An investigation of cost variation across health care settings and the implications for economic evaluation

Richard Grieve

Thesis submitted for the Degree of Doctor of Philosophy to the Faculty of Economics of the University of London

Department of Public Health and Policy
London School of Hygiene and Tropical Medicine
University of London

2005

Abstract

This thesis is concerned with the estimation of costs in economic evaluation. The thesis reviews the theoretical and applied literature on costing and highlights that studies generally ignore cost variation across health care settings. The thesis aims to assess why costs vary across health care settings, and the implications for economic evaluations.

The study uses microeconomic theory to pose hypotheses for cost variation across health care settings and uses a consistent methodology to collect costs across a range of health care settings. The analysis uses multilevel models (MLMs) to test hypotheses concerning cost variation. Statistical theory suggests that MLMs accommodate the hierarchical structure of the data and may therefore be more appropriate than ordinary least squares (OLS) models for identifying reasons for cost variation across settings. The use of MLMs and OLS models for analysing reasons for cost variation are compared. The OLS models find that both patient and higher-level covariates are associated with length of hospital stay (LOS) and total cost, but these models overestimate the precision of the higher-level variables. By contrast, the MLMs show that none of the higher-level variables are associated with LOS, and the national level of spending on health care is the only higher-level variable associated with total cost.

The empirical investigation also illustrates that using OLS regression analysis to report cost-effectiveness can lead to inaccurate estimates. By contrast, the MLMs recognise the structure of the data and accurately quantify mean incremental cost-effectiveness and the associated levels of uncertainty.

The thesis concludes that ignoring cost variation across health care settings can lead to inaccurate estimates of cost and cost-effectiveness. Basing decision-making on inaccurate information can move the allocation of health care resources away from the target of allocative efficiency. This thesis presents a methodology for improving the conduct of cost analyses that future economic evaluations can adopt.

Acknowledgements

I am deeply indebted to my supervisor Charles Normand who has provided me with immense support, both intellectual and practical. At each step along this journey, Charles offered wise thoughts always spiced with humour and encouragement, and proved an excellent mentor. I am also very grateful to John Hutton who gave me the confidence and the opportunity to start a PhD. Sincere thanks also go to Kara Hanson and Lilani Kumaranayake for their constructive advice on economic theory and econometrics. Both Kara and Lilani were generous with their time and always had imaginative thoughts to share. I also thank the other members of my advisory group, Diane Elbourne and Jan van der Meulen for their support and candid advice at each stage of the thesis.

I am very grateful to Simon Thompson and Richard Nixon at the MRC Biostatistics Unit for their help with the multilevel modelling, and to James Carpenter for his statistical advice and practical help with using MLwiN. Numerous friends at LSHTM and in the health economics community have helped me during the thesis. I would particularly like to thank Damian Walker, Alec Miners, Jon Karnon, Andrew Hutchings, Barbara McPake and Paul Jacklin for commenting on individual chapters. I am also grateful to both Nat Fenwick and Amanda Bristow for their help with references and proofreading.

Last but certainly not least, I thank George for providing a welcome distraction and Amanda for her love, fortitude and encouragement.

To Amanda.

Abbreviations

CEAC	Cost-effectiveness acceptability curve
CI	Confidence interval
COV	Coefficient of variation
DEA	Data envelopment analysis
DRG	Diagnosis-related group
GDP	Gross domestic product
GLM	Generalised linear model
GLMM	Generalised linear mixed model
HSR	Health services research
ICU	Intensive care unit
ICER	Incremental cost-effectiveness ratio
INB	Incremental net benefit
LOS	Length of stay
LRAC	Long-run average cost
LYG	Life-years gained
MLM	Multilevel model
NICE	National Institute of Clinical Excellence
OER	Official exchange rate
OLS	Ordinary least squares
NMB	Net monetary benefit
PPP	Purchasing power parity
QALY	Quality-adjusted life year
RCT	Randomised controlled trial
SD	Standard deviation
SE	Standard error
SFA	Stochastic frontier analysis
SRAC	Short-run average cost

Table of contents

Publications arising from this thesis

Grieve R, Porsdal V, Hutton J, Wolfe C (2000). A comparison of the cost-effectiveness of stroke care provided in London and Copenhagen. *Int J Technol Assess Health Care.* 16: 684-95.

Grieve R, Hutton J, Bhalla A, Rastenytë D, Ryglewicz D et al. (2001a). A comparison of the costs and survival of hospital-admitted stroke patients across Europe. *Stroke* 32: 1684-1691.

Grieve R, Dundas R, Beech R, Wolfe CDA (2001b). The development and use of a method to compare the costs of acute stroke across Europe. *Age and Ageing* 30: 67-72.

Grieve R, Nixon R, Thompson SG, Normand C (2005). Using multilevel models for assessing the variability of multinational resource use and cost data. *Health Econ* (in press).

List of Tables

List of Figures

List of appendices

Chapter 1: Introduction

1.0 Rationale for the thesis

The last decade has seen important developments in the use of economic evaluation in policy-making. Governments in Canada, Australia, the UK, the Netherlands, Sweden, Finland and Portugal now use economic evaluations to decide which health care technologies to provide (Maynard and Kanavos 2000). The number of published economic evaluations has grown rapidly (Pritchard 2004), and the use of these evaluations in decision-making can potentially lead to the more efficient use of scarce health care resources. However, this objective will only be achieved if these studies use appropriate methodologies for measuring outcomes and costs.

The methods used in economic evaluations have been severely criticised (Birch and Gafni 1992, Drummond et al. 1993, Hutton 1994, Gerard et al. 1999, Birch and Gafni 2002, Drummond and Sculpher 2005). In particular, the methods used in published studies for measuring and analysing costs have been shown to be inadequate (Graves et al. 2002, Barber and Thompson 1998). Unless these methods are improved, the use of economic evaluations will not improve the efficiency of resource allocation.

1.01 Economic evaluation and efficiency

The main purpose of economic evaluation is to provide information that can lead to the more efficient use of health care resources (McGuire 2001). In the context of this thesis there are three different forms of efficiency that warrant consideration: technical efficiency, productive efficiency and allocative efficiency. Technical efficiency defines the minimum resource inputs required to produce a given output. Technical efficiency is necessary but insufficient for productive efficiency, which also depends on the relative factor prices of the different factor inputs. To achieve productive efficiency the costs of producing a technically efficient output are minimised, and so the combination of factor inputs chosen depends partly on their relative factor prices.

Technical and productive efficiency are pre-requisites for achieving allocative efficiency. This 'higher level' notion of efficiency requires that the marginal benefits and marginal costs of different alternatives, each produced in an efficient way, are compared (McPake et al. 2002). Health care decision-makers can use the notion of allocative efficiency to choose which health care programmes to provide. The aim of this comparison is to identify those health care programmes that produce the most units of health gain from the given budget. This comparison might suggest that instead of using certain resources to produce for example stroke care, those resources could be redeployed to produce paediatric care, leading to a larger gain in health for the same cost. This would move the allocation of resources towards the target of allocative efficiency[1].

For cost-effectiveness analyses (CEA) to provide the information required to move resources towards the goal of allocative efficiency, they have to meet certain important requirements (McGuire 2001). For example, each CEA has to report the effectiveness of each health care programme by measuring outcomes on the same scale. Some commentators have argued that the quality-adjusted life year (QALY) is the most appropriate measure of outcome for CEA (Gold et al. 1996)[2]. Importantly for this thesis, a second requirement for CEA is that the costs of each health care programme must reflect opportunity costs (Birch and Gafni 2002). Opportunity costs are the costs of providing each health care programme in a way that is productively efficient (McPake et al. 2002)[3]. If an economic evaluation compares the costs and outcomes of a new intervention delivered in a way that is productively efficient, to an existing service delivered in a productively inefficient way, then the cost-

---

[1] Economists disagree about whether CEA are appropriate for this purpose (see Birch and Gafni 1992, Johannesson and Weinstein 1993, Birch and Gafni 1993). For the use of CEA to move resource allocation towards allocative efficiency, assumptions such as perfect divisibility of resource inputs, and constant returns to scale have to be made (Birch and Gafni 1992).

[2] It is argued that only under very restrictive assumptions is using QALYS as the outcome measure consistent with the goal of allocative efficiency (McGuire 2001).

[3] There are various ways in which the definition of opportunity cost can be applied to the use of economic evaluations. For example, opportunity cost may refer to the value foregone when the NHS does not provide certain health services (Birch and Gafni 2002. However, as this thesis is concerned with cost variation across settings the more relevant definition of opportunity cost is the cost of efficient production. Variation in costs observed across settings may arise because costs in some settings reflect productive inefficiency, and hence intervention costs do not represent opportunity costs. If the costs of either or both the health care alternatives under comparison are not opportunity costs, then using the resultant CEA may fail to move resource allocation towards allocative efficiency.

effectiveness of the new intervention may be overstated. Its implementation would then lead to further allocative inefficiency.

As Tan-Torres Edejar et al. (2003) state:

"It is not useful for policy-makers to know the cost-effectiveness of interventions undertaken in a technically inefficient manner.." (Tan-Torres Edejar et al. 2003, p47).

1.02 Measuring opportunity costs for the decision-making context

It is desirable to measure the opportunity costs of each health care programme for the particular decision-making context. If CEAs are to be used to set national priorities then costs should represent the opportunity costs of providing each health care programme in the country concerned. It is now common for CEAs to collect costs alongside multicentre RCTs. These studies can measure resource use for each patient and calculate a mean cost for each health care programme. These evaluations can potentially estimate opportunity costs for a particular decision context, for example for a national decision-making agency such as NICE.

However, multicentre CEAs may fail to measure opportunity costs because they ignore systematic cost variations across health care settings. In this thesis, a health care *setting*[4] is defined as a health care provider in a particular geographical location. These health care providers may be located in different geographical locations in the same country or in different countries. This thesis uses the term 'cost variation across settings' to refer to *systematic* variation in resource use and/or unit costs across health care providers[5]. This form of cost variation may arise because of for example, differences in incentives to cost-minimise across health care settings. The problem is that even multicentre CEAs typically only measure unit costs in a single health care setting, and may only collect resource use data for a few patients in each health care setting. In these circumstances, it is difficult to establish which health care settings are

---

[4] Throughout the thesis the term 'setting' is used interchangeably with the terms 'centre' or 'health care firm'. Each of these terms refers to a particular health care provider or group of providers.

[5] The term 'cost variation' used in the thesis refers generally refers to the product of the resources used (e.g. the length of hospital stay), multiplied by their unit cost (e.g. the cost per hospital day). Where more specific reference is made to variation in resource use or unit costs this is made explicit. The costs referred to are generally health service costs, but where appropriate broader societal costs are considered.

efficient, and therefore which health care settings' costs represent opportunity costs. Studies cannot identify efficient production unless they measure costs in several health care settings, for example by comparing the costs of providing a particular technology in different geographical locations.

Multinational economic evaluations can measure costs in several settings, however they tend to measure costs in one country and transfer them to another country (Schulman et al. 1996, Johannesson et al. 1997). Even if the costs estimated in one country represent opportunity costs, they are unlikely to represent opportunity costs in a different country, because of variations in for example, factor prices[6].

Thus, multicentre studies whether they are conducted on a national or an international basis have failed to identify reasons for cost variation across settings, and the difficulties this poses for the estimation of opportunity costs.

Economic evaluations that measure costs in several health care centres have demonstrated wide variations in resource use and unit costs across the settings concerned, both within and between countries (Coyle and Drummond 2001, Johnston et al. 1998, Willke et al. 1998). However, these studies do not assess *why* these cost differences occur. In particular, these studies have not provided an empirical investigation of which factors are associated with cost variation. Instead, commentators have listed *a priori* reasons why costs may vary between health care settings in an *ad hoc* manner (O'Brien 1997, Mason 1997). For example, costs in a particular health care setting may be relatively high because the setting faces higher factor prices or has higher levels of productive inefficiency (O'Brien 1997). If the setting produces health care in a less efficient way[7], then the costs of health care

---

[6] Factor prices may vary across international health care settings because for example restrictions in the labour market for health care professionals mean that the wages of health care professionals vary across countries. If health care firms aim to cost-minimise then theory suggests that where firms face differences in relative factor price they will adjust the mix of factor inputs used. As both factor use and factor price are components of unit cost, these between-country differences can lead to differences in unit costs. Firms may also adjust the resources used in providing each health care programme, according to the factor prices they face. Theory suggests that resource use and unit costs are also correlated. Thus if factor prices differ across countries then resource use, unit costs and total costs may vary across international health care settings.

[7] For example, because there is less incentive for the health care decision-makers to cost-minimise.

programmes in this setting do not represent opportunity costs. However, there is limited empirical support for these *a priori* reasons for cost variation.

A thorough investigation is required that uses the relevant literature to identify *a priori* reasons for cost variation, collects costs in a range of health care settings and identifies which of the *a priori* reasons for cost variation are the most important. This investigation can provide guidance on the settings that economic evaluations should collect costs from to ensure that the programme's costs represent relevant opportunity costs.

1.03 Identifying *a priori* reasons for why costs may vary systematically across health care settings.

Insights from microeconomic theory can provide *a priori* reasons for why the costs observed may vary across settings. Production function theory states that firms in different health care settings can all produce a technically efficient output, but with varying combinations of factor inputs. Cost function theory suggests that if firms in different settings face different relative factor prices, then they will substitute those inputs with relatively low factor prices for those inputs with relatively high factor prices to achieve productive efficiency. However, firms in different contexts may vary in the extent to which they achieve technical or productive efficiency.

To understand why health care firms may not choose efficient combinations of factor inputs an understanding of the health care context is important. Health care firms may be constrained from achieving efficient production by contextual factors such as local labour market regulations (Elliott 2003). For example, even if a hospital recognises that it is efficient to reduce the use of doctors' time, local labour market regulations may prevent this. The incentives and regulations provided by national decision-makers may also explain why the rate of diffusion of new technologies varies across countries. These contextual factors may partly explain why systematic cost variations are observed across health care settings.

While some of the factors driving systematic variations in costs may operate at a national or centre-level, the health services research literature suggests that variations in individual patient characteristics across health care settings, may also be important

(Phelps and Mooney 1993). If the case-mix of patients attending a particular health care setting is more complex than average, then this may lead to higher resource use and unit costs.

Empirical evidence is required to assess which of these *a priori* reasons are important for explaining variations in costs across health care settings. Before systematic reasons for cost variation can be identified, certain measurement and analysis issues that arise when comparing costs across health care settings need to be recognised.

1.04 Measurement issues that arise when comparing costs across health care settings

A study that attempts to compare costs across health care settings will encounter measurement issues. Most of these issues arise whether the study compares costs within or across countries. Previous attempts to compare costs across health care settings have not used a consistent costing methodology (Schulman et al. 1998, Willke et al. 1998); this makes it difficult to assess whether the observed cost differences are due to systematic variations in cost or inconsistencies in the costing methodology. In studies that have used aggregated datasets to assess cost variation, it is difficult to assess whether the costs are measured in the same way in each setting (Carey 2000). For example, the costs of hospital overheads are included in some settings but not others. The guidelines for economic evaluation suggest that measuring resource use and unit costs separately can assist with interpreting cost differences between settings (Drummond et al. 1997a). However, this level of disaggregation may be insufficient; it is difficult to interpret reasons for unit cost variation if the study uses aggregated measures of unit cost. Instead, if a disaggregated costing method is used differences in the individual components of unit costs, the factor prices and factor use, can be compared across health care settings. Price and volume indices can then be constructed to examine why unit costs vary by health care setting. These indices have proved helpful in other areas for interpreting factor use and factor prices differences across settings, and the implications for productive efficiency (Danzon and Chao 2000, van Ark et al. 1999).

A measurement issue that arises when comparing costs across international settings, is that costs have to be converted from local currencies into a common currency. A common approach is to use conversion factors based on purchasing power parity (PPP) indices, these indices aim to adjust for international differences in factor prices (Kanavos and Mossialos 1999). Studies estimating costs in different countries generally use Gross Domestic Product (GDP) PPP indices, that attempt to adjust for differences in the price of goods and services consumed in the entire economy. However, if these general indices are used to convert costs in a particular sector they may fail to adjust for differences in input prices (Wordsworth and Ludbrook 2005).

Unexplained variations in costs across settings could relate to unmeasured differences in case-mix. This problem arises in studies assessing levels of inefficiency across health care settings. To estimate levels of inefficiency these studies require data from a large number of health care firms and therefore tend to rely on aggregated datasets (Newhouse 1994)[8]. Such datasets contain routinely collected data on costs, case-mix and outputs and encompass the whole of hospital production. Routine measures such as the Diagnosis Related Group (DRG) system of case-mix classification may fail to recognise differences in the case-mix of patients across hospitals (Iezzoni et al. 1996). It is therefore unclear how much reported variations in efficiency levels across hospitals represent actual efficiency differences, or variations in the case-mix of patients (Cowing and Holtmann 1983).

A further measurement issue is whether the study has sufficient patients and health care settings to identify systematic reasons for cost variation. Some previous studies of cost variation have not sampled enough patients or health care settings to identify systematic variations across health care centres (Willke et al. 1998). These studies may have attributed unexplained variations to random variations when they reflect systematic differences that the study failed to detect.

[8] Even the smallest hospital production studies use cross-sectional datasets from more than 20 health care firms.

1.05 The choice of statistical techniques for identifying reasons for cost variation and analysing multicentre cost-effectiveness.

Studies could address the main measurement issues raised by collecting disaggregated data for sufficient patients and centres. However, decisions made at the analysis stage still determine whether the investigation makes an appropriate assessment of the reasons for cost variation.

Ordinary least squares (OLS) regression analysis and multilevel models (MLMs) are two techniques that the analyst can use to identify factors associated with cost variation across settings. Previously, economic evaluations have used OLS regression models to identify reasons why costs vary between settings (Willke et al. 1998, Coyle and Drummond 2001). However, OLS models assume that individual observations are independent. This assumption does not appear plausible when analysing cost data from several settings, where there are *a priori* reasons for expecting differences in costs. For example, factor prices may vary across health care settings, and the cost data may therefore be clustered within health care settings. In this context, the use of OLS regression models to identify reasons for cost variation could lead to incorrect inferences. MLMs can acknowledge the hierarchical structure of data (Goldstein 1995) and may therefore be more appropriate for analysing variations in costs across health care settings. However, they have rarely been used in health economics (Rice and Jones 1997), and their use for analysing costs and cost-effectiveness has not been explored.

In summary, economic evaluations commonly fail to identify reasons for cost variations across health care settings. Some studies suggest that there are wide cost variations particularly across international health care settings (Schulman et al. 1998, Willke et al. 1998) but these studies fail to use appropriate methodologies. In particular, these studies do not use *a priori* reasoning to pose hypotheses for systematic variations in cost, they ignore measurement issues, and the analytical methods used do not recognise the clustering in the data. The importance of this gap in the literature was recognised by a commissioning brief from the NHS Health Technology Assessment Programme (NHS R&D Programme 1998), and recent WHO guidelines on generalisability (Murray et al. 2000).

An empirical study is needed to thoroughly assess the reasons why costs observed may vary across health care settings. This assessment needs to use *a priori* reasoning to identify factors that may be associated with systematic cost variation. The importance of these *a priori* reasons in explaining cost variations then needs to be tested. Such an empirical investigation has to tackle the measurement and analysis issues raised. Finally, once the study identifies reasons for systematic variations, the findings can inform the conduct and interpretation of economic evaluations.

1.1 Aims and objectives

The overall aim of this thesis is to assess why costs vary across health care settings, and the implications for the methodologies used in economic evaluation. The specific objectives are:

1. To assess how economic evaluations currently consider cost variation across settings.
2. To generate hypotheses for why costs may vary across health care settings.
3. To identify which factors are associated with variation in resource use and cost using MLMs and OLS regression models.
4. To compare the use of OLS regression models to MLMs for analysing multicentre cost-effectiveness data.

The way the thesis considers each of these objectives is described briefly below:

1. To assess how economic evaluations currently consider cost variation across settings.

The literature review considers current practice in economic evaluation and examines both methodological guidelines and empirical studies. The review focuses on how the design and analysis of economic evaluations currently considers cost variation across settings. It highlights that there are important areas of omission in current recommended practice. For example, the guidelines provide scant advice for the analyst conducting a multicentre study on how to select the centres for cost

measurement. The guidance does not consider *a priori* reasons from the economics literature on why factor prices, factor use, and resource use may differ across health care settings. Moreover, there is little advice on how studies should identify systematic variations when analysing cost data, and how they should interpret differences in cost-effectiveness across health care settings.

The thesis therefore identifies several gaps in the methodological literature on economic evaluation that the thesis can inform by identifying reasons for systematic variations in costs across health care settings.

2. To generate hypotheses for why costs may vary across health care settings.

The thesis identifies *a priori* reasons for systematic cost variation by reviewing relevant strands from the health economics, microeconomics and health services research literatures. Costs may vary systematically across health care settings because for example, health care firms face different incentives to cost-minimise, and this may lead to variations in levels of productive inefficiency. Observed costs may also vary because of systematic differences in case-mix. The identification of *a priori* reasons for cost variations across health care settings informs the design and interpretation of the empirical investigation.

The empirical investigation extends an observational study that measures the costs and outcomes of stroke care across 13 centres in 10 different European countries. The study collects information on why costs may vary systematically across the different European centres. For example, the study gathers information on, factor prices, the incentives for each provider to cost-minimise, and the level of health care spending in each country. These data are used to establish whether the reasons for cost variation suggested in the literature review are supported by the empirical investigation.

The empirical investigation takes a consistent, disaggregated approach to cost measurement in each setting. The study measures the use and price of factor inputs, and compares these parameters across the health care centres. The thesis considers the choice of currency conversion factor, by developing a technology specific measure of PPP based on factor inputs used to produce stroke care and their associated factor prices.

3. **To identify which factors are associated with variation in resource use and cost using MLMs and OLS regression models.**

The empirical investigation recognises that multinational cost data may have a hierarchical structure with patients clustered within health care centres in different countries. In this context, the use of MLMs to identify systematic reasons for international cost variation is attractive. MLMs recognise the structure of the data, and may make inferences that are more correct. The thesis compares the use of MLMs with OLS regression analysis for identifying reasons for systematic cost variations, to see whether the choice of technique has an effect on the study's results.

4. **To compare the use of OLS regression models and MLMs for analysing multicentre cost-effectiveness data.**

Once a comprehensive assessment of the reasons for cost variation is undertaken, it is then possible to consider the implications for the conduct of economic evaluations. The thesis uses data from the costing study to generate a multicentre cost-effectiveness dataset. The analysis compares MLMs to OLS regression models for estimating incremental cost-effectiveness. The MLMs are extended to include covariates that adjust for systematic cost variations across health care settings.

The discussion section of the thesis considers more generally the implications of the investigation of cost variation across settings for the design and analysis of economic evaluations.

1.2 Structure of the thesis

The thesis has two main sections: Chapters 2-5 cover the literature review and Chapters 6-9 the empirical investigation.

The literature review starts by examining the relevant economic evaluation literature. Chapter 2 considers the methodological guidelines for economic evaluation with a focus on those aspects that are relevant for an investigation of cost variation across settings. Chapter 3 reviews how economic evaluations have traditionally considered

cost variations across settings, with a focus on multinational studies. Chapter 4 reviews the relevant microeconomics, health economics and health services research literatures. The chapter begins by examining insights from the production and cost function literature, then moves onto consider contextual and patient factors that may explain systematic variations in costs across health care settings. The literature review also highlights measurement issues that arise when comparing costs across health care settings. Hypotheses are posed regarding potential reasons for cost variation. Chapter 5 considers some of the techniques for identifying reasons for cost variation across settings. The chapter discusses issues that arise when using techniques for measuring efficiency, OLS regression analysis or MLMs for analysing cost variation. The review focuses on the issues that arise when using MLMs to analyse variations in cost and cost-effectiveness across health care settings. Chapter 5 concludes by offering an overall critique of the literature reviewed and provides the conceptual framework for the empirical investigation.

Chapter 6 begins by considering how the conceptual framework applies to the empirical investigation and introduces the case study for the empirical study. The chapter describes the methods used to collect information on the patient, centre and national factors potentially associated with cost variation across settings. The methodology used to collect resource use data is described. The resource use data are presented and used in conjunction with the information on patient and contextual factors to pose hypotheses for resource use and cost variation. Chapter 7 describes the methods used to collect unit costs and explains how the empirical study addresses some of the measurement issues raised. It also describes how price and volume indices are calculated and used to analyse unit cost differences across the study settings. Unit costs for each health care setting are presented, together with the price and volume indices. The choice of conversion factor for translating costs to a common currency is considered, and the total costs are presented using different conversion factors.

In Chapter 8 OLS regression analyses and MLMs are used to identify factors associated with systematic variations in resource use and costs across health care settings. The implications of the choice of technique are discussed. Chapter 9 compares OLS regression analyses and MLMs for CEA using a multicentre dataset

developed from the cost data. Each of the techniques is used to estimate overall incremental cost-effectiveness. This chapter illustrates how using a MLM with covariates can recognise systematic cost differences across health care settings, when analysing cost-effectiveness.

Chapter 10 summarises the main findings from both the literature review and the empirical investigation. The main themes to emerge from the thesis are discussed alongside the limitations of the approach. Chapter 11 presents the conclusions, recommendations for further research and the policy implications that arise from the thesis.

1.3 Overall contribution of the thesis

The overall contribution of the thesis is to raise awareness of current problems concerning the way costs in economic evaluations are measured and analysed. In particular the literature review highlights that previous studies have tended to disregard cost variation across health care settings. Ignoring this cost variation may lead to inaccurate estimates of the cost and cost-effectiveness of health care services.

It is therefore a primary objective of this thesis to examine why this cost variation exists, and to examine the implications for economic evaluation. This thesis identifies *a priori* reasons for why costs may vary across health care settings. These reasons are grouped into patient factors (e.g. casemix) and contextual factors (e.g. the level of health care infrastructure in the setting concerned).

Previous studies of cost variations across settings have used highly aggregated datasets, and the results have been difficult to interpret because measurement issues pervade these studies. The cost differences observed may simply reflect methodological inconsistencies in for example, the methods used to estimate unit costs. This thesis contributes to research in this area by using a disaggregated dataset to tackle the measurement issues raised. The methodology developed disentangles reasons for cost variation across international health care settings. Price and volume indices originally developed on time series data, are used to examine the role of price

and volume differences in explaining cost variation across health care settings. This disaggregated approach therefore extends traditional cost function studies (see for example Feldstein, 1967), and enables the thesis to identify a range of reasons for cost variation across settings.

A key contribution of the thesis is to apply a method of analysis-- multilevel modelling that recognises that multicentre cost data are hierarchical. Although MLMs have been recommended for use in health economics (Rice and Jones, 1997), few studies have followed this approach. The empirical section of this thesis demonstrates that MLMs are more appropriate than OLS regression analyses for analysing reasons for cost variations across health care settings. In particular, the MLMs appropriately estimate the precision of the higher level variables associated with total cost. By contrast the thesis demonstrates that OLS regression analysis overestimates the precision of these higher-level variables and leads to incorrect inferences about which factors are associated with cost variation. The thesis also demonstrates that using OLS regression analysis in multicentre studies can lead to inaccurate estimates of mean incremental cost-effectiveness, and the associated uncertainty.

The thesis concludes that ignoring cost variation across health care settings can lead to inaccurate estimates of cost and cost-effectiveness. Basing decision-making on inaccurate information can move the allocation of health care resources away from the target of allocative efficiency. While the focus of the empirical contribution is on observational cost data collected in a multinational context, the underlying methodological concerns raised by ignoring cost variation across health care settings apply more generally. This thesis presents a methodology for improving the conduct of cost analyses that future economic evaluations can adopt.

Chapter 2: Economic evaluation guidelines on costs and cost variation across settings

2.0 Introduction

As the demand for health care continues to outstrip the supply, information is required to assist policy-makers allocate resources in an efficient and equitable way (Maynard and Kanavos 2000). Economic evaluation provides a framework for presenting information on the costs and effectiveness of health care interventions, so that decision-makers can allocate resources in a way that maximises the population's health given resource constraints (Drummond et al. 1997a). Several countries have moved towards the statutory use of economic evaluation in setting health care priorities[9]. Yet for economic evaluations to make a useful contribution, these studies have to be conducted on a sound basis in accordance with economic and statistical principles. Otherwise, using these studies may fail to move the allocation of scarce health care resources towards the goal of allocative efficiency; decision-makers may become disillusioned with economic evaluations and return to using other ways to set health care priorities (Donaldson et al. 2002, Hutton and Maynard 2000).

2.01 Improving the methodological quality of economic evaluations

In an attempt to improve the quality of economic evaluations, methodological guidelines have been developed that give advice on how to conduct these studies. The guidelines consist of formal requirements issued by reimbursement agencies and more methodological advice published in academic journals. In part, these guidelines reflect recent methodological and empirical developments, and emphasise areas of agreement

[9] Economic evaluations are now used by governments in Canada, Australia, the UK, the Netherlands, Sweden, Finland and Portugal to inform decisions about which health care technologies are to be publicly provided (Hutton and Maynard 2000).

amongst health economists (Johnston et al. 1999). For example, there has been recent progress in the measurement and valuation of health outcomes (Brazier et al. 1999), and the development of methods to represent uncertainty (Briggs and Gray 1999, Fenwick et al. 2004). By contrast, less research resources have been devoted to improving costing methods in economic evaluation (Graves et al. 2002) and in particular to the issue of cost variation across health care settings. However, the issue of cost variation would appear important as many studies now rely on cost data collected alongside multicentre RCTs. In these studies, there are *a priori* reasons for expecting costs to differ across settings, because of for example variations in factor prices.

The prior belief that costs vary across health care settings raises certain issues for the conduct of CEA. These issues include: in which centres should the study collect resource use and unit cost data? If the study is multinational, should the analysis pool cost data across different countries? How can a national decision-maker apply a multinational measure of cost-effectiveness to a particular country? Economic evaluations are required to consider these issues; otherwise, the results may not be relevant for decision-making. Decision-makers have stated that they may not use published evaluations if they do not believe that the results apply to their local context decision context (Drummond et al. 1997b, Weatherly et al. 2002).

2.02 Purpose of the review of methodological guidelines

The main aim of this chapter is to assess how methodological guidelines suggest economic evaluations should consider systematic variations in costs across settings. This review covers study design, analysis and presentation of results in economic evaluation. The review highlights where there are omissions or disagreements in the methodological literature relating to the issue of cost variation. The review also identifies aspects of the guidance that can be used to ensure methodological consistency when measuring costs across health care settings. The findings from the review can therefore inform the empirical investigation for the thesis. The empirical investigation needs to apply a standard costing method across different health care settings so that any differences in observed costs across health care settings can be

attributed to systematic variations rather than methodological differences in the way the costs are measured.

2.1 Methodology used in literature review

A structured review was conducted of the guidance issued for economic evaluations. The focus of the review was on methodological guidance that relates to cost variation across health care settings. The literature review covered advice for economic submissions to government agencies and general guidance published in peer-reviewed journals aimed at academic researchers[10].

2.11 Data sources and search strategy

The literature search located guidance from a range of sources. The Medline, BIDS, HEED, NHS NEED and EMBASE databases were searched using the search terms 'costs and cost analysis', 'economic eval*', 'transferability', 'generalisability', 'generalizability', 'multinational', and 'multicentre' over the years 1990-2004. This strategy was supplemented by screening the bibliographies of recently published articles reviewing the methodology used in economic evaluations, reviewing the bibliographies of papers retrieved, searching library catalogues for relevant books, reviewing websites of institutions producing or using economic evaluations, and based on recommendations from colleagues working in this area. In addition to including published papers, the review also included 'grey literature' in particular conference papers and working papers. Relevant electronic journals were searched to identify recent articles that were 'in press'.

Inclusion and exclusion criteria

Titles, abstracts (if available) and full papers were examined to identify relevant sources for the purposes of this review. Sources were required that discussed aspects of study design, analysis or presentation of results related to cost variation across

[10] Although it is recognised that these are not formal guidelines, they are included in the general term 'guidelines' used throughout this thesis.

settings[11]. The review only included those guidelines published or produced after 1990. Previous reviews of costing methods found guidelines prior to 1990, tended to ignore methodological issues (Wolstenholme 2001). This review was therefore limited to those guidelines that incorporated recent developments in the field. The review was also limited to English language sources.

The findings from the review are reported in each of the three key methodological areas: study design (section 2.2), data analysis (2.3) and presentation and interpretation of results (2.4).

2.2 Study Design

The guidelines reviewed mostly include sections on framing and designing the study, inclusion of the relevant resource use items, and collection and measurement of resource use and unit costs. Each of these areas is reviewed below.

2.21. Framing and designing the economic evaluation

The guidelines emphasise that the study question has to be clearly defined. In particular, it is important to identify the relevant decision-maker, as this may determine many of the key methodological standpoints, as Torrance et al. (1996) point out:

> "...understanding the decision context will guide the choice of audience and the perspective of the study." (Torrance et al. 1996, p54.)

Jonsson and Weinstein (1997) suggest that the perspective taken to the evaluation may differ across countries in a multinational CEA. For example while in England and Wales NICE recommends that CEA should exclude the costs related to lost

[11] Questions such as which form of economic evaluation to use are not relevant, as the issues raised by cost variation across settings apply whichever type of economic evaluation is chosen. For more general reviews on methodology in economic evaluations see the books by Drummond et al. (1997a) or Gold et al. (1996).

income, this may inappropriate for measuring costs in Germany as sickness funds have to meet the costs associated with lost income as well as health care costs.

The decision context also has implications for the 'vehicle' chosen for the economic evaluation; economic evaluations may use national RCTs, multinational RCTs or decision-analytical models[12]. In some decision contexts methodological guidelines have expressed a preference for CEA based on national RCTs (CCOHTA 1997). Nevertheless, whichever study design is preferred there are general issues to address about cost variation across settings.

Most of the guidelines emphasise that analysts should carefully define the health care programmes under evaluation. So for example, the type of surgery or the dose and the duration of the drug regimen should be clearly stated. The comparators chosen should include current routine practice, which may vary depending on the decision context, for example the country concerned (Byford and Palmer 1998). The evaluation should define what 'routine practice' means, for example if patients are routinely managed following a stroke, the evaluation should explain what resources use is involved. This can help a decision-maker understand what the baseline is for the analysis, and assess whether the results are likely to apply to their particular context.

2.22 Including the relevant items of resource use

The guidelines emphasise that all relevant items of resource use should be included in a cost analysis (Johnston et al. 1999, Drummond et al. 1997a). The items to include depend on the methodological standpoints taken including the study's perspective and time-frame. There are also empirical issues to consider, including the quantitative importance of the resource use items concerned, and their association with health outcomes (Johnston et al. 1999).

---

[12] Economic evaluations may also be based on observational studies, or use a combination of these different study designs, e.g. a model may use effectiveness data from a national RCT, but cost data from an observational study.

a) Perspective or viewpoint

A primary consideration is whether the theory of welfare economics provides the basis for the study[13]. The welfarist position provides a strong basis for including a wide range of programme costs irrespective of the agency upon which they fall (Johnston et al. 1999). From a welfarist perspective the costs to any public sector agency, the costs of lost production, the costs to patients and their carers and any other costs to society should be included. However, there is widespread disagreement about whether welfare economics should provide the basis for economic evaluation, and others recommend taking an extra-welfarist position and only including those costs that are relevant from the perspective of the decision-maker (Tsuchiya and Williams 2001). It is still possible that from an extra-welfarist standpoint the societal perspective is taken, for example the CCOHTA guidelines suggest taking the widest possible decision-making perspective, i.e. the societal perspective (CCOHTA 1997). This allows decision-makers to use the results from the perspective most relevant to them. As Brouwer et al. (2001) state:

> "Taking other perspectives like the health care budget perspective is therefore to be discouraged as being too narrow and not recognising that budgets are arbitrary divisions in how resources are organised." (Brouwer et al. 2001, p70).

Nevertheless, other guidelines do recommend taking a narrow viewpoint. For example, recent NICE guidance for submissions to the technology appraisal process suggests that a health and personal social service perspective (PSS) is appropriate (NICE 2004). Taking this perspective ignores all costs falling on other public service providers, patients, and on society from lost production. This has implications for ensuring consistency in cost analysis conducted or used in different settings. The boundaries between different agencies may vary across settings. For example, in some countries the health budget pays for the cost of nursing home care, whereas in others the social care budget or the patient may bear the cost. Taking a narrow perspective leads to inconsistent criteria being applied to the inclusion of resource use across

---

[13] Welfare economic theory rests upon economic models of individual behaviour. It is based on the assumption that individuals are the best judge of their own well-being and provides the theoretical basis for cost-benefit analysis. The theoretical basis of cost-effectiveness analysis is less clear, and has been subject to considerable debate in the health economics literature (McGuire 2001).

different contexts. One way to address this problem is to take a societal perspective to costing across all settings.

b) Time-frame

The choice of resource use items for inclusion in the cost analysis also depends on the study's time-frame. However, there is no consensus about which time-horizon is most appropriate. Torrance et al. (1996) recommend taking a long-term time-horizon, while the NICE guidelines suggest that the time-horizon should be sufficient for the evaluation to examine the full impact of the interventions on costs and outcomes (NICE 2004). The NICE guidelines suggest that when the evaluation is of technologies for chronic diseases, such as cancer, diabetes or ischaemic heart disease, this requires a lifetime time-horizon (NICE 2004). If the time-horizon is too short, the analysis may exclude important costs and outcomes associated with the intervention. This could be an important issue when collecting costs across several settings. If some settings treat the disease early and intensively, whereas in other settings, high-cost interventions are provided much later, then the time-horizon must be sufficient to avoid the inconsistent inclusion of costs.

c) Quantitative importance

Johnston et al. (1999) suggest that if there is no difference in a particular cost between the interventions concerned, then the analyst could exclude these costs. However, as Drummond and Davies (1991) point out in a multicentre study, even if costs are the same in both the treatment and control groups over all the centres included, there may be cost differences between the two groups within individual health care settings. It is therefore unwise to neglect context-specific cost differences when deciding on which resource use items to include.

d) Association with health outcome

Although the focus of the thesis is on costs and cost variation, decisions about which resource items to include also need to consider patient outcomes. The costs included in an economic evaluation should be opportunity costs. Opportunity costs are the

costs of technologies that are produced efficiently, where productive efficiency refers to the minimum cost of producing a given output or outcome. Thus, if resource use that does not improve outcomes could be identified, the CEA could exclude these costs from the overall costs of the health care programmes. Including these costs would mean that the costs reported no longer reflect productive efficiency and opportunity costs. Some guidelines have suggested that the relevant outcomes are those that relate to health rather than utility, and therefore the relevant resource use items to include are those that improve health rather than utility (Culyer 1990). The guidelines do not explain how to identify those resource inputs that are associated with improvements in health outcome. To understand the relationship between resource inputs and health outcomes the study could estimate the likely production function for the disease and interventions in question (Brouwer et al. 2001). However, this may be difficult especially in a multinational context (Koopmanschap et al. 2001). The relationship between resource use and health outcome may differ across settings, because of for example, differences in case-mix. Rather than just deciding on a standard list of resource use items for inclusion, studies should try to recognise these issues when identifying the relevant resource use to measure.

2.23 Collection and measurement of resource use

The guidelines all agree that it is necessary to present separate estimates of resource use and unit cost to help decision-makers understand whether the resource use and unit costs observed in the study setting will differ from their own (NICE 2004, Luce et al. 1996, Drummond et al. 1997a, CCOHTA 1997, Drummond and Jefferson 1996). However, there is a lack of definitive guidance on exactly how to measure resource use. Issues that are relevant for assessing cost variation across settings include: the source of resource use data, the numbers of patients to measure resource use for, the numbers and types of setting to include, and the level of aggregation to use. Each of these issues is now considered in turn.

a) Source of resource use data

An RCT, observational study, routine database or expert opinion can provide the source of resource use data. A cost-effectiveness model can use information from any

of these sources. Some of the guidelines state that the preferred source of data for resource use measurement is a pragmatic RCT (Mason 1997). Collecting data from pragmatic RCTs has the advantage of producing unbiased estimates of the differences in resource use between the interventions concerned, and it allows costs to reflect routine clinical practice. By contrast, Mason (1997) criticises explanatory RCTs[14] for being atypical of routine practice because of protocol driven costs, and the inclusion of atypical patients and centres.

A multicentre RCT has advantages over other forms of study design in that it can explore differences in resource use across centres, and has the potential to produce results that are more generally applicable. Drummond and Davies (1991) suggest that before the results from an economic evaluation alongside a multinational RCT can be applied to a national context, the pooled international estimates of resource use should be adjusted to each local context. Drummond and Davies (1991) demonstrate how expert opinion can be used to adjust resource use parameters collected in an international evaluation and make them more nationally relevant. However, Luce et al. (1996) highlight the potential inaccuracies of using expert opinion to estimate resource use and recommend using it as a last resort.

Mason (1997) considers the problem of using observational data to compare resource use between treatment groups and discusses the bias that may arise according to the selection of patients. In a multicentre observational study, the selection of patients for each treatment arm may also vary by centre leading to differences in the degree of bias across the centres concerned. Mason (1997) does though acknowledge that observational studies are more likely to represent routine clinical practice than RCTs, and studies that use this study design may produce more generalisable cost estimates. Observational studies can therefore be used to establish why costs of routine practice may vary across health care settings in different geographical locations.

[14] Explanatory RCTs use highly regulated protocols, and involve blinding of patients and health care professionals.

b) Number of patients to include in the resource use measurement

In an economic evaluation based on a RCT the study's power to detect differences in cost or cost-effectiveness between the interventions concerned partly depends on how many patients are included in the measurement of resource use. However, the guidelines are not prescriptive about recruiting sufficient numbers to detect these differences. Instead, the guidelines highlight the difficulties faced by economic analysts when trying to detect differences in economic endpoints, as sample sizes are usually determined by power calculations for the main clinical endpoint (Johnston et al. 1999, Briggs and Gray 1999, Gray et al. 1997). As there is often greater variability in costs than outcomes amongst patients, more cases are usually needed to detect differences in economic endpoints (Gray et al. 1997). Power calculations are also problematic because they require data on cost differences between interventions, and on the distribution and variability of costs, these data are usually unavailable *ex ante* (Johnston et al. 1999).

In addition to variability in costs amongst patients receiving a health care programme within a particular setting, there may be differences in resource use and costs across settings. This may have additional implications for the study's ability to estimate accurately and precisely the incremental costs of an intervention for a particular decision context. The systematic differences in resource use across settings may be particularly large in an international study. In light of these systematic differences, it might be tempting to base power calculations on conducting country-specific analyses. However, if multinational economic evaluations are to recruit sufficient patients to detect country-specific differences then this would drastically increase the costs and duration of these studies.

There is little advice in the guidelines on how many patients to sample when collecting data on resource use. This may depend on several factors including the size of the differences the study is aiming to detect, the sampling variation across patients, and particularly in an international context, any systematic variations that exist across patients and health care settings.

c) Number and types of setting to include

In economic evaluations based around RCTs a potentially important issue is: from how many and which centres should resource use data be collected? None of the guidelines offer specific advice on the optimal number or location of sites for measuring resource use data. Coyle et al. (1998) suggest that:

"...cost data from a single site participating in the study would be acceptable if it is demonstrated that the cost structure of the site is typical of all other sites within the study" (Coyle et al. 1998, p140).

The authors do not explain why obtaining cost data from a 'typical site' is desirable, or how it is possible to judge whether the cost structure is typical without measuring costs in all sites. Baladi et al. (1996) offer slightly more stringent criteria for centre selection and state that rather than being representative of the study sites, the centres(s) selected for resource use measurement should be representative of the setting in which the technology is going to be implemented; they state:

"..for example if a technology is to be deployed in secondary level hospitals, costs should be derived from this particular hospital group and not estimated from tertiary teaching hospitals." (Baladi et al. 1996, p4).

They go onto recommend using costs based on routine clinical practice that are derived from a number of different institutions, rather than using costs from a particular institution for a specific purpose.

While Drummond and Davies (1991) suggest that collecting data from all centres in a multicentre study could lead to more generalisable results they recognise that problems arise when measuring resource use in different settings. They cite instances in multinational studies where there are systematic differences in resource use across settings. In these circumstances, the analyst can either pool the results or report separate cost estimates for each setting. The problem with pooled results is that they may not represent the opportunity costs of the health care programmes in any particular jurisdiction. If the analysis reports separate costs for each setting, then there may be insufficient cases to assess differences in cost and cost-effectiveness between

the programmes concerned. Using country-specific results would also question the purpose of doing a multinational study in the first place (Drummond and Davies 1991). Some of the guidelines suggest that there are circumstances where all the study centres should collect resource use data. For example, Johnston et al. (1999) recommend:

"If centres are likely to differ in terms of their economic characteristics, then resource use should be collected from all centres.." (Johnston et al. 1999, p4).

The guidelines by Johnston et al. (1999) offer the most insight into the issue of site selection. They state that although it might be possible in the analysis to adjust for any observed heterogeneity in costs across the centres, crucial decisions made at the design stage regarding the centres included in the study determine the heterogeneity observed. Whilst Johnston et al. (1999) agree with previous guidelines that the centres chosen for resource use and unit costing should be 'representative' of where the technology is going to be implemented, they also point out that a 'representative' centre needs defining.

Johnston et al. (1998) assess whether costs collected in a trial setting are representative of the programmes' costs in a more general health care context. They suggest 'economic factors' that might be associated with cost variation across settings, which include whether a centre is urban or rural, has high or low occupancy rates, and is a teaching or non-teaching hospital. However, the guidelines do not offer a theoretical basis for these suggestions.

To understand what constitutes a representative centre this thesis argues that insights from theory can identify the factors associated with cost variation across settings. None of the guidelines apply the notion of opportunity costs when deciding on which centres the resource use (and costs) should be collected in. Some centres may be producing the health care programmes concerned in an efficient manner, whereas in other centres, resources may be wasted. If the costs included are to represent opportunity costs then the costs from the efficient centres are those that are relevant. A study identifying efficient production would need to collect costs in several centres,

and conduct a thorough assessment using a consistent methodology to identify reasons for any variation in costs between settings.

d) Level of aggregation in resource use measurement

The guidelines highlight that there is wide variation across studies in the level of aggregation used when measuring resource use and unit costs. A gross costing or 'top down' approach involves using an aggregated measure of resource use such as the length of hospital stay (LOS) and combining this with a similarly aggregated measure of unit cost, such as the cost per hospital bed-day. By contrast, a micro costing or 'bottom-up' approach requires measuring the individual resource use items (e.g. individual blood tests) consumed during the hospital stay for each patient and then valuing these using the appropriate unit cost (e.g. cost per test). The guidelines do not offer any clear advice on which method is preferable. Indeed some of the guidelines ignore the issue completely (NICE 2004, Commonwealth Department of Human Services and Health 1995).

The level of aggregation used in resource use measurement has implications for interpreting variability across settings. Measuring resource use at a highly aggregated level makes it difficult to assess why costs vary across health care settings. For example, just measuring hospital length of stay (LOS) does not enable a study to examine differences in the intensity of resource use across settings. Certainly, it is important for a study that aims to identify reasons for cost variation across international settings to use a disaggregated approach. This can help the study to describe differences across settings in the use of factor inputs and may improve understanding of why costs vary across settings. A study conceals these differences if it only measures resource use at a highly aggregated level.

