629
25	Cost savings/	6959
26	Exp “cost of Illness”/	52112
27	Health care costs/	20932
28	Drug costs/	10226
29	Health expenditures	11814
30	Exp economics, pharmaceutical/ or exp economics/ or exp economics, medical	434969
31	Exp”Fees and Charges”/	7674
32	Exp Budgets/	10867
33	(high adj cost).tw.	4938
34	(low adj cost).tw.	14369
35	Cost utility.tw.	1626
36	(fiscal or funding or financial or finance).tw.	58679
37	(health?care adj cost).tw.	612
38	(cost adj estimate).tw.	119
39	(cost adj variable).tw.	27
40	(unit adj cost).tw.	461
41	(economics$ or pharmacoeconomics$ or price$ or pricing).tw.	26902
42	Or/20-41	492126
43	19 and 42	703

Appendix 4.2: Search Strategy for SCOPUS (Medline, Embase and Science Direct)

Search performed in SCOPUS on April, 16th, 2011. 883 hits exported from SCOPUS to Refworks

((**statin** OR **simvastatin** OR **pravastatin** OR **lovastatin** OR **fluvastatin** OR **atorvastatin** OR **rosuvastatin** OR **hmg** OR **co-reductase inhibitor** OR **hydroxymethylglutaryl-coa reductase inhibitors** OR **anticholesteremic agents** OR **lipid lowering**) AND (TITLE-ABS-KEY(**cost** OR **cost-effectiveness** OR **cost-utility** OR **cost-benefit**)) AND (**coronary disease** OR **heart disease** OR **cerebrovascular disorders** OR **stroke**))

Appendix 4.3: Search strategy for Academic Search Complete (Business Source Premier and CINAHL)

Search performed in Academic Search Complete on April, 16th, 2011. 305 hits exported from Academic search complete to Refworks

(**statin** OR **simvastatin** OR **pravastatin** OR **lovastatin** OR **fluvastatin** OR **atorvastatin** OR **rosuvastatin** OR **hmg** OR **co-reductase inhibitor** OR **hydroxymethylglutaryl-coa reductase inhibitors** OR **anticholesteremic agents** OR **lipid lowering)** and **(cost** OR **cost-effectiveness** OR **cost-utility** OR **cost-benefit) and (coronary disease** OR **heart disease** OR **cerebrovascular disorders** OR **stroke**)

Limiters - Exclude MEDLINE records
Search modes - Boolean/Phrase

Appendix 4.4: Search strategy for Health Economics Evaluation Database (HEED)

Search performed in HEED on April, 19th, 2011. 226 hits exported from HEED to Refworks

((**statin** OR **simvastatin** OR **pravastatin** OR **lovastatin** OR **fluvastatin** OR **atorvastatin** OR **rosuvastatin** OR **hmg** OR **co-reductase inhibitor** OR **hydroxymethylglutaryl-coa reductase inhibitors** OR **anticholesteremic agents** OR **lipid lowering)** and **(cost** OR **cost-effectiveness** OR **cost-utility** OR **cost-benefit)** and **(coronary disease** OR **heart disease** OR **cerebrovascular disorders** OR **stroke**))

Appendix 4.5: Search strategy for JStor

Search performed in JStor on April, 19th, 2011. 569 hits exported from JStor to Refworks

((anticholesteremic agents OR **lipid lowering** OR **statin** OR **simvastatin** OR **pravastatin** OR **lovastatin** OR **fluvastatin** OR **atorvastatin** OR **rosuvastatin** OR **hmg** OR **coa-reductase inhibitor)** AND **(cost** OR **cost-**

effectiveness OR **cost-utility** OR **cost-benefit)** AND **(coronary disease** OR **heart disease** OR **cerebrovascular disorders** OR **stroke))**

Search was restricted to articles and reviews only (no pamphlets or 'miscellaneous') The reason was that the search engine provides a high number of non-relevant hits (e.g. front and back matters of journals, provided that the search terms appear on them) This restriction limited the number of hits from 1366 to 569. Secondly, all search terms were applied to 'full text' as Jstor only stores abstracts for about 10% of their papers listed.

Appendix 4.6: Search strategy for Pubmed

Search performed in Pubmed on April, 19th, 2011. 843 hits exported from Pubmed to Refworks

#1	statin OR simvastatin OR pravastatin OR lovastatin OR fluvastatin OR atorvastatin OR rosuvastatin OR hmg OR co-reductase inhibitor OR hydroxymethylglutaryl-coa reductase inhibitors OR anticholesteremic agents OR lipid lowering	69648
#2	cost OR cost-effectiveness OR cost-utility OR cost-benefit	480968
#3	coronary disease OR heart disease OR cerebrovascular disorders OR stroke	1141994
#4	#1 and #2 and #3	843

Appendix 4.7: Search strategy for Web of knowledge

Search performed in Web of Knowledge on April, 19th, 2011

- 676 hits exported from Web of Science to Refworks
- 292 hits exported from Biosis Previews to Refworks
-

Topic=(**statin** OR **simvastatin** OR **pravastatin** OR **lovastatin** OR **fluvastatin** OR **atorvastatin** OR **rosuvastatin** OR **hmg** OR **co-reductase inhibitor** OR **hydroxymethylglutaryl-coa reductase inhibitors** OR **anticholesteremic agents** OR **lipid lowering)** AND Topic**=(cost** OR **cost-effectiveness** OR **cost-utility** OR **cost-benefit)** AND Topic**=(coronary disease** OR **heart disease** OR **cerebrovascular disorders** OR **stroke**)

Web of knowledge hosts the databases Web of science, Medline and Biosis Previews. The search was conducted separately for Web of Science and Biosis Previews and subsequently exported to RefWorks. As Medline was covered by previous searches, this database was excluded here.

Appendix 4.8: Search strategy for CRD (includes DARE, HTA and NHS-EED)

Search performed in CRD on April, 19th, 2011. 177 hits exported from CRD to Refworks

(statin OR **simvastatin** OR **pravastatin** OR **lovastatin** OR **fluvastatin** OR **atorvastatin** OR **rosuvastatin** OR **hmg** OR **co-reductase inhibitor** OR **hydroxymethylglutaryl-coa reductase inhibitors** OR **anticholesteremic agents** OR **lipid lowering)** AND **(cost** OR **cost-effectiveness** OR **cost-utility** OR **cost-benefit)** AND **(coronary disease** OR **heart disease** OR **cerebrovascular disorders** OR **stroke)**

Appendix 4.9: Search strategy for Wiley Online Library

Search performed in Wiley Online Library on April, 20th, 2011. 537 hits exported from Wiley Online Library to Refworks

statin OR **simvastatin** OR **pravastatin** OR **lovastatin** OR **fluvastatin** OR **atorvastatin** OR **rosuvastatin** OR **hmg** OR **co-reductase inhibitor** OR **hydroxymethylglutaryl-coa reductase inhibitors** OR **anticholesteremic agents** OR **lipid lowering in Abstract** OR **statin** OR **simvastatin** OR **pravastatin** OR **lovastatin** OR **fluvastatin** OR **atorvastatin** OR **rosuvastatin** OR **hmg** OR **co-reductase inhibitor** OR **hydroxymethylglutaryl-coa reductase inhibitors** OR **anticholesteremic agents** OR **lipid lowering** in Publication Titles AND **coronary disease** OR **heart disease** OR **cerebrovascular disorders** OR **stroke** in FullText AND **cost** OR **cost-effectiveness** OR **cost-utility** OR **cost-benefit** in FullText

Search was limited to title and abstract (Statin search terms) and full text for other search terms as the Wiley search engine searches in all fields otherwise (including references). Hence, without this limiter the search would have resulted in more than 15000 hits.

Appendix 4.10: Search strategy for Cochrane Library

Search performed in Cochrane Library on April, 21th, 2011

- 71 hits exported to RefWorks from *Cochrane Database of Systematic Reviews(CDSR)*
- 24 hits exported to RefWorks from Cochrane Central Register of Controlled Trials
- 81 hits exported to RefWorks from Cochrane Methodology Register
- 2 hits exported to RefWorks from Database of Abstracts of Reviews of Effects (DARE)
- 17 hits exported to RefWorks from Health Technology Assessment Database (HTA)
- 216 hits exported to RefWorks from NHS Economic Evaluation Database (NHS EED)

(statin OR **simvastatin** OR **pravastatin** OR **lovastatin** OR **fluvastatin** OR **atorvastatin** OR **rosuvastatin** OR **hmg** OR **co-reductase inhibitor** OR **hydroxymethylglutaryl-coa reductase inhibitors** OR **anticholesteremic agents** OR **lipid lowering)** and **(cost** OR **cost-effectiveness** OR **cost-utility** OR **cost-benefit)** and **(coronary disease** OR **heart disease** OR **cerebrovascular disorders** OR **stroke)**

Appendix 4.11: Systematic reviews, meta-analyses, opinion pieces, etc. on the cost-effectiveness of statins which were hand-searched for relevant references (in alphabetical order of the first author)

ID	Reference
R1	ARA, R., RAFIA, R., WARD, S.E., WIERZBICKI, A.S., REYNOLDS, T.M., REES, A. & PANDOR, A., 2009. Are intensive lipid-lowering regimens an optimal economic strategy in patients with ACS? An acute and chronic perspective. *Expert Review of Pharmacoeconomics & Outcomes Research,* 9(5), pp. 423-433.
R2	BROWN A D & GARBER, A.M., 1998. Cost effectiveness of coronary heart disease prevention strategies in adults. In: *Aspects of Hypertension Management*; Mallarkey, G., eds., ADIS International Ltd; Auckland, New Zealand
R3	COUKELL, A.J. & WILDE, M.I., 1998. Pravastatin - A pharmacoeconomic review of its use in primary and secondary prevention of coronary heart disease. *PharmacoEconomics,* 14(2), pp. 217-236.
R4	FARMER, J.A., 1998. Economic implications of lipid-lowering trials: current considerations in selecting a statin. *American Journal of Cardiology,* 82(6A), pp. 26M-31M.
R5	FRANCO, O.H., PEETERS, A., LOOMAN, C.W. & BONNEUX, L., 2005. Cost effectiveness of statins in coronary heart disease. *Journal of Epidemiology & Community Health,* 59(11), pp. 927-933
R6	GREENHELD, W., WILSON, J., BAYLISS, S. & HYDE, C., 2008. *The clinical and cost-effectiveness of intensive versus standard lipid lowering with statins in the prevention of cardiovascular events amongst patients with acute coronary syndromes: a systematic review (Structured abstract).* West Midlands Health Technology Assessment Collaboration.
R7	GROVER, S.A., 1999. The cost effectiveness of preventing cardiovascular diseases. *Canadian Journal of Cardiology,* 15, pp. 114G-116G.
R8	GUMBS, P.D., VERSCHUREN, M.W.M., MANTEL-TEEUWISSE, A., DE WIT, A.G., DE BOER, A. & KLUNGEL, O.H., 2007. Economic evaluations of cholesterol-lowering drugs: a critical and systematic review. *PharmacoEconomics,* 25(3), pp. 187-199.
R9	HAY, J.W., YU, W.M. & ASHRAF, T., 1999. Pharmacoeconomics of lipid-lowering agents for primary and secondary prevention of coronary artery disease. *PharmacoEconomics,* 15(1), pp. 47-74.
R10	JACOBSON, T.A., 1997. Preventing coronary heart disease in the managed care era: improving the cost-effectiveness of lipid-lowering therapy with HMG-CoA reductase inhibitors. *The American Journal of Managed Care,* 3, pp. 29-41.
R11	JACOBSON, T.A., SCHEIN, J.R., WILLIAMSON, A. & BALLANTYNE, C.M., 1998. Maximizing the cost-effectiveness of lipid-lowering therapy. *Archives of Internal Medicine,* 158(18), pp. 1977-1989.
R12	KORTT M A & ARMSTRONG.,E.P., 1998. Cholesterol-lowering therapy interventions: a pharmacoeconomic assessment. *Disease Management and Health Outcomes,* 4(4), pp. 193-203.
R13	KREUZER, J. & KUBLER, W., 2001. Secondary prevention after cardiac infarct; therapeutic efficiency--cost-benefit ratio. *Internist,* 42(5), pp. 713-719.
R14	LINDGREN, P. & JÖ¶NSSON, B., 2009. From 4S to IDEAL: The health economics of the statin trials. *European Journal of Cardiovascular Prevention and Rehabilitation,* 16(2), pp. 138-143.
R15	MAC NEIL, P., 1998. Economic aspects of hypercholesterolemia treatment with HMG-CoA reductase inhibitors: a review of recent developments. *Canadian Journal of Cardiology,* 14, pp. 14A-16A.
R16	MALHOTRA, H.S. & GOA, K.L., 2001. Atorvastatin: An updated review of its pharmacological properties and use in dyslipidaemia. *Drugs,* 61(12), pp. 1835-1881.
R17	MORRIS, S., MCGUIRE, A., CARO, J. & PETTITT, D., 1997. Strategies for the management of hypercholesterolaemia: a systematic review of the cost-effectiveness literature (Brief record). *Journal of Health Service Research & Policy*, 2(4), pp. 231-250
R18	NEYT, M., DE LAET, C., VAN BRABANDT, H., FRANCO, O. & RAMAEKERS, D., 2009. Cost-effectiveness of statins in the primary prevention of cardiovascular disease: a systematic review and economic analysis for Belgium. *Acta Cardiologica*, 64 (1), pp. 1-10