2.23 Collection and measurement of unit costs

Some of the issues that arise when measuring resource use also apply when estimating unit costs. For example, if the study measures resource use at a highly aggregated level, then highly aggregated unit costs are also required. The following section discusses this and other issues arising in the measurement of unit costs across settings.

a) Sources of unit cost data

Unit cost data can be taken from previously published costing studies, from national databases or by estimating costs specifically for the study. Some of the guidelines are quite prescriptive and suggest that evaluations use national estimates of unit costs and append these to the guidelines (CCOHTA 1997, Commonwealth Department of Human Services and Health 1995, Oostenbrink et al. 2000). In England, the Department of Health (DoH) and Personal Social Services Research Unit (PSSRU) both provide national cost estimates for hospital and community care (DoH 2002, Netten and Curtis 2002). These unit costs vary widely across settings, and estimates of the costs and cost-effectiveness of health care programmes could incorporate this variability. However, there are concerns about the quality of these unit cost estimates (Dawson and Street 1998, Bliss 1999). In particular, UK NHS reference costs use financial rather than opportunity costs. These reference costs take a highly aggregated approach to costing, so it is difficult to identify which resource inputs are included and whether they are measured consistently across health care settings. Thus, it is difficult to interpret whether any observed differences in unit costs reflect methodological inconsistencies or actual cost differences.

b) Estimation of opportunity and marginal costs

Some of the guidelines emphasise that unit costs should aim to measure opportunity costs (Drummond et al. 1997a, Luce et al. 1996, CCOHTA 1997). These guidelines point out that although in general market prices can represent opportunity costs, this might not be the case in health care, as many factor inputs are not delivered via a perfectly competitive market (Luce et al. 1996). Thus prices might not be available, and even where prices are available they may not represent opportunity costs. Luce et al. (1996) suggest that prices can be adjusted to make them representative of opportunity cost and in the United States cost-to-charge ratios have been used for this purpose (Luce et al. 1996). However, the extent to which health care prices diverge from opportunity costs is likely to differ across geographical health care settings, because of for example, different labour market structures.

As opportunity costs are not routinely available or easily calculated, economic evaluations may fail to use them (Walker et al. 1997). Indeed the guidelines offer little advice on how analysts should measure opportunity costs, as Birch and Gafni (2002) point out when offering a critique of previous NICE guidelines:

"...a general problem that underlies many aspects of the guidelines relates to the limited attention given to the concept of opportunity cost...the solution to the problem of using market prices that do not reflect opportunity costs is to use other data which also do not reflect opportunity costs..." (Birch and Gafni 2002, p187).

The most recent guidelines from NICE state that they prefer unit costs to reflect the financial costs to the NHS and PSS, rather than the opportunity costs. If economic evaluation aims to move resource use towards allocative efficiency then opportunity costs are required. Thus, the NICE position is inconsistent with economic theory. Using financial costs in economic evaluations and decision-making may lead to inefficient resource allocation.

Another important example of where adopting different methodological standpoints can lead to variation in cost estimates comes in the debate about how to measure marginal costs. As economic evaluations usually aim to measure small changes in the mix of services provided, the relevant concept to use is marginal rather average cost (Goddard and Hutton 1991, Baladi et al. 1996, Jacobs and Baladi 1996)[15]. However, there is no clear consensus on how to measure marginal costs.

Jacobs and Baladi (1996) state that marginal costs can be approximated by average costs. This assumes that marginal costs are constant and that total costs only include those items that may be regarded as variable for the time-period concerned. Such an approach requires identifying those factors of production that are fixed rather than variable. This assessment is subjective, may change over time, and would be particularly susceptible to local factors (Dawson and Street 1998, Drummond et al. 1997a, Walker et al. 1997). One way of limiting this problem is to take a long-run

[15] Marginal cost is the additional cost of a one-unit increase in output, whereas average costs reports the total cost divided by the total quantity produced.

perspective to cost measurement, in which case all cost items including overheads and capital costs should be included in total costs, and then long-run average costs approximate marginal costs. Following this approach would make the range of items included in unit costs more consistent across settings. However, to ensure comparability, studies that collect unit costs across settings also have to use the same methodology to measure this common range of cost items.

c) Allocation of overheads and capital costs

If the health care interventions under evaluation use different quantities of fixed inputs such as durable equipment or buildings, the CEA has to allocate the costs of these fixed inputs to the individual patients included in the treatment groups concerned. It is particularly difficult to apply the same methodology across settings when allocating the costs of overheads and capital to individual patients (Graves et al. 2002). Although attempts have been made to try and standardise the way overheads are allocated (Jacobs and Baladi 1996), many economic evaluations rely on the attribution techniques used by finance departments in the institutions concerned which may differ according to the health care setting.

d) Aggregated versus disaggregated unit costs

Just as the guidelines did not define the appropriate level of aggregation for measuring resource use, they are also reticent about how aggregated the measure of unit costs should be, except to say that one will determine the other. Clearly, adopting a highly aggregated approach to unit costing makes it difficult to ensure resources are valued using marginal and opportunity costs (Walker et al. 1997). As Reid et al (2003) highlight using aggregated unit costs, for example DRGs ignores any variations in the intensity of care over an individual patient's hospital stay. It is also difficult to ensure that the same unit costs are included in each health care setting (Jonsson and Weinstein 1997). For example, unit costs based on Diagnosis Related Groups (DRGs) may include capital costs in one setting, but these may be excluded in another. The case-severity of patients included within a particular DRG may vary across health care settings within or across countries. Any study comparing costs

across different settings that uses a highly aggregated approach may struggle to disentangle measurement differences from systematic differences in unit costs.

e) Number and characteristics of centres

The guidelines disagree about whether studies should collect unit costs from all centres or countries participating in a multicentre study (Commonwealth Department of Human Services and Health 1995, Baladi et al. 1996). The guidelines argue that the centres selected should be 'representative' for example Baladi et al. (1996) explain that:

> "...a site selection bias would result from the use of estimates derived from institutions that may not reflect the cost structure that prevails in the chosen perspective..." (Baladi et al. 2001, p4).

While some of the guidelines highlight the desirability of just using a national or regional estimate of average unit costs (Commonwealth Department of Human Services and Health, 1995), the selection and measurement of a 'representative' unit cost is not straightforward. As Brouwer et al. (2001) state:

> "...if one needs a national or regional average for the cost per unit of a service it is not easy to give a simple rule of thumb on the number of observations (sites) needed in order to get a robust estimate... the best provisional advice may be to use at least three to five observations for each specific organisational setting and to make overall estimates as a weighted average using the prevalence of the specific settings as well. "
>
> (Brouwer et al. 2001, p82).

The way in which the centres chosen for unit costing are currently selected appears arbitrary or based on convenience (Pang 2002). As Glick et al. (2002) highlight centre selection is rarely driven by empirical or theoretical rationale:

> "The countries selected might be ones that enrol many patients in the trial, ones that represent the spectrum of economic development among countries that participated in the trial, ones in which the countries' regulators require a submission for reimbursement, ones for which unit costs are readily available, or ones in which the study sponsor wishes to make economic claims." (Glick et al. 2002, p518)

The lack of consideration given in empirical studies to addressing this issue is not surprising given that previous guidelines have failed to offer clear advice on this issue. For example, Coyle et al. (1998) argue that if in a multicentre study detailed resource use data have been collected it may not be necessary to estimate unit costs in each centre. However, the authors do not offer any justification for making this recommendation.

One concern about measuring unit costs in many health care centres is the additional research resources associated with approach. As Reed et al. (2003, p397) point out:

> "The costing methods must strike a balance between the high cost of data collection in a clinical trial, the availability of cost data, and the collection of sufficient resource use data to capture variation in costs between patients"

While the opportunity costs of devoting research resources towards more extensive cost measurement should be recognised, the value of collecting this additional information must also be assessed. If a study only estimates unit costs in a single centre leading to an inaccurate estimate of the cost-effectiveness of the intervention then the value of acquiring more appropriate unit cost estimates may outweigh the additional research resources.

2.3 Data Analysis

A number of issues arise when analysing cost data that relate to cost variation across settings and these are discussed below:

2.31 Calculation of total costs per patient

Traditionally economic evaluations have not measured costs on an individual patient basis. Total costs per patient were calculated based on an 'average' cost over the patients included, which gave no indication of the sampling variation, surrounding the mean estimate. Recently, there has been an increase in the number of economic evaluations undertaken alongside multicentre RCTs (see for example, Schulman et al.

1996, Glick et al. 1998). This study design allows resource use to be measured for each patient and multiplied by the appropriate unit costs to give the total costs per patient. Mean costs can then be reported with stochastic measures of uncertainty as recommended by the guidelines (Manning et al. 1996). Even in CEAs based on models, recent guidelines advocate the use of probabilistic sensitivity analysis (NICE 2004). This requires a measure of the mean cost per patient and its standard error, to assess the impact of sampling uncertainty on the study's results.

While it is desirable to estimate total costs per patient, decisions at the design stage about how to collect resource use and unit costs, may determine how exactly this is done[16]. In particular, where an economic evaluation is based on a multinational trial, depending on how the study is designed the options available at the analysis stage include using:

1. Mean trial-wide measures of resource use, with a single or average set of unit costs.
2. Mean trial-wide measures of resource use data with country-specific unit costs.
3. Country-specific resource use and unit costs.

The guidelines do not offer advice on which of these options is preferable. The first two options would not provide an estimate of opportunity costs in any particular country or decision-context (Baladi et al. 1996). In option one neither the resource use nor the unit costs are likely to represent the levels of factor inputs or factor prices used in producing the health care programme in an efficient way in any particular country. Coyle et al. (1998) advise against taking this approach except under certain circumstances:

[16] Clearly, there are other issues that arise when analysing costs including how to deal with missing data. To consider these issues in more depth readers are referred to Manning et al. 1996, Briggs and Gray 1999, and Briggs et al. 2003.

"...cost data from a single country would be acceptable if it was demonstrated that the cost structure of the site is typical of all other sites within the study" (Coyle et al. 1998, p140.)

However, the authors do not provide guidance on how to assess whether or not a site is 'typical'.

The CCOHTA guidelines suggest transferring the results of a multinational trial to a Canadian context by assigning Canadian unit costs to trial-wide resource use data (CCOHTA 1997) (option 2). In option two, although the unit costs used may be relevant to each decision context if they represent opportunity costs, the resource use measure may not be applicable. Jacobs and Baladi (1996) argue that to make a study applicable to the Canadian health care system it is insufficient to take a multinational RCT and 'Canadianize it' by simply using Canadian price weights. They state that the analyst has to justify that the patterns of resource use for all the centres in the trial are the same as those in Canada. The analysts could do this using data just from Canadian centres if these are included in the study, or by using separate observational studies or expert opinion to adjust the results to the Canadian context. This approach fails to recognise the insight from economic theory that decision-makers may choose to adjust the combination of factor inputs used according to the local levels of factor prices (Raikou et al. 2000). Across international health care settings factor prices may vary widely because of for example, differences in the characteristics of the labour market.

Theory would therefore suggest that the resources used are correlated with the unit costs and that these parameters should not be measured in separate locations. Option three is therefore most likely to provide relevant estimates of resource use and unit costs, that approximate opportunity costs. However, the problem with this approach is that there may be insufficient cases in the particular country to provide robust estimates of costs and cost-effectiveness (Drummond and Davies 1991, Brouwer et al. 2001).

In general, there is a lack of advice in the guidelines on how international studies should calculate costs. This gap in the literature could partly reflect a lack of evidence on cost variation across health care settings (see Chapter three).

2.32 Currency conversion factor.

If resource use and unit cost data are collected for the country of interest, then the results can simply be presented in the local currency. However, if a multinational study requires a single measure of cost-effectiveness (as implied by option 1) it is necessary to convert prices into an appropriate reference currency (e.g. US dollars). Similarly if a study wants to compare costs across multinational health care settings this requires conversion of country-specific costs into a reference currency. There are various currency conversion factors available for this purpose including official exchange rates (OER), or purchasing power parities (PPP) based on general measures of GDP or medical prices. None of the more formal guidelines for economic evaluation offer guidance on the most appropriate method for currency conversion. This is a potentially important issue for any study aiming to compare costs across countries, and so this thesis reviews the health economics literature on currency conversion factors as part of the empirical investigation (Chapter 7).

2.33 Dealing with skewed data

CEAs alongside RCTs can estimate total costs per patient and then report mean costs per patient for each group. These costs can be presented with appropriate measures of precision to represent sampling variation (Barber and Thompson 1998, Thompson and Barber 2000). The guidelines recommend using statistical tests to report whether any differences in costs between treatment alternatives are statistically significant (Briggs and Gray 1999, Thompson and Barber 2000). The study can test for statistical significance using standard parametric methods such as t-tests. However, cost data are notoriously skewed (Gray et al. 1997, Briggs and Gray 1998, Briggs and Gray 1999, Barber and Thompson 1998), with a small number of patients accounting for a relatively high proportion of total costs; the assumption of normality required for parametric tests may not be valid for cost data. While transforming the data may gain extra power for detecting significant differences in costs between treatment

alternatives (Gray et al. 1997), recent guidelines recommend reporting mean costs on the untransformed scale (Thompson and Barber 2000, Briggs and Gray 1999). Thompson and Barber (2000) also point out that standard t-tests are reasonably robust when sample sizes are large, even in the presence of skewed cost data. Another approach that does not make such strong distributional assumptions is to use the non-parametric bootstrap (Efron and Tibshirani 1993).

While there is a consensus in the methodological literature on some issues such as the reporting of mean costs, the guidelines do not offer advice on how the study should deal with any systematic variations in mean costs across settings. Using data from a multicentre study Nixon and Thompson (2005) present the use of a multilevel model (MLM) that recognises mean differences in costs and cost-effectiveness within UK health care settings. While the MLM described acknowledges variation across centres it does not examine whether there are systemic reasons for the variation observed. In multinational economic evaluations, there may be large systematic differences in costs across settings, because of differences in for example, relative factor prices. However, the guidelines reviewed do not suggest how the analysis should identify any systematic differences in mean costs across settings. In addition, the distribution of the cost data may vary across the countries concerned, but again it is unclear how the analyst should tackle this.

2.4 Presentation and interpretation of results

The relevant issues that arise when presenting and interpreting the results of cost analysis and CEA are discussed below.

2.41 Resource use, unit costs and total cost

The guidelines acknowledge the importance of presenting costs disaggregated into measures of resource use and unit cost for each of the technologies in question (Drummond et al. 1997a, Luce et al. 1996, CCOHTA 1997). This helps those using the evaluation to assess whether the results apply to their local context. However, whether this level of disaggregation is sufficient for identifying reasons for cost

variation across health care settings is questionable. Unit costs may vary across settings because of differences in factor use or factor price. However, the guidelines stop short of recommending that unit costs are reported in a more disaggregated way that would allow differences across settings in these components of unit cost to be recognised.

2.42 Cost-effectiveness and sampling variation

Recent methodological developments in this area have shown that the incremental cost-effectiveness ratio (ICER) is generally an inappropriate summary statistic for CEA. The problem is that the standard error of the ICER is often intractable and summarising the sampling variation surrounding the ICER using confidence intervals leads to problems of interpretation. Instead, incremental net benefits (INB) and cost-effectiveness acceptability curves (CEAC) have become the preferred measures of cost-effectiveness (NICE 2004). Where a CEA is based on a RCT, Hoch et al. (2002) suggest that net benefits (NB) can be estimated on the cost scale as net monetary benefits (NMB) where:

$NMB_i = \lambda E_i - TC_i$

λ is the societal willingness to pay for a unit of health gain, E_i is the observed effect for an individual *i*, TC_i are the total costs for individual i. The mean INB is then the mean NMB for the treatment group minus the mean NMB for the control group.

The current methodological guidelines for submissions to the NICE technology appraisal programme recommend that if, the mean INB of a new intervention is positive at a threshold of £30,000 per QALY, then there is strong evidence in favour of its adoption (NICE 2004). However, it is important to present the uncertainty surrounding the cost-effectiveness estimate as this highlights the need for further research (NICE 2004). The advantage of using the INB is that as it is a linear expression, it does not suffer from the problems of the ICER, and the standard error of the INB can be estimated directly. The sampling variation surrounding the cost-effectiveness estimate can therefore be reported using stochastic measures of

uncertainty such as CEACs and 95% confidence intervals around the INB. Clearly the cost-effectiveness of a particular technology depends on the value used for λ, the societal willingness to pay for a unit of health gain. CEACs make this sensitivity explicit by plotting the probability that the intervention is cost-effective, given the data, against the value of the ceiling ratio λ (Fenwick et al. 2004). However, simply presenting a CEAC does not highlight any systematic variations in cost or cost-effectiveness across health care settings; in effect the observed variation is all attributed to inter-patient variability, rather than any systematic differences across settings.

Recent guidelines recommend that analysts should consider other forms of uncertainty (Briggs and Gray 1999). These forms of uncertainty can be defined as: methodological uncertainty, uncertainty related to patient characteristics, uncertainty related to extrapolation and uncertainty related to generalisability.

2.43 Methodological uncertainty

Much of this chapter has highlighted methodological issues that arise when estimating costs, and highlighted areas where there is a lack of methodological agreement, or lack of advice from the literature. To limit methodological variability across studies and enhance comparability, some guidelines have defined a reference case that stipulates particular methodological standpoints (Gold et al. 1996, NICE 2004). Other guidelines have stopped short of prescribing the use of particular methods, and have acknowledged that important areas of disagreement still exist (Johnston et al. 1999, Drummond et al. 1997a). Mason (1997) highlights that it may be difficult to use a consistent methodology across health care settings, particularly when measuring costs. For example, if highly aggregated cost estimates are used in different international settings, then the methods of cost allocation may differ leading to problems comparing results. For any study assessing cost variation across settings, it is necessary to minimise the methodological inconsistencies across the settings concerned.

2.44 Variability according to defined patient characteristics

The cost-effectiveness of an intervention may vary according to particular patient characteristics. Recent guidelines recommend the use of sub-group analysis to present cost-effectiveness results for different patient groups (NICE 2004). Hoch et al. (2002) extended the net-benefit approach to include the use of patient covariates. This allows the INB of an intervention to be reported for different patient sub-groups. In a multicentre study, patient characteristics may differ across settings, and it may be important to allow for these inter-patient differences, when assessing cost-effectiveness.

2.45 Variability according to the method chosen for extrapolating long-term cost-effectiveness

There may be important uncertainty about the method to use for extrapolating results. For instance if a model has been used to extend the results from a trial, to present cost-effectiveness over a longer time-horizon, the method chosen for extrapolation may be an important determinant of cost-effectiveness. Clearly, the choice of extrapolation method may vary according to the health care setting, as for example the relationship between short and long-term outcomes may vary across international health care settings. Where there is uncertainty about the assumptions used in making such an extrapolation, the guidelines suggest using sensitivity analyses (Briggs and Gray 1999).

2.46 Generalisability

Briggs and Gray (1999) define generalisability as:

> "...the extent to which the results of a study, as they apply to a particular patient population or setting and/or a specific context, hold true for another population and/or in a different context." (Briggs and Gray 1999, p7.)

The guidelines reviewed often highlight generalisability as a secondary issue to be addressed after the internal validity of the study's results has been established (Drummond et al. 1997a). Within the methodological literature on generalisability

there are three clear strands: the first refers to generalisability across different geographical settings (Glick and Cook 2003), the second relates to generalisability from studies based on RCTs to routine practice, (Mason 1997, Drummond and Davies 1991) and the third relates to generalisability over time (Sculpher et al. 1997). The first form of generalisability is the one most relevant for this thesis; the methods used to analyse variation in cost and cost-effectiveness across geographical locations may have implications for tackling this form of generalisability.

O'Brien (1997) and Mason (1997) suggest barriers to generalising results across different locations, which include: differences in the demography and epidemiology of disease, clinical practice and conventions, incentives and regulations, relative price levels, consumer preferences and opportunity costs and availability of alternative treatments. However, the authors do not provide a theoretical basis to justify why these particular factors may limit generalisability.

Jian et al. (1998) highlight the importance of understanding cost variation before making general conclusions about the efficiency of health care interventions:

"To identify the most efficient way of providing immunisation services.. a rough quantitative notion of the relative importance of factors contributing to cost variation is useful.." (Jian et al. 1998, p7).

However, a problem with assessing cost variation and its implications for generalisability is that there is a lack of evidence in the economic evaluation literature on why costs vary between different places (Goree et al. 1999).

Most of the guidelines focus on tackling issues relating to generalisability in the *reporting* of the results. For example, Byford and Palmer (1998) emphasise the importance of describing treatment alternatives and reporting separate estimates of resource use and unit costs, to make results transparent so that they can be interpreted in a more general context. However, these suggestions do not themselves make the results of a CEA generalisable to a different context. O'Brien (1997) emphasises the

importance of carefully scrutinising the relevance of the comparators used, the practice patterns and the price weights before applying the results to a different context.

Some of the guidance suggests dealing at the *analysis* stage with the problems posed by cost variation across settings. For example, although Jacobs and Baladi (1996) highlight the importance of trying to avoid bias when selecting the site for costing, they then suggest adjusting for this bias in the analysis. They recommend deriving this measure of bias by comparing the unit costs collected from the site in question with those used in national databases. Briggs (2000) suggests that where a study is aiming to produce a national estimate of cost-effectiveness it may be appropriate to use national estimates of unit costs. A model could then incorporate the variation in unit costs across settings as part of the overall estimate of uncertainty surrounding the study's results. Bryan and Brown (1998) suggest that it may be possible to make results from a study more locally applicable by re-analysing cost-effectiveness using local unit costs.

All these approaches assume that national unit cost databases of appropriate quality are available. They also ignore differences in resource use across settings which may be more important than unit cost differences in determining cost variation. As Johnston et al. (1999) point out it may not be possible to adjust for site selection biases at the analysis stage. For example, an appropriate national unit cost database may not be available. In a trial-based evaluation, the choice of study sites for collecting resource use and unit costs may be of crucial importance in determining the study's results, and the extent to which it can consider cost variation across settings.

As Briggs and Gray (1999) concede if suitable cost data are not available it may be difficult to use standard statistical techniques to address generalisability issues. Instead, analysts may rely on one-way sensitivity analysis to assess the applicability of results to different contexts. This would seem analogous to the situation before RCTs were used for collecting resource use data, when sampling variation was assessed using deterministic sensitivity analysis. The inadequacies of relying solely

on deterministic sensitivity analyses have been clearly highlighted (Manning et al. 1996). The fundamental criticism of using deterministic analyses to assess sampling variability, namely that the ranges used are at the discretion of the analyst, also applies when using this method to try to improve generalisability. Suitable datasets are required so statistical techniques can be used to analyse reasons for cost variation and the implications for the generalisability of results.

2.5 Discussion

This review of the methodological guidelines for economic evaluation had two main purposes. Firstly, the review aimed to identify gaps or areas of disagreement in the literature relating to the assessment of cost variation across settings. Secondly, the review tried to identify the key aspects of study design for an empirical investigation of cost variation across health care settings.

Ideally, an economic evaluation should use costs that represent efficient production of the health care technologies under evaluation i.e. opportunity costs. However, costs may vary across health care settings, and one reason for this is that health care firms in different geographical locations may have different incentives to cost-minimise. However, this review found little evidence to suggest that the guiding principles for economic evaluation considered cost variation across settings, and identified health care settings where the costs observed represented departure from opportunity costs. If the costs used do not represent opportunity costs then the study may produce misleading estimates of the relative cost-effectiveness of each health care programme. The issue of how to ensure that the costs used in economic evaluations approximate opportunity costs was not addressed by many of the guidelines including the recent methodological guidance issues by NICE (NICE 2004).

The conduct of an economic evaluation is in three main stages, the study design, data analysis and interpretation of the results. This review finds that at each stage there are omissions or disagreements in the guidelines about issues pertaining to cost variation across settings. At the design stage, while the guidelines highlight the appeal of using pragmatic clinical trials as a vehicle for measuring resource use data, there are still unresolved issues. For example, it is unclear how the number and type of centres for measuring resource use and costs should be determined. While there is some agreement that the centres chosen for resource use or unit cost measurement should be 'representative' or 'typical' of the context where the technology is going to be deployed, there is no clear definition of what constitutes a 'representative' or 'typical' centre. Recent guidelines highlight that more information is required on why costs

vary across centres. In particular, if costs vary because some centres are inefficient whereas others are relatively efficient, then it is inappropriate to use costs that are representative of all the centres concerned.

The guidelines generally treat the choice of centres for collecting resource use and unit cost data as separate issues, and in some instances, recommend that analysts use different sources. For example, some of the guidance states that in a multinational study it is acceptable to rely on pooled measures of resource use, and only estimate unit costs specifically for the country concerned. If resource use and unit costs are measured in different settings this contradicts economic theory. Cost function theory suggests that relative factor prices (a component of unit costs) in a given setting are correlated with relative factor use (a component of unit cost and resource use). The only approach that allows the data structure suggested by theory to be maintained, is to estimate separate unit costs, resource use and cost-effectiveness in each centre or country. In general the guidelines do not recommend taking this approach, although Halliday and Darba (2003) do suggest differences across countries should be recognised in the presentation of results. Estimating unit costs and resource use in each country would increase the costs of a multinational study and presenting cost-effectiveness results for each country would reduce the study's power to detect differences in costs between treatment groups. Subsequent chapters in this thesis examine conceptual and empirical reasons for cost variation and aim to inform the choice of settings for collecting resource use and unit cost data.

Some of the guidelines advocate the use of highly aggregated or national unit costs, as these have the potential to ensure that costs are 'representative' of the jurisdiction concerned. However, it is unlikely that these costs meet some of the other methodological standards recommended by the guidelines. In particular, it is difficult to ensure that the items included in the measure of unit costs are consistent across settings and that the unit costs represent opportunity and marginal costs. If costs are measured at a highly aggregated level it may be difficult to assess whether any differences in unit costs between health care settings, reflect differences in methodology or real differences. As Sculpher and Drummond (2005) highlight:

"... the higher the degree of aggregation, the greater the chance of hiding various methodological flaws." (Sculpher and Drummond 2005, p18)

It therefore seems more appropriate to use a disaggregated approach in the empirical investigation to compare costs across health care settings.

The guidelines fail to provide any clear advice on how the results should be *analysed* to consider reasons for cost variation across settings. In a multicentre study, either national or international, it is unclear how any systematic differences in mean costs across the centres should be identified.

Analysts have made recent progress in the methods used to analyse and present uncertainty in CEA. This progress has informed the most recent guidelines that have recommended using CEACs and INBs to present the results of CEA (NICE, 2004). These techniques are useful for summarising the sampling variation across the patients included in the study. However, in an economic evaluation based on a multicentre RCT there may be wide *systematic* variations in costs and cost-effectiveness across patients recruited from different centres. The guidelines suggest that there has been a relative lack of development in methods for dealing with this form of uncertainty, beyond recognising that results may not be generalisable across settings. Studies may be unable to assess systematic cost variation across settings if at the design stage, the study does not use an appropriate methodology for collecting resource use and unit cost data. If the study recruits insufficient patients, or measures costs using an inconsistent methodology across health care settings, then the study will struggle to distinguish between methodological differences, sampling variation or systematic differences in costs across the study settings.

A study of cost variation across settings can therefore inform methodological development in this area. This chapter also identifies methodological principles to inform the design of the empirical investigation. The study's perspective should be sufficiently broad, and the time-horizon long enough to capture a consistent range of

costs in each setting. The review also suggests that taking a disaggregated approach to costing could prove particularly helpful. Taking an aggregated approach would make any differences in unit costs across settings difficult to interpret. These differences could be due to inconsistencies in methodology, as opposed to systematic differences in factor use or factor price. Data from a multinational observational study or RCT could be used for comparing resource use and costs across international settings, but if data are collected from an observational study, it is particularly important to collect detailed information on patient factors, to minimise selection bias when comparing costs across countries. The review did not identify any guidance on how reasons for cost variation across settings should be analysed.

2.6 Conclusions

Concerns about the quality of economic evaluations have led to the development of guidelines aimed at improving the quality of these studies. The guidelines have generally reflected areas of agreement and recent methodological developments. The purpose of the more formal guidelines has been to standardise methods across different disease areas and interventions rather than across health care settings. For some of the policy-making guidelines, it is clear that economic evaluations are required that are sound and relevant for the national decision-making context. However, there are certain key areas of omission that a study assessing cost variation across geographical health care settings can usefully inform: covering the study design; analysis and interpretation of results (see Table 2.1). The lack of guidance may partly reflect, a lack of empirical work in the area of cost variation, and the next chapter reviews the relevant studies. There would seem scope for methodological development in this area that could inform subsequent guidelines, and help economic evaluations provide a sounder basis for health care decision-making.

Table 2.1: Areas where a study assessing cost variation across settings could inform the conduct of economic evaluations.

Study Design
• No centres for resource use or unit cost measurement?
• Choice of centres for resource use or unit cost measurement?
• Appropriate level of aggregation in resource use or unit cost measurement?
Study analysis
• How to pool resource use, unit cost or total cost data across settings?
• How to analyse cost data to allow for differences in the mean and the distribution of costs across settings?
• How to adjust the cost parameters in a model to estimate cost-effectiveness in different settings?
Presentation of study results
• How to present context specific results in multicentre studies?
• How to use statistical methods to make results more representative of the decision-context?
• How to consider the extent to which the results are generalisable?

Chapter 3: Evidence from the economic evaluation literature on cost variation across settings

3.0 Introduction

The previous chapter highlighted gaps in the methodological literature relating to cost variation across settings. This may reflect a lack of empirical evidence on reasons for cost variation. The aim of this chapter is to examine how published economic evaluations consider cost variation across settings by reviewing the relevant empirical evidence. Economic evaluations conducted alongside multicentre RCTs can collect resource use data from different locations; these studies can therefore consider cost variations across health care settings. In particular, if RCTs recruit patients from different countries, the accompanying economic evaluation may be required to present separate estimates of cost-effectiveness for different national decision-makers. These studies may therefore have an incentive to examine whether costs vary across international settings. In a cost-effectiveness analysis alongside a national multicentre RCT, if the results are intended for a national decision-making agency (for example NICE in England and Wales) then there may be less incentive to present setting-specific results. However, there may still be cost variations across settings in a national context, and this review considers some examples of studies that have considered this.

Methodology used in the literature review of empirical studies

This aspect of the literature review builds on the previous chapter by considering examples of how variation in costs, and cost-effectiveness across health care settings has been considered in the empirical literature. Rather than attempting to quantify the numbers of empirical studies that have used different methods, the purpose of the review was to provide examples of studies taking different methodological

standpoints in relation to the issue of cost variation across health care settings. These examples were identified by searching the PubMed and HEED databases using the search terms 'multicentre', 'multinational', 'costs and cost analysis' and 'econ eval' for the years 1990-2004. The bibliographies of the sources retrieved and from those articles included in the methodological review in Chapter 2, were reviewed to identify other relevant articles.

The articles retrieved were examined to see whether they raised methodological issues related to the issue of cost variation. Studies were included that raised relevant methodological issues that may apply more generally to studies with similar overall designs. The study designs considered were multinational economic evaluations alongside multinational RCTs, multinational economic evaluations based on models and systematic reviews, and economic evaluations based on national multicentre RCTs. The review was therefore structured around these different forms of study design and included examples of economic evaluations in each category. Within each category only studies that met the criteria for a full economic evaluation- costs and outcomes measured for two or more alternatives (Drummond et al. 1997), were included.

The first section of the review examines how economic evaluations alongside multinational RCTs consider cost variation across settings. Section two reviews studies that use cost-effectiveness models to see how they deal with cost variation across international health care settings. Section three reviews economic evaluations conducted alongside national multicentre RCTs. Section four describes the main methodological issues that arise, before section five discusses the gaps identified by the literature review.

3.1 Multinational Economic evaluations based on multinational RCTs

An increasing number of studies estimate the effectiveness of health care interventions using multinational RCTs (Cook et al. 2003) and many of these studies

now incorporate a multinational CEA (Reed et al. 2004, Lopez et al. 2003, Doyle et al. 2001, Rutten-van Molken et al. 1998, Schulman et al. 1996, Glick et al. 1998, Mark et al. 1995, Johannesson et al. 1993). These evaluations have the potential to measure the costs of the technologies concerned across a range of countries, and assess variation in costs and cost-effectiveness across international settings. However, as the previous chapter argues, decisions taken at the design stage, in particular the approach taken to unit cost and resource use measurement, may determine whether the study can investigate variations in costs across settings. The approaches taken to resource use and unit cost measurement fall into three broad categories:[17]

1. Studies that combine mean trial-wide measures of resource use, with a single or average set of unit costs.
2. Studies that combine mean trial-wide measures of resource use with country-specific unit costs.
3. Studies that use country-specific resource use and unit costs.

The following sections consider each of these approaches in turn using examples from the multinational economic evaluation literature.

3.11 Studies that combine mean trial-wide measures of resource use, with a single or average set of unit costs.

Some international CEAs have taken resource use estimates pooled over all the centres in the study and combined these with a single set of unit costs (Schulman et al. 1996, Glick et al. 1998, Mark et al. 1995, Johannesson et al. 1993). This study design does not allow for the analysis of variations in resource use or unit costs across countries. It is unclear whether these unit costs are those associated with efficient production, and therefore represent opportunity costs. In the economic evaluation of the FIRST study Schulman et al. (1996) combined trial-wide resource use measures taken from European and US patients with unit costs from a single American teaching hospital. The study reported average costs per patient for each of the technologies

[17] A fourth category of study allows outcomes to vary across settings. However, the role of outcome variation is outside the scope of the thesis.

concerned, and then incremental costs. The analysis ignored any cost differences across the international health care settings included in the study, and the unit costs were unlikely to apply to the European centres. The study could have collected unit costs from centres in each country and reported country-specific measures of cost-effectiveness.

In an economic evaluation conducted alongside a multinational Phase III RCT Lopez et al (2003) compared the overall costs associated with different antibiotics using resource use data pooled over trial centres in South America and Mexico. Unit cost data were taken from secondary sources and it was unclear whether they represented unit costs in any of the countries included in the RCT.

The 4S (Scandinavian Simvastatin Survival Study) (1994) recruited patients to an international RCT comparing simvastatin to placebo for secondary prevention of coronary heart disease. The trial included patients from Sweden, Norway, Finland and Iceland. The CEA collected highly aggregated unit costs from four hospitals in Sweden (Johannesson et al. 1997). The study estimated total costs per patient by combining these unit cost data with resource use estimates across all patients included in the multinational trial. The study presented estimates of cost-effectiveness that pooled costs across all the countries, and did not report any variation in the incremental costs or cost-effectiveness of the intervention across the countries concerned.

Studies have also collected unit costs from several countries, calculated an average set of unit costs, and then combined these average unit costs with trial-wide measures of resource use (Glick et al. 1998). An international economic evaluation of tirilazad mesylate for patients with subarachnoid haemorrhage, recruited patients from 11 countries covering centres in Australia, Europe and the United States (Glick et al. 1998). The study estimated average unit costs using data from six of these countries, and combined these data with trial-wide resource use estimates. In the initial CEA, the authors reported a single cost-effectiveness measure pooled across all the study centres (Glick et al. 1998). Although an ordinary least squares (OLS) regression

analysis showed that differences in resource use led to cost variations across countries, the authors did not assess whether there were any international differences in incremental costs, or incremental cost-effectiveness (Glick et al. 1998).

3.12 Studies that combine mean trial-wide measures of resource use with country-specific unit costs.

A number of studies have collected unit costs for each country included in a multinational RCT and then combined these with trial-wide measures of resource use (Henderson and Brown 1999, Jansen et al. 2001, Casciano et al. 2001, Lorenzoni et al. 1998). Some studies found that while unit costs varied widely across countries, estimates of incremental cost-effectiveness were more stable. In a coronary prevention study Henderson and Brown (1999) found that unit costs, in particular the costs of ICU, CCU and CABG[18], varied widely across countries, but the study used the same unit costs for the intervention in each setting and there were no significant differences across countries in the relative cost-effectiveness of the new treatment.

In a study, comparing recombinant tissue plasminogen activator to streptokinanse for preventing coronary heart disease, Lorenzoni et al. (1998) combined trial-wide resource use with country-specific unit costs. Despite the observed variations in unit costs, the ICERs did not vary appreciably across the countries concerned. While the unit costs of the interventions differed by up to five-fold across the settings, it was unclear whether the study used a consistent costing methodology across all countries. Drummond and Davies (1991) demonstrated the importance of a consistent costing method using as an example a multinational economic evaluation of new procedures for cataracts. Although this study reported wide variations in unit costs across countries, these were based on professional fees that varied widely across settings. For example, the fee for enucleation ranged from $2000 to $6000. These differences reflected the inconsistent costing methodology used. In some centres, the fee may represent the opportunity costs whereas in others the fee may include profit to the provider.

[18] ICU: Intensive care unit, CCU: coronary care unit and CABG: coronary artery bypass graft.

Other studies found that cross-country variations in unit costs led to important variations in cost-effectiveness (Hull et al. 1981, Jansen et al. 2001, Henderson and Brown 1999). Hull et al. (1981) found that the cost-effectiveness of different ways to diagnose deep vein thrombosis, varied by country according to differences in unit costs. In a study estimating the cost-effectiveness of different strategies for treating cardiovascular disease, Brown et al. (2001) collected unit costs from several different centres. Even though a consistent method was used to measure unit costs in each setting, unit costs varied widely within and across countries. For example, the unit costs of a PTCA varied from 1380 to 2700 Euros within the UK and from 1850 to 4000 Euros across the countries concerned.

A multinational economic evaluation comparing Terbinafine and Itraconazole as interventions for toenail onychomycosis included centres in Finland, Germany, Iceland, Italy, the Netherlands and the UK and found that the cost-effectiveness of the intervention varied across countries (Jansen et al. 2001). Terbinafine was relatively cost-effective in all countries except Finland where Itraconazole was the most cost-effective option. The main reason for this was the relative price of the interventions; the relative price of Terbinafine was higher in Finland, than elsewhere.

Each of these studies ignored any variations in resource use across the countries concerned, and any interrelationships between resource use and unit costs. The methodological issues this raises are considered in section four.

3.13 Country-specific resource use and country-specific unit costs

Several studies have used unit costs and resource use for each country (Willke at al. 1998, Rutten-van Molken et al. 1998, Jansen et al. 1997, Stalhammer et al. 1999, Berger et al. 1998; Reed et al 2003). Some of these studies have collected resource use and unit cost data for each country but then failed to report country-specific measures of cost-effectiveness (Reed et al 2004; Rutten-van Molken et al 1998). For example, Rutten-van Molken et al. (1998) compared the relative cost-effectiveness of formeterol to salmeterol for patients with asthma. The study showed that unit costs

varied widely across countries, for example the costs of a visit to the emergency room varied from $16 (Italy) to $346 (Spain) which was partly caused by inconsistencies in the costing method used across the centres. Regression analysis demonstrated that the costs for the control group were different across the countries concerned, however, the study did not examine whether the incremental costs of treatment varied by country. The authors highlighted that relative prices differed across the countries and this could potentially have an impact on the relative cost-effectiveness of the interventions. For example, an intervention that reduced clinician visits would appear more cost-effective in Sweden where physician costs were relatively high compared to France where physician costs were relatively low.

Several studies found that relative cost-effectiveness varied by country (Willke et al. 1998, Stalhammer et al. 1999, Berger et al. 1998). Stalhammer et al. (1999) used data from a multinational trial to assess the cost-effectiveness of omeprazole and ranitidine for management of patients with gastro-oesophageal reflux disease. The base-case analysis used trial-wide resource use and country-specific unit costs. A sensitivity analysis used country-specific resource use and unit costs. While the pooled result suggested that omeprazole was the least-cost option there were important variations across countries; in France and Ireland ranitidine was the lowest cost option, whereas omeprazole was the least-cost option in Italy and Spain.

Willke et al. (1998) examined the impact of using country-specific resource use and unit costs as opposed to pooled estimates. The authors extended the CEA of tirilazad mesylate for patients with subarachnoid haemorrhage (Glick et al. 1998). From the 11 countries originally recruited to the study, unit costs were collected in five countries. The initial analysis pooled resource use, unit costs and outcomes across the five countries and estimated that the mean ICER was $45,892 per death averted (Glick et al. 1998). Willke et al. (1998) used OLS regression models to estimate separate mean ICERs for each country. These regression models were initially based on trial-wide resource use data and unit costs for each country. Allowing unit costs to vary across countries meant that the mean ICER for the intervention ranged from $46,818 in country 1 compared to $69,145 in country 4 (Table 3.1). However, when country-

specific resource use data were combined with country-specific unit costs, the ICERs varied considerably, in country 5 the intervention was dominant (better outcomes, lower costs) whereas in countries 2-4 the cost-effectiveness ratio exceeded $90,000 per death averted. Thus, once the study used country-specific resource use and unit costs the countries fell into two groups: those where the intervention was on average cost-effective (countries 1 and 5) and those where the intervention may not be cost-effective (countries 2, 3, and 4). The initial analysis that used average measures of resource use, and presented a pooled estimate of cost-effectiveness led to misleading conclusions (Glick et al. 1998).

Table 3.1: ICERs ($ per death averted) for tirilazad mesylate compared to placebo for treating subarachnoid haemmorhage

Country	trial-wide resource use country-specific unit cost	country-specific resource use country-specific unit cost
1	46,818	5,921
2	57,636	91,906
3	53,891	90,487
4	69,145	93,326
5	65,800	***
Average	45,892	45,892

*** Intervention has lower costs, improves outcomes and therefore dominates placebo, but the paper did not report ICER. Source: Willke et al. (1998).

Willke et al. (1998) considered cost variation across international settings more carefully than the other studies reviewed. However, while this study demonstrates the importance of considering variation in resource use as well as unit costs across countries, it suffers from important limitations. A problem with presenting country-specific estimates is that the study loses power to detect differences in costs and cost-effectiveness between the treatment and control groups. By just presenting mean estimates this study fails to report sampling uncertainty. This is important as although

the between-country differences in the mean estimates appear large, they could just reflect random variations.

The study does not consider whether there were any systematic reasons for the differences in costs and cost-effectiveness across the countries concerned. There may be differences in the characteristics of the providers or patients across these countries that lead to the observed differences in incremental costs, and incremental cost-effectiveness. Assessing the reasons for cost variation would allow the cost-effectiveness results to be stratified based on factors that define international differences in patterns of care across the centres. This would make the results more useful, for decision-makers not represented by countries included in the study. There may be circumstances where none of the study centres' costs are likely to approximate opportunity costs in a different decision context. In these circumstances, further primary data collection may be required.

3.2 Multinational Economic evaluations based on models and systematic reviews

Another form of study design that has the scope to consider cost variation across settings is when an economic evaluation is conducted using a model or a systematic review. Cost-effectiveness models have been recommended for transferring cost-effectiveness data to different decision contexts (Drummond and Davies 1991). For example, information on effectiveness and resource use might be available from a clinical trial in a different location to the decision context. A study may then collect specific unit costs for the study context and use a model to combine these unit costs with more general information on resource use and the effectiveness of the interventions. However, a problem in these modelling studies is that the costing methods used may be inadequate, in particular any correlation between resource use and outcomes, or resource use and unit costs may be lost if different elements of cost-effectiveness data are transferred from different contexts. Another problem with assessing cost variation in model-based economic evaluations is that these studies

may not use a consistent methodology in each health care setting (Barbieri et al, 2005).

There are examples in the literature of model-based CEA that use a pooled estimate of resource use from a multinational RCT, alongside country-specific unit costs. For example, Casciano et al. (2001) evaluated Doxezosin for controlling blood pressure and found that cost-effectiveness varied between the UK and Italy because of differences in unit costs.

Other studies have used country-specific resource use and unit costs. Menzin et al. (1996) used a model to extend results from a US trial of rhDNase in adults with cystic fibrosis. The model assessed the cost-effectiveness of the intervention in France, Germany, Italy and the UK. Although originally the analysis used US resource use patterns from the RCT alongside country-specific unit costs, this led to inaccurate estimates of cost-effectiveness in each of the European decision-contexts. Instead, the authors used evidence from local observational databases to adjust the overall resource use estimates from the trial for each local context. The adjusted results were used in the model to provide country-specific estimates of the relative cost-effectiveness of the intervention. The country-specific estimates suggested that the intervention was more cost-effective in Germany where morbidity costs were higher, than in Italy and France where morbidity costs were lower. However, the study design did not consider differences in the way costs were measured in each country. For example, it was unclear whether the same resource use items were included in each setting. This makes it difficult to assess whether the differences in cost-effectiveness across the study settings were due to systematic differences or methodological inconsistencies.

Jansen et al. (1997) found that meloxican was relatively cost-effective compared to diclofenac for treating osteoarthritis across all the countries included in the study. However, the cost savings were greater in France than in the UK or Italy. Heaney et al. (2000) compared four different antidepressants and found similar results across the

12 countries included in the study, even though there were inter-country differences in resource use and unit costs.

De Pouvourville and Tasch (1993) used a model to estimate the relative cost-effectiveness of misoprostol compared to NSAIDs[19] for treating gastrointestinal disease in Belgium, France, the UK and the United States. Although there were differences in clinical practice and costs across the countries included, this did not affect the relative cost-effectiveness of the respective interventions. Drummond et al. (1991) considered the methodological issues that arose when using a decision-analytical approach to estimate the cost-effectiveness of misoprostol in different countries. However, the study used Delphi panels to estimate resource use, which clearly limited the internal validity of the results. Unit cost data were taken from different sources in each country; the study used hospital charges for centres in Belgium and the US, whereas in France and the UK specific costing studies were conducted. This may partly explain the wide international variations in the unit costs.

As the authors themselves state:

".. the non-comparability of hospital finance systems is likely to be a serious impediment to cross-national comparison." (Drummond et al. 1991, p678).

There was international variability in both the duration and the intensity of resource use, with the US having shorter LOS but higher resource use during that stay. The intervention was more cost-effective in the US than in the other countries. However, whether robust conclusions can be drawn based on studies that take resource use data from expert opinion is doubtful.

3.21 Systematic reviews of cost-effectiveness studies in different countries

If the economic evaluation is conducted alongside a systematic review this raises further issues about how to incorporate cost variation across settings. An economic evaluation based solely on a systematic review requires that data are pooled across different studies. While this may be judged appropriate for the outcome data

[19] NSAIDS: nonsteroidal anti-inflammatory drug.

methodological differences across these studies may lead to methodological problems with pooling cost and cost-effectiveness data. Jefferson et al. (1996) considered these issues when they reviewed economic evaluations of influenza vaccination. The review identified 10 economic evaluations in this area conducted across different geographical locations. The review considered methodological differences across studies including the choice of currency conversion factor, and inconsistencies in the methods used to measure costs. The study concluded that the choice of conversion factor made little difference to the results. Of greater concern was that in four out of the ten studies reviewed, different resource use items were included which partly explained the variation in the results. Of those studies that measured similar items of resource use there was considerable variability in the levels of resource use, in particular there were differences in hospital LOS that led to differences in the study's conclusions.

In a systematic review, Späth et al. (1999) assessed whether published economic evaluations of adjuvant therapy for breast cancer could provide useful information for setting health care priorities in France. Although six studies met the review's inclusion criteria the authors found that, none of these studies could inform French decision-makers. The studies did not identify the resources use included or report the unit costs. The authors conclude that if secondary data cannot be transferred from the study setting to the decision-context, then a local economic evaluation is required.

3.3 Economic evaluations based on national multicentre RCTs

When economic evaluations are conducted alongside RCTs in a single country or jurisdiction they usually still recruit patients from several different centres (see for example UK Small Aneurysm Trial Participants, 1998). The same decisions about how and where to measure resource use and unit costs have to be faced as in a multinational study. However, there are a number of important differences in this context compared to the multinational setting. The most important difference is that these studies are more likely to be tailored to a national decision-making agency, for example NICE. Here it is more relevant to present a national estimate of the relative

cost-effectiveness of the intervention, rather than an estimate of the cost-effectiveness for each setting. However, even when it is desirable to present national estimates, the choice of centres for measuring resource use and unit costs could still be important in determining the results.

A common approach in national multicentre economic evaluations is to measure resource use for each patient and combine this with an overall measure of unit costs (see for example MRC Laparoscopic Hernia group, 2001; Longworth et al. 2001; Elbourne et al, 2002). These unit costs may be national estimates, averages across all the study centres, or specific estimates from a particular centre (UK Small Aneurism Trial Participants, 1998; Longworth et al. 2001; MRC Laparoscopic Hernia group, 2001). Studies then presents an overall cost per patient for each treatment arm and then an incremental cost pooled over all the treatment centres. This approach does not allow any differences in resource use or unit costs across the centres to be identified and explicitly considered in the analysis. It is therefore difficult to assess whether the costs used represent the efficient provision of health care programmes.

Following the recommendations of Drummond et al. (1997a) studies tend to report separate resource use and unit costs across all the centres (see Graves et al. 2002 for a review), and then use sensitivity analysis to test the impact of changing the unit costs used. For example, Coast et al. (1998) compared a hospital at home scheme with conventional care. In the main analysis, average unit costs were taken from a single national source (Netten and Dennett 1996), and then in the sensitivity analysis the robustness of the conclusions was tested by using different estimates for the cost per hospital day. Sculpher et al. (1993) took a similar approach when comparing the cost-effectiveness of abdominal hysterectomy with transervical endometrial resection for patients with menorrhagia. Resource use data were taken from a national RCT with unit costs collected from one hospital and applied globally. In the sensitivity analysis, high and low estimates of hospital per bed-day costs across all hospitals in the UK were used to test the robustness of the results.

In these examples, the issue of cost variation across settings is relegated to a one-way sensitivity analysis, which can limit the scope for fully addressing the impact of cost variability. To assess the impact of cost variation the studies only changed estimates of unit cost and ignored any interrelationship between resource use and unit cost. The analyst has completion discretion over the range of estimates used in the sensitivity analysis. The methodology does not allow the study to fully investigate cost variation across settings and the implications for the stability of the results.

3.4 Methodological issues the review raises concerning cost variation across settings.

The empirical studies reviewed were primarily concerned with providing estimates of cost-effectiveness, rather than addressing methodological issues related to cost variation. However, the review raised some important methodological issues for any study that attempts to identify systematic reasons for cost variation across settings.

3.41 Ignoring interrelationships between factor price and factor use

Raikou et al. (2000) highlighted the problems with using location-specific resource use but relying on average unit costs. They contrast this approach with one whereby resource use data for each patient are combined with setting-specific unit costs. While studies that use an average measure of unit costs may presume that the results would be similar compared to approaches that estimate unit costs for each centre, economic theory suggests that these approaches may generate different results. If relative factor prices differ across health care settings for example geographical locations, then individual health care firms may choose different combinations of factor inputs. Raikou et al. (2000) show that in this case using average unit costs would overstate costs compared to using setting-specific unit costs. Using average unit costs fails to allow for the substitution of those inputs that are relatively low cost for those that are relatively high cost.

Raikou et al. (2000) used simulated data to illustrate the likely differences between approaches using average and centre-specific unit costs. One situation where the two approaches led to similar results was when the change in resource inputs was highly stochastic. However, the authors conclude that in general there would be differences between the two approaches unless health care centres do not respond to differences in unit costs in a manner consistent with economic theory. To understand whether health care centres do respond to differences in unit costs in the way suggested by economic theory, the authors argue strongly that empirical studies are required; they state that:

". These findings highlight the need for more caution in multicentre studies. They also emphasise the need for more and better information on the production process in the health care sector. Little is known about the degree of factor substitution in this sector." (Raikou et al. 2000, p197).

Goree et al. (1990) also emphasise the importance of recognising the potential relationship between resource use and unit costs, they state:

"Drummond and Davies (1991) suggest that resource quantities and unit prices should be reported separately so that costs can be recalculated for settings where price levels differ. ....this solution ignores the inseparability of prices and quantities .." (Goree et al. 1999, p570).

Economic theory would therefore suggest that it is more appropriate to collect resource use and unit costs in each health care setting, although there is a need for studies that consider whether health care firms do adjust factor use according to differences in factor price in the manner predicted by theory.

3.42 Reasons for observed variations in unit costs across settings

Several of the studies reviewed attempted to identify reasons for unit cost variation (Schulman et al. 1998, Johnston et al. 1998, Hutton 2001). Schulman et al. (1998) examined the reasons for unit cost variability across countries. In a further reanalysis of the study of tirilazad mesylate for the treatment of subarachnoid haemorrhage unit costs were estimated in seven out of the 11 countries included in the RCT. Where unit costs were unavailable in a particular country they were estimated using an imputation

method based on the relative unit costs in that country compared to the others. To improve methodological consistency across the countries, the study only included 'variable costs' and excluded overheads, and capital costs. However, this definition of variable costs is arbitrary, and other costs (e.g. staff costs) may remain fixed in the short-term. Also, items that are defined as overheads in some countries may be included as consumables in another. A generalised linear regression model was used to try and establish reasons for variability in unit costs. The results showed that the type of procedure explained 44%, and country 11% of the overall variation in unit costs. However, for certain procedures there were wide variations in unit costs across countries. The study did not examine why there may be international variations in unit costs across countries. The analysis did not estimate the effect of differences in case-mix, and the intensity of resource use across countries.

Johnston et al. (1998) examined factors that may be associated with differences in unit costs between trial and non-trial centres providing a breast cancer screening program in the UK. The study examined what characteristics differed between trial and non-trial centres, and found that the trial centres were larger, had lower uptake rates, and had more highly skilled nursing staff. The study estimated the effect of these centre characteristics on unit costs. The results suggested that having a higher level of qualified staff and a lower uptake rate of the screening service led to higher unit costs for the trial centres. By assessing differences in the characteristics of the trial and non-trial centres the study provided an assessment of how 'representative' the unit costs were from the trial centres.

By identifying the centre characteristics that were associated with unit costs, the study was able to adjust unit costs to be more generally applicable. This study may therefore have identified unit costs that were more representative of the opportunity costs for the decision context (the UK NHS).

Hutton (2001) conducted a detailed assessment of reasons for unit cost variation in a developing country context. The study used data collected alongside an antenatal trial that included centres in South Africa, Thailand and Cuba. The results suggested that

there were wide differences in unit costs across the study settings that were associated with differences in resource productivity, occupancy levels, staffing patterns, exchange rates, input mix and the size of health care facilities. However, the study had insufficient health care centres to identify which centres were the most productively efficient, and hence had the unit costs closest to opportunity costs.

3.43 From how many centres should unit costs be collected?

The number of centres or countries chosen for unit costing is often arbitrary. Unit costs are frequently only collected in a sub sample of study centres and it is unclear whether these unit costs represent opportunity costs. In a multicountry study, Glick et al. (2002) collected unit cost data in four countries (Belgium, France, Spain and the UK) for 47 different hospital diagnoses. The authors tried to estimate what the error in the cost estimates would be if unit costs were only collected from a sub sample of the countries. The unit costs collected from one or more of the countries were omitted and instead the analysis imputed these unit costs based on the costs from the remaining countries. The study assessed how many countries it was appropriate to collect the unit cost data from based on the level of error in the unit costs estimated; this error decreased according to the number of countries used for the imputation. The results showed that when the imputation was based on the results for only one country, the imputation error was high (87% of the mean unit cost estimate), whereas when the unit cost imputation was based on data from three countries, the error was correspondingly lower (68% of the mean unit cost estimate). Clearly, the imputation error even with using three countries to impute unit costs is still large and may have a bearing on the accuracy of the cost-effectiveness measure. The number of countries included in the analysis was relatively small for making any definitive conclusions about the number of countries required for estimating accurate unit costs. In addition, the study failed to identify reasons for the inter-country cost variation and the implications for measuring opportunity costs.

3.44 From which centres should unit costs be collected?

Goree et al. (1999) highlighted that as well as the number of centres, the characteristics of the centres are also important in determining unit costs. The authors

compared three methods for collecting unit cost data: one based on selecting the most convenient hospital, one based on selecting a sample of convenient hospitals and the third involved selecting a sample of hospitals by stratifying for factors potentially associated with cost variation. The costs per patient for different ways of inducing labour for pregnant women were reported for each of these strategies. The results showed that the median costs per patient for each strategy varied by about 30% according to the method used for collecting unit costs. In some cases, the ranking of the relative costs of the respective treatments varied according to the methodology used. The results confirm that the choice of methods for estimating unit costs can have an important impact on the absolute and relative costs of different treatment strategies.

Goree et al. (1999) put forward a research agenda to try to address this gap in the methodological literature. They suggest that a better understanding of cost variation across hospitals is needed. They recommend that unit costs are collected from as many hospitals as possible. Multivariate analyses can then identify the important cost drivers, for example teaching status, urban or rural location and labour and capital mix may be associated with unit costs. The results from these studies could provide further guidance on the number or type of hospitals required for unit costing.

3.45 Reasons for resource use variation

Hutton (2001) examined reasons for resource use variability across health care settings using data from a multinational RCT. This study found that there were wide variations in resource use both within and across countries. The study used OLS regression methods to assess factors associated with resource use variation. The results found that these variations were associated with differences in case-mix, clinical practice and accessibility. However, most of the resource use variation was left unexplained.

In a multinational study, Bennett et al. (1995) investigated the use of adjunct therapy for bone marrow transplantation using data from hospitals in Paris and New York.

The study found very different patterns of resource use across these settings, and attributed these differences to variations in practice patterns. However, as the study only included two countries it was not possible to undertake statistical analyses of the reasons for international resource use and cost differences.

Holmes et al. (1997) compared the resource use following interventions for cardiogenic shock between US and non-US countries. The results showed that diagnostic and therapeutic procedures were used more frequently in the US than other countries, however no reasons were offered for the observed variations.

3.46 Statistical Methods to analyse differences in costs and cost-effectiveness

Multicentre studies that collect individual patient data can use statistical techniques to analyse variations across health care settings. In a re-analysis of the 4S cost-effectiveness study Cook et al. (2003) used statistical techniques to examine whether the cost-effectiveness of simvastatin compared to placebo varied across study centres. The study reported the incremental cost-effectiveness of simvastatin for each country using ICERs and incremental bet benefits (INB) on the cost scale[20] (Table 3.2). The cost-effectiveness results showed that the overall incremental cost per additional survivor was $59,201. The mean ICERs ranged from $47,032 to $129,474 across the countries. The mean INBs were reported with 95% confidence intervals to summarise the sampling uncertainty. The results showed that with a ceiling ratio (λ) of $75,000 per additional survivor, the mean INB was positive in three of the countries, but negative in Finland and Iceland. Cook et al. (2003) used tests for heterogeneity to examine whether there were significant interactions between treatment and country.

[20] Also known as net monetary benefits (NMB).

Table 3.2: Incremental cost-effectiveness of simvastatin compared to placebo

Country	N	ICER	Mean INB[1] (95% CI)
Denmark	355	47,032	1,208 (-2,648 to 5,065)
Norway	511	43,526	1,371(-1,546 to 4,288)
Sweden	845	58,208	542 (-1,626 to 2,711)
Finland	433	120,363	-850(-3,414 to 1,714)
Iceland	79	129,474	-637(-7,359 to 6,086)
Overall	2223		525(-826 to 1,876)

[1]λ=$75,000 per additional survivor. Source: Cook et al. (2003).

The tests for interaction showed the mean INB did not differ significantly across the countries. The study concluded that the hypothesis that there was significant variation in the cost-effectiveness of simvastatin across the countries, could be rejected. Instead, the study suggested it was acceptable to pool these measures of cost-effectiveness across the countries concerned.

A major problem in using these tests for interaction is that there may be insufficient cases to detect whether significant interactions exist. The authors concede that even in this CEA which had a relatively large sample size (n=2,223), the power of the tests to detect an interaction in the INB across countries was only 12.1%. The conclusion that there were no significant interactions and that it was acceptable to pool the results was limited by the study's lack of statistical power. Also, the results were based on unit costs collected in one country. Depending on the interrelationship between unit costs and resource use, the cross-country differences in the INB may have been underestimated.

In a national multicentre economic evaluation, Coyle and Drummond (1998) tried to establish which factors were responsible for resource use and cost variation. They

used data from two trials comparing different methods of radiotherapy, one for patients with head and neck cancer the other for patients with cancer of the bronchus. The study measured unit costs and resource use in each centre. The results showed that in both studies there was significant variation in costs amongst patients in the intervention arm. The authors used OLS regression analysis to identify reasons for this variation. The results suggested that variables summarising overtime payments to staff, clinical practice differences and patient case-mix were significant. The use of OLS regression analysis meant the study could not reach definitive conclusions about which factors were most important. The OLS regression analysis assumed that each patient observation was independent, whereas patients may be clustered within health care centres. In their discussion the authors emphasise the exploratory nature of the work and that further research is required to consider the potential importance of centre characteristics in explaining cost variation. They suggest that:

"it may be worthwhile that certain centre characteristics are recorded to assist the interpretation of multicentre data.... Given the results of this study collection of unit cost data from at least a sample of treatment centres is necessary for all studies based on a number of treatment centres.." (Coyle and Drummond 1998, p162).

In general, economic evaluations have not used statistical methods to assess variations in resource use, unit costs, total costs and cost-effectiveness across health care settings. This meant the studies reviewed could not make definitive conclusions about the factors driving cost variation across settings.

3.5 Discussion

The purpose of this chapter was to consider how economic evaluations have considered cost variation across settings. The review found many examples of studies that have taken an international approach to economic evaluation. These studies covered different decision-making contexts and therefore had reason to consider cost and cost-effectiveness variation across settings. However, these studies failed to adequately assess this cost variation because of limitations in the studies' design. In particular, these studies did not measure unit costs in all centres using a consistent methodology. Unless unit costs are measured in the same way in each setting it is difficult to identify systematic differences in for example factor prices, or levels of technical efficiency across settings. Studies have generally included insufficient centres to identify reasons why costs vary across settings. These studies have failed to assess whether cost differences exist across centres because of variations in efficiency or because of other factors for example case-mix differences or random variation. It is important to identify those centres that produce care efficiently to ensure that the unit costs used in an evaluation represent opportunity costs.

Another approach is where economic evaluations have measured setting-specific unit costs and combined these with general measures of resource use. Some of these studies found that there were important variations in unit costs that led to differences in the cost-effectiveness of the intervention. However, as Raikou et al. (2000) point out combining setting-specific unit costs with general measures of resource use may not be appropriate. Economic theory suggests firms adjust their use of factor inputs according to differences in factor prices. Thus, it is theoretically incoherent to collect unit costs and resource use from different settings and ignore possible correlation between the different parameters.

A few multinational economic evaluations have demonstrated that cost variation across settings can lead to differences in cost-effectiveness. The Willke et al. (1998) found international variations in resource use and that using country-specific measures of resource use led to wide variations in cost-effectiveness. The pooled

results suggested the intervention was universally cost-effective whereas the country-specific results suggested that the centres could be divided into two groups based on cost-effectiveness. The study did not establish why these two groups had such different estimates of relative cost-effectiveness.

Johnston et al. (1998) found that even within a particular decision context, for example a national health care system, there were important variations in cost. The choice of centres for cost measurement in national economic evaluations should be based on the principle of opportunity costs. The different forms of economic evaluation reviewed have neglected this and made arbitrary decisions about which centres to include in the cost analysis.

None of the studies examined provided a comprehensive assessment of which factors were associated with resource use and cost variation. While commentators have provided a suggested list of factors (O'Brien 1997, Mason 1997) these did not feature in the empirical literature. It has been suggested that studies should collect data on covariates that capture variation amongst patients and centres (Brown et al 2004). However, it is not clear how these covariates should be identified when designing the study.

Methodological failings limited the studies' scope for identifying reasons for cost variation. Often cost data were not collected in sufficient centres to examine reasons for cost variation. In both the model-based and the RCT-based evaluations, there were methodological inconsistencies in cost measurement across health care settings, especially where a highly aggregated approach was taken. Even in those studies that used a consistent methodology across settings only a small number of settings were included in the cost analysis (less than 7), an aggregated approach was taken to resource use and unit cost measurement, and an inappropriate statistical methodology was used.

To examine which factors are associated with cost variation studies are required that use insights from economic theory, and collect disaggregated cost data using a consistent costing methodology across a range of settings. However, these are necessary but insufficient criteria for identifying reasons for cost variation; an appropriate statistical methodology is also required. The study by Cook et al. (2003) presented a statistical methodology for assessing cost variation across settings using tests for interaction. However, this method had insufficient power to detect differences across setting even in a study with over 2000 cases. Coyle and Drummond (1998) used OLS regression methods to analyse variation in costs across settings. This approach assumes that the individual observations are independent. However, this assumption may be implausible in multicentre studies as patients may be clustered within settings. There is scope for methodological development in this area and Chapter 5 reviews different techniques for assessing cost variation.

3.6 Conclusions

While commentators suggest that costs are likely to vary across health care settings (O'Brien 1997, Mason 1997), this review found few empirical studies that have carefully examined cost variation across settings (Goree et al. 1999, Hutton 2001). Many of the studies reviewed suffered from serious methodological limitations that hindered cost comparisons across health care settings. Studies are needed that use appropriate methodologies that can identify reasons for cost variation. This methodology should be based on economic theory, use a disaggregated and consistent approach to cost measurement in sufficient health care settings, and rely on an appropriate statistical method. The seriousness of this gap in the literature is highlighted by a commissioning brief from the NHS Health Technology Assessment Programme (1998), and recent initiatives from the WHO that aim to analyse unit cost data across many different countries (Murray et al. 2000). The next chapter considers how relevant strands from microeconomic theory can inform methodological developments in this area.

Chapter 4: Review of theoretical and empirical evidence for cost variation across settings

4.0 Introduction

The previous chapter highlighted gaps in the economic evaluation literature on why costs may vary across health care settings and the implications for the design, analysis and interpretation of economic evaluations. This chapter reviews relevant theoretical and empirical literature to provide insights into the reasons for cost variability across health care settings. The review is divided into four sections; the first section considers production and cost functions, the second section differences in the health care context, the third section differences in patient factors, and the fourth section the measurement issues that arise when comparing costs across health care settings. For each of these areas the concepts are outlined, before the evidence on cost variation across health care settings is assessed. The aim of this chapter is to generate hypotheses for why costs may vary across health care settings, and to provide a theoretical framework for the empirical investigation.

4.1 Review of theoretical concepts on production and cost functions

This chapter begins by reviewing the key concepts underpinning the production and cost function literature. The production function specifies the relationship between the inputs used and the level of output produced. The production function has been described as:

> "..... the core concept in the economic theory of production. It is critical to know how much output can be produced with certain inputs, and to know if there are alternative ways of producing the same product" (Heathfield and Vibe 1981, p4).

The production function specifies a purely technical relationship between the quantity of inputs used and the level of output produced, under the assumption that the firm aims to minimise the costs of producing a given level of output. Where one homogenous output (q) is being produced from various inputs (v_1.....v_n) then the production function may be stated as: $q=f(v_1....v_n)$. Knowing the technical relationship between inputs and outputs is a necessary but not sufficient condition for cost minimisation, which also requires knowledge of relative factor prices.

In the most basic production function, the factors of production are divided into two groups labour (v_1) and capital (v_2) producing a one dimensional, homogenous output (q). In the short-run capital is assumed to be fixed, whereas labour is variable, but in the long-run the levels of both factors can be varied.

The production functions provide a way of defining the alternative production methods and specifying the range of technical possibilities that are open to producers. The notion that there are different ways of producing the same good is often ignored in industries such as health care where providers may subscribe to what Fuchs has termed the 'monotechnic' view (Folland 1997). In fact different quantities of various care inputs: doctors' time, nurses' time, therapeutic interventions, and capital may be combined to produce a particular level of output.

There is a wide literature documenting the variations in resource use that exist across health care settings, this literature is often termed the 'practice variations literature'

(McPherson et al. 1982, Wennberg 1984, Phelps and Mooney 1993). Much of the literature comes from the US where large administrative databases have been used to compare resource use across different geographical areas (see the review by Phelps and Mooney 1993). The results have suggested that for some procedures the reasons for medical intervention seem clear and the level of intervention is similar across regions. For example, intervention for hip fracture or acute myocardial infarction (MI) is similar across different geographical areas (Phelps and Mooney 1993). However, variations appear much larger for non-surgical admissions to hospital (Phelps and Mooney 1993; Phelps and Parente 1990, Wennberg 1984, McPherson et al. 1982). Phelps and Mooney (1993) point out that hospital care for cardiac and mental illness seem to dominate lists of hospital admissions with high variability across regions. Beech et al (1996) documented wide variation in the LOS in the acute hospital following stroke across centres in Western European countries. Wolfe et al. (1999) showed that in centres with different ways of managing stroke care there were also variations in outcomes post-stroke.

However, these studies tend to simply document that such variations exist without detailing *why* they might occur. One reason why variations in resource use might occur is if substitution is occurring. Often clinical practice variations are greatest when various alternatives exist. For example, within orthopaedics the rate of hospitalisation for hip fractures has a small coefficient of variation (COV)[21], whereas for fracture of the forearm the COV is much higher, in part because more alternatives exist (e.g. ambulatory surgery, or direct treatment in an emergency room, Wennberg (1984). To assess reasons for variations a broad range of production processes have to be considered.

4.11 Isoquants, factor substitution and technical efficiency

Isoquants (meaning equal quantities) show the possibility of factor substitution in producing a given level of output. The slope of the isoquant illustrates the extent to which factor substitution is possible. The slope is the ratio of marginal products, so a change in the marginal productivity of labour or capital will change the slope of the

[21] In health care, there may be a lack of substitution because of regulations or incentives in the health care setting that prevent substitution taking place, these are reviewed in section 4.2.

isoquant, which will have implications for the input mix chosen. The isoquant represents a frontier of technically efficient production. Isoquants are commonly drawn with a negative slope and convex to the origin, to illustrate the diminishing returns from using a particular factor input. This arises when factors of production are imperfect substitutes for one another. As the proportion of say labour relative to capital increases, then assuming that the same level of output must be maintained, the amount of labour that can be traded-off for a given amount of capital decreases. However, if the factors of production are perfect complements, then one factor cannot be substituted for another.

In certain areas of health care, it may be possible to choose different input mixes while still producing a technically efficient level of output. However, different factor mixes may also be signs of different levels of technical efficiency which can lead to differences in total cost.

Some studies have used a production function framework to analyse differences in resource use or more specifically the mix of resource inputs used to produce a particular good. Jensen and Morrisey (1986) analysed the effect of physician inputs on the level of health care provided. They found that physicians had a strong positive influence on the productivity of other inputs and were substitutes for other resources. This study illustrated the importance of case-mix adjustment: after adjusting for case-mix the marginal product of labour increased relative to capital. The study also demonstrated that there was a high potential substitutability of nurses for medical staff. This theme has been followed up by a number of studies reviewed by Richardson and Maynard (1995) who suggested that between 30% and 70% of the tasks currently carried out by doctors could be done by nurses. However, the review criticised the design of the studies, in particular little account was taken of the outcome of care provided. Also, none of these studies evaluate why substitution may or may not occur.

Parente (1989) examined substitution of outpatient for inpatient care for various clinical conditions, across different regions in the Rochester, New York area. The study found that outpatient care for knee procedures was negatively correlated with

the use of inpatient care across regions. However, for most of the other clinical areas, the use of inpatient and outpatient care was positively correlated, indicating that the services were complements rather than substitutes (Table 4.1). Similarly, Phelps and Mooney (1993) found positive correlation between the uses of particular procedures that could *a priori* be defined as substitutes. The only evidence of substitution was in the use of ICU and non-ICU bed-days for patients with MI, angina or chest pain.

Table 4.1: Correlation of inpatient and outpatient resource use

	total	urban	rural
Procedure			
Cataracts	0.38	0.43	0.00
Glaucoma	0.44	0.40	0.99
Tonsillectomy and adenoidectomy	0.09	0.18	-0.17
Varicose veins	0.48	0.81	-0.22
Haemorrhoids	-0.09	0.13	-0.34
Knee procedures	-0.25	-0.72	0.12

Source: Parente (1990)

In their seminal paper using a production function framework to analyse the substitutability of various health care inputs, Cowing and Holtmann (1983) found substantial levels of substitution between nursing and professional workers, nursing and general workers, nursing and administrative workers and professional and administrative labour. While this study found some evidence inputs were complements, hospitals generally appeared to have good flexibility in substitution between different labour inputs, even in the short-run. More recently, Carr-Hill et al. (1995) examined the impact of changing the nursing skill-mix on the output of nursing care. They found that reducing the number of untrained nurses and increasing the number of trained nurses, led to higher outputs, for a given level of input. Despite the stated importance of understanding production processes in order to understand

the behaviour of health care firms, there are very few studies that have estimated the degree of factor substitution in health care (Raikou et al. 2000).

4.12 Evidence on technical efficiency

An important reason for assessing production processes is to estimate the extent to which firms are failing to achieve technical efficiency, that is they are failing to maximise output from a given set of inputs, or failing to use the minimum inputs required to produce a given output. In an early work, Feldstein (1967) used the residuals of the production function as a measure of technical efficiency. Thus a hospital with a residual of zero from the production function, had an average level of technical efficiency, whereas a hospital with a residual greater than zero was more technically efficient. This pioneering study demonstrated wide variation in technical efficiency across hospitals in the UK NHS. Valdmanis (1990) used Data Envelopment Analysis (DEA) to estimate technical inefficiency in not-for-profit and for-profit hospitals in the US. This study indicated that for-profit hospitals achieved higher levels of efficiency. A similar study by Register and Bruning (1987) found no significance difference in efficiency levels between for-profit and not-for-profit hospitals.

More recently stochastic frontier analysis (SFA) has been used to measure efficiency. SFA attempts to counter the principle problem of DEA, which is its sensitivity to unobservable shocks[22]. Wagstaff (1989b) in a study on Spanish hospitals found that 90% of observed inefficiency was due to random shocks, leaving only 10% due to actual inefficiency. Zuckerman et al. (1994) used SFA to analyse a group of US hospitals and found efficiency levels of between 80 and 88%, which were similar to the estimates Valdmanis (1990) found using DEA techniques. However, SFA struggles to distinguish between technical and productive efficiency. As Jacobs (2000) points out this is an important limitation as different forms of inefficiency may call for different policy responses. In a DEA study Maniadakis and Thanassoulis (2000) used DEA to estimate the impact of the UK NHS reforms on technical efficiency, and concluded that the incentives put in place by the 'internal market' had

[22] Chapter 5 provides a review of these different techniques.

little impact on technical efficiency. The SFA by Folland and Hofler (2001) did provide separate estimates of technical and productive efficiency but concluded that the estimates of technical efficiency were heavily dependent on the methodology used. Each of these studies used highly aggregated datasets to estimate efficiency, and encountered measurement issues, these are considered in section four below, and in Chapter 5.

4.13 Productive efficiency

The point of productive efficiency can be found with reference to the firm's isoquants and to the isocost line (also known as the budget constraint). The isocost line represents all the combinations of factor inputs that can be purchased for a constant total cost. The cost-minimising point on the isoquant is where the isoquant is tangental to the isocost line. Health care firms in various settings who are producing a similar product may face similar isoquants but different relative factor prices. Different relative factor prices may arise because of differences in the characteristics of the local markets for factor inputs (see section 4.23). Theory suggests that firms who face different relative factor prices may choose different combinations of factor inputs. The extent to which firms respond to changes in relative factor prices is measured by the elasticity of substitution σ:

$$\sigma = \frac{\text{relative change in } (V_1/V_2)}{\text{relative change in } (P_1/P_2)}$$

where V_1/V_2 denotes the technically and productively efficient ratio of factor inputs and P_1/P_2 the ratio of input prices. σ can be between 0 and ∞. When $\sigma = 0$ the factors of production are perfect complements, and where $\sigma = \infty$ the factors of production are perfect substitutes.

Different forms of the production function make different assumptions about σ. For example, the Cobb-Douglas production function assumes that $\sigma = 1$, the constant elasticity of substitution function assumes that σ is constant for the same level of output, and the translog production function allows σ to vary across different levels of

output. The implications of these assumptions are considered further in the section on measurement issues (4.3).

For each level of output (q_0,q_1….q_n) there is a ratio of factor inputs at which costs are minimised. The cost function summarises the relationship between total costs (TC), outputs (q) and input prices (p_1,p_2): $TC=h(q,p_1,p_2)$. The total cost function can then be used to derive average and marginal costs. Just as the isoquant represents a frontier of efficient production with respect to output, so the cost functions represent points of efficient production with respect to cost (Call and Holahan 1983).

4.14 Empirical evidence on factor substitution and productive efficiency

There has been little empirical work on the extent to which substitutions of factor inputs occur in response to differing factor prices in the health care sector (Raikou et al. 2000). Feldstein (1967) estimated the elasticities of substitution across different groups of factor inputs in the UK NHS, and concluded that providers spent too much on doctors and drugs. However, the Cobb-Douglas functional form may have been too restrictive. Lavers and Whynes (1978) used the more flexible translog production function but also concluded that in general the UK NHS did not adopt the combination of factor inputs required to achieve productive efficiency. The study found that the ratio of doctors to beds was too high and the ratio of nurses to beds was too low. Maniadakis and Thanassoulis (2000) used DEA to estimate different levels of efficiency before and after the 1991 NHS reforms; they found that any improvements observed in efficiency were attributable to gains in productive efficiency. In a study of the relative efficiency of public and private health care facilities in Nigeria, Wouters (1993) found that public health care facilities were productively inefficient. However, this study was relatively small for assessing inefficiency. Those firms that 'appeared' technically inefficient were deleted from the dataset, and then a cost function with the Cobb-Douglas functional form was used to estimate productive efficiency.

All of the studies reviewed use aggregated datasets. This makes it difficult to assess case-mix and quality of care differences across the studies, and identify systematic

differences in productive efficiency; this issue is considered again in section 4.3 and in Chapter 5.

Phelps and Mooney (1993) suggest that while providers should be influenced by the costs of using different combinations of factor inputs, it may be that providers do not have the information available to allow relative costs to influence their decisions. In a before-and-after study, looking at the impact of introducing information on costs, doctors significantly reduced the number of laboratory tests they requested in response to having information on the costs of tests. When this information was withdrawn the number of tests requested increased.

4.15 Consideration of the short-run versus the long-run

In order to understand why costs differ across settings it is important to distinguish between firms taking a short-run compared to a long-run perspective.

The cost function in the short-run

In the short-run it is assumed that the firm is only able to vary one input, usually labour, while the other (capital) remains fixed. For example, in Figure 4.1 the firm can reduce short-run average costs (SRAC) by increasing labour input until the cost minimising point on $SRAC_1$ is reached at point A. If the firm is operating on the $SRAC_2$ at point B, the firm may be underutilising the fixed input. Increasing the labour input to point C, by increasing labour input would mean the fixed input was more fully utilised and the SRAC falls. Different levels of under-utilisation may exist across health care settings depending on incentives to cost-minimise, and the extent to which inflexibilities exist in the various factor markets. In the short-run it is assumed that the level of capital is fixed, it is not possible to reduce the level of capital to cost-minimise.

The impact of under-utilised health services has been widely quoted in the literature as having a significant impact on unit costs (e.g. WHO 1979). However, the evidence for this rests largely on descriptive data. For example, Berman (1986) concluded that differences in utilisation rates and programme management 'probably' accounted for

most of the variation in unit costs across settings. The limited econometric work on the impact of under-utilisation on unit costs comes from studies whose estimates of technical inefficiency include a component that can be categorised as under-utilisation[23]. For example, Zuckerman et al (1994) show that their measures of efficiency are inversely related to occupancy rates. The elasticity of inefficiency with respect to the occupancy rate is about 20%. This implies that a 10% increase in the occupancy rate would reduce inefficiency by 2%.

The cost function in the long-run

In the long-run it is assumed that the firm can adjust the level of capital as well as labour to try and cost-minimise. For example, if the firm in Figure 4.1 starts at point A then to cost-minimise in the long run it has to increase the level of capital and move to the cost minimising point on the next short run average cost curve ($SRAC_2$), at point C.

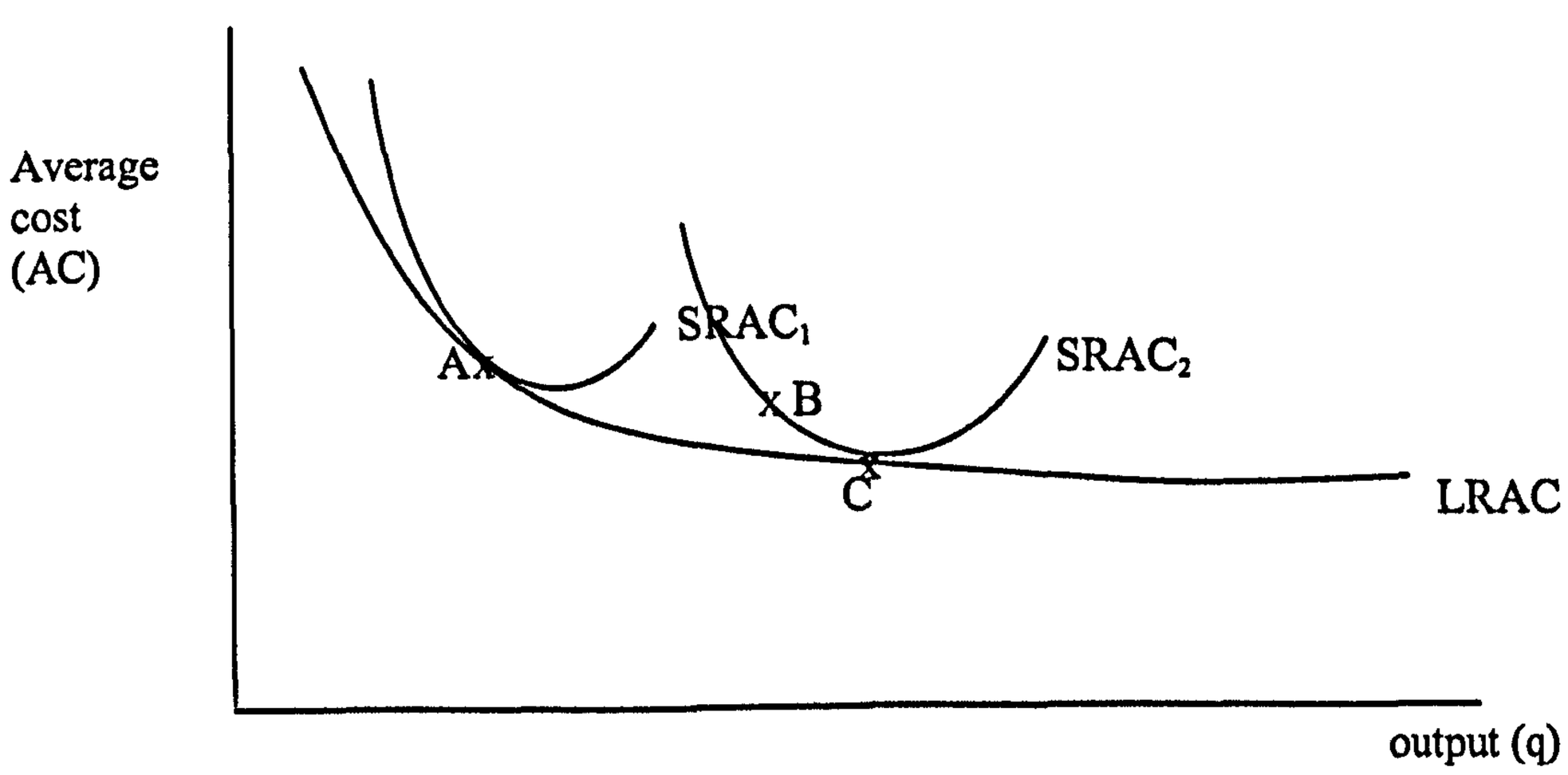


Figure 4.1: Short-run and long-run average total cost curves

The long-run average cost curve (LRAC) joins all the short-run cost minimising points, thus showing the least cost combinations for different levels of output. If economies of scale exist, then the LRAC is downward sloping. Health care firms in

[23] Strictly speaking underutilisation should not be counted as part of technical inefficiency, technical inefficiency refers to waste and failing to minimise the inputs required to produce a given output, whereas underutilisation is a short run concept and refers to the reduction in the cost per unit of output that would occur if output was increased, so that the labour available would be more fully utilised.

different locations may vary in the extent to which they can achieve economies of scale.

Evidence on whether health care firms take a short-run or long-run perspective. The cost function literature defines the short-run perspective as being where one group of inputs (labour) is variable, while another input set (capital) is fixed. Clearly, this represents an unrealistic and highly simplified dichotomy when applied to health care. In the health care setting, many labour inputs may be regarded as fixed, whilst certain capital inputs may be regarded as variable (Dawson 1994). There is a continuum from the short-run to the long-run, along which each input in turn may change from being regarded as fixed to variable. As Gravelle and Rees (1982) explain, what this simplified definition of fixed and variable really relates to, is the cost of adjusting the level of the input concerned. Although in the short-run, the costs of increasing say the number of beds on a hospital ward may exceed the expected increase in revenue, these adjustment costs may be justified over the long-run. Health care firms in different contexts may vary according to whether they display short-run or long-run behaviour. If it cannot be assumed that all firms are operating from a long-run perspective and hence able in principle to adjust all factor inputs according to differences in factor prices, this makes it difficult to interpret cost differences between settings as being attributable to variations in productive efficiency.

In the context of a competitive market situation, Cowing and Holtmann (1983) found that in the US the production of for-profit hospitals met the short-run conditions for cost-minimisation. They also found though, that the same hospitals did not meet the long-run conditions for cost-minimisation. Bilodeau et al. (2000) conducted a similar study but using cost data on not-for-profit hospitals in Canada. This study used data from all the hospitals in Quebec over a 12-year period and concluded that these hospitals also met the conditions for short-run but not for long-run cost-minimisation. Bilodeau et al. (2000) discuss whether this was because government rather than private sector decision-makers made investment decisions. However, the authors point out that their results are similar to those of Cowing and Holtmann (1983) and conclude that hospitals tend not to cost-minimise in the long-run irrespective of their ownership status.

4.16 Economies of scale

A range of econometric studies have looked at the role of economies of scale in explaining differences in hospital costs (see review by Aletras 1999). The more reliable studies have found that the hospital of average size (one with 200-300 beds) faces constant returns or diseconomies of scale. Söderland et al. (1995) found that the total number of beds in the hospital was not significantly associated with hospital costs. However, when the analysis was conducted at the speciality level, the cost per episode was significantly associated with the number of beds. The use of highly aggregated data may therefore conceal the relationship between volume and output.

4.2 Contextual factors

The principles of the production and cost function literatures have been applied to health care delivery, to examine under what circumstances health care providers deliver services in a way consistent with cost-minimisation. This next section examines what contextual factors appear to be associated with differences in cost. This section focuses on the labour market for health care inputs to consider why differences in factor prices and factor inputs may arise. The review highlights that in the context of health care production, there be may restrictions on efficient production that differ across settings and these could be associated with cost variation.

4.21 Incentives

Much of the applied literature on cost and production functions has focused on analysing whether those health care providers with greater incentive to cost-minimise exhibit more efficient behaviour (Valdmanis 1990, Cowing and Holtmann 1983, Sloan 2000). The evidence suggests that even if providers have clear incentives to profit-maximise, it is unclear whether health care providers' behaviour is consistent with cost-minimisation in the long-run. Cowing and Holtmann (1983) compared the behaviour of for-profit and not-for-profit providers in the US and showed that in the short-run for-profit providers were aiming to cost-minimise. Not-for-profit hospitals

were using a higher quantity of inputs in the production process they were also producing a higher quality product than their for-profit counterparts[24]. Sloan (2000) highlighted the difficulty in adjusting for patient and provider factors when trying to estimate the effect of incentives on costs. For example, levels of teaching, research and case-mix may all vary according to the hospitals' ownership status, and may limit the comparability of costs. Following a comprehensive review of the literature on the effect of ownership on costs, Sloan (2000) concluded that:

"Overall the empirical evidence demonstrates no systematic differences in efficiency between profit and not-for-profit hospitals" (Sloan 2000, p1168).

The way in which providers are reimbursed may differ across health care settings, which may also influence the use of health care. The Copenhagen case study found that introducing fees paid to providers for delivering certain services increased their use (Krasnik et al. 1990). In the US, managed care has been shown to improve technical efficiency in hospitals (Miller and Luft 1997). One reason for this may be the lower rate at which technology is diffused in a managed care rather than a fee for service (FFS) setting. However, these studies have again struggled to recognise differences in case-mix and the quality of care across health care providers.

Many countries have moved towards prospective payment systems for hospital reimbursement in an attempt to improve productive efficiency. In the USA, prospective payment systems based on Diagnosis Related Groups (DRGs) have been introduced to try and reduce costs. DRG information is used to categorise inpatients according to diagnosis and resource use, and the reimbursement rate per case is set to the average for the DRG. Providers have an incentive to reduce costs below the DRG reimbursement rate as they can keep any cost 'savings' that arise. Variants on the DRG system have been introduced in many EU countries in an attempt to reduce hospital costs. However, it is uncertain whether introducing a prospective payment system does reduce costs (Ellis and McGuire 1986). Often the DRG reimbursement only covers hospital services, so hospital providers have a clear incentive to shift costs onto community providers, or the patient themselves. Another problem is that

[24] This raises measurement issues that will be discussed in the next section.

providers have an incentive to misclassify the diagnosis to try to receive additional reimbursement. Instead of offering financial incentives, governments may introduce targets to try and reduce costs. However, evidence from the UK suggests that such policies have had limited impact on efficiency (Jacobs and Dawson 2003).

International studies suggest that overall health care expenditures are likely to vary according to incentives within the health care system to reduce health care costs. For example, Hurst (1991) found that under a system where physicians were paid by fee for service (FFS) and providers were retrospectively reimbursed, expenditure was generally higher than one where there were fixed budgets for hospitals and clinicians reimbursed by capitation fees or salaries. At a more micro level, McClellan and Kessler (1999) demonstrated that differing incentives across 16 countries were important in explaining the relative uptake of hi-tech interventions for cardiovascular disease.

The introduction of user charges for health services may reduce the demand for health services and the extent to which patients in different health systems are required to pay for a given service may be important in explaining differences in resource use across settings. However, a review of studies examining the effect of user charges on the demand for health services found that the price elasticities of demand for health services were generally relatively inelastic (Folland and Stano 1990). In most European countries, patients have to pay a sizeable contribution towards social care services such as nursing homes or home help unless their assets are too low, or they fall below a needs-based threshold (Hantrais and Letablier 1996). Depending on the extent to which these social services may substitute for hospital services, higher copayments for social services may lead to higher use of hospital care.

4.22 Technological progress and access to technology

Firms may be constrained in their attempts to cost-minimise by the technological progress in their health care system. Technological progress occurs in an industry if new capital is developed or workers become more highly skilled. This may lead to

higher factor productivity, meaning that the same output can be produced with lower levels of factor inputs. While technology in general may be cost lowering, in health care, new technology often improves outcomes but at higher cost (Bradford et al. 2001). Health care firms in different settings may have different levels of access to new technology, which explains why a higher output is produced from apparently similar levels of factor input. For example, a new pharmaceutical intervention, or highly trained health care professional may be available in one setting, but not others. This may partly reflect differences in the societal willingness to pay for a unit of health gain, which may depend upon national income, or relative preferences for health care versus other good and services (O'Brien 1997).

Bradford et al. (2001) highlighted the importance of the 'learning curve' in determining efficiency in production using as examples coronary bypass surgery (CABG) and balloon angioplasty (PTCA). The study found the average level of inefficiency for the established procedure, CABG was lower (8.7%) than for the new procedure, PTCA (38%). The authors go onto suggest that where hospitals are at different places on the learning curve this can have an important impact on the efficient use of a technology.

The rate of uptake of new technologies may differ across international health care settings and this may hinder attempts to compare technical or productive efficiency. Dervaux et al. (2004) tried to compare technical efficiency between French and US hospitals, but concluded that there were substantial differences in the technologies used that meant problems arose when comparing estimates of technical efficiencies based on a general measure of technology.

In a study comparing the use of new technologies following myocardial infarction (MI) across 16 countries, McClellan and Kessler (1999) reported wide variations across countries in the use of 'high tech' costly interventions such as CABG and PTCA. The rate of adoption of these technologies varied according to the level of regulatory control on the introduction of new technologies in different countries. As McClellan and Kessler (1999) comment:

"countries such as the United States with relatively weak supply-side controls have relatively high growth rates.." (McClellan and Kessler 1999, p37).

4.23 Labour market characteristics

The cost function literature tends to assume that although the level of technology and factor prices are determined exogenously, the firm can chose the factor-mix consistent with cost-minimisation. An issue this raises is whether firms in different health care settings do face different relative factor prices. As labour costs form an important component of total health care costs in most industrialised countries (OECD 2001), particular attention is now given to the labour market. In a perfectly competitive labour market, the wage rate would be determined by the intersection of the supply and demand for labour. Theory usually depicts the supply curve for labour as being upward sloping, with respect to price. The demand curve is usually downward sloping and depends on the marginal productivity of labour, and the price of the final product. As Figure 4.2 illustrates, if firms pay a wage (W_l) lower than the equilibrium wage (W_e) then workers will not supply their labour, and there will be vacancies equal to L_e-L_l. These vacancies would act as a signal for these firms to increase their wages, and supply would increase accordingly until an equilibrium wage and quantity is reached at E.

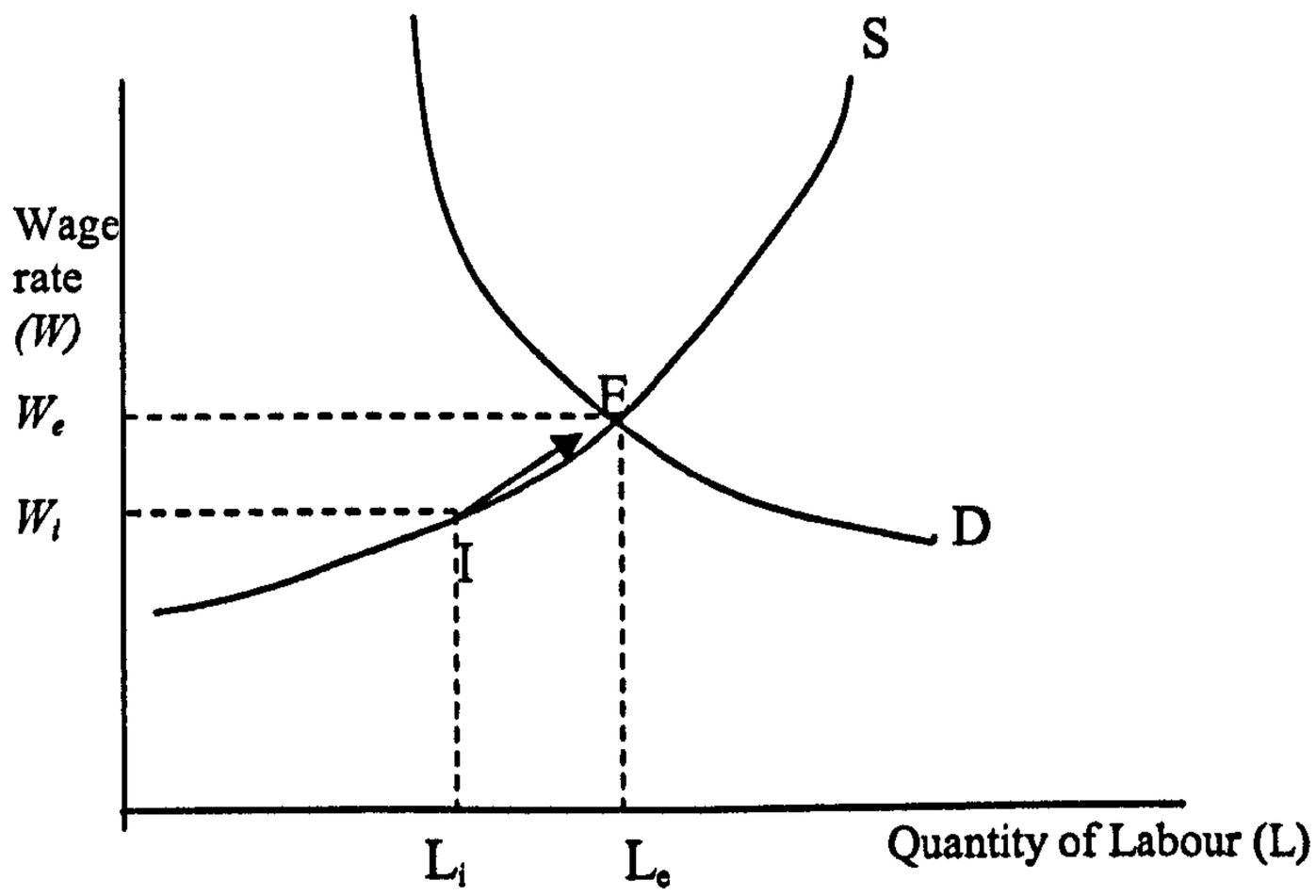


Figure 4.2: Achieving equilibrium in a competitive labour market

However, this model of the perfectly competitive labour market may not exist in health care. Labour markets in this sector are characterised by shortages of particular factor inputs in certain regions, which suggest that the market has rigidities and does not reach an equilibrium. For example, in the UK labour market the overall demand for nurses exceeds the supply, but in rural areas in the UK, the reverse is true (Elliott 2003). There is wide variability in the supply of doctors across Western Europe. For example, the number of qualified doctors per 1000 of the population is 1.6 in the UK, compared to 5.5 in Italy (OECD 2001). This could reflect differences in the regulation of the labour market for physicians between these two countries.