R19	THOMPSON, D. & OSTER, G., 1992. Cost-effectiveness of drug therapy for hypercholesterolaemia: a review of the literature. *PharmacoEconomics,* 2(1), pp. 34-42
R20	PERRAS, C. & BALADI, J.F., 1998. A clinical and economic review of HMG-CoA reductase inhibitors in coronary heart disease - summary. Canadian Coordinating Office for Health Technology Assessment (CCOHTA). Technology Overview: Pharmaceuticals Issue 12
R21	PERRAS, C. and BALADI, J., 1997. HMG-CoA reductase inhibitors: a review of published clinical trials and pharmacoeconomic evaluations - nonsystematic review. Canadian Coordinating Office for Health Technology Assessment (CCOHTA). CCOHTA Report 5E.
R22	PLOSKER, G.L. & LYSENG-WILLIAMSON, K., 2007. Atorvastatin: A pharmacoeconomic review of its use in the primary and secondary prevention of cardiovascular events. *PharmacoEconomics,* 25(12), pp. 1031-1053.
R23	RECKLESS, J.P.D., 2000. Cost-effectiveness of statins. *Current Opinion in Lipidology,* 11(4), pp. 351-356.
R24	SCHWARTZ, J.S., 1999. Comparative economic data regarding lipid-lowering drugs. *American Heart Journal,* 137(5), pp. S97-S104.
R25	SKREPNEK, G.H., 2005. Cost-effectiveness of HMG-CoA reductase inhibitors in the treatment of dyslipidemia and prevention of CHD. *Expert Review of Pharmacoeconomics and Outcomes Research,* 5(5), pp. 603-623.
R26	SMITH, D.G., 2003. Pharmacoeconomics of lipid-lowering drugs. *Current Atherosclerosis Reports,* 5(1), pp. 67-72.
R27	VAN DER WEIJDEN, T., KNOTTNERUS, J. A., AMENT, A.J.H.A, STOFFERS, H.E.J.H. & GROL, R.P.T.M., 1998. Economic evaluation of cholesterol-related interventions in general practice. An appraisal of the evidence. *Journal of Epidemiology and Community Health,* 52, pp. 586-594
R28	WARD, S., LLOYD JONES, M., PANDOR, A., HOLMES, M., ARA, R., RYAN, A., YEO, W. & PAYNE, N., 2007. A systematic review and economic evaluation of statins for the prevention of coronary events. *Health Technology Assessment,* 11(14), pp. 1-160.

Appendix 4.12: Original research articles meeting the inclusion criteria of the systematic literature review on the cost-effectiveness of statins in the primary and secondary prevention of CVD

(in alphabetical order of the first author)

ID	Reference
18	ALONSO, R., FERNANDEZ DE BOBADILLA, J., MENDEZ, I., LAZARO, P., MATA, N. & MATA, P., 2008. Cost-effectiveness of managing familial hypercholesterolemia using atorvastatin-based preventive therapy. *Rev. Esp. Cardiol.*, 61(4), pp. 382-393.
19	ANNEMANS, L., MARBAIX, S., WEBB, K., VAN GAAL, L. & SCHEEN, A., 2010. Cost effectiveness of atorvastatin in patients with type 2 diabetes mellitus - a pharmacoeconomic analysis of the collaborative atorvastatin diabetes study in the Belgian population. *Clinical Drug Investigation,* 30(2), pp. 133-142.
67	ARA, R., PANDOR, A., STEVENS, J., REES, A. and RAFIA, R., 2009. Early high-dose lipid-lowering therapy to avoid cardiac events: a systematic review and economic evaluation. *Health Technology Assessment,* 13(34), pp. 1-74.
20	ARAUJO, D.V., BAHIA, L., SOUZA, C.P.R. & PAVÃ£O, A.L.B., 2007. Cost-effectiveness and budget impact analysis of rosuvastatin and atorvastatin for LDL-cholesterol and cardiovascular events lowering within the SUS scenario. *International Journal of Atherosclerosis,* 2(3), pp. 189-194.
1	ASHRAF, T., HAY, J.W., PITT, B., WITTELS, E., CROUSE, J., DAVIDSON, M., FURBERG, C.D. & RADICAN, L., 1996. Cost-effectiveness of pravastatin in secondary prevention of coronary artery disease. *American Journal of Cardiology,* 78(4), pp. 409-414.
37	BERGER, K., KLOSE, G. & SZUCS, T.D., 1997. Economic aspects of drug therapy exemplified by pravastatin. A socioeconomic analysis of cholesterol synthase enzyme inhibition in coronary heart disease patients. *Medizinische Klinik,* 92(6), pp. 363-369.
2	CARO, J., KLITTICH, W., MCGUIRE, A., FORD, I., NORRIE, J., PETTITT, D., MCMURRAY, J. & SHEPHERD, J., 1997. The West of Scotland coronary prevention study: economic benefit analysis of primary prevention with pravastatin. *BMJ,* 315(7122), pp. 1577-1582.
58	CARO, J.J., HUYBRECHTS, K.F., KLITTICH, W.S., JACKSON, J.D. & MCGUIRE, A., 2003. Allocating funds for cardiovascular disease prevention in light of the NCEP ATP III guidelines. *American Journal of Managed Care,* 9(7), pp. 477-489.
59	CDC DIABETES COST-EFFECTIVENESS GROUP, 2002. Cost-effectiveness of intensive glycemic control, intensified hypertension control, and serum cholesterol level reduction for type 2 diabetes. *JAMA,* 287(19), pp. 2542-2551.
65	CHAN, P.S., NALLAMOTHU, B.K., GURM, H.S., HAYWARD, R.A. & VIJAN, S., 2007. Incremental benefit and cost-effectiveness of high-dose statin therapy in high-risk patients with coronary artery disease. *Circulation,* 115(18), pp. 2398-2409.
60	CHAU, J., CHEUNG, B.M., MCGHEE, S.M., LAUDER, I.J., LAU, C.P. & KUMANA, C.R., 2001. Cost-effectiveness analysis of applying the Cholesterol and Recurrent Events (CARE) study protocol in Hong Kong. *Hong Kong Medical Journal,* 7(4), pp. 360-368.
39	DAVIES, A., HUTTON, J., O'DONNELL, J. & KINGSLAKE, S., 2006. Cost-effectiveness of rosuvastatin, atorvastatin, simvastatin, pravastatin and fluvastatin for the primary prevention of CHD in the UK. *British Journal of Cardiology,* 13(3), pp. 196-202.
64	DRUMMOND, M.F., MCGUIRE, A. & FLETCHER, A., 1993. *Economic evaluation of drug therapy for hypercholesterolemia in the UK.* University of York, CHE discussion paper No 104.
23	FRANCO, O.H., DER KINDEREN, A.J., DE LAET, C., PEETERS, A. & BONNEUX, L., 2007. Primary prevention of cardiovascular disease: cost-effectiveness comparison. *International Journal of Technology Assessment in Health Care,* 23(1), pp. 71-79.
15	GANZ, D.A., KUNTZ, K.M., JACOBSON, G.A. & AVORN, J., 2000. Cost-effectiveness of 3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitor therapy in older patients with myocardial infarction. *Annals of Internal Medicine,* 132(10), pp. 780-787.
62	GLICK, H., HEYSE, J.F., THOMPSON, D., EPSTEIN, R.S., SMITH, M.E. & OSTER, G., 1992. A model for evaluating the cost-effectiveness of cholesterol-lowering treatment. *International Journal of Technology Assessment in Health Care,* 8(4), pp. 719-734.

24	GREVING, J., VISSEREN, F., DE WIT, G. & ALGRA, A., 2011. Statin treatment for primary prevention of vascular disease: whom to treat? Cost-effectiveness analysis. *BMJ (Clinical research ed.),* 342, pp. d1672.
3	GROVER, S.A., COUPAL, L., PAQUET, S. & ZOWALL, H., 1999. Cost-effectiveness of 3-hydroxy-3-methylglutaryl-coenzyme A reductase inhibitors in the secondary prevention of cardiovascular disease: forecasting the incremental benefits of preventing coronary and cerebrovascular events. *Archives of Internal Medicine,* 159(6), pp. 593-600.
4	GROVER, S.A., COUPAL, L., ZOWALL, H. & DORAIS, M., 2000. Cost-effectiveness of treating hyperlipidemia in the presence of diabetes : who should be treated?. *Circulation,* 102(7), pp. 722-727.
16	GROVER, S.A., COUPAL, L., ZOWALL, H., ALEXANDER, C.M., WEISS, T.W. & GOMES, D.R.J., 2001. How Cost-Effective is the Treatment of Dyslipidemia in Patients with Diabetes but without Cardiovascular Disease? *Diabetes Care,* 24(1), pp.45-50
61	GROVER, S.A., HO, V., LAVOIE, F., COUPAL, L., ZOWALL, H. & PILOTE, L., 2003. The importance of indirect costs in primary cardiovascular disease prevention: Can we save lives and money with statins? *Archives of Internal Medicine,* 163(3), pp. 333-339.
22	GROVER, S.A., COUPAL, L. & LOWENSTEYN, I., 2008. Preventing cardiovascular disease among Canadians: Is the treatment of hypertension or dyslipidemia cost-effective? *Canadian Journal of Cardiology,* 24(12), pp.891-898
5	HAMILTON, V.H., RACICOT, F.E., ZOWALL, H., COUPAL, L. & GROVER, S.A., 1995. The cost-effectiveness of HMG-CoA reductase inhibitors to prevent coronary heart disease. Estimating the benefits of increasing HDL-C. *JAMA,* 273(13), pp. 1032-1038.
25	HEART PROTECTION STUDY COLLABORATIVE GROUP, 2009. Statin Cost-Effectiveness in the United States for People at Different Vascular Risk levels. *Circulation.Cardiovascular Quality & Outcomes,* 2(2), pp. 65-72.
52	HEART PROTECTION STUDY COLLABORATIVE GROUP, MIHAYLOVA, B., BRIGGS, A., ARMITAGE, J., PARISH, S., GRAY, A. and COLLINS, R., 2006. Lifetime cost effectiveness of simvastatin in a range of risk groups and age groups derived from a randomised trial of 20,536 people. *BMJ,* 333(7579), pp. 1145.
57	HJALTE, K., LINDGREN, B., PERSSON, U. & OLSSON, A.G., 1989. Lipid lowering therapy: Cost estimates in Sweden. In: Lewis, B., Assman, G., eds. *The social and economic context of coronary prevention. Current medical Literature*: Proceedings of the international task force for prevention of coronary heart disease.
47	JOHANNESSON, M., BORGQUIST, L., JONSSON, B. & LINDHOLM, L.H., 1996. The cost effectiveness of lipid lowering in Swedish primary health care. The CELL Study Group. *Journal of Internal Medicine,* 240(1), pp. 23-29.
6	JOHANNESSON, M., JONSSON, B., KJEKSHUS, J., OLSSON, A.G., PEDERSEN, T.R. & WEDEL, H., 1997. Cost effectiveness of simvastatin treatment to lower cholesterol levels in patients with coronary heart disease. *The New England Journal of Medicine,* 336, pp.332-336
14	JONSSON, B., COOK, J.R. & PEDERSEN, T.R., 1999. The cost-effectiveness of lipid lowering in patients with diabetes: results from the 4S trial. *Diabetologia,* 42(11), pp. 1293-1301.
13	JONSSON, B., JOHANNESSON, M., KJEKSHUS, J., OLSSON, A.G., PEDERSEN, T.R. & WEDEL, H., 1996. Cost-effectiveness of cholesterol lowering. Results from the Scandinavian Simvastatin Survival Study (4S). *European Heart Journal,* 17(7), pp. 1001-1007.
26	KHOURY, H., WAGNER, M., MERIKLE, E., JOHNSON, S.J. and ROBERTS, C., 2009. Cost-effectiveness of atorvastatin in the primary prevention of major cardiovascular events in patients with type 2 diabetes in Canada. *Canadian Journal of Diabetes,* 33(4), pp. 363-374.
27	KONGNAKORN, T., WARD, A., ROBERTS, C.S., O'BRIEN, J.A., PROSKOROVSKY, I. & CARO, J.J., 2009. Economic evaluation of atorvastatin for prevention of recurrent stroke based on the SPARCL trial. *Value in Health,* 12(6), pp. 880-887.
21	LINDGREN, P., ERIKSSON, J., BUXTON, M., KAHAN, T., POULTER, N., DAHLOF, B., SEVER, P., WEDEL, H., JONSSON, B. & ANGLO-SCANDINAVIAN-CARDIAC OUTCOMES TRIAL INVESTIGATORS, 2010. The economic consequences of non-adherence to lipid-lowering therapy: results from the Anglo-Scandinavian-Cardiac Outcomes Trial. *International Journal of Clinical Practice,* 64(9), pp. 1228-1234.
51	LINDGREN, P., GRAFF, J., OLSSON, A.G., PEDERSEN, T.J., JONSSON, B. & IDEAL TRIAL, I., 2007. Cost-effectiveness of high-dose atorvastatin compared with regular dose simvastatin. *European Heart Journal,* 28(12), pp. 1448-1453.
17	MARTENS, L.L. & GUIBERT, R., 1994. Cost-effectiveness analysis of lipid-modifying therapy in Canada: comparison of HMG-CoA reductase inhibitors in the primary prevention of coronary heart disease. *Clinical Therapeutics,* 16(6), pp. 1052-1062.
28	MORRIS, S., 1997. A comparison of economic modelling and clinical trials in the economic evaluation of cholesterol-modifying pharmacotherapy. *Health Economics,* 6(6), pp. 589-601.

29	MORRIS, S. & GODBER, E., 1999. Choice of cost-effectiveness measure in the economic evaluation of cholesterol-modifying pharmacotherapy. An illustrative example focusing on the primary prevention of coronary heart disease in Canada. *PharmacoEconomics,* 16(2), pp. 193-205.
7	MULS, E., VAN GANSE, E. & CLOSON, M.C., 1998. Cost-effectiveness of pravastatin in secondary prevention of coronary heart disease: comparison between Belgium and the United States of a projected risk model. *Atherosclerosis,* 137(Suppl), pp. S111-6.
50	NAGATA-KOBAYASHI, S., SHIMBO, T., MATSUI, K. & FUKUI, T., 2005. Cost-effectiveness of pravastatin for primary prevention of coronary artery disease in Japan. *International journal of cardiology,* 104(2), pp. 213-223.
63	NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELENCE (NICE), 2008 (reviewed 2010). *NICE clinical guideline 67 lipid modification, cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease; London, UK*
44	NHERERA, L., CALVERT, N.L., DEMOTT, K., HUMPHRIES, S.E., NEIL, H.A.W., MINHAS, R. & THOROGOOD, M., 2010. Cost-effectiveness analysis of the use of a high-intensity statin compared to a low-intensity statin in the management of patients with familial hypercholesterolaemia. *Current Medical Research and Opinion,* 26(3), pp.529-536
38	OBERMANN, K., MATTIAS, G., SCHULENBURG, J.M. & MAUTNER, G.C., 1997. Economic analysis of secondary prevention of coronary heart disease with simvastatin (Zocor) in Germany. *Medizinische Klinik,* 92(11), pp. 686-694.
8	PERREAULT, S., HAMILTON, V.H., LAVOIE, F. & GROVER, S., 1998. Treating hyperlipidemia for the primary prevention of coronary disease. Are higher dosages of lovastatin cost-effective? *Archives of Internal Medicine,* 158(4), pp. 375-381.
42	PEURA, P., MARTIKAINEN, P.P., SOINI, E., HALLINEN, T. & NISKANEN, L., 2008. Cost-effectiveness of statins in the prevention of coronary heart disease events in middle-aged Finnish men. *Current Medical Research and Opinion,* 24(6),pp.1823-1832
9	PHAROAH, P.D. & HOLLINGWORTH, W., 1996. Cost effectiveness of lowering cholesterol concentration with statins in patients with and without pre-existing coronary heart disease: life table method applied to health authority population. *BMJ,* 312(7044), pp. 1443-1448.
54	RAIKOU, M., MCGUIRE, A., COLHOUN, H.M., BETTERIDGE, D.J., DURRINGTON, P.N., HITMAN, G.A., NEIL, H.A.W., LIVINGSTONE, S.J., CHARLTON-MENYS, V. & FULLER, J.H., 2007. Cost-effectiveness of primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes: results from the Collaborative Atorvastatin Diabetes Study (CARDS). *Diabetologia,* 50(4), pp. 733-740.
55	RAMSEY, S.D., CLARKE, L.D., ROBERTS, C.S., SULLIVAN, S.D., JOHNSON, S.J. & LIU, L.Z., 2008. An economic evaluation of atorvastatin for primary prevention of cardiovascular events in type 2 diabetes. *PharmacoEconomics,* 26(4), pp. 329-339.
30	ROSEN, V.M., TAYLOR, D.C., PAREKH, H., PANDYA, A., THOMPSON, D., KUZNIK, A., WATERS, D.D., DRUMMOND, M. & WEINSTEIN, M.C., 2010. Cost effectiveness of intensive lipid-lowering treatment for patients with congestive heart failure and coronary heart disease in the US. *PharmacoEconomics,* 28(1), pp. 47-60.
31	SCUFFHAM, P.A. & CHAPLIN, S., 2005. A cost-effectiveness analysis of fluvastatin in patients with diabetes after successful percutaneous coronary intervention. *Clinical therapeutics,* 27(9), pp. 1467-1477.
56	SCUFFHAM, P.A. & CHAPLIN, S., 2004. An economic evaluation of fluvastatin used for the prevention of cardiac events following successful first percutaneous coronary intervention in the UK. *PharmacoEconomics,* 22(8), pp. 525-535.
32	SCUFFHAM, P.A. & KOSA, J., 2006. The cost-effectiveness of fluvastatin in Hungary following successful percutaneous coronary intervention. *Cardiovascular Drugs & Therapy,* 20(4), pp. 309-317.
46	SIGVANT, B., HENRIKSSON, M., LUNDIN, F. & WAHLBERG, E., 2011. Asymptomatic peripheral arterial disease: is pharmacological prevention of cardiovascular risk cost-effective? *European journal of cardiovascular prevention and rehabilitation : official journal of the European Society of Cardiology, Working Groups on Epidemiology & Prevention and Cardiac Rehabilitation and Exercise Physiology,* 18(2), pp. 254-261.
43	SLEJKO, J.F., PAGE, R.L.,2ND & SULLIVAN, P.W., 2010. Cost-effectiveness of statin therapy for vascular event prevention in adults with elevated C-reactive protein: implications of JUPITER. *Current Medical Research & Opinion,* 26(10), pp. 2485-2497.
41	SOINI, E.J., DAVIES, G., MARTIKAINEN, J.A., HU, H.X., TUNCELI, K. & NISKANEN, L., 2010. Population-based health-economic evaluation of the secondary prevention of coronary heart disease in Finland. *Current Medical Research & Opinion,* 26(1), pp. 25-36.
40	SPAANS, J.N., COYLE, D., FODOR, G., NAIR, R., VAILLANCOURT, R., GROVER, S.A. & COUPAL, L., 2003. Application of the 1998 Canadian cholesterol guidelines to a military population: health benefits and cost effectiveness of improved cholesterol management. *Canadian Journal of Cardiology,* 19(7), pp. 790-796.

11	SZUCS, T.D., BERGER, K., MÄRZ, W. and SCHÄFER, J.R., 2000. Cost-effectiveness of pravastatin in secondary prevention in patients with myocardial infarction or instable angina in Germany. An analysis on the basis of the LIPID trial. *Herz,* 25(5), pp. 487-494.
45	SZUCS, T.D., BERTEL, O., DARIOLI, R., GUTZWILLER, F. & MORDASINI, R., 2000. Pharmacoeconomic evaluation of pravastatin in coronary secondary prevention in patients with myocardial infarct or unstable angina pectoris. An analysis based on the LIPID Study. *Praxis,* 89(18), pp. 745-752.
10	SZUCS, T.D., GUGGENBERGER, G., BERGER, K., MÄRZ, W. & SCHÄFER, J.R., 1998. Pharmacoeconomic evaluation of pravastatin in the secondary prevention of coronary heart disease in patients with average cholesterol levels. An analysis for Germany based on the CARE study. *Herz,* 23(5), pp. 319-329.
49	SZUCS, T.D., KLOSE, G. & DUSING, R., 2004. Cost-effectiveness of atorvastatin for the prevention of coronary disease. An analysis of the ASCOT study. *Deutsche Medizinische Wochenschrift,* 129(25-26), pp. 1420-1424.
33	TAYLOR, D.C.A., PANDYA, A., THOMPSON, D., CHU, P., GRAFF, J., SHEPHERD, J., WENGER, N., GRETEN, H., CARMENA, R., DRUMMOND, M. & WEINSTEIN, M.C., 2009. Cost-effectiveness of intensive atorvastatin therapy in secondary cardiovascular prevention in the United Kingdom, Spain, and Germany, based on the treating to new targets study. *European Journal of Health Economics,* 10(3), pp. 255-265.
34	TONKIN, A.M., ECKERMANN, S., WHITE, H., FRIEDLANDER, D., GLASZIOU, P., MAGNUS, P., KIRBY, A., MULRAY, S., DENTON, M., SALLABERGER, M., HUNT, D., SIMES, J. & LIPID STUDY, G., 2006. Cost-effectiveness of cholesterol-lowering therapy with pravastatin in patients with previous acute coronary syndromes aged 65 to 74 years compared with younger patients: results from the LIPID study. *American Heart Journal,* 151(6), pp. 1305-1312.
48	TROCHE, C.J., TACKE, J., HINZPETER, B., DANNER, M. & LAUTERBACH, K.W., 1998. Cost-effectiveness of primary and secondary prevention in cardiovascular diseases. *European Heart Journal,* 19(Suppl C), pp. 59-65.
53	TSEVAT, J., KUNTZ, K.M., ORAV, E.J., WEINSTEIN, M.C., SACKS, F.M. & GOLDMAN, L., 2001. Cost-effectiveness of pravastatin therapy for survivors of myocardial infarction with average cholesterol levels. *American Heart Journal,* 141(5), pp. 727-734.
12	VAN HOUT, B.A. & SIMOONS, M.L., 2001. Cost-effectiveness of HMG coenzyme reductase inhibitors; whom to treat?. *European Heart Journal,* 22(9), pp. 751-761.
35	WAGNER, M., LINDGREN, P., MERIKLE, E., GOETGHEBEUR, M. & JÖNSSON, B., 2009a. Economic evaluation of high-dose (80 mg/day) atorvastatin treatment compared with standard-dose (20 mg/day to 40 mg/day) simvastatin treatment in Canada based on the Incremental Decrease in End-Points Through Aggressive Lipid-Lowering (IDEAL) trial. *Canadian Journal of Cardiology,* 25(11), pp. e362-e369.
36	WAGNER, M., GOETGHEBEUR, M., MERIKLE, E., PANDYA, A., CHU, P. & TAYLOR, D.C., 2009b. Cost-effectiveness of intensive lipid lowering therapy with 80 mg of atorvastatin, versus 10 mg of atorvastatin, for secondary prevention of cardiovascular disease in Canada. *Canadian Journal of Clinical Pharmacology,* 16(2), pp. e331-45.
66	WARD S, LLOYD JONES M, PANDOR A, HOLMES M, ARA R, RYAN A, et al., 2007. Systematic review and economic evaluation of statins for the prevention of coronary events. *Health Technology Assessment*, 11(14)