A perfectly competitive labour market would predict that in the long-run the level of wages would reach the same equilibrium wage across the health care settings concerned. However, international labour markets for health professionals appear to be in a permanent state of disequilibria, as wage rates differ across countries. Cromwell and Mitchell (1986) found that differences in wages explained up to 33% of the observed variation in costs across US hospitals. Fuchs and Hahn (1990) suggested the US had higher health care costs than Canada due to higher physician

fees. In Eastern Europe, the wages of health care professionals are much lower than in the West (ILO 2002). There has been a fall in real wages for public sector workers in ex-Soviet Union countries that have moved towards market economies (ILO 2002).

An important reason why the conditions for a competitive labour market fail in health care is that there are insufficient buyers and sellers. In many countries, wages in the public sector are set by negotiation between one buyer (the government) and one seller for example, a national union acting on behalf of its members. Thus, the market may be characterised by having a monopolistic and monopsonistic structure. Wage rates may differ between countries depending on differences in the relative bargaining powers of these two parties. In some countries, there may be several buyers, for example, different health insurance firms and individual health care professionals may negotiate wages. This market characteristic may lead to differences in wages within a country or between countries with different characteristics.

In some health care settings, the wage paid may only be a small consideration to the worker deciding whether to enter or remain in the particular labour market. Rosen (1986) highlights the importance of compensating differentials in understanding disequilibia in the labour market. Differences in the costs of living, the working conditions and the local environment may be important considerations when setting wages. In Eastern European countries, health care professionals may receive 'under the counter' payments from the patient that partly compensate them for their low official wage (Thompson and Witter 2000). In addition, important state benefits may be provided for the employee including subsidies for childcare, housing, food and health care. Many employees also work in the private sector to increase their state income. In Lithuania, a survey of health care workers highlighted that flexible working hours were an important determinant of job satisfaction, and this policy may encourage workers to reduce the hours they work in the state sector (ILO 2002). In the UK, it has been demonstrated that the supply of health care professionals is quite inelastic with respect to price, and that many other factors including job satisfaction are important in determining nurses' decisions about whether to remain in the labour market (Elliott et al. 2003). In addition, the variability in vacancies for GPs across the UK was found to be independent of their wage, which is broadly similar across the

country (Elliott 2003). This indicates that factors other than the wage rate are important in determining where GPs choose to supply their labour.

Barriers to entering and exiting the health care labour markets in different countries may also lead to differential factor prices. In some countries, for example the UK, doctors may be contracted to work for a five-year period, and in other countries the state may require training costs to be repaid by those leaving the labour market. While there is some evidence of labour market mobility across countries, for example in the UK more than 20% of doctors were trained elsewhere (Elliott 2003). In general, requirements to meet professional standards may differ across countries and may limit the transfer of labour.

The perfectly competitive labour market also assumes that there is perfect information. However, in health care, the marginal productivities of different health care workers are often unknown. Similarly, the value of the product is often unknown as prices for many health care services are unavailable. This makes it difficult to set a wage that equals the marginal product. For a hospital looking to recruit health care workers from a different country, it may be difficult to assess what their wage should be, based on their likely marginal productivity. The extent to which the training and skills learnt in one context may be transferable to another may be unknown, and limit the transfer of workers between different settings.

Labour markets and factor substitution

Cost function theory suggests that when a factor price is relatively high in a particular setting, it may improve productive efficiency to use another input instead. Wagstaff (1989a) reviews econometric work conducted on the UK labour market and highlights the work by Gray et al. (1986) and Lindsay (1980) who both suggest that as the wage rates in the health care sector rises relative to the price of capital, there is a substitution of capital for labour. Lindsay (1980) also looked at the net emigration of physicians from the UK, and found that it was related to current and lagged values of rate of return to employment in medicine. However, in many instances, labour market restrictions may inhibit change from taking place. For example, professional bodies

may have strict guidelines that limit the substitutability of different labour inputs (Elliott 2003). Different labour markets have different regulations regarding the number of hours different health care professionals are allowed to work. For example, under the new contracts in the UK NHS junior doctors can work up to 48 hours per week, whereas in Germany regulations mean that doctors are not allowed to work more than 40 hours per week. Another restriction is that health care professionals usually have to receive an accreditation, which for some professions may be difficult to acquire in particular countries. Governments in Eastern Europe have only recently given accreditation to professions such as occupational therapists, and training colleges are rare. The supply of these inputs is therefore limited nationally and may constrain their use in local health care settings.

Although production and cost function studies often presume that health care firms are acting from a long-run perspective, and therefore all factor inputs are variable, inflexibilities in the labour market may mean that firms in some settings fail to take a long-run perspective. Even for nursing inputs which due to the shorter training periods and relative flexibility of contracts, may be thought more adjustable, Carr Hill et al. (2003) reported a complete lack of flexibility in the deployment of nurses in 30 UK hospitals.

In summary, national labour market characteristics may determine the level of labour inputs deployed in local health care settings. Local decision-makers may have limited influence over the price and use of labour inputs. The price of labour inputs may differ across countries, as restrictions and inflexibilities in international labour markets mean that wages never reach equilibrium.

4.26 Health care budget

Several studies have highlighted that health care costs are correlated with national income (reviewed by Hutton 2001). For example, Barnum and Kutzin (1993) found a positive correlation between average health care costs and GNP per capita. These

findings reflect that labour costs are an important source of health care costs and wage rates are highly correlated with per capita income.

Those countries with higher levels of national income tend to spend a higher proportion of their income on health care, that is health care has been shown to have the characteristics of a 'normal' good (Gerdtham et al. 1992). The national level of spending on health care has implications for the factors of production that may be available to an individual health care firm. For example, the numbers of health care professionals that are trained and available to the labour market may depend on the national budget, and may be stipulated by a national decision-maker. Similarly, national decision-makers may decide whether or not to reimburse a new technology and this may depend on the budget available for health care.

Phelps and Mooney (1993) discuss 'Roemer's Law' that suggests demand for services follow supply, whereas economic theory would suggest that supply follows demand. Differences in the resources available for health care may be an important determinant of the use of specific health services across settings in different countries. For example, in Western Europe the number of acute bed-days available per 1000 people ranges from a low of 2.2 in Turkey to a high of 7.0 in Germany (OECD 2001). This suggests that higher resource use of health services in Germany may partly reflect a greater availability of hospital resources. However, simply observing utilisation does not establish whether utilisation levels reflect supply or demand factors.

4.27 Role of clinical decision-makers

While in part, the use of factor inputs may be determined by the overall context within which the health care firm operates, the role of local decision-makers, in particular clinicians may be important. Reviews of the evidence on clinical practice variations have found that there are large differences in the way patients are managed across settings after adjusting for differences in patient factors (Folland and Stano 1990). There is considerable debate in the literature about which factors drive these 'unexplained variations'. Wennberg (1984) suggests that they reflect physicians'

uncertainty in diagnosis and treatment of specific conditions. However, Evans (1990) rejects the idea that the variation is explained by physician uncertainty and suggests instead that it reflects professional disagreement about what is best practice. Phelps and Mooney (1993) suggest that this variability reflects a lack of information available to the individual clinician about the effectiveness and cost-effectiveness of different procedures.

4.28 Clinical guidelines

In an attempt to reduce variability in clinical decision-making and to move towards best practice, various countries have introduced clinical guidelines. However, it is unclear whether these guidelines are based on evidence of clinical effectiveness let alone productive efficiency (Eccles and Mason 2001). The guidelines on best practice may differ, particularly across international health care settings. Even if the guidelines recommending 'best practice' are consistent, the adherence to these guidelines may differ across health care settings leading to differences in resource use. There is little empirical evidence available on the extent to which guidelines are adhered to in different settings and the implications for observed resource use variation (Eccles and Mason 2001).

4.29 Differences in the characteristics of individual health care providers

Johnston et al. (1998) assessed why unit costs varied between trial centres and centres routinely screening for breast cancer in the UK. They found that the centres included in the trial differed in several ways to non-trial centres. Some of these differences were associated with unit cost variations, these included the location of the centre (urban or rural), and the throughput of screening cases. Cowing and Holtmann (1983) and Huttin and de Pouvourville (2001) found that hospitals involved with teaching and research had significantly higher unit costs than non-teaching hospitals.

The level of access to health care workers with different levels of expertise may vary across settings, and this may lead to differences in resource use and costs. The study by Roos and Roos (1982) found that teaching hospital status and level of medical

experience reduced resource use. In the context of stroke care, the specialisation of the doctor is important, as there is evidence that neurologists provide more tests and investigations than general physicians do, but their patients have better outcomes (Mitchell et al. 1986). Other studies have shown that teaching hospital status leads to higher costs, but it is unclear whether outcomes are better than in non-teaching hospitals (Johnston et al. 1998).

4.3 Patient factors

4.31 Case-mix

An important reason why the resources used to manage a particular patient group may vary across health care settings is that the case-mix of the populations concerned may differ. Case-mix may be defined as the characteristics of a single patient or group of patients seeking treatment (Söderland et al. 1996). Concepts such as diagnosis and severity, physical dependence and age, may all be regarded as part of case-mix (Söderland et al. 1995). Comparison of resource use or unit cost between health care settings needs to carefully adjust for case-mix, yet this may prove problematic when using the production function framework to pose hypotheses for cost variation, as such datasets have to analyse output and costs measured at a highly aggregated level. This issue is considered further in section 4.4 and in Chapter 5.

4.32 Income and socioeconomic status

Patients' access to and use of health services may partly depend on their income and socioeconomic group (SEG). In the UK, there is conflicting evidence on the effect of income and SEG on resource use. Davey Smith et al. (1990) found that those patients with the lowest incomes and in the poorest socioeconomic groups had the lowest use of health services and the worst health status. By contrast, O'Donnell and Propper (1991) found that those with lowest income made more use of health services. The difficulties in interpreting correlations between income and health are discussed by Deaton (2002) who suggests that even if access to health services is similar across

SEG the utilisation of those health services may differ leading to more effective care in those from higher SEGs. International comparisons of resource use have failed to develop consistent measures of SEG that can be applied across countries, and have ignored the role of SEG in explaining resource use differences (Beech et al. 1996).

4.4 Measurement issues

The literature on production and cost functions can be used to guide an empirical investigation into which factors are associated with cost variation across health care settings. However, considering cost variation across settings raises certain measurement issues. These measurement issues include: the assumptions made when estimating production and cost functions, and how best to measure case-mix, factor inputs, outputs, factor prices and random variation. These issues are considered in the following section. Chapter 5 consider the question of which technique is appropriate for analysing cost variation.

4.41 Assumptions that are commonly made when estimating production and cost functions

Certain assumptions are routinely made when estimating production and cost functions. Neoclassical production functions often make the assumption of homogeneity, which means that the elasticity of scale with respect to output is constant. For example, a production function that is homogeneous of degree one, exhibits constant returns to scale at all levels of output (Doll and Orazem 1973). The main advantage of estimating homogenous production functions is that the slopes of the isoquants depend only on the input proportions and are independent of the scale of production. The implication for estimating a cost function is that the input proportions required for cost-minimisation, only depend on factor prices and not the level of output. This makes cost functions based on homogenous production functions much easier to estimate. Empirical studies commonly make another restriction, that the production function is homothetic (Gravelle and Rees 1982). A production function is homothetic if it can be written as an increasing transformation of a linear homogenous function (Gravelle and Rees 1982). This is a more attractive property than

homogeneity as it allows for the advantages described in estimating the cost function, without making the restriction of constant scale elasticity.

Within the group of production functions that are homothetic, a key determinant of the properties of the production function is the assumption made about the elasticity of substitution of factor inputs, α. The most flexible form of production function allows α to vary with respect to output. More recently, studies of hospital production have favoured using the fully flexible transcendental logarithmic form (Cowing and Holtmann 1983, Bilodeau et al. 2000, Rosko 2000). Such flexibility comes at a cost though, as it requires a relatively large number of parameters to be estimated. As Newhouse (1994) points out, an analyst may decide to estimate a function with 100 input and output groups; this number of groups does not appear excessive to cover the whole of hospital production. However, using n=100 in the translog production function would require 5000 parameters to be estimated. In most hospital cost or production datasets there are insufficient degrees of freedom to estimate this number of parameters. Across the whole of the US for example, there are only around 5,000 general hospitals. The problems posed by using the translog is clearly apparent in the studies by Cowing and Holtmann (1983) and Conrad and Strauss (1983), that both used translog production functions and relied on very crude measures of case-mix and output to adjust for differences across hospitals. As Breyer (1987) states:

> "an accurate reflection of patient heterogeneity seems to demand more variables than can be handled in a flexible functional form model.. in existing studies either flexibility of the functional form or precision of the case-mix measure was sacrificed entirely" (Breyer 1987, p151).

Some production function studies in health economics have assumed instead that the elasticity of substitution is constant with respect to output. Amongst the constant elasticity of substitution production functions, a further restriction is made by the Cobb-Douglas production function, which assumes a unitary elasticity of substitution between factor inputs. Some analysts have concluded that the Cobb-Douglas provides an adequate approximation to a more flexible functional form (Reinhardt 1970), whereas other studies in health care have disputed the assumption that the elasticity of substitution is unitary (Jensen and Morrisey 1986).

However, it is clear is that the more restrictive the assumptions made about the production function, the more degrees of freedom are available in the analysis, thus allowing potentially more careful adjustment for case-mix and output differences between settings. This highlights that given the existing size of the dataset, analysts have to decide upon the trade-off between making highly restrictive assumptions about the technology or about the homogeneity of the inputs and output measures used[25].

4.42 The difficulties in measuring case-mix

The assumptions made about production, are therefore important in determining the case-mix measures used. However even, if the most restrictive form of the production function is used, large numbers of health care firms and patients are required to estimate the relationships concerned. For example, in one of the smallest production and cost function studies reviewed, Wouters (1993) used the highly restrictive Cobb-Douglas production and cost functions to estimate efficiency of different health care providers in Nigeria. However, even this study included a total of 86 hospitals and relied on administrative data. Administrative datasets only contain measures of case-mix collected as part of routine practice.

Most case-mix classification systems that are routinely used for administrative purposes are based on DRGs. As DRGs are routinely recorded they are available for many health care providers in the USA and have proved attractive measures of case-mix adjustment for production and cost function studies requiring data on large numbers of health care providers (Valdmanis 1990, Vita 1990, Vitaliano 1987). However, studies using routinely collected case-mix measures covering the full range of hospital production may have inadequately adjusted for case-mix differences across settings. For example, Newhouse (1994) points out that using DRGs to adjust for case-mix differences across hospitals assumes that within a DRG any cost variations across hospitals are random. If DRGs do not fully adjust for differences in case-mix

---

[25] This issue is considered further in Chapter 5.

across particular hospitals then cost variations may reflect unmeasured case-mix differences.

Iezzoni et al. (1996) divided case-mix measures into two broad categories: general measures of case-mix that are routinely available and disease-specific measures collected from medical records. Within each of these broad categories, there are many different case-mix measures of varying degrees of sophistication (see Iezzoni 1997a for a comprehensive review). For example, in a UK context Söderland et al. (1996) compared measures of case-mix adjustment routinely available, including DRGs[26], Health Care Resource Groups (HRGs) and a specialty classification, to predict variation in LOS and total cost. The results showed that while DRGs and HRGs generally explained a reasonable proportion of the variations in LOS, in some specialties for example geriatrics these measures left a high proportion of the variance unexplained (Söderland et al. 1996).

Disease specific case-mix measures have now been developed for many diagnoses (see Iezzoni 1997b). For some diagnoses, these measures do not explain any more heterogeneity than routinely collected measures. For example, Iezzoni et al. (1996) used 11 severity measures including both routinely collected and disease specific measures recorded from medical records. The study assessed how much variation in LOS following pneumonia each measure explained both amongst patients and across hospitals. The results showed that none of the measures explained substantial variation in LOS within or across hospitals. For all measures and both comparisons, the adjusted R^2 was less than 0.20. The authors suggested that the case-mix measures examined may have been insufficiently sensitive to differences in patient characteristics, they also highlight that differences in provider characteristics were not considered and may well have been important.

A number of studies found that disease specific measures of case-mix can explain more of the variation in resource use than routinely collected measures. Vos et al. (1994) found that for patients with acute MI more detailed clinical measures of case-

[26] Within the general term DRGs there are many different versions.

mix detected differences in resource use across hospitals for patients who fell into the same DRG. Escarce (1993) found that 43% of the observed variation in cataract surgery rates was explained by including economic and sociodemographic variables.

In the context of stroke care, a range of case-mix measures have been developed and shown to partly explain variations in resource use, cost and outcome across centres. These measures include the sub type of stroke, whether or not the patient is continent at hospital admission, and whether or not the patient is paralysed at the time of maximum impairment from the stroke (Wade 1994, Davenport et al. 1996, Beech et al. 1996, Wolfe et al. 1999). Beech et al. (1996) used these measures of case-mix adjustment to compare LOS across a range of European centres. The study found that there were still wide variations in resource use after case-mix adjustment, which may be partly explained by unmeasured patient characteristics. However, this study did not consider the use of alternatives to hospital care or contextual factors such as the level of health care infrastructure in each of the countries concerned.

The approach used to measure costs in particular the level of aggregation, has implications for understanding the role of case-mix in explaining cost variation. For example, if a highly aggregated approach is used, and number of hospitalisations are the resource use measure and cost per hospitalisation the unit cost, then patient characteristics may be important in determining unit costs. Söderland et al. (1995) used an aggregated approach to unit costing and found that case-mix explained 77% of the variations in costs per case across NHS providers in the UK. However, if instead a disaggregated approach is taken then the patient-mix may be more important in determining resource use than unit cost.

In summary, the choice of case-mix measure is a key issue for any study attempting to estimate reasons for cost variation across settings. Highly aggregated measures of case-mix have been found to be inadequate for adjusting for differences in case-mix across different hospital providers. An empirical study of cost variation across settings should consider using disaggregated measures of case-mix. However, it should be recognised that for certain clinical areas, even using detailed case-mix measures may

still leave high levels of unexplained variation in resource use and costs across health care settings.

4.43 Aggregation of factor inputs

The diagrammatic representation of a production function often divides inputs broadly into labour and capital. When estimating production functions the problem of having sufficient degrees of freedom to estimate parameters, also restricts the numbers of parameters used to represent inputs and outputs. Hence, when estimating the inputs to the health care production function highly aggregated measures of input may be used such as the total number of employees in each hospital. When comparing the use of inputs across health care settings, this broad definition is unlikely to be sufficiently detailed as for example workers are likely to have different levels of skill and are therefore likely to have different marginal productivities. Firms may also differ in the quality of inputs they use in producing health services. For example, in the stroke literature, care in a specialised stroke unit improves outcomes for a given level of resource use, compared to care in general medical wards (Stroke Unit Trialists' Collaboration 1999). Comparison of costs across settings therefore needs to consider differences in the quality of factor inputs.

Measures of unit cost may also have to recognise differences in the quality of inputs[27], Berry (1973) found that quality enhanced hospital services cost 16% more per inpatient day than basic hospital services in the USA. Other studies have suggested though that improved quality of care led to lower unit costs particularly if it leads to reduced morbidity (Folland et al. 1997). Studies that compare costs across health care settings using highly aggregated datasets struggle to adjust for quality of care differences which may instead be attributed to differences in efficiency (Valdmanis 1990, Aletras 1999).

4.44 Aggregation of outputs

The same problem of aggregation exists when comparing outputs across settings. Using an aggregated measure of output, may lead studies to ignore differences in

[27] This depends on the level of aggregation at which unit costs are measured. Highly aggregated unit costs e.g. costs per episode are likely to vary according to the quality of care provided. More disaggregated measures such as the cost of a blood test may be less susceptible to quality differences.

patient outcomes. The multi-product cost function literature has tried to compare both the quantity and quality of health care produced (Healey et al. 2000, Butler et al. 1995, Cowing and Holtmann 1983). However, studies, which have attempted to compare output across health care providers, have often struggled to measure the quality of care provided and have often relied on aggregated measures of throughput for example the annual number of inpatient days (Folland et al. 1997). Using a narrow measure of throughput is likely to be inconsistent with health care providers' objectives which may for example be to maximise health from the resources available (Hurst 1991). If factor inputs fail to increase throughput but do improve health status, then their marginal productivity would be underestimated. This problem is likely to be greatest in production or cost functions, which take a flexible functional form and therefore include coarser classifications of case-mix and output (Newhouse 1994). While measurement and adjustment for differences in the quality of care is generally difficult, the use of patient-level measures of case-mix and outcome, may prove useful for considering the quality of care when comparing costs across settings, even if their use is not feasible within a cost function framework derived from a well-behaved production function.

4.45 Accounting for random variation

Another measurement issue that arises when comparing resource use across settings is that the differences observed in resource use may reflect random rather than systematic differences (Diehr et al. 1990). Variability in resource use across settings would be observed because of chance even if each health care provider was using identical production processes. McPherson et al. (1982) looked at differences in the rates of surgical procedures across several countries and found that random variation only accounted for 1-4% of total variance in Canada, and on average about 15% of variation in England and Wales.

4.46 Consistent measurement of factor inputs, factor prices, unit costs or total costs

In Chapter 2 it was emphasised that resource use and unit costs need to be measured across health care settings using a consistent methodology. This also applies to the

measures of factor inputs, factor prices, output and total cost covered in this chapter. This issue is often neglected by the cost and production function literature which tends to use highly aggregated administrative datasets to measure parameters across a large number of health care providers (Newhouse 1994). Different methods may be used in each health care setting for example to allocate costs. Any measurement inconsistencies may hinder attempts to identify systematic reasons for cost variations.

4.47 Exchange rates

To compare costs across international health care settings, local costs have to be converted into a common currency. This raises the question of which conversion factor is most appropriate. One alternative is to use the Official Exchange Rate (OER). However, a problem with using OERs is that they do not reflect the relative price of non-traded goods such as health care. Instead, Purchasing Power Parity indices (PPPs) may be used. PPPs aim to eliminate price differences across countries for both traded and non-traded goods and services (Kanavos and Mossialos 1999). There are different PPP indices that are available for converting goods and services produced or consumed in different sectors of the economy. The choice of conversion factor for comparing costs across international health care settings can have an important impact on the cost variations observed (Hutton 2001). This issue is considered further in the context of the empirical investigation (Chapter 7).

4.5 Discussion

The aim of this chapter was to use conceptual and empirical literature to establish hypotheses for why costs may vary across health care settings. The literature reviewed considered four main areas: the cost and production function literature, context factors, patient factors and measurement issues that arise when comparing costs across health care settings. The main hypotheses to emerge from each of these areas are considered briefly below.

The production function literature highlights that there may be different ways of producing the same output. In health care, clinical practice variations may be observed but these might not reflect differences in technical efficiency. The practice variations literature generally fails to identify reasons why clinical practice may differ across settings. In particular, the role of alternatives to hospital care is routinely ignored. Differences observed in resource use may indicate that each health care setting is producing care efficiently but faces differences in the technology available or in relative factor prices. Alternatively, variations across settings may reflect differences in technical, productive or scale efficiency. There is only limited evidence to suggest that differences in efficiency in the hospital sector may arise according to incentives to cost-minimise.

Contextual factors that are likely to have an influence on production are often exogenous to the individual health care firm. The literature review suggested that factor prices may differ across settings, because of inflexibilities in international markets for good and services. This is likely to occur in the labour market for health care professionals which, even within a particular country, is known to suffer from rigidities and inflexibilities and remain in a state of disequilibrium (Elliott 2003). While individual countries may impose a bilateral monopoly and effectively establishing a uniform price across the country, international prices of labour inputs may differ even after adjustment using an appropriate conversion factor. Any comparison of factor prices across countries should recognise that such differences may remain because of compensating differentials; these include the possibility of

earning extra income from private practice, and satisfaction with the local working and living conditions[28].

Restrictions in the labour market may lead to differences in relative factor prices across health care settings. Insights from cost function theory would suggest that in response to different factor prices firms would choose differing factor inputs to maximise productive efficiency. However, the national availability of factor inputs and clinical guidelines may limit the power of individual health care providers to adjust factor use according to differences in factor price. National levels of income or the level of spending on health care may determine the national supply of particular inputs, or the availability and diffusion of new technologies. The extent to which clinical guidelines are followed, or health care professionals maintain control over the inputs used in production, will vary across settings and may be another important determinant of resource use and cost.

The role of patient factors alongside contextual factors in explaining cost variation also warrants careful consideration. The literature has focused on either group of factors and studies have in general failed to consider both groups of factors together. Iezzoni et al. (1996) highlight that including a range of patient and provider factors is important when comparing resource use across settings.

This section of the literature review has posed some clear hypotheses as to why costs may vary across different health care settings (see Table 4.2 below). However, there are some challenges in applying the production and cost function literature to health care, and in testing these hypotheses. In particular, the studies reviewed struggled to separate *systematic* differences in inputs and outputs across health care settings from differences in the way these parameters were *measured*. A particular problem is that there is important heterogeneity in the case-mix of patients across settings even within particular DRGs. This patient-level heterogeneity may drive variability in resource use, factor inputs and outputs across settings. In production and cost function studies the less restrictive the assumptions made about the technology the higher the level of aggregation used in the measurement of case-mix and output. The greater the level of

[28] These issues are considered further in the discussion section of Chapter 7.

aggregation the more likely it is that cost differences across settings reflect unmeasured differences in patient case-mix, rather than differences in efficiency. It would seem most appropriate for studies identifying reasons for cost variation to measure case-mix at a patient level using disease-specific measures.

4.6 Conclusions

This chapter has considered concepts from the production and cost function literatures to provide a framework for identifying reasons for cost variation across health care settings. The review has identified contextual and patient factors that may be important in explaining resource use and cost variation. To test these hypotheses certain measurement issues have to be addressed, and the use of disaggregated data is appealing for this purpose. The challenges these measurement issues pose are considered further in the next chapter that reviews the techniques used to measure cost variation.

Table 4.2: Summary of hypotheses from literature for understanding why resource use, unit costs and total costs vary across settings

Area	Hypotheses with examples
1. Production/ cost function	
Care alternatives	Observed differences in factor inputs may reflect substitutions e.g. alternatives to inpatient hospitalisation that may lead to reductions in LOS.
Technical efficiency	Technical efficiency may vary because of differences in incentives to cost-minimise
Factor prices and productive efficiency	Differences in factor prices, may lead to differences in factor use.
Scale efficiency	Economies of scale may be fully realised by hospitals with approx. 300 beds
2. Contextual factors	
Access and diffusion of new technology	Differences in national income, incentives or constraints may lead to different rates of technology diffusion and access to new technology.
Incentives	Variation in method of hospital reimbursement, use of patient copayments may effect resource use.
Labour market restrictions	Factor price, and unit cost differences may reflect local wage negotiations, and immobility in the labour market. Levels of factor input may be determined by national restrictions.
National spending on health care/ level of health care infrastructure	Higher national spending on health care may lead to improved health care infrastructure, increased use of new technologies, and better access to highly trained health care professionals
Guidelines	Levels of factor inputs, and resource use may be set by national guidelines, and differ by location.
Clinical uncertainty	Disagreement about best practice between local decision-makers may lead to variations in resource use.
Characteristics of providers	Teaching hospitals may have better access to new technologies and highly trained personnel.

3. Patient factors	
Case-mix	More complex case-mix associated with higher resource use and/or unit costs. In stroke care more complex case-mix represented by cases incontinent or paralysed at admission.
Socioeconomic status	Unclear whether lower SEG is associated with higher or lower resource use.
4. Measurement issues	
Case-mix	Case- severity measures in aggregated production function studies inadequate, so observed differences inefficiency could reflect case-mix differences.
Random variation	Random variation potentially important at patient or setting level.
Quality of inputs and outputs	Aggregated measures fail to adjust for differences in quality of inputs or outputs e.g. stroke units may use better quality inputs leading to better quality outputs, ignored by aggregated studies.
Variability in costing methods	Variation in methods used to allocate overheads, may explain differences in unit costs.
Choice of conversion factor	Level of international cost variation observed may depend on choice of conversion factor.

Chapter 5: Techniques to evaluate cost variation

5.0 Introduction

The previous chapter highlighted theoretical and empirical reasons why costs may vary across different settings. It was hypothesised that costs may vary because of differences in the use of factor inputs or resource use that may reflect differences in factor prices but also differences in technical or productive efficiency. Levels of technical or productive efficiency may vary according to the health care context, in particular according to the characteristics of the labour market and the method used to reimburse health care providers in the country concerned. Any study attempting to assess reasons for cost variation across health care settings encounters measurement issues in particular, it may be important to understand the role of case-mix, the quality of care and random variation in explaining observed cost differences.

The aim of this chapter is to consider the appropriateness of some of the techniques available for analysing cost variation. This review covers the use of techniques to measure efficiency- including stochastic frontier analysis (SFA) and data envelopment analysis (DEA). These techniques have mainly been used in observational studies that *describe* associations. For example, studies have used these techniques to examine the association between the method of hospital reimbursement and the level of productive efficiency (Hollingsworth 2003). For analysing cost data in economic evaluation techniques such as ordinary least squares (OLS) regression models and multilevel models (MLMs) may be attractive as they can be used to *evaluate* the cost-effectiveness of an intervention recognising cost variation across health care settings. Examples of the use of these techniques are presented in health economics, but also in other areas of research such as health services and education research. The pros and cons of using each of these techniques for assessing the reasons for cost variation are examined.

This chapter is split into five main sections: section one provide an overview of techniques for measuring efficiency, section two reviews the use of OLS regression models for explaining cost variation; section three discusses in some depth the use of MLMs for both observational and evaluative studies. In section four, the implications of using each of these techniques for assessing cost variation are discussed. Section five summarises the main findings from the entire literature review and considers the implications for the empirical investigation.

5.1 Methods for estimating technical and productive efficiency

There has been much interest from national governments in using measures to identify inefficient performance (Smith 1995). For example, in the UK the Department of Health used performance indicators to identify those provider units that warrant further investigation (Audit Commission 1999). The methods for measuring health care efficiency can be divided into four main categories: OLS regression models, deterministic cost frontiers, stochastic frontier analysis (SFA) and data envelopment analysis (DEA). Each of these methods have been used to estimate efficiency using observational datasets.

5.11 OLS regression models

OLS regression models were traditionally used to estimate levels of technical and productive efficiency across different hospitals. In his pioneering work Feldstein (1967), used a cost function of the form:

$$y_j = \beta_0 + \Sigma\beta_k x_{jk} + u_j ; \qquad (1)$$

where y_j is the cost for an individual hospital j, and x_{jk} is the proportion of hospital j's patients in the kth case-mix category. Thus, the cost for treating a patient in case-mix category k is $\beta_0 + \beta_k$. From this model, Feldstein (1967) defined an efficiency index as:

$$y_i^* = \frac{y_j}{\hat{y}_j} = \frac{y_j}{y_j + \hat{u}j} \qquad (2)$$

Where $\hat{y}_j$ and $\hat{u}_j$ are the OLS fitted values and residuals. In this model, the residuals are used as measures of efficiency. Those hospitals with positive residuals have higher than average costs and are defined as performing in a relatively inefficient manner, whereas those hospitals with negative residuals are categorised as having above-average levels of efficiency. Central tendency measures produce an analysis of performance relative to *average* performance, not compared to the best performance amongst providers. The residuals are assumed to be distributed symmetrically around the cost function and they do not define a *frontier* of efficient production. While the residuals are used to define efficiency, they are also likely to include measurement error, or random variation. Despite these acknowledged problems, studies during the 1970s and early 1980s tended to follow Feldstein's approach and used these 'non-frontier' OLS models to estimate efficiency (Hurst 1977, Culyer et al. 1978, Culyer and Drummond 1978, Steele and Gray 1982).

5.12 Deterministic cost frontier (DCF)

Deterministic cost frontiers attempt to tackle the problem that simply using residuals from an OLS model to define efficiency does not identify a frontier of efficient production (Aigner and Chu 1986). The DCF is defined as:

$$y_j = \beta_0 + \Sigma\beta_k x_{jk} + u_j ; \qquad \text{where } u_j \geq 0 \qquad (3)$$

In contrast to the previous model, the error term is constrained to be positive, and therefore the costs for each health care setting are above the deterministic cost frontier. The degree of inefficiency is therefore indicated by the size of u_j. One estimator for the model is corrected OLS (Schmidt and Lovell 1979, Greene 1980). The drawbacks of this method are that it treats the most efficient hospital as 100% efficient, and the whole of the error term is assumed to reflect inefficiency. This ignores random noise due to measurement errors and any unobservable heterogeneity (Wagstaff 1989b).

5.13 Stochastic Frontier Analysis (SFA)

SFA has been used to measure differences in efficiency across health care providers[29] and recognises that costs may vary randomly across providers. Compared to the OLS models, an additional error term v_j is introduced (model 4). Thus, the overall error surrounding the estimates is split into two components: v_j is a two-sided error term and captures measurement error and unobserved heterogeneity, and u_j measures inefficiency relative to the stochastic frontier:

$$y_j = \beta_0 + \Sigma\beta_k x_{jk} + v_j + u_j \qquad \text{where } u_j \geq 0 \qquad (4)$$

v_j may be high if for example there are unexpected expenditures on hospital repairs or a temporarily high and unobservable level of disease severity. SFA mainly uses structural cost functions that are based on the dual of the production function (See section 5.23, and also Diewert 1982).

An important aspect of SFA is that particular distributions have to be assumed for the error terms, for example, Aigner et al. (1977) assume that v_i is normal and u_i is half normal. Newhouse (1994) criticised SFA for making these assumptions about the distributions of the residuals, as they are untestable. Commentators have criticised SFAs based on cross-sectional data for relying on skewness in the distribution of the residuals to estimate inefficiency, as Wagstaff (1989b) states:

> "Inefficiency is reflected in, and only in, skewness in the residuals, absence of skewness is construed as evidence of absence of inefficiency." (Wagstaff 1989b, p664)

This may be an important problem, especially as aggregated datasets are routinely used for SFA in health care. In these datasets, it is difficult to adjust for differences in case-mix and the quality of output across health care providers (Newhouse 1994). There may be omitted variables such as more detailed measures of case-mix that could potentially lead to skewed residuals. In these circumstances, SFA could incorrectly attribute unmeasured case-mix differences across providers to differences in efficiency (Wagstaff 1989b, Dor 1994). Using the level of skew in the residuals to

[29] SFA usually measures the sum of both technical and productive inefficiency, though analysts sometimes assume zero productive inefficiency and attribute all the observed inefficiency to technical inefficiency (Jacobs 2000).

define efficiency, may also lead to difficulties concerning the choice of functional form for the cost data. For example, Street (2004) illustrated a situation where the original analysis found that the residuals were skewed, and part of this skew was attributed to inefficiency. However, once the data were transformed using a logarithmic transformation the residuals were approximately normally distributed. In this situation, it is not possible to identify inefficiency across the provider units.

While SFA allows for stochastic or random variation, the efficiency estimates for individual firms are routinely presented as point estimates. Street (2004) found that once sampling variation across hospital units was acknowledged, it was difficult to make firm conclusions about the relative inefficiency of different hospitals.

More recent applications of SFA have used panel, rather than cross-sectional data (Koop 1997). In this case assumptions are no longer required about the distribution of the error term. However, these models have routinely assumed that inefficiency remains constant over time which may not be plausible in an industry such as health care, where technological development is ongoing. Battese and Corelli (1995) relax the assumption of constant efficiency over time; their model makes the weaker assumption that the rate of change in efficiency is common across health care units.

5.14 Data envelopment analysis (DEA)

DEA is a non-parametric, deterministic technique, that aims to measure efficiency for each firm relative to the most efficient unit(s) that form the efficiency frontier (or envelope) and are assumed to be 100% efficient (Farrell 1957). Efficiency in DEA is defined as the weighted sum of outputs to the weighted sum of inputs (Hollingsworth and Parkin 1998). DEA can measure relative levels of technical, productive, allocative or scale efficiency. The main problem with DEA is that it does not allow for measurement error or random fluctuations. The main advantage of DEA is that unlike SFA it does not have to make parametric assumptions that may not be supported by the data. In general, this means that DEA does not require as many health care units as SFA. However, as DEA still requires sufficient observations to make plausible assumptions about the underlying production technology and to avoid the results being driven by outlying observations. DEA has therefore mainly been used on

administrative datasets with many observations (see reviews by Jacobs 2000 and Hollingsworth 2003)[30]. There are many recent examples of the use of DEA in the cost function literature that use aggregated datasets (Hollingsworth and Parkin 1998, Söderland and van der Merwe 1999, Morey and Dittman 1996), but only a few examples of DEA that aim to inform the economic evaluation literature (Hutton 2001, Johnston and Gerard 2001, Valdmanis et al. 2003).

5.15 Comparisons of DEA and SFA

Recent studies have compared DEA with SFA, to see how sensitive estimates of technical efficiency are to the method used. Studies using cross-sectional data have found that these estimates are highly sensitive to the choice of technique and have suggested that it is not possible to make robust conclusions about the levels of technical efficiency from the use of either method (Jacobs 2000, Wagstaff 1989b). Jacobs (2000) found that where there are high levels of random noise the different techniques are likely to diverge in their estimates of efficiency. The more recent comparisons of the techniques using panel data have found much less variation according to the methodology used; Linna (1998) used SFA and DEA methods on panel data and concluded that the correlation between the results from the different methods was high.

A consensus is emerging in the literature that suggests panel data estimates are useful for highlighting changes in efficiency over time. Analysis of these data may also help establish whether outliers identified in cross-sectional data represent one-off data anomalies or consistent differences in levels of efficiency.

5.16 General problems with using techniques of efficiency measurement to identify reasons for cost variation

In health care a range of studies have used either DEA or SFA to assess which provider characteristics are associated with differences in efficiency (Hollingsworth 2003, Linna 1998, Vitiliano and Toren 1994, Rosko 1999). A central problem for these studies is how to make plausible assumptions about the underlying production

[30] Most of the DEA studies reviewed have datasets with at least 20 observations at the provider level.

technology while correctly estimating the effect of patient and centre-level characteristics on observed levels of inefficiency.

Breyer (1987) suggests that studies measuring efficiency can be divided into two broad groups depending on their general approach to this problem: those studies that specify a behavioural cost function and those that estimate a structural cost function. The models that used behavioural cost functions included a range of variables to try and explain cost variation across settings (see for example Evans 1971, Lee and Wallace 1973 and Lave and Lave 1970). Wagstaff (1989a) has criticised these behavioural cost functions:

> "The observation by Evans (1971) that the absence of incentives to minimise costs means cost functions need to be interpreted as 'behavioural' rather than 'technological' has tended to result in an 'anything goes' attitude to model specification." (Wagstaff 1989a, p14).

Wagstaff (1989a) highlights that analysts have often ignored economic theory when deciding on which explanatory variables to include in a behavioural cost function. For example the stock of beds, case flow, occupancy rates and LOS were all included in cost functions, despite theory suggesting that they would be correlated with each other.

During the 1980s, there was a move away from the behavioural cost function approach and more studies used structural cost functions (see for example Conrad and Strauss 1983, and Cowing and Holtmann 1983). A structural cost function is defined as one that describes the minimum cost of providing a given output, as a function of an exogenous vector of factor prices (Breyer 1987). This implies that the explanatory variables included should only relate to output and price. Using structural cost functions also limits the choice of functional form in SFA, as the cost function should be the dual of the underlying production function (Diewert 1982).

Often though the form of the production function is unknown, and this has led analysts using SFA to specify the cost function with a flexible functional form (Breyer 1987). As Newhouse (1994) points out this flexibility comes at a price though, as using this functional form increases the number of parameters that have to

be estimated[31]. This makes anything other than crude adjustments for differences in case-mix and the quality of outputs across providers difficult (see Chapter 4), even when the datasets concerned include data for hundreds of provider units.

The analyst is therefore faced with a choice between making assumptions that may adequately represent the characteristics of the underlying technology or specifying cost functions that are able to detect differences in patient characteristics, inputs and outputs across health care settings. Studies estimating efficiency using DEA or SFA have tended to take the former approach and relied on using large datasets, with highly aggregated measures of case-mix, inputs, outputs and costs that encompass the whole of hospital production (Caves et al. 1980). Using these measures clearly leads to measurement issues, for example costs may be measured using different methodologies across health care settings and case-mix measures may be too crude to detect differences in case-severity across hospitals. Thus, any attempt to establish reasons for cost variability that uses aggregated datasets may struggle to correctly identify systematic reasons for cost variation.

In a comprehensive review of the recent literature using both parametric and non-parametric studies Hollingsworth (2003) identified 188 published papers on efficiency measurement in health care. The review suggested that 25% of these studies used DEA in combination with regression analysis in a 'two stage' approach to identify reasons for differences in efficiency. Under this approach the analysis initially uses DEA to calculate mean efficiency scores for each health care provider. These scores are then used as a dependent variable in a regression equation to identify those provider characteristics associated with differences in efficiency. This approach ignores any variability in efficiency within each health care setting. In addition, this study design only allows for aggregated measures of case-mix to be used in the analysis.

---

[31] Although DEA studies avoid the problems associated with specifying a particular functional form the technique still requires large numbers of observations to avoid outlying observations having a disproportionate impact on the results. Therefore, DEA has seldom been used on datasets with fewer than 20 observations at the provider level (see Hollingsworth 2003).

As SFA and DEA both tend to rely on aggregated datasets to measure efficiency these techniques tend to ignore the hierarchical structure of cost data drawn from different settings. In this context, patients may be clustered within health care settings, and ignoring this clustering may mean that these analyses make incorrect inferences (see section 5.21 for a further explanation).

Most of the studies reviewed were conducted within a single country and did not have to consider the methodological issues that arise when comparing costs across health care settings in different countries. Adam et al. (2003) demonstrated the use of an OLS cost function to compare costs across countries. However, the study suffered from methodological shortcomings: it paid little attention to variation in factor prices, it used highly aggregated data, and did not consider clustering within countries, and the implications for the significance of the variables included in the cost function.

The techniques for measuring efficiency are not appropriate for identifying reasons for cost variation, using datasets of the size routinely used in economic evaluations. However, some of the issues discussed during the development of these measures, notably the choice of explanatory variables, and the interpretation of residual variation may also apply to other techniques for identifying reasons for cost variation.

5.2 OLS regression models for identifying reasons for resource use and cost variation.

The previous section illustrated the problems with using OLS regression models for comparing efficiency across health care providers using observational datasets of hospital costs. OLS regression models have also been used in observational studies in particular clinical areas such as stroke and coronary heart disease to identify which patient characteristics are associated with cost variation (Lipscomb et al. 1998, Adams-Dudley et al. 1993). However, these models can also be used in an evaluative context to address particular methodological issues. For example, OLS regression models have been used to identify which costs are important to measure in a CEA based on a RCT. Whynes and Walker (1995) used OLS regression analysis to

establish whether the costs of colorectal cancer could be approximated using a 'reduced list' of resource use categories. Studies have also used OLS regression analysis to tackle the same methodological issue in the areas of mental health (Knapp and Beecham 1993) and adult intensive care (Elliott and Buxton 1998). In an evaluative context, Hoch et al. (2002) used OLS models to estimate how cost-effectiveness varied according to patient characteristics. However, none of these studies examined why costs or cost-effectiveness vary across health care settings.

A few studies have used OLS regression analysis to estimate variation in cost-effectiveness and more specifically costs across health care settings using data from multicentre RCTs. Willke et al. (1998) used an OLS regression model to present a different measure of cost-effectiveness for each country in a multinational study. Hutton (2001) used OLS models to estimate the variation in unit costs and total costs between centres in different countries. Coyle and Drummond (1998) used an OLS model to identify factors that drove cost variation amongst health care centres in the UK. Each of the costing studies reviewed used patient-level cost data, and here an OLS model may take the form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i ; \quad \varepsilon_i \sim Normal(0,\sigma^2) \tag{5}$$

where y_i is the cost for the ith individual; x_i is a patient-level explanatory variable, with associated slope coefficient β_1; β_0 is the intercept and ε_i the error term, which represents unexplained variability between individuals, is assumed to be normally distributed with a mean of zero. The use of these patient-level data means that the model has the potential to adjust for differences in case-mix, and quality of outcome, when assessing cost variation across settings.

5.21 A key statistical assumption in OLS regression models

A key statistical assumption made by these OLS models is that observations across patients are independent and have a common variance. This assumption may be violated when analysing cost variation, as cost data may be hierarchical with patients

nested within centres and centres nested within countries. This concern may arise when analysing data from a RCT or an observational study. In either context patients may be nested, if the characteristics of patients attending the same health care centre are similar compared to those attending a different health care centre. The data may also be clustered within health care settings if there are other cost drivers for example relative factor prices or the level of technological development that operate at a centre or a national level. Thus, costs may have a hierarchical structure as illustrated in Figure 5.1.

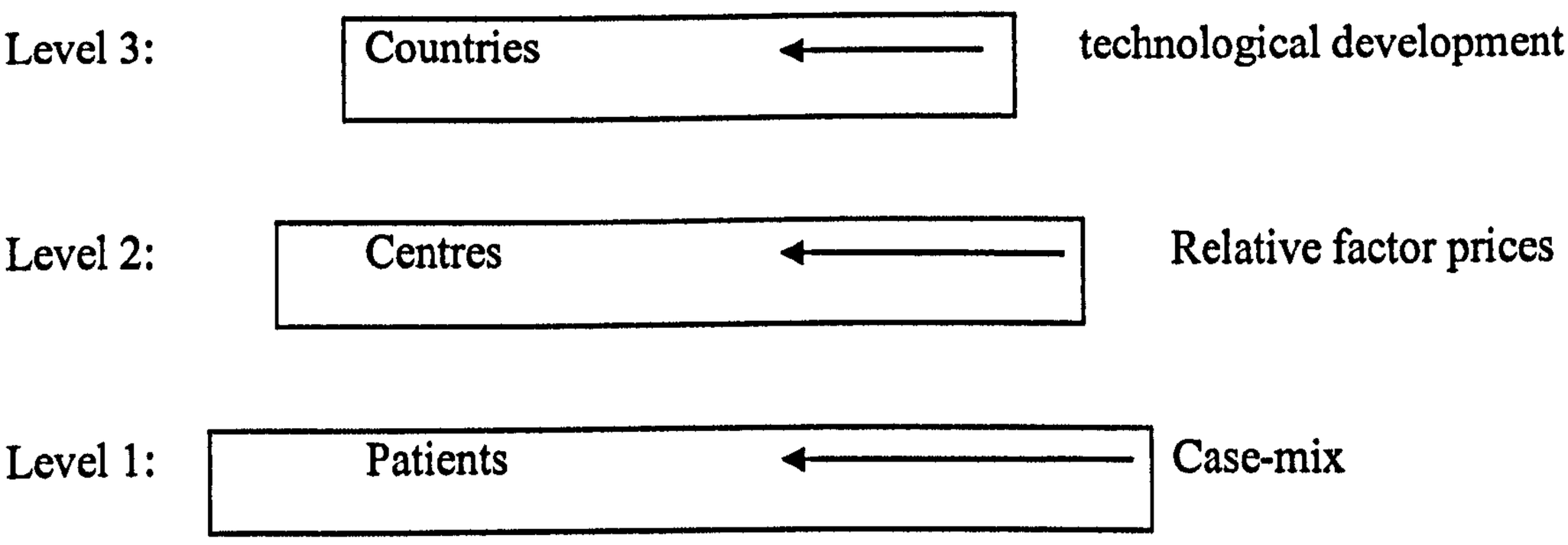


Figure 5.1: Typical hierarchical structure for cost data, with examples of factors that are potentially associated with cost variation at each level of the hierarchy.

The remainder of this section focuses on the use of OLS regression analysis when patients' costs are clustered within health care centres[32]. In this context, the OLS model (5) estimates a single intercept and regression line for the relationship between x and y, ignoring the nesting of patients within heath care centres.