Appendix 4.13: final data abstraction form

1/3 - study-level information

Study ID:_____ **Date of review:** _____

Title: _____

_____

Authors: _____

_____

Year: _____**Journal, publisher:** _____**No. Issue:** _____

Geographic origin of paper: _____**Language:** **E** **G** **other** **(exclude)**

Immediately excludable? (exclusion criterion)
- ☐ Yes, no drug treatment
- ☐ Yes, no economic evaluation
- ☐ Yes, other (specify)
- ☐ No (continue)

What type of paper is this? (exclusion criterion)
- ☐ Systematic review or meta analysis (check refs)
- ☐ Other type of review (check refs)
- ☐ Opinion piece
- ☐ Other (specify)_____
- ☐ Original research article (continue)

Type of economic evaluation (exclusion criterion)
- ☐ Cost analysis
- ☐ Cost-minimization analysis
- ☐ Cost effectiveness analysis
- ☐ Cost-utility analysis
- ☐ Cost-benefit analysis
- ☐ Other (specify)_____
- ☐ Unclear (why?)_____

Any other reason for exclusion? (exclusion criterion)
- ☐ **No**
- ☐ **Yes (why?)**_____

_____

_____

_____

_____

_____

Primary source of funding:
- ☐ Research Council
- ☐ Government
- ☐ University
- ☐ Industry
- ☐ Other (specify)_____
- ☐ Unclear

Name primary source of funding?

_____

_____

_____

_____

_____

_____

Was study multinational?
- ☐ Yes
- ☐ No

General study design:
- ☐ Primary modelling
- ☐ Secondary modelling
- ☐ Other (specify)_____
- ☐ unclear

If primary modelling, what was the study design?
- ☐ Not applicable (secondary modelling)
- ☐ Pragmatic RCT
- ☐ RCT
- ☐ Cohort Study
- ☐ Case-Controll Study
- ☐ (Controlled) before-after-study
- ☐ Other (specify)_____
- ☐ Unclear

If secondary modelling, what was the study design?
- ☐ Not applicable (primary modelling)
- ☐ Markov-Model
- ☐ Decision tree
- ☐ Lifetable
- ☐ Other (specify)_____
- ☐ Unclear

Method of effect calculation
- ☐ CHD risk reduction
- ☐ Cholesterol reduction
- ☐ Other (specify) _____
- ☐ Unclear

Scope of assessment
- ☐ Coronary Artery Disease (CAD)
- ☐ Cerebrovascular Disease (CD)
- ☐ Peripheral Vascular Disease (PVD)
- ☐ CAD and CD
- ☐ CAD and PVD
- ☐ CAD, CD and PAD
- ☐ Unclear

Timing? _____

Comments on costing methodology

_____

_____

_____

Inflation adjustment?
- ☐ Yes
- ☐ No (not performed)
- ☐ No (not applicable)
- ☐ Unclear

Currency conversion
- ☐ Yes
- ☐ No (not applicable)
- ☐ Unclear

Adjustment method

_____

_____

_____

Conversion method / comments

_____

_____

_____

Final data abstraction form

2/3 – QHES-assignment

Study ID:_____

Overall QHES score: _____ **QHES-score from external source?** ☐ No → continue whith QHES assignment ☐ Yes → Source: _____

QHES – Assignment

QHES criterion	Yes	No	Score	Comments
Was the study objective presented in a (1) clear and specific, and (2) measurable manner?			7	
(1) Were the perspective of the analysis stated (societal, third-party payer, etc.) (2) Were reasons for the selection of the perspective stated?			4	
Were variable estimates used in the analysis from the best available source (i.e., randomized control trial - best, expert opinion - worst)?			8	
If estimates came from a subgroup analysis, were the groups pre-specified at the beginning of the study?			1	
Was uncertainty handled by either - statistical analysis to address random events, or - sensitivity analysis to cover a range of assumptions?			9	
Was incremental analysis performed between alternatives for resources and costs?			6	
Was the methodology for data abstraction (including the value of health states and other benefits) stated?			5	
(1) Did the analytic horizon allow time for all relevant and important outcomes? (2) Were benefits and costs that went beyond 1 year discounted (3% to 5%) and (3) was justification given for the discount rate?			7	
(1) Was the measurement of costs appropriate and (2) Was the methodology for the estimation of quantities and unit costs clearly described?			8	
(1) Were the primary outcome measures(s) for the economic evaluation clearly stated and (2) were the major short-term, long-term, and negative outcomes included?			6	
- Were the health outcomes measures/scales valid and reliable? - If previously tested valid and reliable measures were not available, was justification given for the measures/scales used?			7	
Were the (1) economic model (including structure), (2) study methods and analysis, (3) and the components of the numerator and denominator displayed in a clear, transparent manner?			8	
Were the (1) choice of economic model, (2) main assumptions, # (3) and limitations of the study stated and justified?			7	
Did the author(s) explicitly discuss direction and magnitude of potential biases?			6	
Were the conclusions/recommendations of the study justified and based on the study results?			8	
Was there a statement disclosing the source of funding for the study?			3	

Total: _____

Any other comments, (especially if there are any "other" variability factors not covered by other variables)

_____
_____
_____
_____
_____

Final data abstraction form

3/3 - data-level information

Study - ID:_____ **Estimate-ID:** _____

Outcome measure (exclusion criterion)
- ☐ Life years saved (LYS)
- ☐ QALYS
- ☐ Disease episode avoided
- ☐ Surrogate (specify)_____
- ☐ Other (specify) _____
- ☐ Unclear

Measure of cost-effectiveness (exclusion criterion)
- ☐ INMB (full review)
- ☐ ICER (short review)
- ☐ Other (specify) _____

Classification of data-point (inclusion criterion)
- ☐ Subgroup
- ☐ SA-efficiency
- ☐ SA-CVD-risk
- ☐ SA-inclusion of cost
- ☐ SA-intervention cost
- ☐ SA-QALYS
- ☐ SA-treatment duration
- ☐ SA-Discount rate
- ☐ None of the above (end review)

Any comments on subgroup or sensitivity analysis?

ICER: _____ **INCR cost** _____ **INCR effect:** _____ **INMB:** _____

(state estimates in local currencies)

Geographical origin of:

CE-Estimate	
Resource use data	
Unit cost estimate	
Effectiveness data	
Utility values	

Specify primary source of:

resource use data	
unit cost data	
effectiveness data	
utility values	
baseline risk data	

Any other relationship to other paper (e.g. model)? _____

Study perspective

A) As stated by Authors
B) On costs, concluded by reviewer
C) On effects, concluded by reviewer
(study perspective may be altered within sensitivity analysis, and is therefore placed on data-level)

A	B	C	
☐	☐	☐	Patient
☐	☐	☐	Provider
☐	☐	☐	NHS (Health Insurance)
☐	☐	☐	Public Sector
☐	☐	☐	Societal
☐	☐	☐	Other (specity)_____
☐	☐	☐	Unclear

Treatment duration _____
Time horizon _____
Extrapolation beyond follow-up ☐ Yes ☐ No
Discount rate on costs _____
Discount rate on benefits _____

If QALYS were used, what was the method of preference elicitation?
- ☐ Not applicable
- ☐ Standard Gamble
- ☐ Discrete Choice
- ☐ Time Trade OFF (TTO)
- ☐ Person Trade Off (PTO)
- ☐ Visual Analogue Scale (VAS)
- ☐ EQ-5D
- ☐ Healthy Years Equivalent (HYE)
- ☐ Health Utility Index (HUI)
- ☐ SF (36, 20, 6D, 12, specify)_____
- ☐ Any other (specify) _____
- ☐ Unclear

If QALYS were used, what do preference values reflect?
- ☐ Not applicable
- ☐ Patient values
- ☐ Population values
- ☐ Any other (specify) _____
- ☐ unclear

CHD related medical history
- ☐ No (primary prevention)
- ☐ Yes (secondary prevention
- ☐ unclear

Framingham factors:

TSC (mmol/L)		gender (% female)		Smoker (% in sample)	
HDL (mmol/L)		DM II (% in sample)		Risk category as stated by the authors	
LDL (mmol/L)		Hypertension (% in sample)		Absolute CHD risk (value)	
(mean) age		Mean Systolic BP (mmHG)		Absolute CHD risk (respective period)	

Intervention

Drug name
- ☐ Simvastatin
- ☐ Fluvastatin
- ☐ Atorvastatin
- ☐ Pravastatin
- ☐ Lovastatin

Total daily dose _____mg
Annual drug cost _____
Intervention Efficacy _____
_____

Comparator

Drug name
- ☒ Doing nothing
- ☐ Simvastatin
- ☐ Fluvastatin
- ☐ Atorvastatin
- ☐ Pravastatin
- ☐ Lovastatin

Total daily dose _____mg
Annual drug cost _____

Appendices for Chapter 5

Appendix 5.1: Descriptive statistics of data and study-level covariates

Appendix 5.1.1: Descriptive statistics of patient and disease characteristics – continuous variables

Variable	Obs.	missing	in %	Min	Max	Mean	Std. Dev.
tcl	1193	901	43.03%	4.5	10.35	6.676	1.204
hdl	1147	947	45.22%	0.9	1.55	1.168	0.102
ldl	926	1168	55.78%	1.9	8.2	4.509	1.035
hypert	826	1268	60.55%	0	1	0.317	0.381
sbp	1140	954	45.56%	120	164.7	137.475	13.348
smokers	1141	953	45.51%	0	1	0.291	0.335
diab	1163	931	44.46%	0	1	0.178	0.349

Appendix 5.1.2: Descriptive statistics of patient and disease characteristics – categorical variables

Variable category	Frequency (full sample: 2094)	In %	cummulative
Age_cat			
- <45	322	15.38%	15.38%
- 46 to 55	439	20.96%	36.35%
- 56 to 65	862	41.17%	77.51%
- 66 to 75	299	14.28%	91.79%
- > 75	98	4.68%	96.47%
- Unclear (missing)	74	3.53%	100%
Gender_cat			
- Female	576	27.51%	27.51%
- Male	799	38.16%	65.56%
- Mixed study sample	556	26.55%	92.22%
- Unclear (missing)	163	7.78%	100%
CHD_history			
- No (primary prevention)	1064	50.81%	50.81%
- Yes (secondary prevention)	958	45.75%	96.56%
- Mixed study sample	72	3.44%	100%
Risk_cat*			
- Very low (<10%)	193	9.22%	9.22%
- Low (<20%)	367	17.53%	26.75%
- Medium (<30%)	278	13.28%	40.02%
- High (<40%)	140	6.69%	46.71%
- Very high (> 40%)	106	5.06%	51.77%
- Secondary prevention	958	45.75%	97.52%
- Unclear (missing)	52	2.48%	100%
* defined as 10 year CHD risk estimated using patient risk factors reported above using Framingham risk equation (reference)			

Appendix 5.1.3 Descriptive statistics of intervention and comparator characteristics – continuous variables