By ignoring the data's structure, an OLS model may report biased estimates of the patient-level explanatory variables. In an evaluative context, an OLS regression analysis may therefore report a biased estimate of the effect of treatment on costs: the incremental costs. The bias tends to be worst when the analysis estimates an overall effect but where the study design is 'unbalanced'. For example, in a multicentre study comparing 'treatment' to 'control' an OLS estimate of the overall treatment effect

[32] Health care centres may in this context refer to different hospitals, nursing homes or primary care providers.

may be biased if the characteristics of the patients in the treatment and control groups differ across the centres. However, even if the study design ensures a balanced allocation across the treatment groups and centres, the costs in the control group and the incremental costs of treatment may differ across centres. Here, an OLS regression analysis that does not allow for this heterogeneity and simply reports the overall effect of treatment on costs, could report a biased estimate. To avoid this bias separate OLS analyses could be run for each centre. However, this approach lacks statistical power and does not identify reasons for variation in costs or cost-effectiveness.

5.22 Inclusion of centre-level variables

In either an observational or an evaluative context it may be desirable to include centre or national level variables to try and identify reasons for variation in costs or cost-effectiveness across health care settings. The OLS model (5) may be extended to include a centre-level variable z_i, for example the size of the hospital concerned, measured by the total number of beds. An OLS model may then take the form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i ; \qquad \varepsilon_i \sim Normal(0,\sigma^2) \qquad (6)$$

Here, the centre-level variable is included as if it was measured at a patient level, thus spuriously inflating the amount of information supplied; the precision of the centre-level estimate is therefore overestimated. Take as an example an OLS regression model with total cost per patient (y_i) as the dependent variable with x_i estimating the effect of case-mix, measured at a patient level and z_i estimating the effect of hospital size, measured at a centre-level. Suppose this observational dataset includes 10,000 patient episodes of care in five hospitals. If hospital one has 400 beds and 1,000 patient episodes, then each of the 1000 patient episodes would be given a value of 400 for the number of beds in the hospital[33]. This ignores the clustering of patient episodes within each hospital. The OLS analysis would therefore overestimate the precision of the estimated effect of hospital size (z) on total cost (y) (Figure 5.2).

[33] Similarly, if hospital 2 had 500 beds and 2000 patient episodes each of these patient episodes would be given a value of 500 etc.

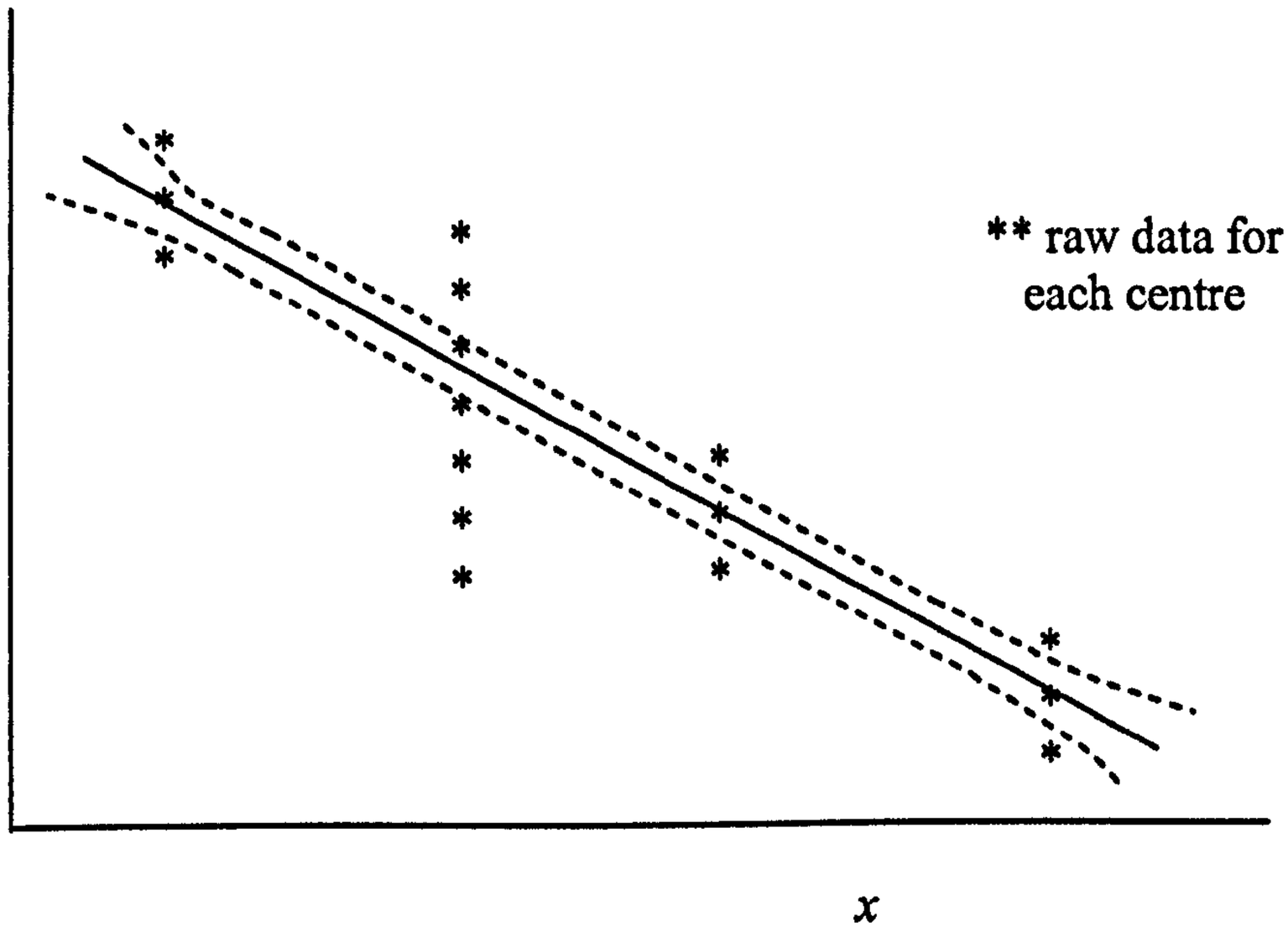


Figure 5.2: OLS estimate of the effect of hospital size*(z)* on total cost per patient episode (*y*) that assumes each patient episode is independent. The dotted lines show 95% confidence intervals.

Rather than including a centre-level covariate, the OLS model could include a dummy variable z_{ij}, for each centre (j). This produces a *fixed* effect of the estimate of the average cost per patient in each centre (model 7):

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij} + \varepsilon_i ; \qquad \varepsilon_i \sim Normal(0,\sigma^2) \qquad (7)$$

This approach provides unbiased estimates of the effects of both the patient covariates and centre *per se*. In an evaluative context, Willke et al. (1998) took a similar approach in their estimation of country-specific cost-effectiveness ratios. To allow for heterogeneity in the cost-effectiveness of a new intervention across countries they included a country by treatment interaction term in the model. However, one disadvantage of this model is that it is not possible to include centre-level covariates alongside dummy variables for centre-effects, thus this model cannot be used to examine, *why* costs may vary across different centres. Another problem is that this model is inefficient, it is necessary to estimate separate parameters for each centre,

and there may be insufficient cases to detect differences in effect sizes across the centres (see section 5.34).

5.3 Multilevel models (MLMs)

MLM is a generic term that encompasses random effects models, random intercept and random slope models, hierarchical models, and random effects meta-analyses (Goldstein 1995). MLMs incorporate the hierarchical structure of data and enable those factors that are associated with cost variation to be modelled directly. The use of MLMs would seem consistent with the findings from the review of economic theory (Chapter four), that suggested factors driving cost variation could operate at different levels of a hierarchy (Figure 5.1). By recognising this hierarchy, MLMs should provide more appropriate estimates of individual and higher-level effects and their variances, than OLS regression models. They could therefore prove useful in either descriptive or evaluative studies.

Much of the pioneering work on the development of MLM has been conducted in the field of education (Goldstein 1992, Goldstein et al. 1993, Goldstein and Spiegelhalter 1996). Here there has been particular interest in separating out the effects of school characteristics such as class size, and pupils' characteristics such as gender, on the level of educational attainment. The results from these models have contributed to the debate over the effectiveness of schools in the UK, and have been used to construct league tables of educational performance (Goldstein 1992, Goldstein and Spiegelhalter 1996). Other applications of MLMs are found in the social science literature, for example, Kreft and de Leeuw (1998) describe a MLM to examine the association of workers' education level with income in 12 different industries. The results showed that the workers within each industry were more similar than workers in different industries. The MLM found that the effect of education on income differed according to each of the industries concerned. The authors reported that using a MLM enabled both industry and individual effects to be accurately estimated and therefore produced more appropriate estimates than using OLS models.

Rice and Leyland (1996) reviewed the use of statistical methods in health services research (HSR) and found that the clustering of observations is commonly ignored. This may reflect the lack of attention given to the development of MLM in this literature, the lack of statistical expertise applied to HSR, and the absence until recently, of suitable software (Localio et al. 2001). There are still plenty of examples of MLMs in the HSR literature for example Goldstein and Spiegelhalter (1996) demonstrated the importance of recognising the hierarchical structure of outcome data when comparing the performance of hospitals, and Rice (2001) illustrated the use of MLM to investigate equity in health care. MLMs have been used to perform meta-analyses that allows for random variation between studies (Thompson 1993). Recent developments in this area have included extending the use of random effects meta-analyses to examine *why* effect sizes may differ across studies (Thompson and Sharp 1999).

While health economists often have to work with clustered data, there are few examples of the use of MLMs in the health economics literature in either descriptive or evaluative studies (Rice and Jones 1997). A recent observational study used a MLM and a OLS regression model to estimate the effect of managed care on health service costs in the US. The results demonstrated that the MLM provided more appropriate estimates of why costs varied between managed care and fee for service providers (Carey 2000). Burgess et al. (2000) used a MLM in an observational study to distinguish between systematic differences in hospital performance and random variation owing to small numbers in particular hospitals. They acknowledge that an important limitation of their study was the lack of patient-level data, which meant the MLM was not able to adjust for case-mix differences across the hospitals.

Scott and Shiell (1997) used a MLM to estimate the effect of changing the way Australian GPs were reimbursed using a dataset consisting of over 4000 consultations nested within 400 GPs. The results showed that drug prescribing was reduced following the change in reimbursement. Cairns and van der Pol (1997) used a MLM to estimate intertemporal preferences for future health. Respondents were asked to identify what future level of benefit made them indifferent between a benefit to be

received in one year's time, and a more distant delayed benefit, for two different periods of delay. Both MLMs and OLS regression models were used to estimate what factors were associated with the discount rates implied from the individual's responses. The results suggested that the standard errors for the explanatory variables were underestimated in the OLS model. In addition, the MLM demonstrated the variation that existed both amongst individuals' responses and across individuals.

Carey (2000) suggests that the lack of suitable, micro datasets has constrained the use of MLMs in health economics, but as Phelps (1995) points out economic evaluations are available that contain data on both patient and institutional characteristics; these datasets may be appropriate for using MLMs to analyse what factors are associated with cost variation. In addition, panel datasets, that allow MLMs to assess variability within and across individuals, are becoming increasingly available for health economic analyses (Jones 2000).

There would seem to be two forms of MLM that are particularly appropriate for analysing hierarchical cross-sectional cost data, the random intercept and the random intercept and slope model. Each of these are considered below:

5.31 Random intercept models

Moving from an OLS regression model to a MLM structure changes the way the unexplained variation, the random error term is modelled. The most basic MLM, the random intercepts model, includes an additional term, which represents the unexplained variation that exists amongst higher-level units. Therefore, in the example of a two level model where individuals are clustered within centres, using subscripts *i* and *j* for the *i*th individual in the *j*th centre, the model may be written:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + \varepsilon_{ij}\,; \qquad \varepsilon_{ij} \sim Normal(0,\sigma^2),\ u_j \sim Normal(0,\tau^2) \tag{8}$$

where β_0 is the mean cost when x_{ij}=0, u_j is a random variable with zero mean and constant variance (τ^2) which applies to all the cases in a particular centre, and e_{ij} is a random error term that represents the unexplained variation for individuals within a

centre. u_j indicates the effect of centre on the average cost per patient, over and above that explained by the set of independent variables. The intercept for the jth centre (previously given as β_0) is now given as a fixed component (β_0) plus a random component (u_j). So in the example below (Figure 5.3), the relationship between case-mix (x) and total cost per patient (y) for the reference centre, is given by the bold line $y_{ij} = \beta_0 + \beta_1 x_{ij}$. The relationships for each of the other centres are given by parallel lines, these differ from the bold line according to the centre level residual, u_j.

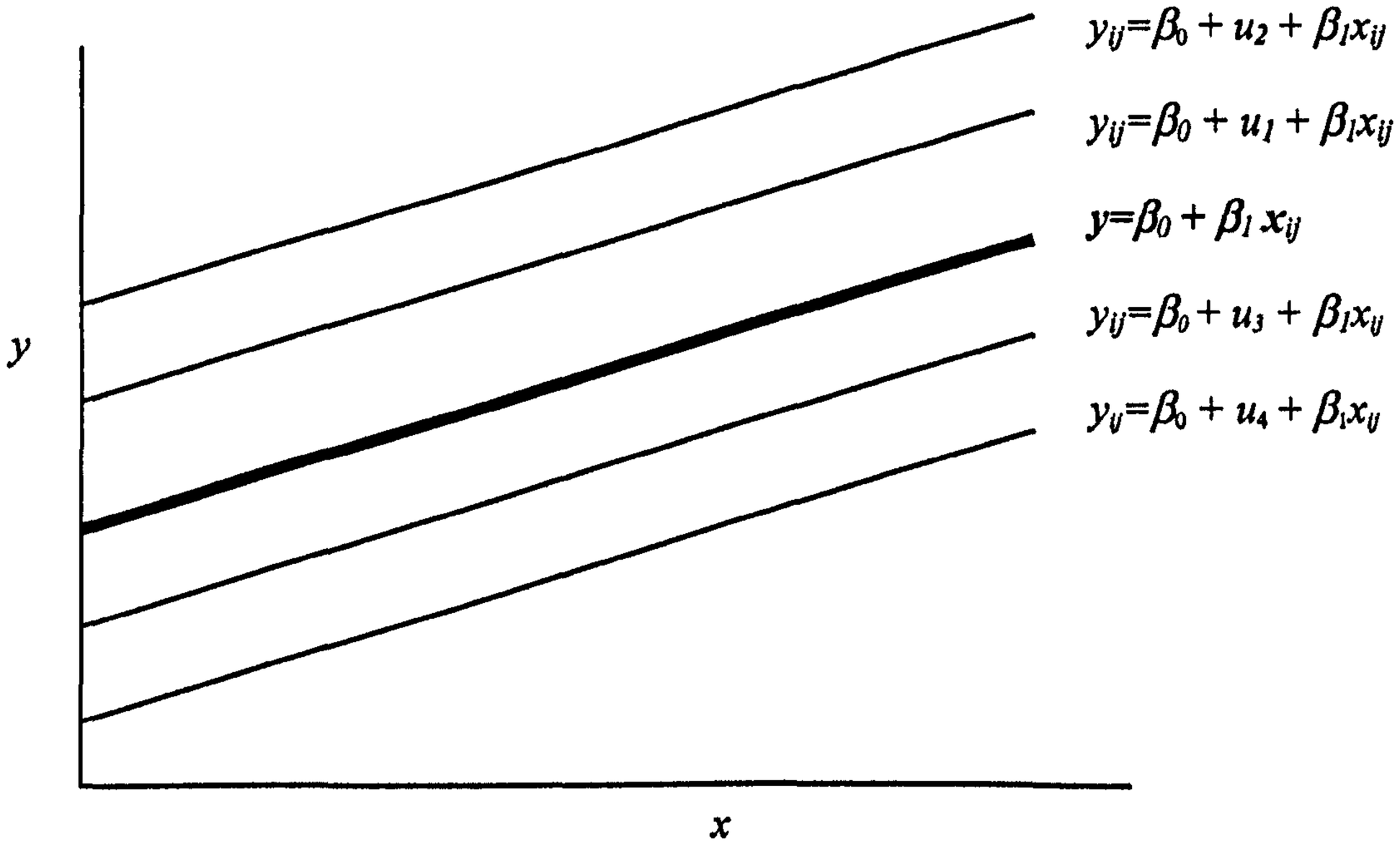


Figure 5.3: Random intercepts model for the relationship between case-mix (x), and total cost per patient (y)

In this model, any variation between the centres, after adjusting for case-mix differences are recognised by allowing each centre to have a different intercept. This improves on the OLS model (5) with a patient-level covariate, by avoiding any bias from ignoring unexplained differences across the centres. The impact this has on the estimated effect of a patient-level covariate depends on the distribution of the data. There may be circumstances where patient-level estimates are similar from both the OLS and random intercepts models. In for example a CEA alongside a RCT, if patients in each centre had similar characteristics in the treatment and control groups and the mean costs in each group did not vary across the centres, then the OLS and MLM estimates of incremental cost would be similar. However, there may be

circumstances, as in Figure 5.4, where the effect for a patient level covariate- x differs between the random intercepts and the OLS estimates. This could be illustrative of a situation in an observational study where for example there is interest in the relationship between age (x) and cost (y). However, one problem may be that if the age of patients differs across the centres, then the OLS analysis estimating the overall effect of age on cost may be inappropriate. Here the OLS estimate of the effect of x on y is the bold horizontal line indicating that when the clustering in the data is ignored, the effect of x on y is approximately zero. However, once the intercepts are allowed to vary by centre as in the MLM, then there is a positive relationship of x with y.

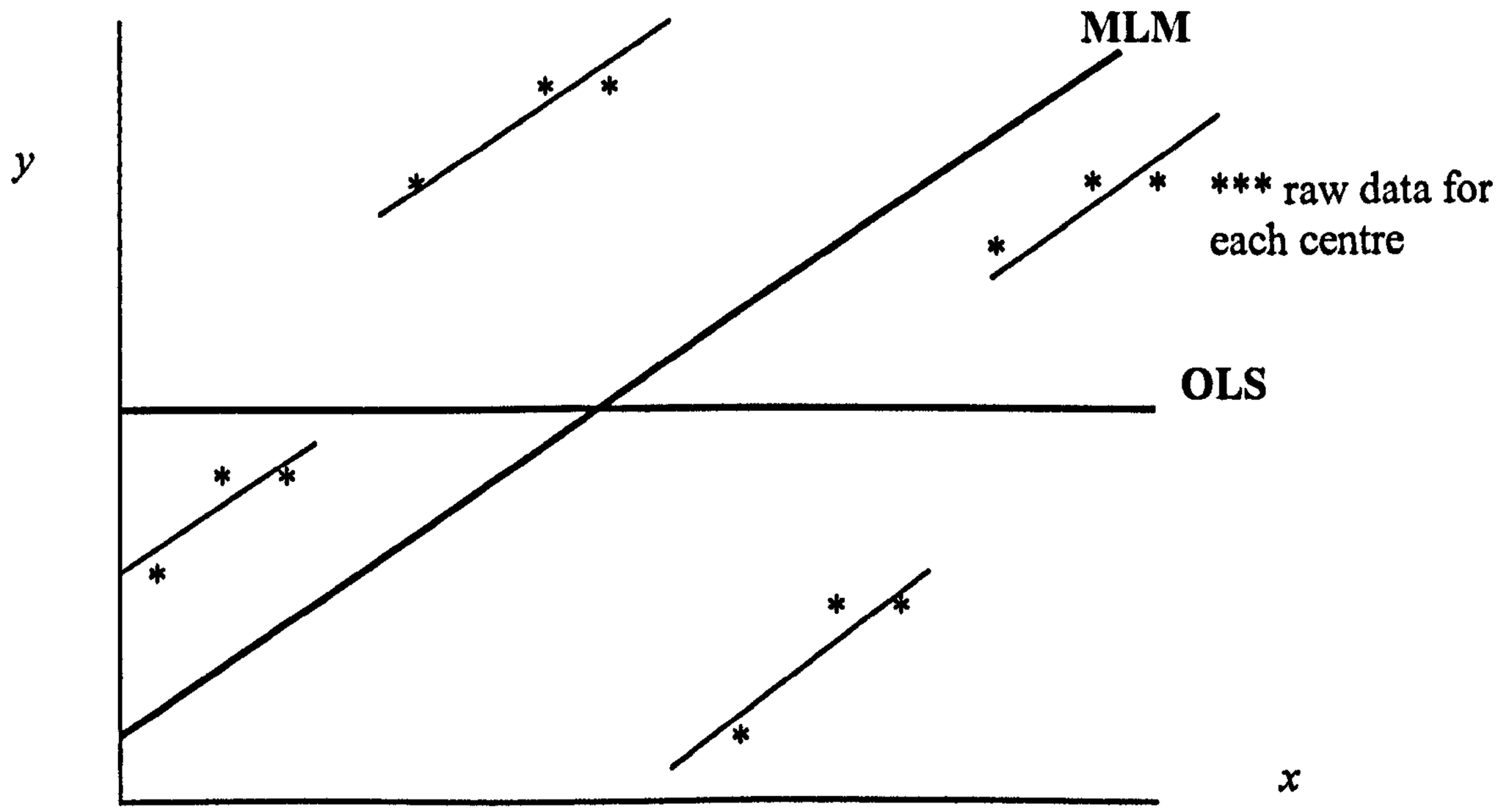


Figure 5.4: A comparison of the effect of a patient-level explanatory variable (x) on cost (y) using an OLS vs MLM (random intercept).

Similarly in an evaluative context the effect of a patient-level covariate, for example, treatment group, may differ between a MLM and a OLS regression analysis. This could happen if either the treatment groups were not balanced in each centre, or if incremental costs differed across centres owing to differences in for example relative factor prices. In these circumstances the estimates from a random intercepts model that adjusts for these differences across the centres would differ from OLS estimates that ignored these variations.

5.32 Random intercepts model with a centre-level covariate

When using data from either observational or intervention studies this MLM can be extended to include additional explanatory variables at the level of the individual or the centre. For example, a centre-level variable, z_j can be added as in the equation below.

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + u_j + \varepsilon_{ij}\,; \qquad \varepsilon_{ij} \sim Normal(0,\sigma^2),\ u_j \sim Normal(0,\tau^2) \qquad (9)$$

In a MLM including a centre-level covariate, the major difference compared to the corresponding OLS model (model 6), is that the MLM correctly recognises that each centre is effectively only providing one data point, and therefore correctly recognises the variance structure. Using the earlier example looking at the effect of hospital size on total cost per episode, a MLM would recognise that hospital size was measured at a centre-level. The MLM would therefore treat the cost data as if there were five data points, rather than 10000 in the five centres and would effectively estimate the effect of hospital size on the average cost per patient in each centre. The level of precision surrounding the estimated effect of hospital size on cost, would be reduced accordingly, so that compared to the OLS model, the confidence intervals surrounding the estimated effect are much wider in Figure 5.5 compared to those in Figure 5.2.

An appealing feature of an MLM is that it allows the effect of including additional explanatory variables on the extent of unexplained variation between centres (τ^2) to be estimated. In addition, the relative degree of dependency amongst observations within a setting can be measured by the intra-class correlation coefficient (ρ) defined as $\rho=\tau^2/(\tau^2+\sigma^2)$. This reflects the strength of 'nesting' within the data hierarchy. For example, in a MLM evaluating the effect of managed care on patient costs, most unexplained variation was found at the level of the patient rather than the health care centre (Carey 2000).

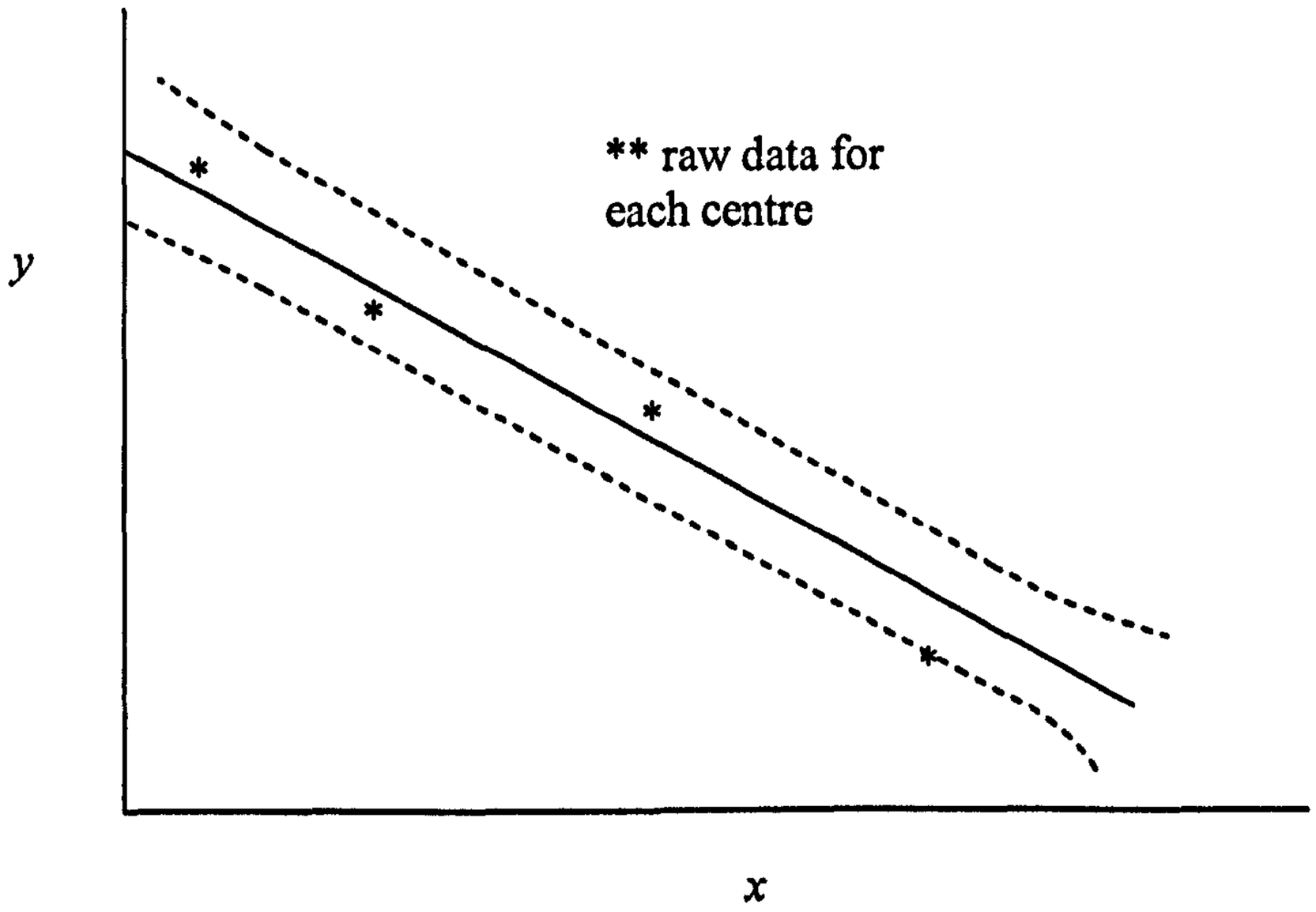


Figure 5.5: Estimated effect of hospital size (z) on total cost per patient (*y*) using a MLM, that recognises patient episodes are clustered within hospitals.

5.33 Random intercept and slope models

A MLM can account for variability in the effect of covariates across higher level units, by using a 'random intercept and slope model' (Leyland and Goldstein 2001). These models are common in the education literature where for example, the aim may be to examine the differential effect of the school attended over time (Goldstein 1995). The pupils' ability at intake can be adjusted for using the random intercept model described above. The differential effect of school attended on attainment over time can then be modelled by allowing the effect of time to vary randomly across the schools- a random intercept and slope model. This requires an additional error term, u_{1j} to be included, which represents a different slope for each school. The effect of time on educational attainment for each school is therefore $\beta_1 + u_{1j}$ where β_1 is the overall effect of time on attainment and u_{1j} is the mean difference in attainment for the individual, *j*th school. u_{1j} is routinely assumed to be normally distributed with a mean of 0 and variance τ_1^2 (model 10). This model also requires the covariance between the

slope and intercepts error terms to be estimated. There is interest in the size of the between-school variance or level of heterogeneity, u_{1j}

$$y_{ij} = \beta_{0j} + u_{0j} + (\beta_1 + u_{1j})\, x_{ij} + \varepsilon_{ij}\,; \qquad \varepsilon_{ij} \sim Normal(0,\sigma^2)\; u_{0j} \sim Normal(0,\tau_0^2) \quad (10)$$

$$u_{1j} \sim Normal(0,\tau_1^2) \qquad \mathrm{cov}(u_{0j}\, u_{1j}) = \tau_{01}$$

This model was used by Carey (2000) when assessing the effect of managed care on costs. Here relationships between cost and both risk status and age were allowed to vary across the health care centres. This model could also be used in the context of a multicentre CEA. Here there may be differences in incremental costs across centres. Including a random slope can recognise this variation across the centres. This form of MLM can also allow for the inclusion of centre-level variables to assess the reasons for this variation in incremental costs or cost-effectiveness across settings. These MLMs can also provide shrinkage estimates for individual centres in the study.

5.34 Shrinkage estimates for individual higher level units e.g. centres

The discussion of OLS models highlighted that when these are used to analyse multicentre data, one approach is to just estimate the overall effects of covariates on the dependent variable (model 6) which can lead to biased estimates. An alternative approach is to use separate OLS regression models for each centre (model 7). So for example, the incremental cost-effectiveness of a new intervention could be estimated separately in each setting. The main problem with this approach is that it leads to an imprecise estimate of cost-effectiveness in each setting i.e. the estimates are statistically inefficient. MLMs offer a further possibility by reporting 'shrinkage estimates' for each higher level unit in this case for each centre. Shrinkage estimators combine the advantages from using the overall mean estimate (greater power and using all information) and the centre-specific estimates (unbiased estimates). These estimators shrink the individual centres' estimates (for example the mean cost in each centre) in towards the overall mean. The estimate for an individual centre therefore uses all the information available and 'borrows strength' from the estimates in other centres. This leads to increased precision and the sharing of information between centres (Rasbach et al. 2002).

The extent to which an individual centre's estimate 'shrinks in' depends on the level of within and between-centre variability. Shrinkage has the greatest impact when a centre's estimate has a high level of within-centre variability and is then 'moved in' towards the mean in those centres with more precise estimates. If the level of between-centre variance is high, the shrinkage estimates have less impact than when the between-centre variance is small. Indeed if there is no between-centre variability, the centre-specific estimates are all the same: i.e. they shrink in completely to the overall mean.

Shrinkage estimators have been used in HSR to provide context specific results, and improve precision (Goldstein and Spiegelhalter 1996). For example in a study estimating immunisation rates in different GP practices, certain individual practices may have very imprecise estimates owing to the small numbers of patients immunised. For these practices a shrinkage estimator would shift the estimated immunisation rate towards the overall mean estimate (Rice and Leyland 1996). The estimates for the smaller practices have 'borrowed strength' from the larger practices.

5.35 Assumptions made in MLMs

While MLMs are appealing for the analysis of hierarchical cost data, certain assumptions have to be made which warrant careful consideration. The most relevant assumptions for the analysis of resource use and cost data are discussed below.

Effects are random not fixed

When considering the use of MLMs versus OLS regression models for analysing cost variation, the relative desirability of fixed versus random effects may be important (Thompson 1993). OLS estimates are always fixed, that is the coefficients estimated only refer to the study sample and not to some wider population from which the data are drawn (Kreft and de Leeuw 1998). In some cases, a fixed effect may be most appropriate, for example, if a RCT compares a high dose drug to a low dose drug, then the estimated effect from the cases sampled may be all the information required. In a MLM, while some coefficients may be fixed, other estimates consist of both a fixed component and a random component representing random variation across the

higher-level units[34]. The use of a random effects estimate infers that the estimates from the centres sampled are representative of the overall population which the study centres are drawn. In some instances, this may be more appropriate; for example, if a MLM is used to estimate the effect of centre on total cost, a random effects estimate suggests that the estimate from these centres represents the overall effect in the whole population. The random effects estimate would potentially be more generalisable than a fixed effects estimate, as it goes beyond those centres included in the study sample.

The specification of random effects does rest on the premise that higher-level units are selected at random from the population concerned. However, in economic evaluation the centres chosen for costing are rarely selected at random. Random effects also assume that the data are exchangeable, i.e. that the data from one centre could be 'exchanged' for the data from another centre (Spiegelhalter et al. 2000). However, if there are systematic cost variations across the health care centres as suggested by economic theory (see Chapter 4), then assuming exchangeability across the centres may not be plausible. Careful consideration of the link between economic and statistical theory would therefore seem warranted when using random effects estimators for analysing cost variation.

Dependent variable is normally distributed

Previous MLMs in health economics (Carey 2000), in health services research (Marshall and Spiegelhalter 2001) and in other sectors such as education (Goldstein 1995) have assumed that the residuals are normally distributed. However, resource use and cost data are usually skewed and the error terms may not be normally distributed (Briggs and Gray 1999). One remedial approach is to use a simple logarithmic transformation, as the distribution of the residual may then be more normally distributed. This method has been used when analysing single-level (Forbes and Dennis 1995) and multilevel cost data (Carey 2000). However, a problem with this approach is that, back transforming the coefficient, provides an estimate of the geometric mean, rather than the arithmetic mean estimate required for policy purposes (Barber and Thompson 1998). More recently, Generalised Linear Models (GLMs)

[34] This characteristic of MLM having both a fixed and a random component often leads to them being termed mixed models.

have been recommended for analysing cost data (Barber and Thompson 2004), as a variety of non-normal distributions can be specified but, unlike data transformations in OLS regression, they make inferences about the arithmetic mean cost directly. Non-parametric bootstrapping has also been recommended for analysing cost data (Barber and Thompson 2000). However, neither GLMs nor non-parametric bootstrapping has been used in hierarchical models for analysing cross-sectional cost data[35].

GLMs are characterised by two essential features: firstly, they have a distribution function (F) that describes the distribution of the independent variable (y) in this case cost per patient. Secondly, they have a link function (g) that describes the scale on which the covariates are related to the mean of this distribution. Suppose y_i is the observed cost for individual i, and μ_i is the expected mean cost from the model $E(y_i)$, then if there is a single covariate x_i representing patient case-mix the general form of the GLM is:

$$g(\mu_i) = \beta_0 + \beta_1 x_i \qquad \text{where } y_i \sim F \quad \text{and} \quad E(y_i) = \mu_i \qquad (10)$$

GLMs are particularly appealing for the analysis of cost data as the focus is always on the arithmetic mean μ_i. In addition, the distribution function F can be chosen from any of the exponential group of distributions (McCullagh and Nelder 1989). This is attractive for the analysis of cost data where the choice of a skewed distribution may be appropriate. Briggs et al. (2005) have shown that choosing the correct distribution for analysing cost data is vital, otherwise the mean estimates may be biased, and the estimate of precision may be statistically inefficient.

The most straightforward of the GLMs is the standard OLS model, that has a normal distribution function and an identity link function. When a patient-level covariate is the only independent variable this model is the same as the OLS model 5. The identity link means that the covariate acts additively on the dependent variable. If for example

[35] Seshamani and Gray (2004) use both these techniques to analyse panel data.

the model includes age (in years) as a continuous independent variable, the model estimates the mean change in total costs as age increases by one year. The main purpose of using GLMs for the analysis of cost data is usually to deal with skewed data where the assumption that the residuals are normally distributed is not appropriate. Here a skewed distribution function may be chosen for example, the Gamma distribution (McCullagh and Nelder 1989). If the Gamma distribution is chosen but the identity link is maintained then the interpretation of the coefficients is the same as in the OLS model 5[36].

Generalised linear mixed models (GLMMs) like MLMs are able to accommodate the hierarchical structure of the data and can specify random intercept or random intercept and slope models. GLMMs share with GLMs, the advantage of being able to make inferences directly about the arithmetic mean, while allowing for non-normal distributions.

Within the group of GLMMs a particularly attractive model for analysing hierarchical cost data is the Gamma model that has a random mean μ_{ij}, and allows for different shape parameters (ϕ_j) across the centres (model 11).

$$y_{ij} \sim Gamma(\mu_{ij}, \phi_j), \qquad \mu_{ij} = \beta_0 + \beta_1 x_{ij} + u_j;$$
$$u_j \sim Normal(0, \tau^2) \qquad (11)$$

This model uses an identity link so that the coefficients can be interpreted on the same scale as the corresponding MLM- the MLM with a patient-level covariate and random intercept (model 8). This model can be easily extended to include higher-level covariates e.g. hospital size, or additional patient-level covariates.

Barber and Thompson (2004) demonstrated that using GLMs rather than OLS models for analysing the effect of patient covariates on LOS and total cost, was more appropriate and fitted the data better. Manning and Mullahy (2001) compared GLMs

[36] Rather than using an identity link, alternative link functions such as the log link can be chosen. If the log link is used the covariates will act multiplicatively on the mean.

with using logarithmic transformations combined with 'smearing' techniques for analysing cost data. They concluded that the choice of method was not clear-cut, and that there were important tradeoffs in terms of bias versus precision, between the different techniques. In the context of performance measurement in education, Carpenter et al. (2003) used a non-parametric bootstrap procedure to analyse hierarchical data on educational performance. They found that in this context the non-parametric bootstrap allowed accurate confidence intervals to be estimated without having to transform the data.

There is a lack of previous research examining the choice of model for estimating factors associated with cost variation, using cross-sectional hierarchical datasets. For these data, it would seem important to compare the use of MLMs that assume a normal distribution with GLMMs that assume skewed distributions, or the non-parametric bootstrap. This comparison could reveal which technique may be more appropriate for dealing with skewed, hierarchical cost data.

Higher-level residuals are independent from the explanatory variable, x_{ij}

MLMs assume that the fixed effects for the independent variables are not correlated with the random effects. If this is the case then the effects estimated are consistent[37] (i.e. reliable), but when x_{ij} and u_j are correlated, then the estimates may be inconsistent (Blundell and Windmeijer 1997). Blundell and Windmeijer (1997) show that even where there is correlation between the x_{ij} and the u_{ij}, then the use of random effects could still be appropriate. In their example, there were approximately 5000 observations and 200 centres. The number of observations per centre ranged from 6 to 95. They concluded that in this instance there were sufficient observations in each group for the MLM to produce consistent estimates for both the fixed and random effects. They suggest that in other circumstances where the number of groups is large, and the numbers in each group are small as is often the case in panel data, then the estimators for the parameters will become inconsistent (i.e. unreliable). In these circumstances, Blundell and Windmeijer (1997) suggest the choice of a fixed effects

[37] Gelman et al. (1998) define consistency as being when $n \to \infty$ the sample mean asymptotes towards the population mean.

specification would produce more consistent results and would be preferable. However, as Rice and Jones point out (1997), there may be considerations other than consistency that govern the choice of a fixed versus random effects model. There may be interest in the parameters associated with the higher-level variables, and then it would be undesirable to just regard such variables as nuisance variables to be adjusted for using fixed effects estimators. The random effects estimate is also more statistically efficient, as it requires fewer parameters to be estimated.

Measurement error

Measurement error is a common problem in econometrics and may pervade the measure of either dependent or independent variables (Gujarati 1988, Jones 2000). Measurement error may occur when for example, there is observer error when recording costs during an interview. Aggregated datasets have been criticised for being especially prone to measurement error, as they often rely on routine or administrative data (Newhouse 1994). In the context of naturally hierarchical data, measurement error often arises as characteristics measured at an individual level are aggregated to give mean estimates for higher-level units. For example, where individual data are missing, or sample data has been collected by for example, only surveying a sub sample of households, then the compositional variable derived at an aggregate level is measured with error (Goldstein and Leyland 2001). The development of approaches for dealing with measurement error in hierarchical models involves adding in reliability measures, however, this work is at an early stage and further research is needed (Goldstein and Leyland 2001).

5.4 Discussion

The aim of the thesis is to evaluate reasons for cost variation across settings and examine the implications for economic evaluation. OLS regression analysis or MLMs are more appropriate techniques for this purpose than DEA or SFA that are used to measure efficiency. Techniques for measuring efficiency tend to require data on large numbers of centres, and usually rely on routinely collected, aggregated costing datasets. Using aggregated datasets to compare costs across health care settings is fraught with measurement issues. In particular, the aggregated case-mix measures used in hospital efficiency studies have been severely criticised for failing to adjust for case-mix differences between different hospitals (Newhouse 1994), and therefore provide inaccurate estimates of inefficiency. The reliance on techniques that require the use of aggregated datasets is not appropriate for addressing the thesis' objectives. Instead, further consideration will be given to those techniques that can analyse cost variation using more disaggregated datasets from observational and evaluative studies.

Economic evaluations now routinely collect patient-level cost data. The development of these patient-level datasets enables OLS regression analyses to estimate how costs (Coyle and Drummond 1998) and cost-effectiveness (Hoch et al. 2002) vary according to patient characteristics. While OLS regression analysis appears a promising technique for analysing cost variation, its use for analysing hierarchical cost data is potentially problematic. OLS regression analysis assumes that individual observations are independent, which is not the case if patients are nested within for example hospitals. There are two strands of *a priori* reasoning to suggest that this clustering exists, firstly patients attending a particular setting may be more similar to one another, than those attending a different setting. Secondly, factors operating at the level of the health care setting, for example factor prices, may differ across settings (see Chapter 4). OLS regression analyses that ignore this clustering of observations at different levels of the hierarchy may provide biased estimates. In this context the problem of bias can arise when analysing data from a RCT or an observational study. Although an RCT can ensure that the treatment groups are 'balanced' there may still be differences in for example the incremental costs of a new intervention across the

study centres. An OLS regression analysis that simply estimated the overall incremental cost-effectiveness across the centres, would ignore this clustering and could therefore lead to biased results.

In the context of a CEA of different health care interventions where data are collected from several health care settings, a MLM may therefore be more appropriate. In particular a MLM that uses a 'random slope' can be used to recognise the clustering in the data, and expose the variation in incremental costs or cost-effectiveness across study centres. By recognising this variation the MLM would avoid the potential bias in the OLS regression analysis. This MLM can also be expanded to include patient and centre-level covariates that may explain why costs and cost-effectiveness may vary across health care settings. A serious problem with using OLS regression analysis to examine the reasons for variation is that by disregarding the hierarchical structure of cost data the significance of higher-level variables may be markedly overstated. Thus OLS regression analyses that have used contextual variables in a cost function without allowing for the clustering of observations within health care centres, may have overestimated the precision of these higher-level estimates.

MLM regression analyses are amenable to the analysis of patient-level data, and do not require vast numbers of higher-level units to evaluate provider characteristics that may be associated with cost variation. The main advantage of using MLMs compared to OLS regression analyses is that they recognise the hierarchical nature of cost data. MLMs can therefore provide more accurate estimates of those factors associated with cost variation and their standard errors. MLM also have the advantage that they can decompose the total variation according to the level of the hierarchy at which it occurs. As Leyland and Goldstein (2001) point out:

"The questions facing researchers concern the degree to which observed differences at the macro-level, typically hospitals or areas- reflect genuine contextual differences at the macro level, or whether they do little more than reflect the composition of those areas in terms of the micro-level, typically the individual." (Goldstein and Leyland 2001, p181)

This would appear to be an important advantage given, that the review of economic theory suggested that factors explaining cost variation would operate at different levels of the data hierarchy (Chapter 4). While MLMs have theoretical appeal for analysing reasons for cost variation, their use in health economics has been limited (Rice and Jones 1997, Jones 2000, Seshamani and Gray 2004). In general, the cost datasets routinely available are highly aggregated, and not amenable to hierarchical modelling, that requires data at more than one level of the data hierarchy. The recent growth in economic evaluations conducted alongside multicentre RCTs provides an opportunity to use MLMs to establish why costs and cost-effectiveness varies across health care settings (Manca et al. 2005).

A number of challenges have been raised concerning the use of MLMs for analysing cost variation. MLMs in other areas have generally assumed that the error terms are normally distributed (Goldstein 1995, Marshall and Spiegelhalter 2001), an assumption that may not be appropriate for the analysis of cost data (Barber and Thompson 1998). A potential solution to this problem would be to use GLMMs. Like MLMs, GLMMs can accommodate the hierarchical structure of data, but they can also allow for non-normal distributions (Barber and Thompson 2004). GLMMs have been used for analysing hierarchical cost data, but in the context of a panel dataset (Seshamani and Gray 2004). However, GLMMs have yet to be used for analysing cross-sectional hierarchical cost data. The use of GLMMs in this context would appear to warrant further investigation. Similarly while the non-parametric bootstrap has been used to analyse single-level cost data (Suri et al. 2001), and multilevel data in the education sector (Carpenter et al. 2003), its use for analysing multilevel cost data needs investigation.

A further issue in the use of MLMs concerns the interpretation of residual variation, in particular the unexplained variation that exists across health care firms. Just as with OLS cost functions, the residuals estimated for each centre, are both above and below the the mean, therefore the residuals cannot be interpreted as measures of efficiency. It is likely that residual variations in cost are likely to reflect measurement errors and random variation, as well as efficiency differences across centres. The recognition of

the role of measurement error and random variation would appear important for any technique, including MLM that aims to compare costs across health care units, even those that use more disaggregated datasets.

There are lessons to be learnt from the debate in the hospital cost function literature about the use of behavioural versus structural cost functions. Although specifying a structural cost function would appear desirable as it is underpinned by production function theory, it has led researchers towards the use of highly aggregated datasets and the ensuing cost functions have been parsimonious with few explanatory variables. By contrast the 'anything goes' approach of the behavioural cost functions can easily be criticised for its disregard of economic theory (Wagstaff 1989a). When analysing the reasons for cost variation using micro datasets it may not be feasible to maintain the theoretical rigour of the structural cost function, but it is still important to use insights from economic theory to decide which explanatory variables to include.

In conclusion, this review found that the use of MLMs is potentially appropriate for analysing factors associated with cost variation and for use in CEA. A MLM can evaluate cost variation in a manner broadly consistent with economic theory, which suggests that reasons for cost variation may operate at different levels. The use of OLS regression analysis may lead to biased estimates, and the significance of contextual variables may be overstated. It is unclear whether OLS regression analyses are an adequate approximation for MLMs in the context of cost and cost-effectiveness analysis, as the lack of suitable, disaggregated datasets has limited the use of MLMs for these purposes. A careful investigation of the use of MLMs compared to OLS regression for analysing the reasons for cost variation is therefore warranted.

5.5 Summary of the central issues emerging from the literature review

5.51 Limitations in the methods currently used in economic evaluations

The literature review highlights that economic evaluations commonly ignore cost variation across settings at both the design and analysis stages. Health care providers in different settings may face different levels of factor prices, use various combinations of factor inputs and differ in the observed costs of producing health care programmes (Chapters 2 and 3). The review suggests that economic evaluations make fundamental decisions at a design stage that limit the consideration of these and other reasons for cost variation. For example, if studies only measure costs in one health care setting, it is unclear whether the costs measured are those of efficient production and therefore represent opportunity costs for the decision context concerned. Unless the costs used in economic evaluations represent opportunity costs, the evaluations may make inaccurate estimates of the relative costs and cost-effectiveness of different health care interventions. The use of inappropriate cost estimates can hinder moves to improve the allocative efficiency of resource allocation. Studies are required that assess reasons why costs vary across health care settings, and identify circumstances where costs may depart from opportunity costs. An assessment of cost variation across settings can provide guidance on the numbers and characteristics of health care settings that should be used in an economic evaluation to ensure that the costs collected represent the opportunity costs for the decision context concerned.

Recent multinational economic evaluations have found that costs may vary widely across health care settings, and that ignoring this variation can lead to inaccurate estimates of cost-effectiveness (Chapter 3). However, these studies have suffered from serious limitations. These studies have not used economic theory to pose hypotheses for why costs may vary, the studies have failed to address measurement issues that pervade the comparison of costs across health care settings, and the analytical methods used have not recognised the hierarchical structure of these cost data.

5.52 A framework for assessing cost variation

Insights from the production and cost function literatures

The production and cost function literatures provide a useful framework for considering why *systematic* variations in cost may occur across health care settings (Chapter 4). The production function literature highlights that health care firms may choose different combinations of factor inputs and still maintain technical efficiency. The cost function literature suggests that firms may adjust the mix of factor inputs used in production according to differences in relative factor prices. Thus, a comparison of observed costs across settings may find that firms across various locations have different factor inputs, and total costs, but each firm is responding to differences in relative factor prices, in a way that is still productively efficient. However, there may be contextual factors that prevent health care firms from achieving productive efficiency.

Consideration of contextual factors

The review used various strands from the literature to identify some of the contextual factors that may be responsible for systematic variations in costs (Chapter 4). Health care firms across international settings may face different incentives for achieving productive efficiency. For example, some countries have introduced prospective reimbursement for hospital services, this provides a stronger incentive for hospitals to minimise costs than retrospective reimbursement. National governments usually try to regulate the overall number of health care professionals employed in each country. The numbers of different health care professionals available to an individual firm may depend partly on the structure of the national labour market. Individual health care centres may also have limited scope for choosing to adopt new health care technologies. National governments may try to control the national levels of public spending on health care, by limiting the rate at which new technologies are introduced. Decision-makers in individual health care settings, for example hospitals may not be able to adjust health care inputs to achieve productive efficiency. Thus, the national level of spending on health care may provide a proxy for the level of health care infrastructure that is available at the national level, but also at the level of the individual health care centre. An investigation of cost variation across health care

settings should use the contextual factors identified in the literature review to pose hypotheses for why costs may vary across health care settings.

Measurement issues

The methodological guidelines for economic evaluation have not carefully considered the measurement issues that arise when comparing or using cost data collected from different health care settings (Chapter 2). For example, the guidelines are not prescriptive about the appropriate level of aggregation to use when measuring resource use or unit costs. However, where studies have taken an aggregated approach to cost measurement, it is unclear whether the same items are included in the unit costs or resource use measured in each health care setting (Chapter 3). An empirical investigation of cost variation across health care settings needs to use a consistent method of cost measurement in different contexts. The review suggests that to reduce methodological inconsistencies between study settings a disaggregated approach to cost measurement is preferable.

The review of the applied cost function literature emphasises the importance of adjusting for differences in case-mix, when comparing costs across health care settings (Chapter 4). Studies that have aimed to estimate differences in technical and productive efficiency across health care settings have required data from many health care units. These studies tended to use routine datasets to compare costs across the entire range of hospital production. For example, the DRG system of case-mix classification has been used, to adjust for case-mix differences that exist across health care providers. Using these highly aggregated measures of case-mix means that observed cost differences across health care settings could reflect unmeasured differences in the patient case-mix. Studies of cost variation across settings therefore need to use appropriate datasets that collect sufficiently detailed measures of patient case-mix to explore the role of differences between patients in explaining cost variation.

The economic evaluations that have considered cost variation across settings have not included sufficient cases or centres to identify systematic reasons for cost variation across health care settings (Chapter 3). The literature review did not find definitive guidance on how to calculate the numbers of patients or centres required to detect

systematic cost differences. However, any study comparing costs across settings should consider whether there are sufficient patients and settings to identify systematic cost differences. Empirical investigations are required to provide guidance on how many centres should be included in economic evaluations, so that systematic reasons for cost variation across settings are identified.

Analysis and presentation of results

The reviews of both the methodological and applied literatures on economic evaluation highlighted that inadequate consideration has been given to the techniques used to analyse variation in costs and cost-effectiveness across health care settings (Chapters 2 and 3). The literature review described alternative statistical methods for analysing resource use, cost and cost-effectiveness variation (Chapter 5), in particular the use of OLS regression analysis and MLMs were considered in detail. While economic evaluations have previously used OLS regression analyses to identify reasons for cost variation, the review found that there may be a fundamental problem with using these techniques for this purpose: OLS regression analysis assumes that individual observations are independent. However, the examination of contextual factors in Chapter 4 suggested that cost data may be clustered within each health care setting as for example patients in a particular health care setting face more similar factor prices, than patients in other health care settings. As MLMs can acknowledge the hierarchical structure of these data, their use may be more appropriate for identifying systematic reasons for variation in costs and cost-effectiveness. MLMs are able to recognise that reasons for cost variation may operate at different levels, for example observed costs may vary between health care settings because of differences in factors that operate at a patient-level (e.g. case-mix) or because of factors that operate at a higher-level (e.g. characteristics of health care providers). If cost data are clustered within health care settings, then cost-effectiveness data may also have a hierarchical structure. MLMs are therefore attractive in an evaluative context for analysing cost-effectiveness data collected from several different locations. However, the use of MLMs in health economics has not been carefully explored, and empirical studies are required to assess whether MLMs should be used to analyse costs and cost-effectiveness.

While the measures of costs and cost-effectiveness presented clearly have to apply to the local decision context, methodological guidelines do not clarify how this should be achieved. In particular, where an economic evaluation has been conducted alongside a multicentre RCT the guidelines do not explain whether the results should be presented for different locations or whether it is acceptable to just present measures of cost and cost-effectiveness that are pooled over all the study centres. An empirical investigation should consider the most appropriate way to present the results of a multicentre cost-effectiveness study. The way the results of an economic evaluation are presented may have an impact bearing on whether they are useful for policy-making and move resource allocation towards allocative efficiency.

5.6 Conclusions

The literature review has assessed how economic evaluations currently consider cost variation across settings, and has identified gaps in the literature. The review has also identified *a priori* reasons why costs may vary across settings. Finally, the literature review has found that there are important measurement and analysis issues that arise when comparing costs across health care settings. An investigation is therefore required that uses *a priori* reasoning and recognises these measurement and analysis issues. This investigation can enhance understanding of why costs vary across health care settings. The next chapter introduces the empirical investigation.

Chapter 6: Introduction to the empirical investigation: conceptual framework, resource use measurement, data description and hypothesis generation.

6.0 Applying the conceptual framework in the empirical investigation.

The literature review identified fundamental gaps in the economic evaluation literature. The methods used in economic evaluation need to be informed by empirical studies that assess why costs may vary across health care settings. The results from these studies could address a range of issues that arise in the design and analysis of multicentre economic evaluations[38]. The empirical investigation in this thesis attempts to address some of the questions raised. These questions include: which centres should costs be collected from in a multicentre economic evaluation? Does using MLMs rather than OLS regression models for identifying factors associated with cost variation lead to different results? How should economic evaluations present results to make systematic variations in costs across health care settings transparent? The empirical investigation attempts to consider these issues by addressing the following specific objectives:

1. To generate hypotheses for why costs may vary across health care settings.
2. To identify which factors are associated with variability in resource use and costs using MLMs and OLS regression models.
3. To compare the use of OLS regression models with MLMs for analysing international cost-effectiveness data.

[38] For a more complete list of issues that could be addressed by empirical studies of cost variation see Chapter 2, page 57.

To address the first of these objectives the empirical investigation has to gather data that can generate hypotheses for why costs may vary across health care settings. The empirical investigation therefore has to measure resource use and unit costs across a broad range of health care settings. The study also has to collect data on the factors that are potentially associated with cost variation across settings. The literature review identified *a priori* reasons why costs may vary systematically across health care settings and grouped these reasons into patient and contextual factors. The empirical investigation therefore collects data on patient factors such as case-mix and on a range of contextual factors (Chapter 6). The contextual factors suggested by the literature review (see Table 4.2 in Chapter 4) include the characteristics of the labour market, the % of GDP spent on health care, the way health care providers are reimbursed and the level of patient copayments. Within this chapter (section 6.4) information from the empirical study on resource use (6.41) and a range of patient (6.42) and contextual (6.43) factors that may be associated with resource use variation. These data are then used to pose about hypotheses about the reasons for resource use variation (6.5).

The literature review suggested that factor prices and factor inputs may vary across health care settings. This may lead to differences in unit costs, and measurement issues when comparing costs across health care settings. Chapter 7 considers the variation in factor prices and factor volumes across the settings included in the study (see section 7.5). Where cost data are collected across international health care settings with differing factor prices, this has implications for the choice of currency conversion factor. Chapter 7 address this specific measurement issue and presents cost estimates using different currency conversion factors.

To provide an assessment of the reasons for systematic cost variations, the empirical investigation has to address the measurement issues raised in the literature review. In particular, the empirical investigation must use a consistent costing methodology across a range of health care settings. The study should take a sufficiently broad perspective to cost measurement and record costs over an adequate time-horizon. The literature review suggested that the empirical investigation should measure costs in a disaggregated way and distinguish between differences in factor inputs and factor prices. The empirical investigation therefore considers the potential role of these

measurement issues in explaining resource use (Chapter 6) and cost (Chapters 7 and 8) differences across the settings included in the empirical investigation.

Even if a detailed approach to measurement is used, this does not mean that the empirical investigation should only measure costs in a few health care settings. Clearly there are trade-offs between avoiding the measurement issues that arise when taking a more aggregated approach and measuring costs in sufficient settings to identify systematic variations in cost. The literature review suggested that where economic evaluations have compared costs across just a few health care settings ($n<7$), they have been unable to identify systematic reasons for cost variation (Chapter 3). An important challenge for the empirical investigation is to avoid inconsistencies in the measurement of costs but also to collect data from sufficient centres to identify systematic variations in costs.

The second objective of the empirical investigation is to consider the use of MLMs to identify reasons for systematic variations in costs (Chapter 8). As previous studies have used OLS regression analysis to assess cost variation, it is important to compare the use of MLMs to OLS regression analyses. This comparison needs to assess whether the use of MLMs rather than OLS regression analyses leads to different results. The empirical investigation also has to consider whether assumptions made when using MLMs in other sectors, namely that the residuals are normally distributed, are plausible in this context.

Finally, as the aim of the thesis is to inform the conduct of economic evaluations, the thesis has to consider the implications of the results of the empirical investigation for the design and analysis of economic evaluations. In particular, the empirical investigation also has to compare the use of OLS regression analysis to MLMs for analysing multicentre cost-effectiveness data. The literature review raised issues concerning the use of MLMs in this context, namely whether the assumption of exchangeability applies when analysing cost data from different health care settings (Chapter 9)

For the empirical investigation to meet the objectives described, a case study is required that that can address the measurement and analysis issues identified. The

next section explains the rationale behind the choice of case study for the empirical investigation.

6.1 Choice of case study for the empirical investigation

To meet the thesis' objectives the empirical investigation has to use an appropriate case study, one that possesses the features defined in the conceptual framework. The case study chosen for the empirical investigation was the EU funded Biomed II stroke study (McKevitt et al. 2000). I was the network health economist on the Biomed II stroke study from 1995-1999, and over this period the work presented in this PhD thesis was commenced[39]. The Biomed II stroke study is an observational study that measures the costs of stroke care for 1757 cases managed in 13 centres in 10 countries. Previous studies have shown that the burden of stroke is large (Bosanquet and Franks, 1998) and that there are important variations in the way stroke care is provided across European centres (Beech et al. 1996). The Biomed II stroke study aimed to measure the resource use, costs and outcomes associated with various ways of managing stroke patients (McKevitt et al. 2000). As part of this study, information was therefore collected on many of the patient and contextual factors that the literature review suggested could be associated with variation in resource use, unit costs and total costs. This study was therefore judged suitable for use in the empirical investigation. The empirical investigation of the thesis extended the Biomed II study to address the measurement and analysis issues outlined in the conceptual framework for the thesis and consider the methodological implications for economic evaluations (Chapter 5).

In the Biomed II study resource use and unit cost data are collected in each centre using a standardised costing methodology (see Chapters 6 and 7). The study included sufficient patients and centres to allow for a thorough investigation of systematic reasons for cost variation as part of this thesis. These features made the Biomed II study more appropriate for the empirical investigation, than many of the economic evaluations based on multicentre RCTs that had been conducted at the

[39] Some of the results have been published in peer-reviewed journals (see the publication list at the beginning of the thesis).

commencement of this thesis or have been conducted since (see Chapter 3). These studies included fewer cases in each centre, only collected unit costs in a small number of centres and did not use a consistent disaggregated approach to cost measurement.

The Biomed II stroke study provides an appropriate context for testing whether using MLMs is more appropriate than using OLS regression analyses for identifying reasons for cost variation (Chapter 8). There are *a priori* reasons for expecting that the costs of stroke care are clustered within health care centres (see Chapter 4), but it is unclear whether the use of MLMs, rather than OLS analysis would lead to different conclusions.

The empirical investigation for this thesis aims to inform the methodological debate about how economic evaluations should be conducted. To consider the specific issue of whether MLMs are more appropriate than OLS regression analysis for analysing multicentre cost-effectiveness data, a multicentre cost-effectiveness dataset is required. As appropriate empirical datasets were not available, the results from the cost analysis are used to extend the observational stroke to generate a cost-effectiveness analysis (Chapter 9). This generated dataset mimics a multinational cost-effectiveness study in this disease area and provides a vehicle for comparing the use of OLS regression analysis versus MLMs for analysing cost-effectiveness.

To summarise, to address the issues raised in the literature review an empirical investigation was required. The empirical investigation in this thesis extends the Biomed II stroke study and provides a case study for examining reasons for cost variation across health care settings. By extending the Biomed II stroke study the empirical investigation addresses measurement issues that arise when comparing costs across settings.

This chapter introduces the methodology used to measure the resource use data, describes these data, and poses hypotheses for resource use variation. Chapter 7 describes the methods used to collect unit costs, analyses why the unit costs may vary across health care settings and presents the unit costs and total costs. Chapter 8 tests the hypotheses for why resource use and total costs may vary across the health care

settings included in this study using OLS regression analysis and MLMs. Chapter 9 extends the results from Chapter 8 to generate a cost-effectiveness dataset. This chapter then compares the use of OLS regression analyses with MLMs for analysing multicentre cost-effectiveness data.

6.2 Chapter overview

The aim of this chapter is to detail the methodology used to measure resource use, to describe the resource use data, and the factors potentially associated with resource use variation, and then to use these data to pose hypotheses for why resource use might vary across different health care settings. Identifying reasons for resource use variation is an important prerequisite for understanding why costs vary across health care settings. The next section of this chapter describes the methodology used to measure resource use and the factors associated with resource use variation. Section 6.4 describes the data and highlights variations across the centres in resource use and in factors potentially associated with resource use variation. Section 6.5 poses hypotheses regarding resource use variation. Finally, section 6.6 discusses the data presented in this chapter and identifies any issues raised by the resource use methodology for the subsequent interpretation of results.

6.3 Measurement of resource use and factors associated with resource use variation

6.31 Centres included in the study

The empirical investigation measures resource use for each patient attending each of the centres included in the study. The term 'centre' refers to the health and community care providers delivering stroke care. Each centre consists of an acute hospital that takes stroke patients as direct admissions, and the majority of these acute hospitals are teaching hospitals. Table 6.1 lists the location of the centres, and the main characteristics of the acute hospitals included in the study. There are some similarities between the centres included; each of the acute hospitals is a publicly

funded hospital, and pays employees by salary rather than according to the volume of patients seen or the number of procedures completed. Each of the acute hospitals takes patients as direct admissions, does not use a triage system for selecting cases, and provides routine acute care to the local community.

However, these centres were not selected because they were similar, but to compare the different ways in which stroke care may be provided across Europe. Previous research found that there were differences in the duration of acute hospitalisation and the main department where patients were managed (Beech et al. 1996). Site visits were undertaken at the start of the Biomed II study to collect more detailed information on the way stroke care is produced in each centre. The site visits found that the centres also differed in the services provided post discharge, for example some centres had additional inpatient care provided in a separate rehabilitation hospital (McKevitt et al. 2000). It was therefore important to collect data that allowed the different production processes in the centres to be described and to generate hypotheses for the reasons for resource use and cost variation.

Table 6.1: Centres included in the study

Centre	Country	Definition of acute Hospital[1]	Main department managing stroke patients	Separate rehabilitation hospital?
Almada	Portugal	Teaching	General Medicine	No
Menorca	Spain	Teaching	General Medicine	No
Florence	Italy	Teaching	Neurology	Yes
Dijon	France	Teaching	Neurology	Yes
Copenhagen	Denmark	Teaching	Neurology	Yes
Kuopio	Finland 1	Teaching	Neurology	Yes
Turku A	Finland 2	Teaching	Neurology	Yes
Turku B	Finland 3	District	General Medicine	Yes
London	UK	Teaching	Elderly Care	No
Warsaw	Poland	Teaching	Neurology	No
Kaunas A	Lithuania 1	Teaching	Neurology	Yes
Kaunas B	Lithuania 2	District	Neurology	Yes
Riga	Latvia	Teaching	Neurology	No

Source: McKevitt et al. (2000)[1].Teaching hospital, defined as a hospital which has an accredited medical school, district hospital defined as hospital without a medical school.

6.32 Patients included in the study

Each centre prospectively recruited, over a one-year period, patients who had suffered a first-ever stroke, defined using the WHO criteria (Hatona 1976), and were admitted to hospital. Patients who suffered a stroke but were not admitted to hospital were excluded from the study. Patients who had a subarachnoid haemorrhage were excluded, as well as 31 patients who refused consent to participate in the study. A total of 1981 patients were registered for inclusion in the study, of whom 134 were missing key case-mix variables and 127 follow-up data. The dataset considered for use in this thesis consisted of those 1757 cases with complete information.

The exclusion of decedents from the resource use and cost analysis

The primary outcome measure in the Biomed II study was survival up to three months post stroke. *A priori* reasoning suggests that over a three month period those patients who died would consume less resources than those patients surviving the stroke[40]. The survival analysis for the Biomed II study showed that there were differences in survival following stroke across the centres after adjusting for differences in patient factors (Grieve et al. 2001a). This raises the question: would any cost variation across the centres simply reflect between-centre differences in survival? More specifically, does higher cost in certain centres simply reflect lower mortality in these centres?

One way of considering this issue would be to include death as a variable in a cost function. However, this would lead to problems of interpretation. Firstly, including death alongside patient-level characteristics (see section 6.34) would be problematic. These covariates such as incontinence, are associated with survival as well as cost and hence there would be multicolinearity amongst the independent variables (Gudjarati 1988). A more fundamental problem is that it would be difficult to interpret the potentially endogenous relationship between death and cost. It may be that those patients who died consumed fewer resources, it may also be true that those patients who had access to fewer resources were more likely to die. Disentangling this association is particularly problematic in a multicentre context where some centres have access to more resources and this may be associated with lower mortality.

---

[40] While there is a growing literature suggesting that in more general populations proximity to death is strongly associated with increase in costs, in stroke care less resources are consumed by those dying in the first three months post stroke, than by those surviving this period (Grieve et al. 2001b).

A potential solution to this endogeneity problem is to use an instrumental variable approach. Here, a set of variables are identified that are associated with the variable in question, and are then used instead of the endogenous variable in the original model (McCellan and Newhouse 1997). A key challenge with this approach is to identify a suitable instrument. This requires finding covariates that are associated with the endogenous variable but which are not independently associated with the outcome variable. So in the context of this investigation, the variables concerned would need to be associated with death but not cost. The covariates collected in the study (e.g. incontinence) were generally associated with both cost and survival and hence it was not possible to identify an appropriate set of variables to act as an instrument and avoid the endogeneity problem.

Instead a pragmatic approach was taken to this problem and patients who died were excluded from the resource use and cost analysis (chapters 6-8). The subsequent analysis of the reasons for resource use and cost variation are explored for those patients who were still alive at three months post stroke (n=1298). In the cost-effectiveness study in chapter nine it is possible to include both decedents and survivors (n=1757) in the analysis (see also page 180).

6.33 Resource use measures

To ensure comparability across the centres it is important to address some of the measurement issues raised in the literature review, in particular it is necessary to adopt a sufficiently broad perspective and measure a wide range of costs. A hospital and community health service perspective was therefore taken to resource use measurement. The use of all hospital and community services was recorded for three months post stroke. Costs to patients and their carers and the costs of lost production were not measured in the study. However, information was collected on whether there was input from the patient's carer, for example with assisting the patient with their activities of daily living. This enabled the analysis to examine whether informal care inputs were either a substitute or a complement to hospital or community services (see below).

The length of hospital stay in each hospital, by ward type were recorded for each patient (intensive care unit [ICU], neurology unit, dedicated stroke specific unit, or general medical ward) in each hospital. The total length of stay (LOS) per patient was calculated by summing the length of stay in each hospital. The use of diagnostic investigations was recorded for each patient from medical records.

The literature review suggested that the quality of care may be associated with the cost of care. Care, which meets the definition for organised stroke care,[41] is associated with better outcomes and reduced LOS, and may be defined as better quality care than routine stroke care (Stroke Unit Trialists' Collaboration 1999). During site visits to each centre, care provided on each ward was categorised according to whether it met the criteria for organised stroke care (Stroke Unit Trialists' Collaboration 1999).

The use of outpatient and community services (hospital clinics, therapy, GP visits, home carers, nursing and residential homes), was recorded during interviews conducted by clinical investigators in each centre with the patients and their carers at three months post stroke. The use of outpatient and community services was then summarised as a categorical variable[42]. This variable identified those patients for whom alternatives to hospital care such as home help or home nursing clearly existed. Similarly, for each patient it was recorded whether there was any care input provided by the patient's family, and a separate categorical variable was defined for family support. These two variables are used to identify those patients for whom an alternative to hospital care existed. The literature review suggested that the availability of care alternatives may be important in determining the duration of hospitalisation. So as well as being resource use measures themselves, these variables could be considered as factors potentially associated with hospital resource use.

---

[41] The criteria for organised stroke care relate to the level of training and education for different members of staff, the extent to which multidisciplinary teamwork is practised, and whether care is led by a clinician with a particular interest in stroke (Stroke Unit Trialists' Collaboration 1999).

[42] where 1=no use of community services, and 2= use of community services.

6.34 Patient factors potentially associated with resource use variation

Hospital-based stroke registers were established for one year during 1996-7. Designated investigators at each centre completed standard forms to record baseline information. Individual patient data were collected on patient characteristics (sex, age, pre-stroke living conditions) and stroke severity measures (incontinence during the first week after stroke, paralysis at hospital admission; stroke subtype [cerebral infarction, intra-cerebral haemorrhage, or unspecified stroke]). These patient characteristics had not previously been shown to differ across European health care settings, and may be associated with differences in resource use, cost and outcome (Beech et al. 1996, Wolfe et al. 1999).

6.35 Contextual factors potentially associated with resource use variation

In this section the term contextual factors refers to factors that may *operate* at the level of the ward, department, hospital, centre, region or national health care system. In general, these factors are also measured at the centre or national-level. For some of these factors however, the best data was measured at a patient-level.

One factor that may be associated with the level of resource use is the patients' geographical access to hospital care. The level of access to hospital may depend on factors such as the density of hospitals within a particular region. For this dataset the best data collected on access to care was the time from stroke onset to hospital admission, which was reported for each patient. Similarly, the level of access to specialist facilities, in particular neurologists input, or the use of inpatient rehabilitation facilities may determine resource use. During site visits to each centre, information was collected on whether neurologists were involved in providing care on each of the wards managing stroke patients. Data were also collected on whether separate inpatient rehabilitation services were available at other hospitals following discharge from the acute hospital. These data were used to define for each patient in the dataset whether they had access to input from a neurologist or utilised a rehabilitation hospital during their stay.

For each hospital, information was collected on the bed-occupancy and the number of beds on the wards where stroke patients were managed. Information was gathered in

each centre on whether there was prospective reimbursement for hospital services through a Diagnosis related group (DRG) system. It was recorded in each centre whether patients were required to make copayments for their acute care, and what the nature of these copayments was. For example, it was noted whether patients were charged a fixed amount or whether the copayments was linked to the LOS. Each of these variables was recorded at the level of the health care centre.

Finally, information was gathered from the literature on the national GDP per capita, and the national public health care expenditure as a % of GDP. These measures provide an indication of the national level of health care infrastructure, as this may be associated with resource use differences across the centres (OECD 2000, OECD 2001, Karaskevica and Tragakes 2001, Karski et al. 2002, Cerniauskas and Murauskiene 2002).

To summarise, data on hospital and community resource use were collected for three months post stroke for each patient in each centre. As survival varied across the centres, the main analysis for the thesis uses those patients surviving up to three months post stroke. The focus of the analysis will be on which factors are associated with resource use variation. Information is therefore recorded on patient characteristics for each patient. Data on contextual factors such as the method of reimbursement for hospital services were collected on site visits to each centre and from the literature review, and are reported at the level of the health care centre or country concerned. The next section describes the key resource use data and the patient, centre and national-level factors that are potentially associated with resource use variation.

Table 6.2: Summary of variables collected in the study that are used to investigate resource use variability.

Groups of variables	Specific Variables	Level measured at	Source
Patient factors	Patient case-mix, age	Patient	Project database
Contextual factors			
	access to neurologists?	Patient	Project database
	rehabilitation hospital?	Patient	Project database
	DRG system?	Centre	Site visits
	Patient copayments	Centre	Site visits
	% GDP on health care	National	Literature
Resource use measures	LOS, number CT scans	Patient	Project database

6.4 Data description: Resource use and factors associated with resource use variation

This section begins by presenting resource use differences across the centres. Following this, data are presented on the patient and contextual factors that may be associated with resource use variation. Differences in resource use and in patient-level factors are compared across the centres using statistical tests. The chi squared (χ^2) test is used to establish if there are statistically significant differences across the centres for categorical variables. For continuous variables ordinary least squares (OLS) regression analysis is used with the resource use or patient factor as the dependent variable and dummy variables for each centre as the independent variables. The corresponding χ^2, F statistics and p values are reported for each measure of resource use and for each of the patient factors.

6.41 Description of resource use differences

In this section the resources used in producing stroke care in the different centres are described. The resource use measures selected are those that are likely to be important determinants of total costs, and also those that demonstrate possible substitutions in the production of stroke care, across the centres.

The first comparison of resource use across the centres uses aggregated measures: the total LOS in hospital, the days in institutions (residential and nursing homes), and the proportion of cases visiting a physician in the outpatient department in the three months post stroke. The comparisons demonstrate that important differences exist in these resource use measures across the centres (Table 6.3). For example, although the average total LOS in hospital is 27.3 days, the mean ranged from 7.8 days in the Spanish centre to 39.3 days in the UK centre. Similarly, the mean LOS in institutional care ranged from 0.2 days in the Danish centre to 18.5 days in the Finnish centre. The proportion of patients who visited outpatients in each centre ranged from 4% to 71%.

Table 6.3: A comparison of mean resource use across centres

Centre	N	Mean (SD)total LOS in hospital [days]	Mean (SD)days in institutional care	% visited outpatients
Portugal	73	13.0(14.2)	3.4(16.3)	71
Spain	32	7.8(6.4)	2.3(12.7)	66
Italy	109	20.2(23.0)	5.2(19.2)	17
France	105	24.2(27.5)	5.2(19.4)	30
Denmark	246	37.1(33.7)	0.2(2.1)	52
Finland 1	37	19.1(20.2)	1.1(6.9)	22
Finland 2	81	36.5(31.1)	7.5(17.9)	32
Finland 3	57	37.4(29.0)	18.5(25.8)	9
UK	73	39.3(33.8)	1.3(7.6)	49
Poland	92	25.5(15.9)	1.6(9.3)	72
Lithuania 2	186	26.5(16.6)	1.6(9.8)	4
Lithuania 1	62	28.0(13.5)	0.6(4.5)	48
Latvia	145	18.7(11.1)	2.8(12.8)	4
ALL	1,298	27.3(25.6)	3.2(13.7)	34
		$F(12,1285)=13.33$ P<0.001	$F(12,1285)=11.92$, P<0.001	$\chi^2(12)=336$, p<0.001

As the resource use measurement was for up to 90 days post stroke, any resource use after this time point was excluded from the analysis. To consider the potential impact this may have on the results, the patients' place of residence at three months post stroke is presented below (Table 6.4). The majority of patients were in their own home by three months post stroke. While further resource use attributable to the stroke may have occurred after this time point, for these patients this would only relate to hospital readmissions or to the use of outpatient or community services. Previous research has shown that both of these items are small components of the overall costs of stroke care (Porsdal and Boysen 1997). For those patients who are still in hospital, nursing or residential homes at 90 days post stroke, the further resource use that is excluded may be an important component of total costs. In the context of this investigation, the concern is whether limiting the time-period would have an impact on the comparison of costs across the centres. The centres where most patients were in hospital or institutional care at three months were in Denmark, Finland (2 and 3) and the UK. These centres were those with the longest LOS, and so limiting the time-frame to 90 days is likely to lead to an underestimation of differences in total costs across the centres.

Table 6.4: Place of residence at three months post stroke (% in each place)

Country	N	Home	Hospital	Nursing home care	Residential home	Sheltered accommodation
Portugal	73	95	1	4	0	0
Spain	32	97	0	0	0	3
Italy	109	93	2	0	0	6
France	105	86	5	2	0	8
Denmark	246	76	22	0	0	2
Finland 1	37	97	3	0	0	0
Finland 2	81	72	14	12	2	0
Finland 3	57	56	5	28	11	0
UK	73	74	15	11	0	0
Poland	92	96	1	3	0	0
Lithuania 1	62	100	0	0	0	0
Lithuania 2	186	98	1	1	0	0
Latvia	145	98	1	3	0	0
ALL	1,298	87	7	5	1	0
				$\chi^2(12)=11.92, p<0.001$		

More detailed comparisons of resource use within the hospital suggest that there is considerable variation both within and across centres in the resources used to produce stroke care. For example, in each hospital patients may stay on general medical, ICU, stroke units, elderly care units, rehabilitation units or neurology wards. The main area for caring for stroke patients varied according to the centre. In Table 6.5 the average number of days patients spent on each ward are presented for each centre. Over all the centres, the mean LOS was highest on the rehabilitation wards, however in certain centres (UK and Poland) the mean stay was highest on dedicated stroke units, whereas in the centres in Portugal and Finland, most days were spent on general medical wards. The relative number of days patients stay on each ward may be an important determinant of the overall cost variability both within and across centres. The results also show that there was variability in the quality of care provided across the centres according to the average number of days the patients received organised stroke care. Organised stroke care has generally been associated with better outcomes and shorter LOS than conventional care (Stroke Unit Trialists' Collaboration 1999).

Table 6.5: Mean days stayed on each ward, and mean (SD) total days

Centre	ICU	Neurology	Stroke uni	Rehabilitati	General medical	Total days
Portugal	0	6.4	0	0	6.6	13.0 (14.2)
Spain	0.1	0	0	0	7.8	7.9(6.3)
Italy	0	0	6.3*	9.7*	4.1	20.2(23.0)
France	0.4	9.0*	0	14.0*	0.9	24.2(27.5)
Denmark	0.4	6.3*	9.6*	12.2*	8.6	37.1(33.7)
Finland 1	0	9.4*	0	9.6*	0	19.1(20.2)
Finland 2	0	14.4*	0	18.7	3.4	36.5(31.1)
Finland 3	0	0.2	0	11.2	25.9*	37.4(29.0)
UK	0.7	0	19.3*	13.8*	6.2*	39.3*(33.8)
Poland	2.5	0	22.1*	0.9	0	25.5(15.9)
Lithuania 1	0	17.6	0	9.6*	1.2	28.0(13.5)
Lithuania 2	0.3	16.0	0	7.1*	3.2	26.5(16.6)
Latvia	3.3	12.8	0	1.4	1.1	18.7(11.1)
ALL	0.7	8.0	5.0	8.6	4.9	27.3
	F=8.84[1]	F=31.74	F=39.39	F=9.11	F=23.19	F=13.33
	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001

* the ward meets the criteria for organised stroke care. [1] The F statistics are calculated based on an OLS regression of centre on LOS with n=1,298 and 13 parameters there are 1,285 degrees of freedom.

The final measures of resource use are the mean number of CT scans (Table 6.6) and the proportion of patients having a carotid doppler. Again these show wide variation between the study centres ($p<0.001$).

These resource use measures suggest that there were differences in the way stroke care was produced across the centres. The patient and contextual factors that may be associated with these resource use differences are described in the next section.

Table 6.6: Use of CT scans and carotid doppler investigations

Centre	Mean (SD) CT scans	% having carotid Doppler
Portugal	1.23(0.57)	40
Spain	1.06(0.35)	6
Italy	1.31(0.73)	60
France	1.43(0.68)	82
Denmark	1.17(1.05)	49
Finland 1	1.16(0.60)	0
Finland 2	1.36(0.68)	0
Finland 3	1.18(0.50)	0
UK	1.07(0.25)	14
Poland	1.31(1.10)	100
Lithuania 1	0.97(0.63)	48
Lithuania 2	0.19(0.54)	5
Latvia	1.24(0.52)	48
ALL	1.08(0.82)	40
	F=28.76	$\chi^2(12)=496$
	P<0.001	P<0.001

6.42: Description of patient factors potentially associated with resource use and cost variation

The patient factors measured in this dataset that are potentially associated with resource use and cost are described below.

Patient characteristics

The characteristics of the patients included in the study are presented in Table 6.7. There were differences across the centres in the mean age of the patients, the proportion of patients living alone at home before they had the stroke, and in the proportion of patients who were independent before the stroke.

Table 6.7: Age and Pre-stroke status

Centre	N	Age at stroke Mean (sd)	Residence pre-stroke Home alone (%)	Home with partner (%)	Nurs home (%)	Indepen pre-stroke (%)
Portugal	73	64.5(11.5)	10	88	4	99
Spain	32	72.3(8.4)	28	72	0	88
Italy	109	74.2(11.0)	9	87	4	95
France	105	71.0(17.6)	34	61	5	95
Denmark	246	67.6(13.9)	48	51	0	79
Finland 1	37	72.9(9.4)	38	62	0	91
Finland 2	81	70.1(11.6)	42	58	0	70
Finland 3	57	80.8(7.3)	77	14	9	95
UK	73	71.3(11.8)	37	53	10	81
Poland	92	69.6(11.6)	24	75	1	97
Lithuania 1	62	66.9(10.5)	6	94	0	98
Lithuania 2	186	69.9(11.0)	11	89	1	99
Latvia	145	62.7(10.0))	11	88	1	97
ALL	1,298	69.3(12.7)	11	88	1	91
		F=158 p<0.001	χ^2=283, p<0.001			χ^2=125 p<0.001

Independ: Independent, Nurs: nursing

Stroke severity

The severity of the stroke varied across the centres (p<0.001 for all variables), with the proportion of patients who were incontinent at admission ranging from 12% for the centre in Latvia compared to 42% for the centre in Finland 3 centre (Table 6.8). Stroke is a heterogeneous condition encompassing both cerebral infarctions and intra-cerebral haemorrhages that differ in the way they are managed (Wade 1994). The distinction between different subtypes of stroke is usually made by CT scan. In one of the centres (Lithuania 2) where the CT scan rate was very low, the majority of strokes were therefore unclassified. In the remaining centres where most patients had a CT scan, the proportion of cases with cerebral infarction ranged from 68% to 97%.

Table 6.8: Measures of stroke severity (n=1298)

Centre	Incontinent[1] (%)	Paralysed[2] (%)	Stroke subtype		
			Infarction (%)	Haemorrhage (%)	Unknown (%)
Portugal	25	84	75	15	10
Spain	28	84	81	16	3
Italy	40	81	78	18	4
France	18	69	96	3	1
Denmark	33	70	75	11	14
Finland 1	19	65	97	3	0
Finland 2	35	74	86	12	1
Finland 3	42	60	89	11	0
UK	26	67	89	11	0
Poland	20	95	91	8	1
Lithuania 1	16	68	68	15	18
Lithuania 2	25	89	9	5	87
Latvia	12	82	81	16	3
ALL	26	77	72	11	18
	$\chi^2(12)=52$ p<0.001	$\chi^2(12)=68$ p<0.001		$\chi^2(12)=775$ p<0.001	

[1]During the first week after stroke, [2]At hospital admission.

6.43 Description of contextual factors potentially associated with resource use and cost variation

Details were recorded on differences in access to care, the way stroke care was produced in each centre, incentives to reduce hospital costs and proxies for national levels of health care infrastructure. Each of these contextual factors can be used to offer insights into why there are resource use variations across the centres. While these factors are all grouped under the general heading of contextual factors this is simply to define these factors as operating at the level of the health care centre or country concerned, rather than at a patient-level. The definition of contextual or higher-level factors therefore refers to the level at which these factors were assumed to *operate.* In this dataset, some of these factors such as time from stroke onset to admission were *measured* at a patient-level whilst others such as the method of hospital reimbursement or level of health infrastructure were *measured* at a centre or national-level.

Access to hospital care

One hypothesis from the literature review was that those with better access to care might consume higher levels of health care. In this study all the patients included are admitted to hospital, so by definition, they have been able to access hospital care. However, barriers to care may exist in some of the centres, and there may be a delay to hospital admission. The time between onset of stroke and admission varied across the health care settings (Table 6.9). For example, 14% of the stroke patients in the Polish centre arrived at the acute hospital at least seven days after the stroke. By contrast, the majority of the patients in the centres in Spain, Italy, Finland 2, and Lithuania were admitted to hospital within 6 hours post-stroke.

Table 6.9: Time from stroke onset to admission (n=1,298): % in each category

Centre	<6 hours	6-24 hours	1-7 days	> 7 days	unknown
Portugal	44	40	16	0	0
Spain	56	16	16	6	6
Italy	57	28	14	1	1
France	42	37	17	2	2
Denmark	34	21	19	5	21
Finland 1	46	19	8	0	27
Finland 2	58	23	16	0	2
Finland 3	46	30	16	0	9
UK	45	16	27	0	11
Poland	45	29	12	14	0
Lithuania 1	52	29	13	2	5
Lithuania 2	45	31	19	4	1
Latvia	30	27	29	10	4
ALL[43]	43	27	18	4	7
			$\chi^2(12)=241, p<0.001$		

Access to neurologist and utilisation of rehabilitation hospitals

Almost all of the patients in the Eastern European centres had access to neurological care (Table 6.10). The centres in Italy, France, Denmark, and Finland generally had access to both neurologists and rehabilitation hospitals (Table 6.10). For patients in the UK, Spain and Portugal access to neurologists was either low or non-existent, and in these centres there was no use of separate rehabilitation hospitals. In the UK centre, inpatient rehabilitation was delivered in the acute hospital, whereas in the centres in Spain and Portugal, inpatient rehabilitation was not available for the majority of the

[43] Owing to rounding the numbers in this row do not sum to 100%.

patients. This provides a clear example of stroke care being produced in different ways across centres within Western Europe, and may offer helpful insights into why resource use differs across the Western European centres.

Table 6.10: Access to neurologists and inpatient use of rehabilitation hospitals (n= 1,298)

Centre	% access to neurologists	% transferred to rehabilitation hospital[1]
Portugal	32	0
Spain	0	0
Italy	68	21
France	93	28
Denmark	76	27
Finland 1	100	30
Finland 2	91	48
Finland 3	23	30
UK	0	0
Poland	92	4
Lithuania 1	100	44
Lithuania 2	91	29
Latvia	94	3
ALL	75 $\chi^2(12)=615$ p<0.001	20 $\chi^2(12)=175$ p<0.001

[1] In Poland, Latvia, Spain there were no inpatient rehabilitation hospitals in the local area, and any use of rehabilitation hospitals came from patients being transferred to hospitals in other regions

Alternatives to hospital care

The table below (Table 6.11) shows that on average 19% of patients used community support services (home help, home carer, district nurse) after hospital discharge, and 35% of patients had support or help from their family to assist with activities of daily living. Again there was important variability across the centres, a high proportion of patients in the Polish centre had family support (71%) compared to the UK and Danish centres, where none of the patients had family support in the three months following the stroke.

Table 6.11: Alternatives to hospital care (n=1,298)

Centre	% community support	% family support
Portugal	10	30
Spain	38	41
Italy	5	50
France	37	35
Denmark	27	0
Finland 1	30	43
Finland 2	14	35
Finland 3	37	55
UK	21	0
Poland	21	71
Lithuania 1	3	48
Lithuania 2	14	36
Latvia	6	62
ALL	19	35
	$\chi^2(12)=105$	$\chi^2(12)=298$
	$p<0.001$	$p<0.001$

Reimbursement of inpatient hospital services, bed-occupancy and patient copayments

Most of the hospitals in the study (Table 6.12) received a global budget to provide all inpatient and outpatient services. This may limit the level of stroke care that can be supplied. The extent to which this provides an incentive to minimise costs, depends on many factors including whether the demand for hospital services, for example for inpatient stroke care, exceeds the supply. Under global budgets if demand for inpatient care does exceed supply then there may be a greater incentive to discharge patients from hospital than when there is excess capacity. The table below (Table 6.12) provides some indication of the spare capacity on the ward where most stroke patients in each hospital are managed, by reporting the bed-occupancy. Those centres that have global budgets and low bed-occupancies for example, Finland 1 and 2, and the UK centre may have weak incentives to discharge the patients from hospital.

Diagnosis related groups (DRGs) are a prospective payment system whereby hospitals are reimbursed in advance for the cost of treating patients with a particular diagnosis. Hospitals have an incentive to minimise costs, as if the costs are lower than the DRG reimbursement rate, the savings are kept by the hospital. DRGs may therefore provide a more direct incentive than global budgets for reducing LOS. However, unless DRGs

cover the entire range of disease costs, they may lead to costs being shifted to a different provider. For example, in the French centre in this study, the DRG system does not cover the costs of the rehabilitation hospital, so cost shifting may occur. The incentive to minimise the total costs of hospitalisation may not be as strong as for the other centres reimbursed by DRGs, where the DRG covers the majority of stroke care costs.

Table 6.12: Centre factors: patient copayments, reimbursement of hospital services and bed-occupancy on main ward managing stroke patients in each centre

Centre	Patient copayments? (acute care)	Reimbursement of hospital services	Bed occupancy[1]	No. beds on main ward for stroke patients
Portugal	No	DRG	0.94	28
Spain	No	Global budget	0.81	31
Italy	No	DRG	0.90	4
France	Yes	DRG	0.80	25
Denmark	No	Global budget	0.94	15
Finland 1	No	Global budget	0.75	23
Finland 2	No	Global budget	0.75	28
Finland 3	No	Global budget	0.99	28
UK	No	Global budget	0.80	12
Poland	No	Global budget	0.81	16
Lithuania 1	No	DRG	0.92	40
Lithuania 2	No	DRG	0.87	80
Latvia	Yes	DRG	0.91	78

[1] Calculated over the financial year 1997-1998, by dividing the total number of occupied bed-days by the total number of available bed-days.

Presence of copayments

In the centres in Latvia and France, patients were asked to make a small contribution towards the cost of each day in hospital to cover for example, food costs. The presence of these payments may therefore provide a small incentive for the patients to seek earlier discharge from hospital.

Proxy measure for the national level of health care infrastructure

The literature review suggested that the level of health care infrastructure in the country concerned, could be important in determining the resources used at a local level. Various alternative measures could be used to proxy the level of health care

infrastructure available. Table 6.13 presents two such measures: the GDP per capita, and the proportion of GDP spent on public health care. As the data show GDP per capita varies seven fold across the countries with centres included in this study. GDP per capita is not used as the measure of health care infrastructure in the subsequent analysis of resource use variation, as it is important to include price as an independent variable, and price is strongly correlated with GDP. So using GDP per capita as a measure of national health care infrastructure alongside measures of factor price could lead to problems of multicollinearity. Instead an alternative measure is used, the proportion of GDP spent on public health care. This measure is less likely to be correlated with factor prices. The percentage of GDP spent on health care also varied across the centres included in the study, with those centres in Eastern Europe generally spending a lower proportion of their national income on health care. As health care is a normal good those countries with higher GDP per capita generally tend to have higher levels of health care spending, this may lead to higher resource use and unit costs in these countries.

Another alternative is to use the level of health care infrastructure that is available locally in each health care centre. This could be proxied by for example, the budget of each acute hospital. However, the budget facing the individual hospital would be directly correlated with both the resource use measure (the outcome variable) and the price variable in the subsequent regression equations analysing resource use variation.

While the % of GDP spent on health care was chosen as the measure of health care infrastructure, the choice of this variable had implications for the selection of other variables for the regression analysis. In particular this variable was correlated with measures of the production processes in each centre such as access to neurologists or use of rehabilitation hospitals. Thus, when specifying the relationship between the various factors and resource use (Chapter 8) it will be necessary to choose carefully the variables for inclusion in the regression equations.

Table 6.13: Measures of national health infrastructure for each centre in the study

Centre	National GDP/head (1997/$PPP)	National Public Health expenditure % share of GDP
Portugal	13,840	5.2
Spain	15,720	5.4
Italy	20,060	5.7
France	21,860	7.3
Denmark	22,740	6.8
Finland 1	18,980	5.3
Finland 2	18,980	5.3
Finland 3	18,980	5.3
UK	20,520	5.6
Poland	6,380	4.7
Lithuania 1	4,510	5.1
Lithuania 2	4,510	5.1
Latvia	3,650	3.9

Sources: OECD (2001), World Bank (2000) Cerniauskoas and Murauskiene (2002), Karaskevica and Tragakes (2001), Karski et al. (2002)

6.5 Hypotheses generated about factors associated with resource use variation

The data described in the previous section are now used to generate hypotheses about factors associated with resource use variation. OLS regression models are used with LOS as the dependent variable. The first model includes the patient factors described as independent variables, and describes their association with LOS. The second model includes dummy variables for each centre, apart from the Latvian centre which was taken as the reference centre. The third model considers whether there were still variations in LOS across the centres after adjusting for patient factors by including patient factors alongside dummy variables for each centre.

Table 6.14 presents the regression coefficients from the three models. The data described by model one show that for example, patients who were independent pre-stroke stayed in hospital for on average six days less than patients who were dependent pre-stroke. Patients who were incontinent stayed in hospital for on average 19 more days than patients who were continent pre-stroke. Patients for whom the

delay between onset of stroke and hospital admission was unknown, had a LOS that was on average 10 days longer than those who were admitted within 6 hours.

The results from model 2 suggest that there are important variations in the mean LOS across the study centres ranging from the Spanish centre, where patients had a LOS on average 10 days shorter than in the Latvian centre, to the UK centre where the mean LOS was 21 days longer than in the Latvian centre.

The estimates from model 3 show that including patient factors improved the fit of the model, the adjusted R^2 increased from 0.10 in model 2 to 0.45 (likelihood ratio test, $p<0.001$). This suggests that including patient factors is important in explaining the overall variation in LOS. Allowing for differences in patient factors across the centres, also has a small impact on the relative LOS across the centres. Those centres with more complex than average case-mix had smaller coefficients after adjusting for patient factors. For example, the Finland 3 centre originally had the second highest LOS (model 2) relative to the Latvian centre, but as this centre had more complex case-mix, adjusting for patient factors meant that this centre's LOS after adjustment (model 3) fell compared to other centres. Similarly, the Italian centre had shorter relative LOS after adjusting for patient factors reflecting the relatively complex case-mix in this centre. In general though, while including patient factors reduced the variability in LOS left unexplained by the model, there was still wide variability in LOS across the centres concerned.

The descriptive information previously provided on the way stroke care is produced in the different centres offers some insights into why there may be outstanding variability across the health care centres, after adjusting for differences in patient factors. Those centres that had higher LOS after adjusting for patient factors, were generally the centres with low levels of family or community support (e.g. UK and Denmark), and high levels of input available from neurologists and rehabilitation hospitals (Denmark, Finland 2 and 3). Those centres with low LOS were generally those where rehabilitation hospitals (Spain, Portugal, Latvia), and neurologists (Spain and Portugal) were not available, but there were relatively high levels of community (Spain) and family support (Latvia). Some of the individual centre rankings may not

follow from consideration of the factors listed. For example, the centres in Finland 1 and Italy had low LOS despite having high levels of availability of neurologists and rehabilitation hospitals. Despite those exceptions, a general hypothesis to emerge from this data description is that the level of resource use variation between the centres may partly reflect differences in patient characteristics, but also differences in production processes and resource availability across the centres. Other variables such as the level of family support, the level of relative factor prices, the presence of patient copayments or a DRG system may also be important determinants of resource use. The literature review also suggested that these factors may be associated with resource use. These hypotheses are tested in subsequent chapters.

Table 6.14: Description of different factors association with LOS, coefficients (standard errors).