Variable (in 2010 £-Sterling)	Obs.	missing	in %	Min	Max	Mean	Std. Dev.
Cost_int	1957	137	6.54%	21.48	1917.56	521.59	335.10
Unitcost_int	1738	356	17.00%	0.0012	0.1773	0.05	0.032
Cost_comp	2092	2	0.10%	0	1002.36	26.06	115.37
Unitcost_comp	2083	11	0.53%	0	0.156	0.0037	0.018
Incr_int_cost	1727	367	17.53%	-0.1255	0.1773	0.0461	0.039

Appendix 5.1.4: Descriptive statistics of intervention and comparator characteristics– categorical variables

Variable category	Frequency (full sample: 2094)	In %	cumulative
Intervention (brand name)			
- Simvastatin	1080	51.58%	51.58%
- Fluvastatin	41	1.96%	53.54%
- Atorvastatin	184	8.79%	62.33%
- Pravastatin	256	12.23%	74.56%
- Lovastatin	125	5.97%	80.53%
- Rosuvastatin	60	2.87%	83.40%
- Unclear (missing)	348	16.60%	100%
TDD-intervention (total daily dose)			
- Up to 10 mg	65	3.10%	3.10%
- Up to 20 mg	259	12.37%	15.47%
- Up to 30 mg	554	26.46%	41.93%
- Up to 40 mg	654	31.23%	73.16%
- > 60 mg	206	9.84%	83.00%
- Unclear (missing)	356	17.00%	100%
Comparator			
- Simvastatin	153	7.31%	7.31%
- Fluvastatin	3	0.14%	7.45%
- Atorvastatin	44	2.10%	9.55%
- Pravastatin	19	0.91%	10.46%
- Lovastatin	24	1.15%	11.61%
- Rosuvastatin	15	0.72%	12.33%
- Doing nothing	1834	87.58%	99.90%
- Unclear (missing)	2	0.1%	100%
Comparator_short			
- Doing nothing	1834	87.58%	87.58%
- Simvastatin	153	7.31%	94.98%
- Other	107	5.11%	100%
Active_comparator			
- yes	260	12.42%	12.42%
- no (doing nothing)	1834	87.58%	100%
- unclear (missing)	0	--	--
TDD-comp (total daily dose)			
- 0 mg (doing nothing)	1834	87.58%	87.58%
- Up to 10 mg	44	2.10%	89.68%
- Up to 20 mg	26	1.24%	90.92%
- Up to 30 mg	24	1.15%	92.07%
- Up to 40 mg	155	7.40%	99.47%
- Unclear (missing)	11	0.53%	100%

Appendix 5.1.5: Descriptive statistics of methods on data-level – continuous variables

Variable (in 2010 £-Sterling)	Obs.	missing	in %	Min	Max	Mean	Std. Dev.
DRC	2094	0	0%	0	10%	3.94%	1.66%
DRB	2094	0	0%	0	10%	2.99%	1.81%

Appendix 5.1.6: Descriptive statistics of methods on data-level – categorical variables

Variable category	Frequency (full sample: 2094)	In %	cumulative
Outcome_measure			
- LYS	1319	62.99%	62.99%
- QALYs	775	37.01%	100%
Elicitation			
- N.a. (LYS)	1319	62.99%	62.99%
- TTO	112	5.35%	68.34%
- EQ-5D	313	14.95%	83.29%
- Other choice based method	333	15.90%	99.19%
- Unclear (missing)	17	0.81%	100%
Population			
- N.a. (LYS)	1319	62.99%	62.99%
- Patient	215	10.27%	73.26%
- Population	474	22.64%	95.90%
- Unclear (missing)	86	4.11%	100%
Duration			
- < 5years	66	3.15%	3.15%
- 5 to <10 years	788	37.63%	40.78%
- 10 to <15 years	175	8.36%	49.14%
- 15 to <20 years	44	2.10%	51.24%
- 20 to <25 years	87	4.15%	55.39%
- > 25 years (lifetime)	814	38.87%	94.26%
- Unclear (missing)	120	5.73%	100%
Duration_short			
- < 10 years	854	40.78%	40.78%
- 10 to < 20 years	219	10.46%	51.24%
- >20 years	901	43.03%	94.27%
- Unclear (missing)	120	5.73%	100%
Extrapolation beyond follow up?			
- Yes	1942	92.74%	92.74%
- No	152	7.26%	100%
Horizon			
- < 5 years	17	0.81%	0.81%
- 5 to < 10 years	191	9.12%	9.93%
- 10 to < 15 years	330	15.76%	25.69%
- 15 to < 20 years	132	6.30%	32.00%
- 20 to < 25 years	91	4.35%	36.34%
- > 25 years (lifetime)	1333	63.66%	100%
Horizon_short			
- ≤ 10 years	208	9.93%	9.93%
- ≤ 20 years	462	22.06%	32.00%
- > 20 years	1424	68.00%	100%
Horizon_eq_duration			
- Yes	1328	63.42%	63.42%
- No	766	36.58%	100%
Perspective_reported			
- Health insurance (NHS)	1369	65.38%	65.38%
- Societal	214	10.22%	75.60%
- No perspective reported	511	24.40%	100%
Perspective_cost_concluded			
- Provider	20	0.96%	0.96%
- Health insurance (NHS)	1939	92.60%	93.55%
- Societal	135	6.45%	100%
Perspective_benefits_concluded			
- Patient	2094	100%	100%
Perspective_discrepancy			
- No	1459	69.68%	69.68%
- yes	635	30.32%	100%
Data_class (sensitivity analysis)			
- base case	1125	53.72%	53.72%

- efficiency of intervention	12	0.57%	54.30%
- CVD risk at baseline	49	2.34%	56.64%
- Cost (not intervention)	64	3.06%	59.69%
- Cost of intervention	86	4.11%	63.80%
- QALYs	72	3.44%	67.24%
- Treatm. duration / time horizon	288	13.75%	80.99%
- Discount rate	257	12.27%	93.27%
- Other sensitivity analysis	141	6.73%	100%
Basecase			
- Yes	1125	46.28%	46.28%
- No (sensitivity analysis)	969	53.72%	100%
Source_effectivenes data			
- Literature / meta analysis	652	31.14%	31.14%
- PLAC I/II	38	1.81%	32.95%
- CARE	88	4.20%	37.15%
- WOSCOPS	81	3.87%	41.02%
- 4S	509	24.31%	65.33%
- 4S and WOSCOPS	46	2.20%	67.53%
- EXCEL	120	5.73%	73.26%
- LIPID	23	1.10%	74.36%
- CURVES	12	0.57%	74.93%
- CARDS	28	1.34%	76.27%
- ASCOTT	6	0.29%	76.55%
- HPS	280	13.37%	89.92%
- SPARCL	8	0.38%	90.31%
- TNT	42	2.01%	92.31%
- LIPS	33	1.58%	93.89%
- IDEAL	24	1.15%	95.03%
- STELLAR	62	2.69%	97.99%
- CELL	8	0.38%	98.38%
- other	34	1.62%	100%
Barbieri_1			
- Type C	186	8.88%	8.88%
- Type CR	1033	49.33%	58.21%
- Type CU	113	5.40%	63.61%
- Type CRE	193	9.22%	72.83%
- Type CRU	513	24.50%	97.33%
- Type CREU	56	2.67%	100%
Barbieri_2			
- Type 1	186	8.88%	8.88%
- Type 2	1146	54.73%	63.61%
- Type 3	706	33.72%	97.33%
- Type 4	56	2.67%	100%

Appendix 5.1.7: Descriptive statistics of general study characteristics – continuous variables

Variable (in 2010 £-Sterling)	Obs.	missing	in %	Min	Max	Mean	Std. Dev.
Timing_cont*	67	0	0%	1988	2009	2000	5.7
* timing was also defined as a categorical variable (see table x.8 below)							

Appendix 5.1.8: Descriptive statistics of general study characteristics – categorical variables

Variable category	Frequency (full sample: 67)	In %	cumulative
Language			
- English	61	91.04%	91.04%
- German	6	8.96%	100%
Paper_origin			
- Australia	1	1.49%	1.49%
- Belgium	2	2.99%	4.48%
- Brazil	1	1.49%	5.97%
- Canada	13	19.40%	25.37%
- Finland	2	2.99%	28.36%
- Germany	6	8.96%	37.31%
- Hong Kong	1	1.49%	38.81%
- Japan	1	1.49%	40.30%
- Netherlands	3	4.48%	44.78%
- Spain	1	1.49%	47.27%
- Sweden	7	10.45%	56.72%
- Switzerland	1	1.49%	58.21%
- UK	6	8.96%	67.16%
- UK (England / Wales)	6	8.96%	76.12%
- UK (Scotland)	1	1.49%	77.61%
- USA	12	17.91%	95.52%
- Scandinavian countries	1	1.49%	97.01%
- North America (USA / Canada)	1	1.49%	98.51%
- Multi_country	1	1.49%	100%
Author_group_long			
- No relationships	18	26.87%	26.87%
- Group 1	4	5.97%	32.84%
- Group 2	5	7.46%	40.30%
- Group 3	8	11.94%	52.24%
- Group 4	4	5.97%	58.21%
- Group 5	7	10.45%	68.66%
- Group 6	5	7.46%	76.12%
- Group 7	3	4.48%	80.60%
- Group 8	3	4.48%	85.07%
- Group 9	3	4.48%	89.55%
- Group 10	3	4.48%	94.03%
- Group 11	2	2.99%	97.01%
- Group 12	2	2.99%	100%
Author_group_short			
- No relationships	18	26.87%	26.87%
- Group 1	29	43.28%	70.15%
- Group 2	8	11.94%	82.09%
- Group 3	3	4.48%	86.57%
- Group 4	2	2.99%	89.55%
- Group 5	3	4.48%	94.03%
- Group 6	2	2.99%	97.01%
- Group 7	2	2.99%	100%
Timing_cat			
- 1988	2	2.99%	2.99%
- 1990	1	1.49%	4.48%
- 1991	1	1.49%	5.97%
- 1992	1	1.49%	7.46%
- 1993	2	2.99%	10.45%
- 1995	7	10.45%	20.90%
- 1996	8	11.94%	32.84%
- 1997	4	5.97%	38.81%
- 1998	6	8.96%	47.76%
- 1999	1	1.49%	49.25%
- 2000	1	1.49%	50.75%
- 2002	4	5.97%	56.72%
- 2003	2	2.99%	59.70%
- 2004	2	2.99%	62.69%
- 2005	9	13.43%	76.12%
- 2006	3	4.48%	80.60%
- 2007	7	10.45%	91.04%

- 2008	3	4.48%	95.52%
- 2009	3	4.48%	100%
Multinational			
- No (single country study)	61	91.04%	91.04%
- Yes (multinational study)	6	8.96%	100%
Funding_institution			
- Industry	39	58.21%	
- RC/Government / University	11	16.42%	
- Unclear	17	25.37%	
Funding_manufacturer			
- No manufacturer sponsoring	11	16.42%	
- Bristol Myers Squibb	7	10.45%	
- MERCK	12	17.91%	
- Pfizer	13	19.40%	
- Other (Sandoz / AZ / Novartis)	9	13.43%	
- unclear	15	22.39%	

Appendix 5.1.9: Descriptive statistics of methods variables on study-level – categorical variables

Variable category	**Frequency** (full sample: 67)	**In %**	**cumulative**
General_design			
- primary modelling	6	8.96%	8.96%
- secondary modelling	61	91.04%	100%
Primary_design			
- n.a. (primary modelling)	61	91.04%	91.04%
- RCT	6	8.96%	100%
Secondary_design			
- N.a. (primary modelling)	6	8.96%	
- Markov model	41	61.19%	
- Decision tree	7	10.45%	
- Other	13	19.40%	
Effect_calculation			
- CHD risk reduction	41	61.19%	61.19%
- Cholesterol reduction	26	38.81%	100%
Infl_adjustent			
- n.a.	12	17.91%	
- yes	18	26.87%	
- no	2	2.99%	
- unclear (missing)	35	52.24%	
Adjustment_method			
- n.a.	12	17.91%	
- simple consumer price index	8	11.94%	
- healthcare component of CPI	10	14.93%	
- no adjustment though indicated	2	2.99%	
- unclear (missing)	35	52.24%	
Currency_conversion			
- no	52	77.61%	77.61%
- yes	15	22.39%	100%
Conversion_method			
- n.a.	52	77.61%	
- exchange rates	11	16.42%	
- unclear	4	5.97%	
Scope			
- CAD	18	26.87%	
- CAD and CD	35	52.24%	
- CAD, CD, and PAD	11	16.42%	
- unclear	3	4.48%	

Appendix 5.1.10: Descriptive statistics of study quality indicators – continuous variables

Variable (in 2010 £-Sterling)	Obs.	missing	in %	Min	Max	Mean	Std. Dev.
QHES cont (a)	67	0	0%	27.00	100	59.36	16.33
QHES cont (b)	67	0	0%	40.83	100	69.32	13.89

Appendix 5.1.11: Descriptive statistics of study quality indicators– categorical variables

Variable category	Frequency (full sample: 2094)	In %	cumulative
QHES_cat (a)			
- up to 40 pts	7	10.45%	10.45%
- up to 60 pts	31	46.27%	56.72%
- up to 80 pts	22	32.84%	89.55%
- up to 100 pts	7	10.45%	100%
QHES_cat (b)			
- up to 60 pts	20	29.85%	29.85%
- up to 80 pts	40	59.70%	89.55%
- up to 100 pts	7	10.45%	100%

Appendix 5.2: Logit models to determine candidate variables for regression based missing data imputation.

Appendix 5.2.1: Logit model for predicting missingness in TCL (d_tcl)

Obs	880
LR chi2(14)	155.99
Prob > chi2	0.0000
Pseudo R2	0.8508
Log likelihood =	-13.672772

Explanatory	Coefficient	Standard Error	Z	P>|z|	95% CI lower	95%CI upper
INMB	.0001087	.0001675	0.65	0.516	-.0002196	.000437
incr_cost	.0001818	.0002351	0.77	0.439	-.000279	.0006426
incr_effect*	0	(omitted)				
ldl	2.848177	1.430478	1.99	0.046	.0444929	5.651862
Hypert	-.8160491	3.089475	-0.26	0.792	-6.871309	5.239211
Smokers	2.205601	2.481486	0.89	0.374	-2.658022	7.069224
Diab	5.205515	1.09153	4.77	0.000	3.066156	7.344874
Risk_cat	-.0364881	.0190707	-1.91	0.056	-.073866	.0008898
age_cat	9.040136	3.042482	2.97	0.003	3.076982	15.00329
gender_cat	-13.5595	4.344545	-3.12	0.002	-22.07466	-5.044351
intervention	-.1216281	.5415385	-0.22	0.822	-1.183024	.9397679
tdd_int	.1155115	.4737433	0.24	0.807	-.8130083	1.044031
source_effects	-1.871774	.6713426	-2.79	0.005	-3.187582	-.5559669
effect_loc	8.564003	2.342514	3.66	0.000	3.972761	13.15525
effect_calc	-32.63791	12.5298	-2.60	0.009	-57.19586	-8.07995
timing	-1.398225	.5252533	-2.66	0.008	-2.427703	-.3687474
scope	-.9660147	2.59417	-0.37	0.710	-6.050494	4.118465
_cons	2541.819	996.2151	2.55	0.011	589.2737	4494.365
*omitted because of collinearity						

Appendix 5.2.2: Logit model for predicting missingness in LDL (d_lcl)

Obs	795
LR chi2(14)	377.99
Prob > chi2	0.0000
Pseudo R2	0.8057
Log likelihood =	-45.573484

Explanatory	Coefficient	Standard Error	Z	P>|z|	95% CI lower	95%CI upper
INMB	-.0000233	.0000152	-1.53	0.125	-.0000531	6.47e-06
incr_cost	.0002016	.0000457	4.41	0.000	.0001121	.0002911
incr_effect	0	(omitted)				
hdl	7.713941	4.228739	1.82	0.068	-.5742357	16.00212
tcl	.822659	.4027575	2.04	0.041	.0332689	1.612049
hypert	-4.220478	1.537086	-2.75	0.006	-7.233111	-1.207846
smokers	6.507383	1.826005	3.56	0.000	2.928479	10.08629
diab	-.8737671	1.376445	-0.63	0.526	-3.57155	1.824015
risk_cat	-.4342001	.4188733	-1.04	0.300	-1.255177	.3867765
age_cat	1.4256	.5339265	2.67	0.008	.3791233	2.472077
gender_cat	1.64863	.7275759	2.27	0.023	.2226078	3.074653
intervention	.7144262	.2901476	2.46	0.014	.1457474	1.283105
tdd_int	-2.500186	.7446502	-3.36	0.001	-3.959673	-1.040698
source_effects	-.367101	.146308	-2.51	0.012	-.6538594	-.0803425
effect_loc	-.8106235	.2784211	-2.91	0.004	-1.356319	-.2649281
effect_calc	-1.863518	1.742071	-1.07	0.285	-5.277914	1.550878
Timing	.1152788	.0088053	13.09	0.000	.0980207	.1325368
_cons	4.278564	9.612571	0.45	0.656	-14.56173	23.11886
*omitted because of collinearity						

Appendix 5.2.3: Logit model for predicting missingness in HDL (d_hdl)

Obs	1156
LR chi2(14)	190.63
Prob > chi2	0.0000
Pseudo R2	0.4929
Log likelihood =	-98.063952

Explanatory	Coefficient	Standard Error	Z	P>|z|	95% CI lower	95%CI upper
INMB	-.0004334	.0000807	-5.37	0.000	-.0005915	-.0002753
incr_cost	-.0005195	.0000951	-5.46	0.000	-.0007058	-.0003332
incr_effect	0	(omitted)				
tcl	.2634234	.1850258	1.42	0.155	-.0992205	.6260673
age_cat	-.8988485	.2469747	-3.64	0.000	-1.38291	-.4147869
gender_cat	-1.088738	.3228125	-3.37	0.001	-1.721439	-.4560372
tdd_int	-.4200735	.1622744	-2.59	0.010	-.7381255	-.1020215
effect_loc	-.0796008	.0674007	-1.18	0.238	-.2117039	.0525022
timing	-.1552515	.0659099	-2.36	0.018	-.2844325	-.0260705
_cons	315.0081	133.1114	2.37	0.018	54.11462	575.9017
*omitted because of collinearity						

Appendix 5.2.4: Logit model for predicting missingness in hypertension status (d_hypert)

Obs	870
LR chi2(14)	541.44
Prob > chi2	0.0000
Pseudo R2	0.6935
Log likelihood =	-119.65464

Explanatory	Coefficient	Standard Error	Z	P>|z|	95% CI lower	95%CI upper
INMB	.0000686	.0000707	0.97	0.332	-.00007	.0002071
incr_cost	-.0006677	.0001274	-5.24	0.000	-.0009173	-.0004181
incr_effect	0	(omitted)				
tcl	-8.352446	1.33393	-6.26	0.000	-10.9669	-5.737992
hdl	27.93644	4.224154	6.61	0.000	19.65725	36.21563
ldl	3.970261	1.108581	3.58	0.000	1.797483	6.143039
smokers	-1.13448	.7142796	-1.59	0.112	-2.534442	.2654823
sbp	.0116166	.0252213	0.46	0.645	-.0378163	.0610495
diab	.3329223	.7134978	0.47	0.641	-1.065508	1.731352
risk_cat	.4547893	.18705	2.43	0.015	.0881781	.8214004
age_cat	.5698994	.5308344	1.07	0.283	-.4705168	1.610316
gender_cat	6.137617	1.015317	6.05	0.000	4.147632	8.127602
intervention	1.503188	.3332428	4.51	0.000	.850044	2.156332
tdd_int	1.498864	.2670696	5.61	0.000	.9754172	2.022311
source_effects	.5667746	.097632	5.81	0.000	.3754194	.7581297
effect_loc	-1.078352	.1925064	-5.60	0.000	-1.455658	-.7010466
timing	-2.439659	.3163449	-7.71	0.000	-3.059683	-1.819634
_cons	4877.341	631.6731	7.72	0.000	3639.284	6115.397
*omitted because of collinearity						

Appendix 5.2.5: Logit model for predicting missingness in systolic blood pressure (d_sbp)

Obs	1156
LR chi2(14)	190.94
Prob > chi2	0.0000
Pseudo R2	0.4937
Log likelihood =	-97.910332

Explanatory	Coefficient	Standard Error	Z	P>|z|	95% CI lower	95%CI upper
INMB	-.0004355	.0000816	-5.34	0.000	-.0005954	-.0002756
incr_cost	-.000507	.0000951	-5.33	0.000	-.0006933	-.0003207
incr_effect	0	(omitted)				
tcl	.2822932	.188329	1.50	0.134	-.0868249	.6514113
age_cat	-.8940499	.25166	-3.55	0.000	-1.387294	-.4008054
gender_cat	-1.088431	.3247331	-3.35	0.001	-1.724896	-.4519657
tdd_int	-.4445926	.1783034	-2.49	0.013	-.7940608	-.0951243
source_effects	-.0372382	.0689759	-0.54	0.589	-.1724285	.0979521
effect_loc	-.1243427	.1103292	-1.13	0.260	-.3405841	.0918986
timing	-.134513	.0745366	-1.80	0.071	-.280602	.011576
_cons	274.882	149.112	1.84	0.065	-17.37211	567.1362
*omitted because of collinearity						

Appendix 5.2.6: Logit model for predicting missingness in smoking status (d_sbp)

Obs	1156
LR chi2(14)	178.38
Prob > chi2	0.0000
Pseudo R2	0.4612
Log likelihood =	-104.19084

Explanatory	Coefficient	Standard Error	Z	P>|z|	95% CI lower	95%CI upper
INMB	-.0004828	.0000867	-5.57	0.000	-.0006527	-.000313
incr_cost	-.0005183	.0000969	-5.35	0.000	-.0007082	-.0003284
incr_effect	0	(omitted)				
tcl	.2836354	.1835382	1.55	0.122	-.0760929	.6433636
gender_cat	-1.1206	.3096537	-3.62	0.000	-1.72751	-.5136898
source_effects	-.1210404	.0460367	-2.63	0.009	-.2112707	-.0308101
effect_loc	-.1490165	.082239	-1.81	0.070	-.310202	.012169
timing	.0042349	.0560054	0.08	0.940	-.1055337	.1140035
_cons	-4.790422	112.0651	-0.04	0.966	-224.4341	214.8532
*omitted because of collinearity						

Appendix 5.2.7: Logit model for predicting missingness in diabetes status (d_diab)

Obs	1156
LR chi2(14)	190.94
Prob > chi2	0.0000
Pseudo R2	0.4937
Log likelihood =	-97.910332

Explanatory	Coefficient	Standard Error	Z	P>|z|	95% CI lower	95%CI upper
INMB	-.0004355	.0000816	-5.34	0.000	-.0005954	-.0002756
incr_cost	-.000507	.0000951	-5.33	0.000	-.0006933	-.0003207
incr_effect	0	(omitted)				
tcl	.2822932	.188329	1.50	0.134	-.0868249	.6514113
age_cat	-.8940499	.25166	-3.55	0.000	-1.387294	-.4008054
gender_cat	-1.088431	.3247331	-3.35	0.001	-1.724896	-.4519657
tdd_int	-.4445926	.1783034	-2.49	0.013	-.7940608	-.0951243
source_effects	-.0372382	.0689759	-0.54	0.589	-.1724285	.0979521
effect_loc	-.1243427	.1103292	-1.13	0.260	-.3405841	.0918986
timing	-.134513	.0745366	-1.80	0.071	-.280602	.011576
_cons	274.882	149.112	1.84	0.065	-17.37211	567.1362
*omitted because of collinearity						

Appendix 5.2.8: Logit model for predicting missingness in annual drug cost of the intervention (d_cost_int)