	Model one: Patient characteristics	Model two: Centre dummy variables	Model three: Centre dummys patient factors
Constant term	22.35 (12.6)**	18.70(2.01)**	6.23(13.21)**
Patient factors			
Age	-0.13 (0.05)**		-0.07(0.05)
Independ. pre-stroke	-5.97 (2.37)**		-3.03(2.32)
Living alone	Reference		
Living with others	-7.71 (1.50)**		-4.84(1.56)**
Living in nursing h	-9.85 (4.60)**		-8.50(4.44)**
Incontinent	19.33 (1.56)**		18.77(1.49)**
Paralysed	7.03 (1.55)**		8.59(1.51)**
Ischaemic stroke	Reference		Reference
Haemorrhagic stroke	5.37 (2.13)**		5.82(2.04)**
Unknown stroke type	-1.99 (1.72)		-8.08(2.51)**
Ons to admiss < 6 hr	Reference		Reference
Ons to admiss 6-24 hrs	-0.57 (1.58)		0.64(1.52)
Ons to admiss 1-7 days	-0.93 (1.80)		-0.71(1.73)
Ons to admiss >7 days	-1.12 (3.28)		-0.77(3.17)
Ons to admiss unk.	10.35 (2.64)**		7.98(2.60)**
Centres			
UK		20.59(3.48)**	20.46(3.36)**
Finland 3		18.68(3.79)**	14.38(3.14)**
Denmark		18.38(2.53)**	16.23(2.64)**
Finland 2		17.85(3.36)**	14.06(3.71)**
Lithuania 2		7.83(2.68)**	14.21(3.20)**
Lithuania 1		9.30(3.68)**	11.53(3.35)**
Poland		6.81(3.23)**	5.36(2.95)
France		5.52(3.11)	8.19(2.94)**
Italy		1.47(3.07)	-2.55(2.87)
Finland 1		0.35(4.46)	0.29(4.15)
Latvia		Reference	Reference
Portugal		-5.67(3.48)	-6.11(3.19)
Spain		-10.82 (4.73)	-12.28(4.33)**
F statistic	27.68	13.33	21.6
P value	<0.001	<0.001	<0.001
Adjusted R^2	0.20	0.10	0.45

** p<0.05, *p<0.10, Independ: independent, h: home, hrs: hours, unk: unknown

6.6 Discussion

This chapter described the data collected in the empirical investigation on resource use and factors potentially associated with resource use variation in 13 different European health care settings. These data can be used in conjunction with the findings from the literature review, to pose hypotheses for why costs vary across health care settings (see Table 6.15). In general, for the factors operating at a patient level the hypotheses posed by the data were similar to those suggested by the literature. However, the data provided some specific hypotheses not suggested by the literature review. For example, the literature review suggested that older stroke patients were more dependant and therefore likely to consume more resources[44]. The data in this study suggest that those older patients who survive the stroke may use fewer resources than younger patients, although the effect was small and non-significant.

The data described illustrate the variability across the centres in the patients' resource use. The regression models highlight variability in LOS across the study centres, after adjusting for differences in measured patient factors. This unexplained variability could reflect differences in the way these centres produce stroke care. In particular, there are differences amongst the centres in the levels of input from community support services and patients' carers (Table 6.15). The literature review also suggested that certain ways of managing stroke patients, in particular, neurologist led stroke care, were associated with higher resource use. These literature-based hypotheses were then used in conjunction with the differences described in the data, to pose hypotheses as to why LOS varied across the centres. For example, the reason why certain centres had higher resource use may be that their patients had better access to neurologists or to rehabilitation hospitals (Table 6.15). The general health economics literature suggested that the way hospitals are reimbursed, and level of patient copayments may also be associated with resource use (Chapter 4). The data described showed that these factors differed across the centres included in the study, and so the subsequent analyses consider these contextual factors when analysing resource use and cost variation (see Chapter 8).

---

[44] There is also a growing literature on the effect of proximity to death on resource use (see for example Seshamani and Gray (2004). For these data the hypotheses posed by this literature are not considered as all the patients chosen for inclusion survived up to three months post stroke.

The literature review also suggested that the national level of health care spending could act as a proxy for the level of health care infrastructure in the country concerned. A hypothesis from the data is that those centres in countries with high levels of health care spending have better access to specialised personnel and use more advanced technology which leads to higher levels of resource use in these centres.

The objective of this thesis is to examine why resource use and costs vary across international health care settings. The literature review suggested that when addressing this issue certain measurement issues arise (Table 6.15). To tackle these issues an empirical investigation has to use a database with certain features. In particular, sufficient patients and centres are required to identify reasons for resource use and cost variation. This database was judged to have sufficient patients and centres to allow for systematic reasons for variation to be identified at different levels. Resource use had to be collected in several health care settings using a consistent disaggregated methodology. Information on factors potentially associated with resource use and unit cost variation had to be available. In particular, suitable information on case-mix had to be collected at the level of the patient. The Biomed II stroke study had many of the features required for thoroughly examining the reasons for cost variation and was therefore judged an appropriate dataset for addressing the thesis' question. There are nevertheless, weaknesses in the dataset that should be recognised when interpreting the subsequent results.

In this dataset approximately there were concerns about the potentially endogenous relationship between cost (or LOS) and survival. To address this an instrumental variable approach was considered, but this was rejected because it was not possible to identify an appropriate set of variables to use as the 'instrument'. Instead, the main analysis in the thesis is limited to those patients surviving the stroke. This limits the generalisability of the resource use and cost analysis to those patients surviving up to three months post-stroke.

In addition, information was not collected on health-related quality of life. Hence, it may not be realistic to assume that those centres that use fewer resources are producing stroke care in a more technically efficient way- indeed in some of the

centres using more resources those resources are being deployed to produce organised stroke care, which has been shown to improve outcomes.

A societal perspective was not taken to cost measurement, it is therefore possible that those centres making less use of health and community service resources were using more inputs from patients and their carers. Some information was collected though on whether the patients' family helped the patient with activities of daily living, after their hospital discharge. Using these data, it was possible to compare at least family inputs across the centres, to see whether in some centres they seemed to be a substitute for hospital care.

The Biomed II stroke study only included hospital-admitted stroke patients, and so excluded those patients managed in the community. There are differences in hospital admission rates following stroke across different European centres (Ricci et al. 1991; Tuomilehto et al. 1992), and this may partly explain why there are wide variations in the characteristics of the patients included in each centre. The implication for the methodological focus of this thesis-- to examine the reasons for cost variation across settings-- is that differences in patient characteristics across settings should be carefully considered, when attempting to understand reasons for cost variation.

The resource use data were collected for 90 days post stroke. While the Biomed II stroke study originally intended to collect data on resource use, costs and outcomes for up to one year post stroke, this was only feasible for two of the study centres (Grieve et al. 2000). To assess resource use and cost variability across 13 centres, the data at three months post stroke were used. This underestimates the overall costs of stroke care especially in those centres where a sizeable minority of patients are still in hospital at 90 days post stroke. The data described in this chapter show that in the centres in UK, Denmark and Finland (2 and 3) a sizeable minority of patients were still in hospital at 90 days post stroke. As these centres were those with the highest overall LOS, by censoring the data at 90 days it is most likely that cost differences across the centres will be underestimated.

The numbers of patients (n=1,298) and centres (n=13) included in the study will allow the relative importance of resource use and cost variation at either a patient or centre

level to be examined. However, there are insufficient centres in each country to allow cost variation across centres within a country to be explored. Only in Finland (three centres) and Lithuania (two centres) were there more than one centre per country. There was some variation amongst the Finnish centres, for example the LOS in the Finland 1 centre was approximately half that in the other two centres. However, for a thorough examination of reasons for resource use and cost variation within a country a dataset is required with many more centres and observations within a particular country.

The data described also showed that within each centre, patients took various pathways through the system, varying in their use of hospitals within each centre, and wards within each hospital. It is recognised though, that there are insufficient data to model these different hierarchies within each centre when formally assessing the reasons for resource use and cost variation. The numbers of patients, centres, and countries covered by the database means that the subsequent analysis defines the hierarchy at two levels: patients clustered within centres.

The literature review suggested that differences in technical, productive and scale efficiency across health care settings may be important in explaining cost variation. Previous studies have used data from many provider units to estimate efficiency levels across provider units (see Chapter 4). It is recognised that in this study there are insufficient centres to allow differences in technical and productive efficiency to be estimated.

6.7 Conclusions

In conclusion, the findings from the literature review provided the conceptual framework for the empirical investigation. The case study chosen for identifying reasons for cost variation across settings, was the Biomed II stroke study. It was possible to extend this study to consider reasons for cost variation across settings. In this chapter, the methodology and the resource use data for the empirical investigation are presented. These data show that there are wide variations in resource use following stroke, for example in hospital LOS, across the study centres. The

information collected on patient and contextual factors are used to generate hypotheses for why resource use varies across health care settings. The next chapter considers the measurement and analysis of unit costs and total costs.

Table 6.15: Summary of factors likely to be associated with resource use (RU)

Factors	Hypotheses from literature	Hypotheses from data	Level variable measured at
1. Production, cost function literature			
Care alternatives			
Use of community support	Higher use of community support: shorter hospital stay	Centres with higher use of community support, shorter hospital stay	Patient
Use of family support	Higher use of family support: shorter hospital stay	Centres with higher use of family support: shorter hospital stay	Patient
Technical efficiency	Variations in TE: variations in costs	Bed-occupancy rates differ across centres: differences in TE	Not measured
Scale efficiency	Economies of scale in hospital production realised up to 300 beds	Provider units of similar size no hypotheses generated	Not measured
Factor prices	Differences in factor prices: differences in factor use	See next chapter	See next chapter
2. Patient factors			
Age at stroke	No clear hypothesis	Older: lower RU	patient
Independence pre-stroke	Independent lower RU	Independent: lower RU	patient
Residence pre-stroke	Home alone higher RU	Home alone: higher RU	patient
Incontinence	Incontinent higher RU	Incontinent: higher RU	patient
Paralysis	Paralysed higher RU	Paralysed: higher RU	patient
Stroke sub-type	None	Haemorrhagic stroke: higher RU	patient
Socioeconomic status	No clear hypothesis	No hypothesis generated	Not measured

3. Centre or national-level factors			
Hospital reimbursement by DRG system	Presence of DRG system: lower RU	Centres with DRG system lower hospital RU	Centre
Patient copayments	Patient copayments: lower RU	Centres with patient copayments: lower RU	Centre
Level of health care infrastructure	Higher spending on health care: more RU	Centres with higher spending on health care: more RU	national
Access to neurologists	Higher proportion using neurologists: higher RU	Centres with higher access to neurologists: higher RU	patient
Utilisation of rehabilitation hospital	Better health care infrastructure: more use of rehabilitation hospitals	Centres with higher proportion using rehabilitation hospital: higher RU.	patient
4. Measurement issues			
Case-mix	Inefficiency differences observed in aggregated studies could reflect case-mix	Case-mix collected at disaggregated level anticipated to capture differences	patient
Random variation	Random variation potentially important at patient or setting level	Wide variation within and across centres needs to be considered	Patient and centre
Quality of inputs and outputs	Aggregated measures fail to adjust for differences in quality of inputs or outputs	Centres with higher proportion using rehabilitation hospital: higher RU	Centre
Variability in costing methods	Disaggregated methods required to enhance methodological comparability	RU data collected using consistent disaggregated method	Patient and centre
Choice of conversion factor	Choice of CF may have implications for cost variation observed	See next chapter	NA

RU: Resource use, TE: technical efficiency, CF: conversion factor

Chapter 7: Measurement and analysis of unit cost differences across the centres

7.0 Introduction

The previous chapter described the methodology used to collect resource use data and provided an overview of the data collected for the thesis. The literature review highlighted that to understand variation in unit costs amongst settings it is important to disaggregate unit costs into resource inputs and factor prices (Chapter 2). The review suggested that differences in relative factor prices across countries might explain variation in the combination of factor inputs (e.g. relative levels of input from doctors or nurses) and the resources used (e.g. LOS) to produce health care (Chapter 4). The empirical investigation therefore takes a disaggregated approach to unit costing. This chapter presents the methodologies used to measure unit costs. Price and volume indices are constructed to explore reasons for differences in unit costs across the study centres, and to consider the implications of these variations for differences in total costs per patient. The total costs per patient reported for each centre may also depend on the method used to convert local currencies into a common currency. This chapter highlights the methodological advantages and problems associated with different currency conversion factors.

This chapter is split into six sections. The first section summarises the methodological issues involved in constructing price and volume indices, the second section discusses the suitability of different currency conversion factors, and the third section describes the methodology used to collect and calculate unit costs and total costs. The fourth section details the methods used to construct price and volume indices and a 'technology specific' currency conversion factor. The fifth section presents the results: describing the unit costs, the price and volume indices, and the total costs. The final section discusses why the differences in price levels across the centres concerned

may exist and considers the implications for understanding cost variation across health care settings.

7.1 Volume ratios, price and volume indices

Volume ratios and price and volume indices can all be used to explore why differences in unit costs may be observed across health care settings.

7.11 Volume ratios

A starting point for considering why unit costs may differ across health care settings is to compare the ratio of factor inputs used to produce a given level of output (Barnum and Kutzin 1993). For example, Hutton (2001) compares the number of antenatal care visits per health care professional across health care centres, to try and identify those centres that are more technically efficient. However, this static approach does not consider that firms in different settings, particularly across countries, may face different factor prices and this may be reflected in the chosen factor mix.

7.12 Price indices

Price indices have been widely used to compare prices across different time periods (see for example Boskin et al 1986), but their use in comparing prices across different settings is less well developed. This chapter examines the use of price bilateral price indices for comparing factor prices across countries. These indices compare the prices in each country to those in a reference country. A key issue to consider when developing a price index is how to weight the prices of the different goods or factor inputs (Bernt 2000). Studies in health economics that have used price indices have been criticised for ignoring the issue of weighting, and simply summing up the relative price levels of different good and services: i.e. implicitly giving each component equal weighting (US General Accounting Office 1992). Danzon and Chao (2000) used price indices to compare the price of pharmaceuticals across countries and found that previous conclusions that drugs' prices in the US were higher than

elsewhere, were invalid as they were based on biased and unweighted indices. Their work illustrates the importance of using appropriate volume weights and taking a representative sample of product prices for the sector concerned (Danzon and Chao 2000, Danzon and Kim 1998).

Laspeyres and Paasche price indices both use volumes to weight the price of goods or factor inputs (Call and Holahan 1983). A Laspeyres price index (L_{Price}) can compare the factor prices in a comparator country to those in a reference country. Each set of factor prices is weighted by the volume of that factor input used in the reference country. A Laspeyres price index is therefore defined as in equation 1:

$$L_{price} = \frac{\sum_{i=1}^{m} P_2 q_1}{\sum_{i=1}^{m} p_1 q_1} \quad (1)$$

Where p_1 is the price in the reference country, p_2 the price in the comparator country, q_1 the volumes in the reference country, q_2 the volumes in the comparator country, all for factor inputs i....m.

A problem with the Laspeyres price index is that it ignores any substitution effects; the comparator country is assumed to use the same level of factor inputs as the reference country irrespective of any differences in relative factor prices. Instead, a Paasche index (P_{Price}) can be used, which takes the volume of factor inputs used in the comparator country (q_2) to weight the factor prices in both the reference and comparator countries (equation 2).

$$P_{price} = \frac{\sum_{i=1}^{m} P_2 q_2}{\sum_{i=1}^{m} p_1 q_2} \quad (2)$$

For cross-country comparisons, the use of the Paasche index assumes that faced with the factor prices in the comparator country, the reference country would choose the comparator country's volume of inputs, that is the factor mix chosen would adjust to reflect the relative factor price. In reality, variables other than price may determine the

factor mix (see Chapter 4), so that even if the price of a factor input was relatively high in country A compared to country B, this would not necessarily lead to less use of the factor input.

7.13 The Gerschenkron effect and the Fisher Price Index

As Laspeyres and Paasche price indices use different volumes to weight price differences across countries, they may produce different results. The discrepancy between Laspeyres and Paasche price indices has been termed the Gerschenkron effect (Gerschenkron 1951), and is often summarised by the ratio of the Paasche to the Laspeyres price indices (Van Ark et al. 1999). Economic theory suggests that unless the elasticity of substitution is perfectly inelastic there will be correlation between different factor prices and different factor mixes, and that the Paasche/ Laspeyres ratio will therefore be less than one. This hypothesis is examined for each centre included in the empirical investigation by initially constructing both Laspeyres and Paasche indices and the Paasche/Laspeyres ratio.

A common approach to price index construction is to regard estimates from the Laspeyres and Paasche indices as boundaries for the ideal index (Hill 1999). The Fisher index is then constructed by taking the geometric mean of the Laspeyres and Paasche price indices (Fisher 1922). Apart from taking the middle ground between these two extreme positions, the Fisher price index satisfies certain ideal properties defined by the economic theory of index numbers (Diewert 1999). In light of its desirable properties, the Fisher price index has been recommended for comparing prices between countries (Diewert 1999), and is used in this investigation to compare the relative prices across the centres concerned. The Fisher price indices are also used in the subsequent empirical investigation that assesses the effects of differences in factor price on resource use across the centres (Chapter eight).

7.14 Volume indices

The same indices can be used to compare the volume of factor inputs across different countries, weighting relative volumes by factor prices. So a Laspeyres volume index

(L_{volume}) uses factor prices in the reference country as weights (equation 3), whereas the Paasche index (P_{volume}) uses the comparator country's factor prices as weights (equation 4). The Fisher volume index takes the geometric mean of the two.

$$L_{volume}=\frac{\sum_{i=1}^{m} p_1 q_2}{\sum_{i=1}^{m} p_1 q_1} \qquad (3) \qquad P_{volume}=\frac{\sum_{i=1}^{m} p_2 q_2}{\sum_{i=1}^{m} p_2 q_1} \qquad (4)$$

7.2 Conversion of factor prices to a common currency

To compare costs across countries it is necessary to convert local currencies into a common currency (e.g. US dollars). A potentially important question this raises is: which method of currency conversion is most appropriate? One alternative would be to convert local currencies into US dollars using official exchange rates (OER). However, the problem with using OER is that they are only based on the exchange of traded goods and do not reflect the relative price of non-traded goods such as health care. OER do not reflect differences in overall purchasing power between countries (Kanavos and Mossialos 1999). They may partly reflect macroeconomic deficits or surpluses in the economies concerned. Currency speculation may also lead to a dramatic increase or decrease in a country's OERs.

Purchasing Power Parity (PPP) indices have been developed to try to overcome these problems. The purpose of using a PPP index is to try to eliminate price differences between countries, and to permit international comparisons of the volume of services consumed or produced across countries. For example, cross-national comparisons of output are often conducted using PPPs to convert local measures of Gross Domestic Product (GDP), into a common currency, eliminating differences in price levels between the countries concerned. These GDP PPPs are calculated using prices of a 'basket' of goods and services, chosen to represent overall consumption or production in the countries concerned. The price of this basket of goods is compared between the comparison country, say the UK, and the reference country (say the United States). So for example, if the basket of goods is worth $1 in the United States, the PPP index calculates the price in local currency of buying the same basket of goods and services

in the UK (say £0.65). The ratio of these prices 0.65/1=0.65 ($ GDP PPP) provides a measure of purchasing power in the UK compared to the US. This conversion factor can then be used to translate measures of output across the countries concerned adjusting for general cross-national differences in the prices of goods and services. In the empirical investigation for this thesis it is necessary to adjust for differences in factor prices. However, certain issues arise surrounding the choice of conversion factor, these are discussed below.

7.21 Issues in the choice of currency conversion factor.

Commonly used measures of PPP are the GDP PPPs (Schreyer and Koechlin 2002). These are based on a basket of goods and services that are representative of consumption across the whole economy. These measures of PPP have been used to compare national output and productivity across countries (OECD 1999). However, when comparing costs in a particular sector or industry across countries the GDP PPP index may be less appropriate; the basket of goods and services is not representative of consumption (or production) in a specific sector. In particular, the use of a general basket of goods and services may not be suitable in areas such as health care that are dominated by non-tradeable goods and services. Medical care PPPs have been constructed that attempt to adjust for differences in the relative prices of medical care across countries, and these were available for most of the countries covered by the empirical investigation[45]. However, medical care PPPs have been seriously criticised for their heavy bias towards pharmaceutical services (Danzon and Chao 2000, Kanavos and Mossialos 1999) that makes them unlikely to be representative of health services more generally. Danzon and Chao (2000) also pointed out that the pharmaceutical component of the medical care PPP is based on small, unrepresentative samples for each country, and does not use volume weights at the product level to reflect the relative importance of different products. Kanavos and Mossialos (1999) cautioned against the use of medical care PPPs for making international comparisons and multinational economic evaluations have usually relied on GDP PPP indices to report costs across different countries, despite their limitations (see for example Willke et al. 1998).

---

[45] Medical care PPPs were not available for Latvia and Lithuania.

Another alternative is to construct technology specific PPPs using data from the study. This approach has been taken by other studies making international cost comparisons (Wordsworth and Ludbrook 2005, Hutton 2001). This approach uses a 'basket' of factor inputs that are deployed in producing the output concerned. The factor prices for each of these inputs are then estimated for each country compared to those for a reference country. These price differentials are then weighted according to input volumes. The measure of volume used should not be that of either the reference or the comparison country, but ideally an average of the two (i.e. a Fisher price index). The resultant price index provides a 'technology specific' measure of purchasing power in the comparison country compared to the reference country. Using this specific measure of PPP to convert unit costs to a common currency provides an international cost comparison adjusting for differences in the price of the specific inputs used to produce the output concerned.

Any of these methods of currency conversion assumes that the local prices used for the basket of goods and services reflect opportunity costs. If some of the factor inputs in the 'basket' of goods are not traded in a perfectly competitive market, prices may not be available and even if they are, they may not reflect opportunity cost. This is a potential problem whichever currency conversion factor is used.

Subject to this caveat, GDP PPPs have the potential to represent general opportunity costs in the economy as a whole, as the 'basket' of goods and services is taken to be representative of consumption or production across the whole economy. Converting costs using GDP PPPs can therefore provide a measure of the opportunity costs in each economy after adjusting for general price differences between countries. Using GDP PPPs to convert costs into a common currency therefore appears to be consistent with taking a broad, societal perspective to economic evaluation that emphasises the measurement of costs for a broad range of providers with resource use valued using general measures of opportunity cost (see Chapter 2).

A more specific index, whether it is for medical or stroke care, can only provide a measure of purchasing power for the particular sector in question. Using this conversion factor to convert costs to a common currency is more likely to adjust for

cross-national price differences in the particular sector concerned, and isolate the effect of volume differences. However, the trade-off is that the costs observed will reflect a more narrow definition of opportunity cost, one that only considers the value foregone for the sector or technology in question.

The literature review highlighted the need to maintain consistency with economic theory when estimating costs across countries (Chapter 2), and so the main analysis in this thesis uses GDP PPPs to compare costs (Chapter 8). However, it is recognised that using GDP PPPs means that there may be outstanding cross-national differences in the price of those inputs used to produce stroke care. To see the impact this has on the results, a stroke care PPP is also calculated, based on a subsample of the factor inputs used to produce stroke services in each centre. To examine how sensitive the cost analysis is to the choice of conversion factor, the cost analysis (both in this chapter and chapter 8) is repeated with costs converted using the stroke specific PPP index.

7.22 Conversion factors for traded versus non-traded goods

A final issue is whether different conversion factors should be applied to traded and non-traded goods. In theory, if OER are used for traded goods then prices should be equalised across countries. It therefore seems appropriate to use OER to convert the costs of traded goods, and just use PPPs (whether for stroke care or GDP) to convert the costs of non-traded goods. Indeed this is the approach that has been recommended by the WHO (Tan-Torres Edejar et al. 2003). However, both the World Bank and the OECD include traded and non-traded goods in the basket of goods used to calculate GDP PPPs (World Bank 2000, OECD 2000). It is therefore inconsistent to use these GDP PPPs just to convert the prices of non-traded goods. The empirical investigation uses GDP PPPs for all factor inputs in the base case analysis. By contrast, the stroke care PPPs are just calculated for non-traded goods, so where the cost analysis uses this conversion factor, OER are used to convert the prices of traded factor inputs. For completeness, a final cost comparison uses GDP PPPs for non-traded inputs and OERs for traded inputs (for further details see section 7.57).

7.3 Methodology used in the empirical investigation for measuring unit costs

7.31 Overview

Unit costs were collected for each hospital and community service provider in each centre. A major decision regarding the costing methodology was whether to collect unit costs at an aggregated or disaggregated level. For example, the study could have collected unit costs at an aggregated level by collecting information on the average cost per bed-day or *per diem* cost, and multiplying this through by the length of stay to give the overall cost of hospitalisation. However, the literature review suggested that the ideal unit costs would be those that approximated opportunity cost in each centre and used a consistent methodology across the study centres. Using aggregated unit costs has two main problems. It relies on the methods of cost allocation used by the finance departments in each centre. Using this methodology would have meant cost differences across the centres relate simply to differences in the way unit costs are allocated. In addition, the estimation of opportunity costs requires that the cost chosen represents the value of the physical resource in its next best use (Palmer and Raftery 1999). Using an aggregated unit cost does not reveal the physical resource inputs used, and makes it very difficult to assess whether the unit cost does represent opportunity cost.

7.32 Unit costs of labour inputs

The majority of the costs of stroke care come from the labour inputs used during hospitalisation (Forbes and Dennis 1995) so particular care was taken to approximate the opportunity costs of these inputs. For labour inputs, the study collected disaggregated unit costs using a common methodology in each centre. The input from each grade of health care professional involved in producing stroke care in each centre was estimated. A semi-structured questionnaire was used on site visits to each centre, and members of staff at each hospital were asked about their level of input (direct and indirect) for each ward on which stroke patients were managed (see

appendix 1)[46]. The total annual staff time for each grade of doctor, nurse and therapist on each ward was recorded. To provide an average level of input per occupied bed-day the total annual input for each grade of staff was then divided by the total annual number of occupied bed-days for the ward concerned. This average measure of input (in minutes per occupied bed-day), provides a useful measure for comparing resource intensity across the centres. The measures of labour inputs are used in the construction of price and volume indices to help interpret unit cost differences across the centres (see section 7.41).

The average input for each grade of staff is also used to estimate the labour costs per day. These labour inputs would ideally be valued by a measure of factor price that approximates opportunity costs. A general difficulty with estimating opportunity costs is that they require the health care firm to choose the combination of factor inputs that minimises costs in the long-run. However, as the evidence in Chapter 4 revealed, health care firms do not tend to behave in a way consistent with long-run cost-minimisation. The health care firms may regard the costs of adjusting certain factor as too high or there may be a lack of incentive for decision-makers to take the long-run perspective. At best, health care firms' behaviour is consistent with short-run cost-minimisation. Hence, the factor inputs and factor prices observed may not be consistent with productive efficiency.

A further problem in these centres was that the labour markets for health care professionals were not perfectly competitive and prices were not available for labour inputs. To approximate the opportunity costs, the costs of labour input were derived by dividing the total annual cost to the hospital of employing the relevant mid-grade health professional[47], by the total number of hours worked, to give an average cost per hour[48]. Any additional payments to health care professionals from 'under the counter' payments from patients or private work were excluded. The average cost per occupied bed-day was calculated for each grade of health care professional by multiplying their factor input (in hours per occupied bed-day) by their factor price (cost per hour).

---

[46] Care was taken to try and ensure a consistent definition of each grade of staff across the centres based on levels of training and years of experience.
[47] The costs included the salary, overtime, and employers' National Insurance contributions.
[48] These measure of factor price are used in the construction of the price and volume indices (section 7.41).

The average labour cost per occupied bed-day was estimated by aggregating the average costs per occupied bed-day across all health care professionals.

7.33 Investigation costs

Apart from labour costs, the hospital costs included the costs of investigations, consumables, drugs and overheads. Information on the use of investigations was collected for each patient included in the database. The unit cost of each investigation was taken from the finance departments in each hospital and was established during interviews at each centre. These unit costs were based on the price charged to another public sector provider. This was felt to be the closest proxy to opportunity cost, as it did not include a profit to the organisation, and covered all appropriate cost components: staff time, reagent costs and overheads. During the interviews, it was important to establish that the same cost components were included in each centre.

7.34 Drugs, consumables and overheads

Individual patient data were not collected on the use of drugs, consumables or overheads. These are generally small cost items for stroke patients (Forbes and Dennis 1995). None of the centres included in the study provided high cost stroke drugs such as thrombolytics either routinely or as part of randomised controlled trials. The costs of drugs, consumables and overheads were therefore estimated as average costs per occupied bed-day. These costs per occupied bed-day were calculated by dividing the total annual inpatient expenditure for each category in each department by the total number of occupied bed-days. These costs were added to the staffing costs to give a total cost per occupied bed-day (or unit cost) for the relevant hospital wards in each centre.

7.35 Unit Costs post hospital discharge

To establish the unit costs associated with outpatient visits to clinicians and use of rehabilitation services, interviews were undertaken with a range of providers at each centre to again establish the staffing input used in providing care. These inputs were combined with information on the average cost of employing each member of staff to

give average staffing costs. Information on consumables, drugs and overheads was taken from the finance department in each centre and added to staffing costs to provide an average cost for each contact. For residential and nursing home care it was not possible to interview the providers concerned to estimate detailed costs. For these services, the charge (cost per day) to the National Health Service in each centre was used. Where there were several different service providers, the median cost of the item concerned was taken as the unit cost.

7.36 Calculating total costs per patient

For each patient total costs were calculated over the observation period of three months post-stroke. Investigation costs per patient were calculated by multiplying each patient's use of investigations by the unit cost. Other hospitalisation costs per patient (labour costs, consumables and overheads) were calculated by multiplying the length of stay on each ward by the appropriate cost per occupied bed-day, and summing these costs across each ward and hospital. These costs were added to the total cost of investigations, to give a total hospitalisation cost per patient. Total three months costs per patient were the sum of the hospitalisation, outpatient, community and institutional care costs. All costs were adjusted to a 1998 price base using price indices and converted into US dollars initially using GDP PPP indices (OECD 2000, World Bank 2000).

7.4 Constructing the price and volume indices and currency conversion factors

7.41 Price and volume indices

The inputs used for the construction of the price indices were those deployed on the ward where most bed-days were used by the stroke patients in this study. This provided a consistent way of comparing factor inputs and factor prices across the study centres. To construct price indices, disaggregated data were required on factor inputs and factor prices, and these were available for labour costs, overall the main component of costs per occupied bed-day (see Table 7.1).

Within labour costs, nine factor inputs were used to construct the index, doctors' time (chief doctor, senior level doctor, mid-grade doctor, junior doctor), nurses' time (chief nurse, qualified nurse, unqualified nurse), therapists time (physiotherapist, OT, speech therapist). The French centre was chosen as the reference centre, as this centre has the median level of unit costs. Also, for the French centre each of the factor inputs described was used to produce stroke care. For most of the other centres at least one factor input was not used, and for these centres, the overall index was based only on the factor inputs present. The prices of each factor input were taken as the cost per hour ($ GDP PPP) of employing each grade of staff (see section 7.32). The Laspeyres, Paasche and Fisher price and volume indices were calculated for each centre, compared to the French centre.

7.42 Constructing the stroke care PPP index for currency conversion and using this measure

The Laspeyres, Paasche and Fisher indices were constructed as above using the same labour inputs except that prices were kept in each centre's local currency. The reference centre was again the French centre, where the price of labour inputs was reported in French francs. The resulting price indices gave the number of units of local currency required to buy one francs worth of factor inputs in the reference centre- the stroke care PPP. This provided a measure of purchasing power for each centre compared to the French centre. To allow comparison with GDP PPPs which were reported in US dollars, each of the stroke care PPPs were converted from French francs into US dollars using 1998 OER (5 francs to 1$US) (OECD 2000).

7.5 Results

The results section initially describes the unit costs for the main ward in each centre where stroke patients were managed (a more complete set of unit costs is provided in appendix 2). The following section uses price and volume indices to analyse why differences in unit costs are observed across the centres. The use of the indices is then extended to present the stroke care specific PPPs. Finally, the total costs per patient in each centre are described, using different conversion factors.

7.51 Unit costs of hospital care

The unit costs for the main resource use items are presented in Table 7.1, at the level of aggregation traditionally used in economic evaluations (see also appendix 2). The results show that there were wide variations in unit costs across the centres even after using the GDP PPP conversion factor to adjust for price differences. The average total cost per occupied bed-day ranged from $23 in the Latvian centre to $333 in the Italian centre. In particular, there were wide variations in the average costs per occupied bed-day of labour inputs, these ranged from $3 (Latvia) to $256 (Italy). Across all the centres, the labour inputs were the highest cost component, and constituted 57% of the overall cost per occupied bed-day.

Table 7.1: Unit costs: Average costs per occupied bed-day for each category of inputs (% of average total cost per occupied bed-day) [$ GDP PPP]

Centre	Labour inputs		Consumables		Overheads		total
Latvia	3	12%	8	34%	12	52%	23
Lithuania 2	10	21%	20	41%	18	37%	48
Lithuania 1	10	20%	21	42%	19	38%	50
Poland	39	35%	25	23%	47	42%	112
Finland 3	80	72%	7	6%	25	22%	112
Finland 1	83	53%	29	18%	45	29%	157
France	128	74%	17	10%	28	16%	172
UK	138	71%	17	9%	40	20%	194
Finland 2	145	64%	22	10%	60	26%	227
Denmark	130	56%	27	11%	77	33%	234
Spain	83	35%	54	22%	102	43%	239
Portugal	133	50%	53	20%	78	30%	263
Italy	256	77%	15	4%	62	19%	333
ALL	95	57%	24	15%	57	28%	167

Further insights are provided when the unit costs are disaggregated further, into factor inputs and factor prices. An example of this is given for the level of input from a recently qualified doctor. The results suggested that while there was some variation in the use of this particular factor input there were particularly wide variations in factor price across the centres considered (Table 7.2). The level of input and associated factor price are compared below:

Table 7.2: Factor input and factor price: doctors' time[49]

	Factor input (minutes/ bed-day)	Factor Price (cost per hour $ GDP PPP)
Latvia	18	2
Lithuania 2	13	4
Lithuania 1	19	4
Poland	28	10
Finland 3	10	40
Finland 1	18	19
France	15	29
UK	22	18
Finland 2	NA	NA
Denmark	20	29
Spain	20	31
Portugal	15	33
Italy	23	30

NA: not applicable as this input was not used in Finland 2 centre.

7.52 Price indices

Price indices provide a more comprehensive comparison of factor prices across the centres. The results showed that there were wide variations in the price of labour inputs between the Western and Eastern European centres (Table 7.3). This was after converting prices into US dollars using GDP PPPs which suggested that GDP PPPs did not adjust for differences in these relative prices between the Western and Eastern European countries. The results showed that there were wide variations in the price of labour inputs across these centres whichever measure of volume weighting was used.

[49] For a mid-grade doctor, that is someone who has the experience and training required for them to qualify as a specialist (e.g. neurologist) in the country concerned, but who does not have managerial responsibilities.

Within the Western European centres there were relatively small differences in factor prices, after PPP adjustment, and again the choice of index was not important. The differences in relative prices between the three Finnish centres were as wide as between the other Western European centres.

Table 7.3: Price indices ($ GDP PPP) for labour inputs in each centre compared to the French centre.

Centre	Laspeyres Price index	Rank	Paasche Price index	Rank	Fisher Price index	Rank
Latvia	0.07	1	0.07	1	0.07	1
Lithuania 2	0.17	3	0.16	3	0.16	3
Lithuania 1	0.15	2	0.14	2	0.15	2
Poland	0.39	4	0.38	4	0.38	4
Finland 3	1.18	12	1.21	12=	1.18	12
Finland 1	0.74	5	0.77	5	0.75	5
France	1.00	9	1.00	9=	1.00	8=
UK	0.86	6	0.88	6	0.87	6
Finland 2	1.18	13	1.21	12=	1.19	13
Denmark	1.01	10	0.99	8	1.00	8=
Spain	0.87	7	0.89	7	0.88	7
Portugal	1.00	8	1.00	9=	1.00	8=
Italy	1.08	11	1.08	11	1.08	11

7.53 Gerschenkron effect

The Paasche/Laspeyres ratio is less than one when the Gerschenkron effect exists i.e. there is input substitution according to factor price differences. In the majority of the centres, the Paasche/Laspeyres ratio was one or more, suggesting that the

Gerschenkron effect was not detected. Of the four centres where the Gerschenkron effect was detected, three were in Eastern Europe (Table 7.4).

Table 7.4: Paasche/Laspeyres ratio for price of labour inputs in each centre

Centre	Paasche/Laspeyres ratio
Latvia	1.00
Lithuania 2	0.98
Lithuania 1	0.93
Poland	0.98
Finland 3	1.04
Finland 1	1.05
France	1.00
UK	1.01
Finland 2	1.03
Denmark	0.97
Spain	1.02
Portugal	1.00
Italy	1.00

7.54 Volume indices

The volume indices showed that there was generally less variation across the centres in factor input use compared to factor price, in particular between the Western and Eastern European centres (Table 7.5). There was greater variation across the Western European centres in factor use compared to factor price. In particular, the Italian centre had a much higher level of labour input than the other centres.

Table 7.5: Volume indices for each centre's labour inputs compared to those in the French centre.

	Laspeyres volume index	Rank	Paasche volume index	Rank	Fisher volume index	Rank
Latvia	0.39	1	0.39	1	0.39	1
Lithuania 2	0.78	4	0.73	4	0.75	4
Lithuania 1	0.59	2	0.58	2	0.59	2
Finland 3	0.67	3	0.70	3	0.69	3
Poland	1.19	9=	1.17	9	1.18	9
Finland 1	1.37	12	1.44	12	1.41	12
France	1.00	6	1.00	6	1.00	6
UK	1.22	11	1.24	11	1.23	11
Finland 2	1.19	9=	1.22	10	1.21	10
Denmark	1.15	8	1.12	8	1.14	8
Spain	0.90	5	0.91	5	0.91	5
Portugal	1.13	7	1.13	7	1.13	7
Italy	3.11	13	3.11	13	3.11	13

7.55 Comparison of conversion factors

Table 7.6 presents the units of local currency required to buy a dollars worth of goods and services for each of the conversion factors discussed. The stroke care PPP is presented for the Fisher price index. The GDP PPP index showed that that far fewer units of currency were required to buy one dollars worth of good and services in the Eastern European countries, than suggested by the OER (Table 7.6). Compared to the GDP PPP, the stroke care PPPs were appreciably lower in the Eastern European centres. The prices of the labour inputs used in producing stroke care were relatively low compared to general prices in these economies. The most extreme example was in

the Latvian centre where the stroke care PPP was 1/15th that of the GDP PPP. The divergence between the OER, GDP PPP and stroke care PPP indices was much smaller for the Western European countries.

Table 7.6: Units of each centre's local currency required to buy a US dollars worth of goods and services, for different conversion factors (all 1998 price base).

	OER ($)	GDP PPP ($)	Stroke care PPP ($)
Latvia	0.50	0.30	0.02
Lithuania 2	4.00	1.40	0.19
Lithuania 1	4.00	1.40	0.17
Poland	2.40	1.40	0.44
Finland 3	5.37	5.89	5.76
Finland 1	5.37	5.89	3.68
France	5.00	6.10	5.00
UK	0.60	0.60	0.43
Finland 2	5.37	5.89	5.75
Denmark	5.60	8.30	6.80
Spain	124.70	117.90	85.12
Portugal	151.00	119.40	103.62
Italy	1628	1515	1346

Sources: OECD (2000), World Bank (2000) The French centre was the reference centre for each conversion factor.

7.56 Total costs per patient (GDP PPP)

The mean total 3-month costs per patient, converted using GDP PPP are reported in Table 7.7. The mean cost per patient was $5,340. However, this ranged from $550 in the Latvian centre to $9,600 in the Danish centre. The Danish centre had both high factor prices (this chapter) and resource use (chapter six). Over all the centres, the majority of the costs were inpatient costs, however, in certain centres community care costs (France) and institutional care costs (Finland 3) were important, and appeared to substitute for hospital services. The reasons for variation in total costs across the centres are explored in chapter 8.

Table 7.7: Mean Total 3-month costs per patient (% of total costs) [$/PPP] (n=1298)

Country	Inpatient costs	Outpatient costs	Community care costs	Institutional care costs	Total costs
Latvia	456	1	37	56	550
	(83)	(0)	(7)	(10)	
Lithuania 2	1,357	10	24	47	1,438
	(94)	(1)	(2)	(3)	
Lithuania 1	1,623	9	11	16	1,659
	(98)	(1)	(1)	(1)	
Poland	3,283	31	700	58	4,072
	(81)	(1)	(17)	(1)	
Finland 3	4,711	301	371	1132	6,515
	(72)	(5)	(6)	(17)	
Finland 1	2,745	858	488	64	4,155
	(66)	(21)	(12)	(2)	
France	4,681	119	1,052	352	6,204
	(75)	(2)	(17)	(6)	
UK	8,672	294	223	122	9,311
	(93)	(3)	(2)	(1)	
Finland 2	6,975	1,365	229	468	9,036
	(77)	(15)	(3)	(5)	
Denmark	8,881	307	387	24	9,600
	(93)	(3)	(4)	(0)	
Spain	1,965	245	326	27	2,805
	(70)	(9)	(12)	(10)	
Portugal	3,893	1,451	44	17	5,405
	(72)	(27)	(1)	(0)	
Italy	5,153	164	171	681	6,170
	(84)	(3)	(3)	(11)	
ALL	**4,526**	**323**	**298**	**203**	**5,340**
	(85)	**(6)**	**(6)**	**(4)**	
F statistic	40.17	26.70	13.98	8.76	50.32
P value	<0.001	<0.001	<0.001	<0.001	<0.001

F statistic and p value are calculated based on OLS regression analyses with costs as the independent variables and dummy variables for each centre as independent variables. F statistic is calculated with 13 parameters and 1285 degrees of freedom.

7.57 Comparing total costs per patient in US dollars using different conversion factors

In the base case analysis the total costs of stroke care were converted into US dollars using GDP PPPs. The sensitivity analysis then examined the impact of using different conversion factors. For resource inputs defined as tradeable, OER were used to convert costs into a local currency. For non-tradeable inputs GDP PPPs and then

stroke specific PPPs were used to convert costs into US dollars[50]. Ordinary least squares (OLS) regression analyses were used with total costs per patient as the dependent variables, and dummy variables for each centre as the independent variable, to compare each measure of total costs across the centres. The corresponding F statistics, p values and adjusted R^2 are reported for each regression analysis.

The impact of using each of these three approaches on the mean three month stroke costs for all the cases in the dataset is shown below (Table 7.8). The results show that using OERs to convert the costs of the tradeable factor inputs led to a greater divergence in the overall costs of care across the centres, compared to using the GDP PPP factor for all inputs (Table 7.8). However, using stroke specific PPPs led to a smaller divergence in costs between the Eastern and Western European centres. For example, the mean cost in the Polish centre based on the stroke specific PPP was $5,000, which was only just below the average across all the centres included in the study. The F statistic was lowest for the comparison of total costs across the centres that used the stroke specific PPP. This illustrates that more of the variability in total costs was explained at the centre level when this conversion factor was used, as differences in factor prices were reduced.

Although using the stroke care PPP index led to smaller cost differences between the centres, residual differences in cost remain. The adjusted R^2 was 0.28, which indicates the level of cost variation that was not explained by the dummy variables for centre[51]. The following chapter considers why this outstanding cost variation exists.

---

[50] Consumables and overheads were defined as tradeable and staff inputs as non-tradeable, though in reality some staff inputs may be tradeable between countries, and some non-staff inputs may be produced and consumed locally.

[51] In fact, the level of unexplained variation is higher when using the model that reduces overall cost differences between the centres. This apparently counter intuitive result is due to the level of unexplained variation that exists at an individual level, which increases when this conversion factor is used. This is illustrated by the higher standard deviations that surround the centre-level means when this conversion factor is used.

Table 7.8: Mean total cost (sd) in US$ for each case included in the dataset (n=1,298) for each conversion factor used.

Centre	Tradeables: GDP PPP Non-tradeables: GDP PPP Base case	Tradeables: OER Non-tradeables: GDP PPP	Tradeables: OER Non-tradeables: stroke PPP
Latvia	550(370)	400(215)	1,374(955)
Lithuania 2	1,438(919)	800(515)	3,134(3,401)
Lithuania 1	1,659(711)	918(424)	3,973(2,325)
Poland	4,072(2,303)	2,948(1,673)	6,017(3,552)
Finland 3	6,515(3,215)	6,716(3,317)	6,833(3,366)
Finland 1	4,155(2,565)	4,370(2,687)	5,589(3,535)
France	6,204(5,015)	6661(5,333)	7,499(6,015)
UK	9,311(7,983)	9,310(7,983)	11,841(10,213)
Finland 2	9,036(5,815)	9,344(6,026)	9,484(6,119)
Denmark	9,600(7,706)	10,938(8,789)	12,876(10,299)
Spain	2,805(2,514)	2,755(2,466)	3,425(2,993)
Portugal	5,405(4,353)	4,839(3,941)	5,504(4,396)
Italy	6,170(2,565)	6,080(5,216)	6,302(6,080)
ALL	5,340(5,884)	5,401(6,494)	6,906(7,246)
F statistic	50.32	61.14	42.79
P value	<0.001	<0.001	<0.001
Adjusted R^2	0.31	0.36	0.28

The F statistic is calculated with 13 parameters, and 1285 degrees of freedom.

7.6 Discussion

The literature review emphasised the importance of using a disaggregated approach to costing, so that reasons for variations in unit costs could be explored (Chapter 2). The review also highlighted the importance of using a consistent method to measure unit costs in each health care setting (Chapter 2). This chapter presented the methods used to measure unit costs, and analysed reasons for unit cost differences. The method relied on the construction of price indices that have generally been applied to time series data, and have rarely been applied to comparing factor prices across health care systems (Danzon and Chao 2000). The results showed that the price of factor inputs varied widely between the Eastern and Western European centres. For example, when prices were converted using the GDP PPP conversion factor the price of factor inputs was 15 times lower in the centre in Latvia compared to the French centre. The corresponding differences in the volume of inputs between Western and Eastern European centres were relatively small (2.5 times lower in the Latvian centre compared to the French centre). There were differences in the volume of factor inputs amongst the Western European centres, but these were not explained by differences in relative prices. A technology specific PPP index was constructed which adjusts for differences in the prices of those factor inputs used to produce stroke care. Using this conversion factor reduced, but did not eliminate cost differences across the centres. For example, when the GDP PPP conversion factor was used, the mean total costs per patient were 20 times higher in the French centre than in the Latvian centre, but they were still 6.5 times higher when the stroke specific PPP was used. The outstanding cost differences across the centres may reflect differences in factor inputs, resources use (e.g. LOS) and case-mix; the next chapter consider these issues.

There were large differences in factor prices between the Western and Eastern European centres which exceeded the differences in GDP per capita across those countries (Chapter 6). In particular, there were wide variations in the costs of labour inputs between the centres in Eastern and Western Europe. The large wage differentials suggest that the international labour market is not perfectly competitive in this sector. Labour economics predicts that faced with a relatively low wage in one

country, workers would migrate to another (Call and Holahan 1983), this process potentially leading to wage equalisation across the countries concerned, for workers with the same marginal productivity. These results suggest that restrictions in the labour market for health care professionals mean that the market is in a sustained state of disequilibrium. Governments in Western Europe may put in place regulations to limit the number of health care professionals recruited from Eastern European countries. Employers may lack information about the transferability of skills and the relative productivity of health care professionals from different countries.

In Eastern European countries, wages may only be a small determinant of labour supply. As discussed in Chapter 4, there may be compensating differentials for health care professionals. The working conditions, the local environment and additional 'perks' such as free child care or subsidised housing may all encourage a doctor in Eastern Europe to continue supplying labour in spite of the low wage. One limitation of the measure of factor price used is that it only considers the costs of labour to the public health care providers. In the Latvian centre, the doctors also worked in the private sector to complement their income. In ex-soviet countries 'under the counter' payments from patients may also be prevalent (Thompson and Witter 2000, Ensor and Witter 2001, Ensor 2004). As part of the site visits to each centre, details were asked about these payments, but only in one centre (Latvia) were these acknowledged to exist, and even then they were estimated to be small[52]. It is possible though that the study underestimated the opportunity costs of these factor inputs, in the Eastern European centres and the true differences in opportunity costs between the Eastern and Western European centres may be less than observed in this study. It would appear unlikely though that the exclusion of these broader aspects of opportunity cost would fully explain the price differences observed.