Obs	1023
LR chi2(14)	600.09
Prob > chi2	0.0000
Pseudo R2	0.8593
Log likelihood =	-49.116343

Explanatory	Coefficient	Standard Error	Z	P>|z|	95% CI lower	95%CI upper
tcl	.726719	.7577476	0.96	0.338	-.7584391	2.211877
country_id	-.2816622	.0729227	-3.86	0.000	-.4245881	-.1387362
source_effects	-.3320263	.323819	-1.03	0.305	-.9666998	.3026472
drc	-20.06777	25.96249	-0.77	0.440	-70.95332	30.81779
timing	.3121932	.2199898	1.42	0.156	-.1189788	.7433653
duration	1.584883	.6175211	2.57	0.010	.3745642	2.795202
horizon	-2.46523	.4960683	-4.97	0.000	-3.437506	-1.492954
cur_conv	.8132378	6.585694	0.12	0.902	-12.09449	13.72096
conv_method	0	(omitted)*				
data_class	-.1335485	.0814788	-1.64	0.101	-.2932441	.0261471
incr_cost	-.0002244	.0000791	-2.84	0.005	-.0003794	-.0000693
INMB	-.000273	.0000871	-3.13	0.002	-.0004437	-.0001023
intervention	-3.594461	.822772	-4.37	0.000	-5.207064	-1.981857
tdd_int	-.1106023	.7985756	-0.14	0.890	-1.675782	1.454577
multinational	7.047884	6.490862	1.09	0.278	-5.673972	19.76974
_cons	-619.529	438.5537	-1.41	0.158	-1479.078	240.0204
*omitted because of collinearity						

Appendix 5.2.9: Logit model for predicting missingness in unit cost of the intervention (d_unitcost_int)

Obs	1053
LR chi2(14)	92.62
Prob > chi2	0.0000
Pseudo R2	0.3397
Log likelihood =	-90.004987

Explanatory	Coefficient	Standard Error	Z	P>|z|	95% CI lower	95%CI upper
tcl	.4250457	.2163829	1.96	0.049	.000943	.8491484
Intervention	.6177555	.2122761	2.91	0.004	.2017021	1.033809
country_id	-.1241907	.0342685	-3.62	0.000	-.1913558	-.0570257
drc	-46.91049	19.82648	-2.37	0.018	-85.76969	-8.0513
timing	-.1787341	.0575271	-3.11	0.002	-.2914852	-.0659831
duration	.9724997	.242166	4.02	0.000	.497863	1.447136
cur_conv	.8163021	.7922189	1.03	0.303	-.7364184	2.369023
conv_method	0	(omitted)				
data_class	-.2238029	.097233	-2.30	0.021	-.414376	-.0332297
incr_cost	-.0003556	.0001011	-3.52	0.000	-.0005538	-.0001574
INMB	-.000019	.0000203	-0.93	0.350	-.0000589	.0000209
_cons	353.1313	115.2612	3.06	0.002	127.2236	579.0391
*omitted because of collinearity						

Appendix 5.2.10: Logit model for predicting missingness in the annual drug cost of the comparator (d_cost_comp)

Due to very few missing values, logit models did not converge for d_cost_comp. Therefore, imputation models for the cost of the comparator are based on the same set of explanatory variables as the model for the intervention cost, with the explanatory 'intervention' replaced by the explanatory 'comparator'.

Appendix 5.2.11: Logit model for predicting missingness in the unit cost of the comparator (d_unitcost_comp)

Due to very few missing values, logit models did not converge for d_unitcost_comp. Therefore, imputation models for the cost of the comparator are based on the same set of explanatory variables as the model for the intervention cost, with the explanatory 'intervention' replaced by the explanatory 'comparator'.

Appendix 5.3: Multiple correspondence analysis on 'horizon' 'extrapol' and 'hor_eq_dur'

Number of obs: 2094 / Total inertia: 0.26199 / Number of axes: 2

	Principal Inertia	Percent	Cumulative
Dimension 1	0.20677	78.92%	78.92%
Dimension 2	0.00654	2.5%	81.42%
Dimension 3	$6.24e^{-32}$	0.00%	81.42%
Dimension 4	$6.93e^{-33}$	0.00%	81.42%

	Overall			Dimension 1			Dimension 2		
Categories	**Mass**	**Quality**	**%inertia**	**Coord**	**sqcorr**	**contrib**	**Coord**	**sqcorr**	**contrib**
horizon									
< 5 years	0.003	0.773	0.069	4.960	0.762	0.067	3.402	0.011	0.031
5 to 10 years	0.030	0.799	0.336	3.342	0.798	0.340	0.505	0.001	0.008
10 to 15 years	0.053	0.827	0.008	-0.370	0.698	0.007	0.893	0.128	0.042
15 to 20 years	0.021	0.774	0.019	-0.747	0.493	0.012	3.169	0.281	0.211
20 to 25 years	0.014	0.553	0.015	0.074	0.004	0.000	-4.689	0.548	0.319
> 25 years	0.212	0.764	0.033	-0.382	0.746	0.031	-0.331	0.018	0.023
extrapolation									
No	0.024	0.797	0.399	4.072	0.794	0.401	1.297	0.003	0.041
Yes	0.309	0.797	0.031	-0.319	0.794	0.031	-0.102	0.003	0.003
Hor_eq_dur									
No	0.211	1.052	0.033	0.439	0.964	0.041	-0.747	0.088	0.118
Yes	0.122	1.052	0.059	-0.761	0.964	0.071	1.295	0.088	0.204

Appendix 5.4: Gradually building up a random intercepts model with data and study-level covariates

Appendix 5.4.1: Random intercepts model with patient and disease characteristics

	Univariate model	Bivariate model	
	INMB (2010 £ Sterling)	**ΔC/100** (2010 £ Sterling)	**ΔE**
Fixed part:			
N (countries)	18	18	18
N (studies)	67	67	67
N (data)	2094	2094	2094
Intercept (λ=£30.000)	-25176	725	-0.621
TCL (SE)	4153 (588)***	-28.32 (3.50)***	0.037 (0.018)**
HDL (SE)	-26588 (10341)**	102.5 (55.57)*	-0.629 (0.342)*
SBP (SE)	706 (37.05)***	-3.31 (0.22)***	0.012 (0.001)***
Diabetes (SE)	15261 (1750)***	-24.51 (9.23)***	0.421 (0.005)***
Age_cat (SE)			
<45	Omitted	Omitted	Omitted
46-55	98.73 (1162)	-93.92 (6.72)***	-0.311 (0.037)***
56-65	-5976 (1131)***	-133.2 (6.69)***	-0.644 (0.036)***
66-75	-15240 (1206)***	-167.4 (7.11)***	-1.070 (0.038)***
>75	-6692 (2047)***	-130.3 (11.9)***	-0.663 (0.065)***
Unclear	5207 (8436)	-97.02 (51.2)*	-0.145 (0.280)***
Gender (SE)			
Female	Omitted	Omitted	Omitted
Male	10330 (874)***	-34.0 (5.07)***	0.228 (0.028)***
Mixed sample	6839 (4122)*	-57.6 (25.3)**	0.074 (0.131)***
CVD_history (SE)			
No	Omitted	--	Omitted
Yes	7090 (1589)***		0.219 (0.049)***
Mixed sample	9511 (7656)		0.260 (0.254)
Random part:			
σ^2_{u0j} **(Country)**	6274573	2452	0.117
σ^2_{u0k} **(Study)**	299288384	8661	0.371
σ^2_{e0} **(Data)**	204568400	7077	0.201
VPC - Country	1.23%	13.48%	16.98%
VPC – Study	58.67%	47.61%	53.85%
VPC - data	40.10%	38.91%	29.17%
DIC (benchmark)	46013 (46749)	26993 (28735)	

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level

Appendix 5.4.2: Random intercepts model with patient/disease and intervention/ comparator characteristics

	Univariate model	Bivariate model	
	INMB (2010 £ Sterling)	**ΔC/100** (2010 £ Sterling)	**ΔE**
Fixed part:			
N (countries)	18	18	18
N (studies)	67	67	67
N (data)	2094	2094	2094
Intercept (λ=£30.000)	-24788	701	-0.549
TCL (SE)	4236 (590)***	-30.62 (3.31)***	0.035 (0.018)*
HDL (SE)	-26962 (10506)**	132.23 (53.66)**	-0.567 (0.334)*
SBP (SE)	705 (37.37)***	-3.29 (0.21)***	0.012 (0.001)***
Diabetes (SE)	14941 (1753)***	-20.32 (8.89)**	0.424 (0.055)***
Age_cat (SE)			
<45	Omitted	Omitted	Omitted
46-55	71.28 (1159)	-94.27 (6.62)***	-0.312 (0.036)***
56-65	-5986 (1138)***	-133.8 (6.36)***	-0.645 (0.035)***
66-75	-15263 (1206)***	-167.9 (6.80)***	-1.070 (0.038)***
>75	-6724 (2032)***	-131.6 (11.6)***	-0.664 (0.064)***
Unclear	5369 (8531)	-100.2 (44.69)**	-0.184 (0.276)
Gender (SE)			
Female	Omitted	Omitted	Omitted
Male	10344 (876)***	-33.89 (4.93)***	0.228 (0.027)***
Mixed sample	6772 (2032)***	-48.48 (21.14)**	0.047 (0.130)
CVD_history (SE)			
No	Omitted	--	Omitted
Yes	7032 (1606)***		0.225 (0.050)***
Mixed sample	10196 (7573)		0.2323 (0.262)
Cost_intervention	-8.932 (1.863)***	0.133 (0.010)***	--
Cost_comparator	--	-0.146 (0.024)***	--
Active_comparator			
No (doing nothing)	Omitted	--	Omitted
Yes (statin)	-4729 (2468)*		-(0.313 (0.078)***
Random part:			
σ^2_{u0j} **(Country)**	6801933	2916	0.122
σ^2_{u0k} **(Study)**	284972576	6543	0.345
σ^2_{e0} **(Data)**	202350848	6574	0.200
VPC - Country	1.38%	18.19%	18.29%
VPC – Study	57.67%	40.81%	51.72%
VPC - data	40.95%	41.00%	29.99%
DIC (benchmark)	45990 (46013)	26840 (26993)	

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level

Appendix 5.4.3: Random intercepts model with patient/disease characteristics, intervention/comparator characteristics and methodological characteristics on data-level.

	Univariate model	Bivariate model	
	INMB (2010 £ Sterling)	**ΔC/100** (2010 £ Sterling)	**ΔE**
Fixed part:			
N (countries)	18	18	18
N (studies)	67	67	67
N (data)	2094	2094	2094
Intercept (λ=£30.000)	-36030	789	-0.881
TCL (SE)	4447 (571)***	-33.47 (3.10)***	0.041 (0.019)**
HDL (SE)	-27145 (10219)***	125.32 (51.69)**	-0.585 (0.330)*
SBP (SE)	707 (35.76)***	-3.36 (0.19)***	0.012 (0.001)***
Diabetes (SE)	14525 (1700)***	-21.48 (8.27)***	0.409 (0.054)***
Age_cat (SE)			
<45	Omitted	Omitted	Omitted
46-55	171.8 (1117)	-95.5 (6.01)***	-0.311 (0.036)***
56-65	-5858 (1095)***	-134.7 (5.99)***	-0.643 (0.035)***
66-75	-15155 (1163)***	-169.2 (6.31)***	-1.067 (0.038)***
>75	-6450 (1973)***	-131.3 (10.51)***	-0.653 (0.063)***
Unclear	2810 (8729)	-111.2 (47.52)**	-0.299 (0.286)
Gender (SE)			
Female	Omitted	Omitted	Omitted
Male	10395 (849)***	-34.62 (4.50)***	0.227 (0.027)***
Mixed sample	6851 (4114)*	-54.76 (23.95)**	0.028 (0.140)
CVD_history (SE)			
No	Omitted	--	Omitted
Yes	5788 (1586)***		0.195 (0.050)***
Mixed sample	7541 (7474)		0.230 (0.249)
Cost_intervention	-8.64 (1.81)***	0.128 (0.010)***	--
Cost_comparator	--	-0.143 (0.023)***	--
Active_comparator			
No (doing nothing)	Omitted	--	Omitted
Yes (statin)	-5040 (2425)**		-(0.313 (0.078)***
DRB	-131318 (29557)***	--	-3.674 (0.904)***
Persp_cost_concl.			
Health insurance (NHS)	Omitted	Omitted	--
Provider	35105 (4421)***	-397.6 (23.36)***	
Societal	17744 (3636)***	-175.13 (20.42)***	
Horizon			
< 20 years	Omitted	--	Omitted
>20 years (lifetime)	4394 (1383)***		0.156 (0.044)***
Duration equals horizon			
yes	Omitted	--	Omitted
No (treatment duration< horizon)	9580 (2557)***		0.309 (0.084)***
Base case			
Yes	--	Omitted	Omitted
No		11.59 (5.01)**	0.064 (0.030)**
Barbieri_score_2			
Type 1		Omitted	
Type 2	--	-33.42 (15.37)**	--
Type 3		-62.43 (30.76)**	
Type 4		-93.21 (33.19)***	
Random part:			
σ^2_{u0j} **(Country)**	5436921	3366	0.131
σ^2_{u0k} **(Study)**	306533088	10845	0.334
σ^2_{e0} **(Data)**	188981600	5548	0.196
VPC - Country	1.09%	17.04%	19.82%
VPC – Study	61.19%	54.89%	50.53%
VPC - data	37.72%	28.08%	29.65%
DIC	45846	26425	
(benchmark)	(45990)	(26840)	

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level

Appendix 5.4.4: Random intercepts model fully specified on data-level and general study characteristics on study-level.

	Univariate model	Bivariate model	
	INMB (2010 £ Sterling)	**ΔC/100** (2010 £ Sterling)	**ΔE**
Fixed part:			
N (countries)	18	18	18
N (studies)	67	67	67
N (data)	2094	2094	2094
Intercept (λ=£30.000)	-43548	770	-1.315
TCL (SE)	4333 (565)***	-33.90 (3.14)***	0.040 (0.018)**
HDL (SE)	-27193 (9616)***	130.73 (52.47)**	-0.545 (0.308)*
SBP (SE)	704 (35.76)***	-3.36 (0.19)***	0.012 (0.001)***
Diabetes (SE)	14715 (1677)***	-21.80 (8.27)***	0.417 (0.053)***
Age_cat (SE)			
<45	Omitted	Omitted	Omitted
46-55	284.5 (1119)	-95.62 (6.04)***	-0.307 (0.036)***
56-65	-5798 (1096)***	-134.5 (5.95)***	-0.638 (0.035)***
66-75	-15144 (1158)***	-169.3 (6.21)***	-1.065 (0.038)***
>75	-6548 (1963)***	-131.6 (10.70)***	-0.651 (0.063)***
Unclear	-1121 (7663)	-117.7 (43.67)***	-0.424 (0.240)*
Gender (SE)			
Female	Omitted	Omitted	Omitted
Male	10434 (837.1)***	-34.62 (4.50)***	0.231 (0.027)***
Mixed sample	13315 (3721)***	-39.31 (24.91)	0.251 (0.116)**
CVD_history (SE)			
No	Omitted	--	Omitted
Yes	6481 (1555)***		0.225 (0.050)***
Mixed sample	7085 (6512)		0.349 (0.199)*
Cost_intervention	-8.97 (1.78)***	0.127 (0.010)***	--
Cost_comparator	--	-0.138 (0.023)***	--
Active_comparator			
No (doing nothing)	Omitted	--	Omitted
Yes (statin)	-4483 (2334)*		-0.296 (0.073)***
DRB	-131950 (29615)***	--	-3.372 (0.878)***
Persp_cost_concl.			
Health insurance (NHS)	Omitted	Omitted	--
Provider	33077 (4423)***	-399.6 (23.22)***	
Societal	15238 (3600)***	-175.9 (20.76)***	
Horizon			
< 20 years	Omitted	--	Omitted
>20 years (lifetime)	3748 (1358)***		0.153 (0.043)***
Duration=horizon			
yes	Omitted	--	Omitted
No (treatment duration< horizon)	11093 (2423)***		0.326 (0.076)***
Base case			
Yes	--	Omitted	Omitted
No		11.58 (5.03)**	0.067 (0.029)**
Barbieri_score_2			
Type 1		Omitted	
Type 2	--	-35.19 (15.69)**	--
Type 3		-62.06 (32.57)*	
Type 4		-93.38 (34.99)***	
Author_Grover			
No	Omitted	Omitted	Omitted
Yes	33704 (5909)***	161.7 (45.71)***	1.590 (0.192)***
Random part:			
σ^2_{u0j} **(Country)**	974342	4041	0.074
σ^2_{u0k} **(Study)**	194371968	8876	0.155
σ^2_{e0} **(Data)**	189051856	5547	0.196
VPC - Country	0.25%	21.89%	17.41%
VPC – Study	50.57%	48.07%	36.47%
VPC - data	49.18%	30.04%	46.12%
DIC (benchmark)	45844 (45846)	26423 (26425)	

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level

Appendix 5.5: Individually testing country-level covariates in the three-level bivariate random intercepts model (reduced dataset)

Bivariate model				
	Raw Mean (SD) / Proportion (%)	**ΔC/100** (2010 £ Sterling)	**ΔE**	**DIC (**Benchmark: 22596 (%-change))
GDP	37506 (3693)	0.001 (0.001)	0.000 (0.000)	22598 (0.009%)
THE_GDP	10.83 (2.23)	0.346 (1.471)	-0.024 (0.009)***	22595 (-0.004%)
GOV_EXP_THE	73.48 (11.77)	-0.180 (0.280)	-0.003 (0.002)	22599 (0.013%)
PRIV_EXP_THE	25.73 (11.99)	0.213 (0.275)	-0.003 (0.002)	22600 (0.018%)
SOCSEC_GGE+	12.78 (23.06)	-0.096 (0.159)	0.001 (0.001)	22602 (0.027%)
OOP_PRIV_EXP+	55.59 (15.74)	0.061 (0.205)	-0.000 (0.001)	22602 (0.022%)
CVD_POLICY No Yes Unclear	 4 (23.52%) 12 (70.59%) 1 (5.88%)	 Omitted 37.409 (25.48) 2.563 (23.94)	 Omitted 0.111 (0.119) -0.077 (0.117)	22601 (0.000%)
GPs	26.54 (4.91)	-1.143 (1.411)	0.008 (0.007)	22601(-0.022%)
NURSES	96.37 (23.59)	0.091 (0.145)	0.000 (0.001)	22602 (0.027%)
PHARMACISTS	7.35 (1.59)	3.079 (2.249)	0.034 (0.011)***	22991 (-0.022%)
BEDS	36.24 (8.33)	-0.053 (0.461)	0.001 (0.002)	22602 (-0.027%)
AGE	39.53 (1.53)	-2.063 (2.149)	0.050 (0.013)***	22601 (-0.022%)
URBAN	85.03 (5.45)	-0.034 (0.779)	-0.003 (0.004)	22602 (0.027%)
LIFE_EXPECTANCY	80.18 (0.70)	-8.164 (5.399)	0.020 (0.028)	22597 (0.004%)
CVD_DEATH	134.98 (12.84)	0.399 (0.256)	0.000 (0.001)	22599 (0.013%)
BMI_25+	60.53 (5.20)	0.744 (0.931)	0.009 (0.005)*	22601 (-0.009%)
BMI_30+	26.52 (3.43)	0.907 (1.302)	0.017 (0.008)**	22599 (-0.018%)
MEAN_BMI+	27.37 (0.60)	3.915 (7.537)	0.114 (0.042)***	22597 (-0.027%)
TCL_6.2+	20.72 (3.33)	-0.655 (1.013)	-0.006 (0.006)	22601 (0.022%)
MEAN_TCL+	5.33 (0.13)	-22.124 (27.940)	0.112 (0.156)	22601 (-0.022%)
SBP_140+	39.92 (4.83)	-0.084 (0.818)	0.005 (0.005)	22602 (0.027%)
MEAN_SBP+	126.99 (3.62)	-0.156 (1.111)	-0.008 (0.006)	22601 (0.022%)
GLUCOSE_7+	9.58 (1.65)	2.106 (2.249)	0.033 (0.013)**	22596 (-0.000%)
MEAN_GLUCOSE+	5.52 (0.12)	29.703 (28.329)	0.401 (0.161)**	22596 (-0.000%)

* significant at the 10%-level
** significant at the 5%-level
*** significant at the 1%-level

\+ Eight data points referring to the special administrative region Hong Kong have been dropped due to country-level data missing for this geographic domain

Appendix 5.6: Bivariate random intercepts model fully specified on data, study and country-level (reduced dataset)

	ΔC/100 (2010 £ Sterling)	**ΔE**
N (countries)	17	17
N (studies)	61	61
N (data)	1806	1806
Intercept (λ=£30.000)	261.65	-0.017
TCL (SE)	-32.91 (3.30)***	0.067 (0.016)***
HDL (SE)	233.30 (61.89)***	-0.842 (0.311)***
SBP (SE)	-3.45 (0.21)***	0.012 (0.001)***
Diabetes (SE)	-51.40 (14.67)***	0.673 (0.073)***
Age_cat (SE)		
<45	Omitted	Omitted
46-55	-93.04 (6.75)***	-0.258 (0.034)***
56-65	-128.65 (6.59)***	-0.523 (0.033)***
66-75	-163.23 (7.041)***	-0.853 (0.036)***
>75	-127.41 (11.43)***	-0.568 (0.057)***
Unclear	-114.59 (44.59)**	-0.416 (0.240)*
Gender (SE)		
Female	Omitted	Omitted
Male	-32.55 (5.23)***	0.145 (0.026)***
Mixed sample	-25.02 (28.37)	0.123 (0.130)
CVD_history (SE)		
No	--	Omitted
Yes		0.371 (0.057)***
Mixed sample		0.321 (0.201)
Cost_intervention	0.119 (0.011)***	--
Cost_comparator	-0.131 (0.024)***	--
Active_comparator		
No (doing nothing)	--	Omitted
Yes (statin)		-0.306 (0.069)***
DRB	--	-3.920 (0.843)***
Persp_cost_concl.		
Health insurance (NHS)	Omitted	--
Provider	-403.34 (24.11)***	
Societal	-182.05 (21.27)***	
Horizon		
< 20 years	--	Omitted
>20 years (lifetime)		0.197 (0.060)***
Duration=horizon		
yes	--	Omitted
No (treatment duration< horizon)		0.406 (0.072)***
Base case		
Yes	Omitted	Omitted
No	9.62 (5.94)	0.099 (0.029)***
Barbieri_score_2		
Type 1	Omitted	
Type 2	-37.41 (39.85)	--
Type 3	-60.23 (44.82)	
Type 4	-93.04 (46.48)**	
Author_Grover		
No	Omitted	Omitted
Yes	170.07 (62.28)***	1.439 (0.238)***
4S		
No	--	Omitted
yes		0.441 (0.209)
Scope		
CHD	Omitted	
CHD and stroke	-39.26 (9.30)***	--
CHD, stroke and PAD	-30.16 (37.59)	
unclear	-110.51 (72.56)	
GDP_CAPITA	-0.003 (0.001)***	--
GOV_EXP_THE	-1.029 (0.483)**	--
THE_GDP	--	0.062 (0.013)***
AGE_POPULATION	--	0.070 (0.018)***
Random part:		
σ^2_{u0j} **(Country)**	4010	0.086
σ^2_{u0k} **(Study)**	8688	0.165
σ^2_{e0} **(Data)**	6232	0.159
VPC - Country	23.45%	21.50%
VPC – Study	50.80%	41.25%
VPC - data	36.44%	39.75%
DIC (benchmark)	22571 (22596)	
* significant at the 10%-level / ** significant at the 5%-level / ***significant at the 1%-level		