Other studies that have assessed labour costs in Eastern Europe have also found that wages are much lower than in Western European countries. The International Labour Organisation (ILO) found that within particular countries in Eastern Europe the relative wages of health care workers were low compared to those for colleagues in

[52] Estimates varies from 5% to 10% of the average monthly salary.

other sectors (International Labour Organisation 2002). For example, the ILO considered labour costs in the health sector in Lithuania and commented that:

".. average salaries are very low (in Lithuania) compared to EU levels, and were only 83% of the Lithuania national wage." (ILO 2002, p6).

Lower wages in the Eastern European centres may be a spill over from the incentive mechanisms that previously existed in these countries. Under the Soviet-style health care system, hospital budgets were generally fixed, and based on the number of beds (Sheiman 1994). Although budgets were fixed, costs rose, in response to inelastic demand for new drugs and investigations, bought at Western prices (Van Andel 1997). Salaries for health care workers were paid out of the residual from the budget and fell, both in real terms and relative to professionals working in other sectors (Sheiman 1994, ILO 2002). In addition, there was a general perception in ex-soviet countries that workers in service sectors do not produce such a worthwhile output compared to their counterparts in manufacturing industry (Afford 2004). Labour market economics would predict that health care professionals would retrain and work in other sectors, but again a lack of mobility in the labour market coupled with uncertainty about economic prospects in the countries concerned, may have conspired to keep wages low.

There was relatively little variation in relative factor price across the Western European centres, certainly when compared with pharmaceutical markets (Danzon and Chao 2000, Danzon and Kim 1998) and manufacturing industries (Van Ark et al. 1999). This may indicate that in these countries the labour markets in the health sector are more flexible and competitive, leading to similar prices for labour inputs across countries. In each of these centres, the hospitals concerned were public sector teaching hospitals, and there was evidence of transfer of health care professionals across the centres concerned. Interestingly, within Finland there were important variations in the price of inputs amongst the centres, illustrating that variability in factor prices may be important within as well as across countries.

Previous cross-national comparisons have found that factor substitution according to relative factor prices does occur both in manufacturing industry (van Ark et al. 1999) and in the use of pharmaceuticals (Danzon and Chao 2000). However, this investigation found that the measures of relative price were similar whether they were constructed using Laspeyres or Paasche price indices. This suggests that input substitution according to factor price (the Gerschenkron effect) in the pursuit of productive efficiency is generally not taking place. This may reflect a lack of incentive to produce efficiently in the centres concerned. While there were differences across the Western European centres in the use of factor inputs, these did not appear to reflect differences in factor price. These may instead reflect differences in the choice of technology or differences in case-mix across the centres concerned (see Chapter 6). For example, in the Italian centre, most patients were referred to a hi-tech acute stroke unit immediately post stroke, which had a much higher staff to bed ratio than equivalent wards in other centres.

It should also be noted that the Gerschenkron effect was not detected amongst Western European centres where prices were similar but was detected between the Eastern European centres and the French centre where there were large differences in relative factor prices. Where there are smaller more subtle differences in relative factor prices, it may be more difficult to detect any correlation between relative factor prices and relative factor use. In addition, decision-makers in each centre may not have much control over the use of labour inputs. These may be set by regulations laid down by national governments. However, the local decision-maker may have more control over resource use measures such as LOS, and the level of factor price may be a factor the local decision-maker uses when deciding when to discharge the patient from hospital. The measures of relative price constructed in this chapter will therefore be used in the subsequent analysis of reasons for variations in LOS across the centres (Chapter 8).

The unit costing tried to use a disaggregated and consistent methodology across the study settings. This was feasible for the measurement of labour inputs which formed an important component of the overall costs of stroke care. However, for some of the

resource inputs, in particular consumables and overheads it was not possible to use a disaggregated approach to cost measurement. Detailed data were not collected for these items for each patient, and this information was not available at site visits to each centre. For these cost elements, the method of cost allocation was based on methods used by financial departments in each centre. Attempts were made to try and standardise these methods based on recommendations in the literature that suggested allocating joint costs across different departments to the individual department and then to the individual bed-day, based on measures of relative use (Graves 2002). While cost variation across the centres for these items was generally lower than for labour inputs for certain centres these items comprised a high proportion of the total costs per bed-day, and were important in understanding cost variation. For example, in the Spanish centre, overhead costs were 43% of the cost per hospital bed-day. Overhead costs were 3.5 times higher in the Spanish centre than the French centre, whereas the costs per bed-day of labour inputs were 30% less than in the French centre. However, it is likely that some of the observed variability in the costs of overheads reflects differences in the methods used to allocate these costs.

The method used to calculate a more specific currency conversion factor followed other attempts to construct technology specific PPPs and calculated bilateral price indices whereby prices in each centre are compared to those in a reference centre (Wordsworth and Ludbrook 2005, Hutton 2001). These indices are sensitive to the choice of reference centre (Diewert 1999). Instead, multilateral price indices can be developed whereby the prices in each country are compared to all other centres' prices (Diewert 1999). Multinational studies of cost and cost-effectiveness could use multilateral price indices, to adjust more fully for cross national price differences when estimating costs.

Finally, a general problem with using price or volume indices is that they are commonly based on aggregated measures of input, rather than measures at an individual level. They therefore suffer from the same problems of all such aggregated measures in health care, namely that there is inadequate adjustment for case-mix and quality of inputs and outputs across the firms concerned. All the indices described rest

on the premise that the productivity of the factor inputs in each country is identical. Yet, as described in chapter 6, the case-mix, ward-type and outcomes of stroke care vary within and across the centres concerned, suggesting that there are potential variations in the quality of health care inputs. The indices also provide deterministic measures with no indication of the level of random variation that surrounds the use of factor inputs. Indices that are more appropriate would recognise these issues. However, the level of disaggregation used in constructing these indices already goes beyond the separate reporting of resource use unit costs required in economic evaluation, and has given further insights into why costs vary across health care settings.

7.7 Conclusions

To conclude, unit costs vary across the centres because of differences in factor inputs, but even greater differences in factor price. This has implications for the interpretation of costs converted using GDP PPPs, in that the observed cost differences between Eastern and Western centres partly reflect outstanding differences in factor prices. The variation in labour costs was studied in depth and was an important component of the costs of stroke care in each setting. Amongst the Western European centres there were wide variations in the use of labour inputs; these appear an important reason for the observed variation in the costs per bed-day across these centres. The use of different price indices suggested that factor price differences were not driving observed differences in factor use across the Western European centres. Although the cost measurement attempted to use a consistent method and measure opportunity costs in each setting, the study may have underestimated opportunity costs in the Eastern European centres. The factor price of labour inputs was based on the employees' wage and did not include 'under the counter' payments or 'benefits in kind' (e.g. subsidised housing). While careful attempts were made to use a consistent methodology across the centres, the observed variations in unit costs may partly reflect differences across the centres in the way overhead costs were allocated. Nevertheless, the conclusion that important price differentials exist between the centres in Eastern and Western Europe is sufficiently large to appear robust to the caveats discussed. The next chapter examines the importance of factor price and other

variables in explaining the observed variations in resource use and total cost across the study centres.

Chapter 8: Assessing the variability of multinational resource use and cost data using OLS regression models compared to MLMs

8.0 Introduction

The data described in chapter six showed that there were differences between the resources used and the cost of producing stroke care across the 13 different European centres included in the empirical investigation. The literature review identified factors associated with resource use and cost variation across different geographical health care settings. These factors may be measured at different levels-- at the level of the patient, health care centre, or country concerned. For example, relative price, which is usually measured at the level of the health care centre or country concerned, may determine the relative level of resource use in different centres. Other factors that are associated with resource use differences, such as case-mix may operate at a patient-level (Wennberg 1984, McPherson et al. 1982, Phelps and Mooney 1993). Chapter six showed that there were important variations in patient and contextual factors across the centres included in the study. Chapter seven highlighted that relative factor prices and unit costs vary across the centres. There are therefore *a priori* reasons for using methods of analysis that recognise that factors associated with international resource use and cost variability, operate at different levels.

However, studies in this area have generally used ordinary least squares (OLS) models (Willke et al. 1998, Coyle and Drummond 1998). These models assume that observations across patients are independent and have a common variance. This assumption would seem implausible when using data from different centres, as patients' resource use (or costs) within a particular centre may be more similar than that in different centres, for the reasons outlined. Furthermore, centre or national-level variables included in an OLS model are considered as if they were measured at a

patient-level, thus spuriously inflating the amount of information they supply. By contrast, multilevel models (MLMs) are able to incorporate the hierarchical structure of the data (in this case of patients within centres), and provide more appropriate estimates of patient and centre-level effects. MLMs have been recommended for use in health economics (Rice and Jones 1997), but despite the obvious intuitive appeal of using MLMs to assess multicentre resource use and cost data, they have not yet been used for this purpose. An important question to address is: does using MLMs rather than OLS models matter when analysing multicentre resource use and cost data?

The aim of this chapter is to compare the use of OLS regression models and MLMs for assessing the reasons for multicentre resource use and cost variations. The Biomed II stroke dataset described in chapter six, is used for this purpose. Section 8.1 summarises the methodology used to assess the reasons for international resource use and cost variation, and covers issues such as the choice of model specification, and the conversion factor used. Section 8.2 describes the different statistical models used in the analysis. Section 8.3 presents the results, and section four discusses their implications.

8.1 Overview of methodology

This chapter compares the use of OLS models and MLMs for estimating which factors are associated with length of hospital stay (LOS) and total cost. LOS was chosen as the resource use measure as it is the resource use variable that explains most of the variation in the costs of stroke care (Porsdal and Boysen 1999). The basic structure of MLMs and the rationale for their use in health care and health economics has been previously described (Chapter 5), so only a summary is provided here. In the context of this dataset, individuals (level-1 units) are nested within centres (level-2 units). It is recognised that some of the variables defined at a centre-level are measured at a national level e.g. overall level of health care spending. These could have been considered as a separate level however there were insufficient centres in each country to include country as a separate, third level in the model.

8.11 Inclusion of variables for assessing resource use and cost variation.

Chapter four outlined the variables included in the dataset, that have a clear theoretical rationale for explaining either resource use or cost variability. Some of these variables were measured at the patient-level such as case-mix or access to care. Other factors were measured at the centre or country level: relative factor price, budget constraints, incentives to health care providers, and patients. The decision about which of these variable to include in the model was also based on *a priori* reasoning rather than statistical significance. Certain variables were excluded from the models because there were clear theoretical reasons as to why their inclusion in a model explaining resource use or total cost was inappropriate. For example as Wagstaff (1989a) points out, variables such as bed-occupancy are directly correlated with resource use, and total cost, and were therefore excluded from the analysis. It was recognised that measures summarising the production processes in each centre, such as the access to neurologists and the use of rehabilitation hospitals, would be likely to be correlated with more general measures of health infrastructure such as the percentage of GDP spent on health care (% GDP/health). It was also desirable to choose higher-level variables that may have more general use outside the specific disease area. On this basis the context variables that seemed most appropriate were % GDP on health care, use of copayments, use of DRG system and for the resource use model, relative factor prices. As factor price and the use of substitutes to care is used in the calculation of total cost these variables were excluded from the total cost model.

8.12 Distribution of cost variable

MLMs usually assume that the errors terms are normally distributed. However, resource use and cost data are usually not normally distributed (Briggs and Gray 1999). In this dataset, the resource use and total cost data used in this investigation are clearly not normally distributed (see appendix 4). In addition, the shape of the distribution of the data varies across the centres included in the analysis. For this investigation, the main interest is in the effect of covariates on LOS and total cost per patient. The initial analysis will use OLS models and MLMs that assume the residuals are normally distributed. When using regression models to analyse data it is possible

that although the raw data are skewed the residual is normally distributed after adjusting for covariates. However, given the highly skewed cost data, with distributions differing by centre it appears unlikely that the general assumption that the residuals are normally distributed is justified (appendix 4). Generalised linear models (GLMs) models have been recommended as an alternative for analysing cost data (Briggs and Gray 1999, Manning and Mullahy 2001), as a variety of non-normal distributions can be specified but, unlike data transformations in OLS regression, they make inferences about the mean cost directly (see Chapter 5). A particular type of GLM- the gamma model, can also be specified as having a different shape for different higher level units in the data hierarchy. This made the gamma model a potentially attractive alternative for the analysis of this dataset.

8.13 Conversion factor used to convert local costs into common currency

The previous chapter concluded that converting costs from local currencies into a common currency (US dollars) using GDP PPP, did not adjust for differences in relative price between the Eastern and Western European centres. Nevertheless, as GDP PPPs provide a more general measure of opportunity costs in each of the centres concerned the main cost analysis will use total costs converted using GDP PPPs. However, analysis tests the robustness of the conclusions to the choice of conversion factor, by using total costs converted using the stroke care PPP (derived in Chapter 7) in a sensitivity analysis.

8.2 Statistical models used

An OLS regression model takes the form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i ; \quad \varepsilon_i \sim Normal(0,\sigma^2) \qquad (1)$$

where y_i is the outcome variable for the ith individual, x_i is an explanatory variable, with associated slope coefficient β_1. β_0 is the intercept and ε_i , the error term which represents unexplained variability between individuals, is assumed to be normally distributed with a mean of zero. The OLS model assumes that the variance of the error

term is the same for all individuals. Extra explanatory variables can be included at either a patient or centre-level. Those representing centre-level covariates necessarily take the same value for all individuals in a particular centre.

Moving to a MLM structure changes the way the unexplained variation, the random error term, is modelled. The most basic MLM, the random intercepts model, includes an additional term which represents the unexplained variation that exists across centres. Using subscripts *i* and *j* for the *i*th individual in *j*th centre, the model may be written:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + \varepsilon_{ij}\,; \qquad \varepsilon_{ij} \sim Normal(0,\sigma^2),\ u_j \sim Normal(0,\tau^2) \qquad (2)$$

where β_0 is a fixed quantity applying to all individuals, u_j is a random variable with zero mean and constant variance (τ^2) which applies to all the cases in a particular centre, and e_{ij} is a random error term which represents the unexplained variation for individuals within a centre. u_j indicates the effect of centre on the outcome variable, over and above that explained by the set of explanatory variables. The intercept for the *j*th centre (previously given as β_0) is now given as a fixed component (β_0) plus a random component (u_j). The model can be developed by including additional explanatory variables at the level of the individual or the centre.

MLMs allow the effect of adding centre-level explanatory variables on the extent of unexplained variation between centres (τ^2) to be estimated. In addition, the degree of dependency amongst observations within each centre can be measured by the intra-class correlation coefficient (ρ) defined as $\rho=\tau^2/(\tau^2+\sigma^2)$. This reflects the strength of 'nesting' within the data hierarchy.

A number of alternative GLMs were considered for the analysis. A gamma distribution model with random mean (θ_{ij}) and random shape (ϕ_j) parameters was chosen as it fitted the positively skewed LOS and cost data reasonably well:

$$y_{ij} \sim Gamma(\theta_{ij},\phi_j)\ ,\ \theta_{ij} = \beta_0 + \beta_1 x_{ij} + u_j\ ;$$

$$u_j \sim \text{Normal}(0,\tau^2),\ \phi_j \sim \text{truncated Normal}(\mu,\omega^2) \qquad (3)$$

This 'generalised linear mixed model' (GLMM) is a type of multilevel model similar to model 2 in that it allows for random centre effects (u_j), but it accommodates positively skewed data by the use of a gamma rather than a normal distribution, allowing the gamma distributions to have different shapes in each centre. The shape parameters are restricted to be positive by using a normal distribution truncated at zero.

8.3 Estimation

An OLS regression model was fitted to estimate the effect of the patient-level variables identified in chapter six on the total length of hospital stay (LOS) (model 1). Variables were included if the literature review suggested they may be associated with LOS, and retained in the equation even if they did not reach conventional levels of significance. A series of diagnostic tests were performed for heteroscedasticity, multicollinearity and correct functional form.

The analysis was repeated with normal MLM random intercepts (model 2) and gamma GLMM random intercepts (model 3). The normal model was estimated by restricted iterative generalised least squares in MLwiN (Rasbach et al. 2002), equivalent to restricted maximum likelihood. The gamma models were fitted using Markov chain Monte Carlo methods in WinBUGS (Gilks et al. 1996). The LOS analyses were then repeated also including centre-level variables (models 4-6).

The models including patient and centre level variables were re-fitted with total cost as the dependent variable (models 7-9). The total costs used were those converted from local currencies into US$ using the GDP PPP index (see Chapter seven). As LOS is highly correlated with total cost, both in this dataset and more generally for stroke patients, the same factors were included in the models apart from the variables

for price index and care substitutes. The goodness of fit of the different cost models was compared using log-likelihoods, and plots of deviance residuals (Rasbach et al. 2002).

To test whether the conclusions were robust to the choice of conversion factor, the models were refitted with the costs converted from local currency into US dollars ($) using the stroke specific PPP index (models 10-12).

8.4 Results

8.41 LOS models with patient-level variables

The results from the OLS analysis suggested that most of the patient variables considered were associated with LOS (Table 8.1, model 1). The relationship between these factors and LOS was generally that anticipated by *a priori* reasoning. For example, LOS was higher for the more severe cases, i.e. those in coma, incontinent or paralysed at admission. There was an inverse relationship between age and LOS. As stated in Chapter 6 this appears inconsistent with more general literature that suggests age is positively associated with resource use mainly because of the closer proximity to death. However, in this dataset only patients surviving the stroke were included, and the inverse relationship only arises after adjusting for the other patient factors. So for survivors with a given case-mix, younger patients were more likely to stay in hospital longer than older patients. This may reflect the perception that younger patients have a greater capacity to benefit from inpatient hospital rehabilitation, although it should be noted that the effect of age on LOS was small and non-significant. The utilisation of community support was associated with reduced LOS indicating that this may be a substitute for hospital care. The adjusted R^2 for the OLS model was 0.27, this showed that there was still a large element of unexplained variation.

Compared to the OLS model, the normal random intercepts model (model 2) gave somewhat different coefficients and standard errors, and the estimated significance of the covariates changed accordingly. For example, both age and independence pre-stroke had smaller coefficients and were no longer significant, whereas paralysis had a larger estimated effect. The random intercepts model (model 2) fitted the data better than the OLS model as shown by the lower level of unexplained variation at a patient level (σ^2) and the higher log-likelihood statistic. Model 2 estimated the level of unexplained variation, which existed across the centres. The results showed that the majority of the unexplained variation was among patients, rather than among centres, and the intra-class correlation coefficient (ρ) was 0.16.

Diagnostic testing for the OLS model suggested that severe multicollinearity was unlikely to exist; for example, the pairwise correlation coefficients between the explanatory variables did not exceed 0.6. Heteroscedasticity was detected for the OLS model ($p<0.001$ by the Cook-Weisberg test (Cook and Weisberg 1982)). Heteroscedasticity can be caused by using an incorrect functional form (Gujarati 1988), and here the residuals were found to be non-normally distributed, with the shape of the distribution varying according to the health care centre (see appendix 4). The Ramsey reset test for functional form also suggested that the OLS model was misspecified ($p=0.01$).

The analysis was therefore repeated using a gamma model with intercept and shape parameter varying by centre (Table 8.1, model 3). The estimated effect sizes and their associated standard errors differed somewhat from both the two previous models. For example, the coefficient for age was now larger and significant at the 5% level, whereas the coefficient for community support was no longer statistically significant. The log-likelihood statistic indicated that the gamma model fitted the data substantially better than either of the other two models.

Table 8.1: OLS model, MLM and GLMM estimating the effect of patient-level variables on length of hospital stay (LOS): coefficient (SE)

	Model 1: OLS	Model 2: MLM Normal, with random intercept	Model 3: GLMM Gamma, with random intercept
Constant term	41.7 (5.5)**	29.5 (6.2)**	31.3 (4.8)**
Patient variables			
Age	-0.10 (0.06)*	-0.08 (0.05)	-0.13 (0.04)**
Independent pre-stroke	-6.04 (2.38)**	-3.22 (2.32)	-4.82 (2.22)**
Living alone	Reference	Reference	Reference
Living with others	-8.23 (1.55)**	-5.07 (1.55)**	-0.65 (1.13)
Living in nursing home	-10.85 (4.61)**	-8.49 (4.44)*	4.37 (4.31)
Incontinent	19.29 (1.55)**	18.81 (1.49)**	15.65 (1.63)**
Paralysed	7.34 (1.56)**	8.48 (1.50)**	2.45 (0.81)**
Ischaemic stroke	Reference	Reference	Reference
Haemorrhagic stroke	5.35 (2.13)**	5.80 (2.04)**	6.82 (1.65)**
Unknown stroke type	-2.27 (1.73)	-7.71 (2.45)**	-3.67 (1.13)**
Onset to admiss < 6 hrs	Reference	Reference	Reference
Onset to admiss 6-24 hrs	-0.55 (1.58)	0.59 (1.52)	0.77 (0.89)
Onset to admiss 1-7 d	-0.96 (1.80)	-0.70 (1.73)	1.22 (0.96)
Onset to admiss >7 d	-1.33 (3.27)	-0.85 (3.17)	0.28 (1.50)
Ons. to admiss unknown	10.17 (2.64)**	8.05 (2.60)**	6.39 (2.85)**
Family support	-1.53 (1.39)	2.51 (1.49)*	2.46 (0.81)**
Community support	-4.06 (1.71)**	-5.76 (1.68)**	-0.20 (1.04)
Random effects			
σ^2 (within centres)	522	470	442
τ^2 (between centres)		87 (37)	85 (47)
Log-likelihood	-5896	-5845	-5267

** p<0.05, *p<0.10, Ons: Onset, admiss: admission, hrs: hours, d: days. The log-likelihood statistic presents a measure of how well each model fits the data. R^2 values were not available for the MLMs.

8.42 LOS models with patient, centre and national level variables

The models were re-fitted including centre-level variables (Table 8.2). The results showed that the estimated direction of effect for these variables was that anticipated by theory: proxies for the presence of incentives to discharge patients earlier – a DRG system, patient copayments, or higher relative prices (Fisher price index) – were all associated with shorter LOS, whereas a higher proportion of GDP spent on health was associated with longer LOS. For the OLS analysis (model 4) each of these effects was statistically significant at the 5% or 10% level. Although the adjusted R^2 (0.28) was higher than for the previous OLS model including just patient level variables, there was still substantial unexplained variation.

The normal (model 5) and gamma random intercepts models (model 6) found that these centre-level variables were far from statistically significant. The coefficients for the centre-level variables were generally similar for the three models. So for example, whichever model was used, an increase in the proportion of GDP spent on health care of one percentage point was associated with an average increase in LOS of 6 days. The standard errors were much smaller for the OLS model which does not take into account the hierarchical structure of the data, and thus severely overestimated the significance of these variables. The gamma model had the highest log-likelihood, and therefore fitted the data best.

The variance terms for both multilevel models showed that most of the unexplained variation was, again, at the level of the patient rather than the centre (ρ=0.18 for the normal random intercepts model). The MLMs which included centre and national level variables (models 5 and 6) had a similar extent of unexplained variation across centres, and a similar log-likelihood, compared to the models with just patient-level variables (models 2 and 3). This showed that, once the hierarchical nature of the data was recognised, these higher-level variables did not help explain the variability in LOS across centres.

Table 8.2: OLS model, MLM, GLMM estimating the effect of patient, centre and national-level variables on length of hospital stay (LOS): coefficient (SE)

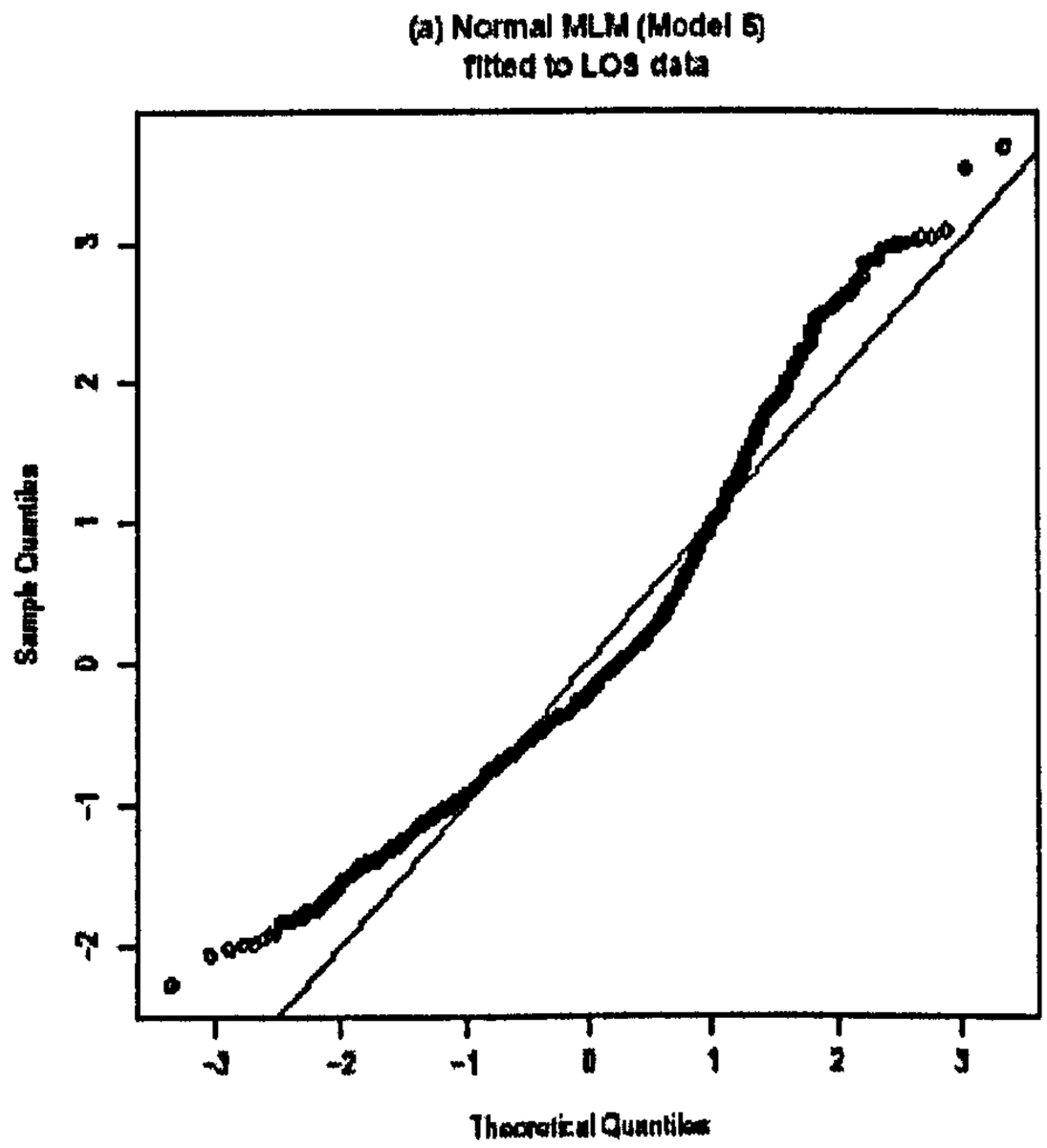
	Model 4: OLS	Model 5: MLM Normal, random intercept	Model 6: GLMM Gamma, with random intercept
Constant term	15.3 (7.8)*	4.7 (24.6)	12.6 (26.1)
Patient variables			
Age	-0.08 (0.05)	-0.07 (0.05)	-0.13 (0.04)**
Independent pre-stroke	-3.84 (2.37)	-3.13 (2.32)	-5.09 (2.19)**
Living alone	Reference	Reference	Reference
Living with others	-6.29 (1.59)**	-4.98 (1.56)**	-0.73 (1.11)
Living in nursing home	-7.03 (4.55)	-8.38 (4.44)*	4.25 (4.19)
Incontinent	19.28 (1.53)**	18.82 (1.49)**	15.57 (1.65)**
Paralysed	7.64 (1.54)**	8.52 (1.51)**	2.31 (0.84)**
Ischaemic stroke	Reference	Reference	Reference
Haemorrhagic stroke	6.03 (2.10)**	5.84 (2.04)**	6.84 (1.63)**
Unknown stroke type	-3.22 (2.17)	-7.82 (2.48)**	-3.74 (1.12)**
Onset to admiss < 6 h	Reference	Reference	Reference
Onset to admiss 6-24 h	0.36 (1.56)	0.64 (1.52)	0.69 (0.88)
Onset to admiss 1-7 d	-0.53 (1.78)	-0.68 (1.73)	1.16 (0.97)
Onset to admiss >7 d	-2.65 (3.25)	-0.87 (3.17)	0.21 (1.53)
Ons. to admiss unknown	6.66 (2.64)**	7.90 (2.60)**	6.12 (2.83)**
Family support	1.72 (1.48)	2.67 (1.49)*	2.46 (0.81)**
Community support	-6.37 (1.73)**	-5.86 (1.69)**	-0.31 (1.04)
Centre and national variables			
% share GDP	5.71 (1.09)**	6.58 (5.13)	5.39 (5.53)
DRG system	-8.11 (1.77)**	-6.16 (7.03)	-4.96 (7.39)
Price index	-10.91 (2.58)**	-10.79 (10.26)	-9.60 (10.93)
Copayment	-3.82 (2.19)*	-4.92 (9.68)	-3.93 (10.30)
Random effects			
σ^2 (within centres)	502	470	369
τ^2 (between centres)		102 (42)	120 (97)
Log-likelihood	-5868	-5845	-5268

** p<0.05, *p<0.10 Ons: Onset, admiss: admission, hrs: hours, d: days

Figure 8.1 shows the deviance residuals from models 5 and 6 in the form of normal plots. If a model is appropriate, deviance residuals should approximately follow a normal distribution and will lie along the line of identity shown in the plots. For the normal MLM (model 5), the residuals show considerable positive skewness, in keeping with the positive skewness of the raw LOS data. The residuals from the gamma GLMM (model 6) show a much better behaviour, indicating the greater appropriateness of this model for these data and confirming the improvement in fit shown by the log-likelihoods.

8.43: Total cost models with patient and centre-level variables

The factors that were associated with LOS, were also associated with total cost (Table 8.3). In the OLS model each of the centre and national level variables was significantly associated with total cost, and the direction of the effect was that predicted by economic theory: the presence of patient copayments or a DRG system were associated with lower mean costs, a higher proportion of GDP spent on health care was associated with higher total costs.



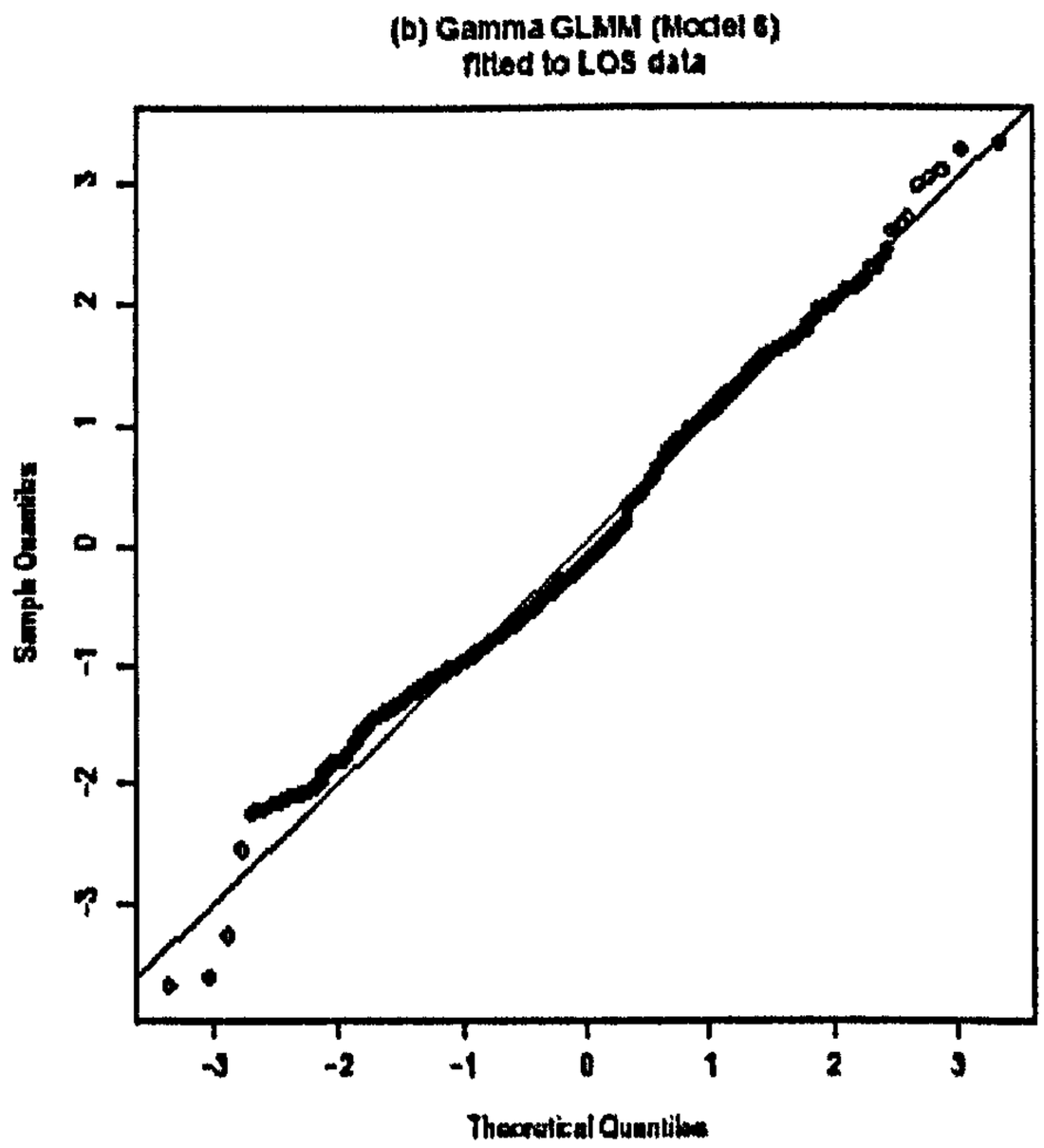


Figure 8.1: Quantiles of the standardised deviance residuals from length of hospital stay (LOS) plotted against the quantiles from a standard normal distribution.

Table 8.3: OLS model, MLM and GLMM estimating the effect of patient and centre and national-level variables on total cost ($/PPP) coefficient (SE)

	Model 7: OLS	Model 8: MLM Normal, random intercept	Model 9: GLMM Gamma, random intercept
Constant	-4470 (1432)**	-6156 (4353)	-2034 (6529)
Patient variables			
Age	-10.0 (10.8)	-6.2 (10.8)	-7.4 (2.6)**
Independ pre-stroke	-916 (478)*	-831 (467)*	-3985 (3679)
Living alone	Reference	Reference	Reference
Living with others	-1344 (316)**	-1309 (309)**	-20 (86)
Living nurs home	412 (917)	0 (892)	994 (602)
Incontinent	3657 (310)**	3561 (300)**	942 (176)**
Paralysed	1709 (309)**	1811 (302)**	56 (54)
Ischaemic stroke	Reference	Reference	Reference
Haemorr. Stroke	985 (424)**	971 (411)**	182 (116)*
Unkn. stroke type	-2140 (385)**	-1110 (497)**	-244 (69)**
Onset-ad < 6h	Reference	Reference	Reference
Onset-ad 6-24 h	-125 (315)	-141 (305)	-5 (62)
Onset to ad 1-7 days	-400 (358)	-500 (348)	39 (64)
Onset to ad >7 days	-740 (652)	-300 (638)	-54 (82)
Onset to ad unkn.	1836 (531)**	2063 (523)**	419 (280)*
Centre and national variables			
% share GDP	2042 (155)**	2200 (755)**	2218 (946)**
DRG system	-1688 (354)**	-1475 (1259)	-1860 (1560)
Price index	NA	NA	NA
Copayment	-1722 (405)**	-1637 (1758)	-1477 (2153)
Random effects			
σ^2 (within centres)	20.5×10^6	19.1×10^6	18.3×10^6
τ^2 (between centres)		$3.6\ (1.5) \times 10^6$	$5.8\ (4.4) \times 10^6$
Log-likelihood	-12761	-12731	-11800

** p<0.05, *p<0.10, NA: not applicable, Independ: independent, Ons: Onset, ad: admission, haemorr: haemorrhagic, unkn: unknown, hrs: hours, d: days. Note that the family support and community support variables (Table 8.3) are excluded from this model, as variables representing care substitutes might be actual components of total cost.

The OLS model again severely underestimated the standard errors of the centre and national level variables, so that their significance was overstated. Model 8, which correctly recognised the hierarchical structure of the data, estimated that the only higher-level variable which was significantly associated with total cost was the % share of GDP spent on health care. Although the presence of patient copayments and a DRG system were each associated with a mean reduction in total cost of approximately $1,500, model 8 found that the standard errors surrounding these variables were large. Model 9 which incorporated the hierarchical structure of the data *and* used a more appropriate functional form which better fitted the data, again found that only % GDP spent on health was associated with total cost. Like the LOS models, the proportion of unexplained variability in total costs amongst patients was substantially higher than that across centres (ρ=0.16).

8.44 Currency conversion factor

In the base case analysis, the costs were reported using the GDP PPP conversion factor. However, using this conversion factor did not completely adjust for price differences across the centres concerned (Chapter seven). Some of the observed relationships between higher-level variables (in particular % GDP/health), and total cost may reflect the relationship with price, rather than resource use. To examine this further, in a sensitivity analysis the costs were converted from local currencies to US$ using stroke care PPPs, which reduced any price differences between the centres. The resulting cost models (10 to 12) therefore looked at the effect of factors (such as %GDP/health) on the overall level of resource use and hence total cost, rather than on resource use *and* factor price and hence total cost (Table 8.4).

The results showed that even when stroke care PPPs were used to convert local costs into US dollars, the same factors were associated with total cost. This suggests that it is mainly through their influence on physical resource use rather than factor price that these variables are associated with total cost (Table 8.4).

Table 8.4: OLS model, MLM and GLMM estimating the effect of patient, centre and national-level variables on total cost ($/stroke care PPP) coefficient (SE)

	Model 10: OLS	Model 11: MLM Normal, random intercept	Model 12: GLMM Gamma, random intercept
Patient variables			
Constant	-6231(3478)	-7,861(5,690)	-3,803(4,768)
Patient variables			
Age	-25.8(13.4)*	-10.3(13.6)	-16.8(6.38)**
Independent pre-stroke	-1205(594)**	-1090(587)*	-1,140(483)**
Living alone	Reference	Reference	Reference
Living with others	-1527(392)**	-1,623(389)**	-156.9(190.1)
Liv in nursing home	407(1139)**	17.3(1121)	1866(997)*
Incontinent	4471(384)**	4458(377)**	2909(370)**
Paralysed	2295(384)**	2304(380)**	270(122)**
Ischaemic stroke	Reference	Reference	Reference
Haemorrhagic stroke	1325(526)**	1242(516)**	715(299)**
Unknown stroke type	-1,576(478)**	-1,757(622)**	446(172)**
Ons to admiss < 6hr	Reference	Reference	Reference
Ons to admiss 6-24 hrs	-154(391)	-195(583)	4.30(144)
Ons to admiss 1-7 days	424(445)	-655(438)	26.9(148)
Ons to admiss >7 days	-614(810)	-590(802)	-166(196)
Ons to admiss unknown	2919(660)**	2,847(658)**	1338(638)**
Centre and national variables			
% share GDP	2503(192)**	2528(812)**	2,292(822)**
DRG system	-2426(439)**	-1558(1,365)	-2,436(1145)**
Price index	NA	NA	NA
Copayment	-1743(503)**	-1,925(1,893)	-684(1,773)
Random effects			
σ^2 (within centres)	$31.6*10^6$	$30.1*10^6$	$17.0*10^6$
τ^2 (between centres)		$4.1(1.8)*10^6$	$5.3(4.0)*10^6$
Log-likelihood	-13,400	-13,027	-12,100

** p<0.05, *p<0.10, NA not applicable, ** p<0.05, *p<0.10 Ons: Onset, admiss: admission, hrs: hours, d: days.

8.5 Discussion

This chapter presented the results of OLS regression models and MLMs for assessing factors associated with resource use and cost variation. The results showed that patient-level variables representing case-severity, access to care, and substitutes to care were significantly associated with resource use and total cost whichever model was used.

Economic theory suggests that the presence of a DRG system, patient copayments and higher relative factor prices would, *ceteris paribus* be associated with shorter LOS. The OLS analysis found these factors were associated with significantly shorter LOS. However, the OLS analysis overestimated the precision of the centre and national-level associations, and therefore made incorrect inferences. The MLMs were more appropriate for analysing the multinational data, and led to different results. In particular, the MLMs showed that, once the hierarchical nature of the data was recognised, none of the higher-level variables predicted resource use and only the level of health care spending was associated with total cost.

The reason for this is as follows (Goldstein 1995). In an OLS regression of, say, cost on % GDP/health, all 1298 patient costs are considered. Thus the regression line is very precisely estimated (Figure 8.2a). In a MLM regression, the correct hierarchical structure of the data is recognised, and essentially the mean cost in each of the 13 centres is regressed on % GDP/health. The resulting regression line is imprecisely estimated (Figure 8.2b). In this example, the regression lines for the OLS and MLM analyses have a similar slope, i.e. the effect of % GDP/health on total cost is similar for each model. However the SE is much larger in the MLM since it involves a regression of 13 rather than 1298 points.

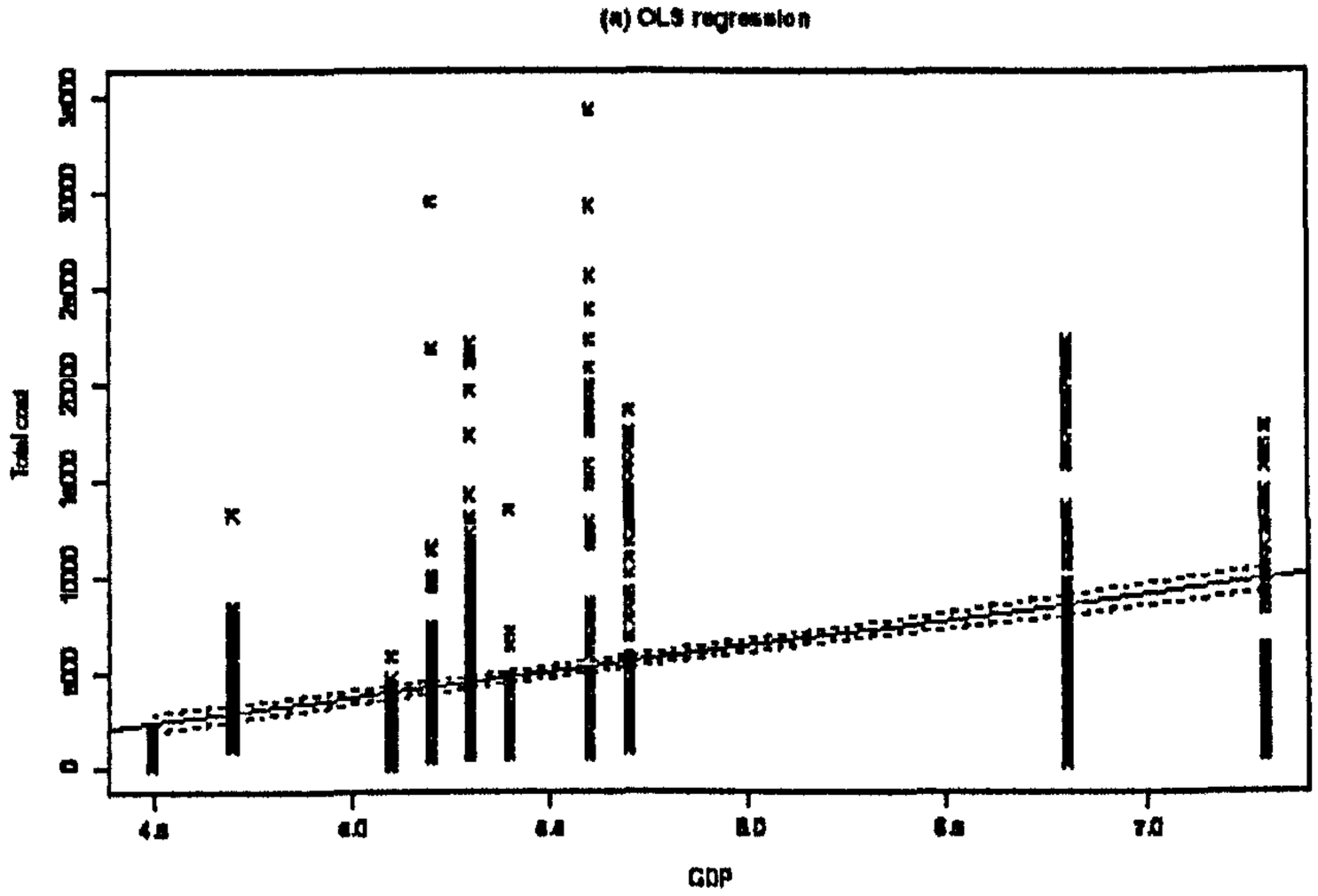


Figure 8.2: The relationship between cost and % GDP/health, based on (a) ordinary least squares regression, (b) multilevel modelling; 95% confidence interval for the regression lines shown.

These results suggest that assessments of international resource use and cost variation should not rely on OLS analyses. While it might appear that only the OLS analysis is producing results that are supported by *a priori* reasoning, the statistical assumptions made are inappropriate. The OLS analysis assumes that the individual error terms are independent, and do not allow for any clustering in the data. Yet economic theory

suggests that where contextual factors such as input prices differ, this will lead, *ceteris paribus* to differences in resource use. Therefore, if centres A and B have different factors prices, the resource use for patients in centre A will be more similar to each other than to individuals within centre B. Thus, the OLS assumption that the error terms are independent would not be supported by economic theory in this context. MLMs acknowledge that hierarchies exist and in this context can allow for individuals to be clustered within centres. MLMs therefore provide a suitable framework for analysing why these differences exist as they estimate the effects of patient and higher-level factors correctly. These models provide a reasonable basis for investigating why costs might vary across centres. By uncovering the reasons for variability based on both economic and statistical rationale this analysis has, at least in this disease area and context, provided a list of reasons for cost variation.

The use of economic theory proved complementary to statistical theory. The choice of variables for inclusion in the models was driven by economic theory, which also helped when interpreting the results. Table 8.5 presents a list of the hypotheses posed by the literature and the data, and tested in the analyses. For example, economic theory suggested that substitutes for hospital care such as family and community support should be included in the analysis and the results showed that the utilisation of these alternatives to hospital care were associated with reduced LOS. The specification of an appropriate statistical model allowed the relationship of each variable with LOS to be correctly estimated and interpreted.

The results suggested that the proportion of GDP spent on health care (% GDP/health) was the only higher-level variable associated with total cost. However, this variable was not associated with the main resource use measure, LOS. The reason why % GDP/health was associated with total cost could be that this variable was correlated with the factor prices used to value the resource inputs. In the base case analysis these factor prices were converted into US$ using the GDP PPP index, which left residual price differences across the centres. To consider whether the association of % GDP/ health and total cost depended on the factor used to convert factor prices into US dollars, the analysis was repeated using a stroke care PPP index. Converting relative

prices using this index left only small differences in relative prices across the countries concerned. The results of this sensitivity analysis suggested that even when the prices used were relatively homogenous, the % GDP/ health variable was still significantly associated with total cost. This suggests that the main way in which % GDP/ health was associated with total costs was through its association with overall physical resource use, rather than through the correlation with relative factor prices. The % GDP/ health variable may therefore provide a reasonable starting point for deciding whether resource use and total costs are similar across centres.

The % GDP/ health was included as a variable to provide a more general measure of health infrastructure in each country. For example, those centres with higher national levels of health care spending generally had more hi-tech equipment available, offered further rehabilitation options to stroke patients, and had neurologists leading care. It is likely the % GDP/ health variable summarises overall differences in the production processes across the centres, and could represent the effects of other variables not included in the analysis that summarise the different models of stroke care that exist across the centres concerned (Table 8.5).

In a database with 13 centres it was not possible to use enough centre-level variables to identify which particular differences in the health care infrastructure across centres were associated with total cost differences. Nor was it possible to identify other centre or national-level variables that were associated with resource use and total cost. Further studies in different disease areas are required to extend this analysis and examine what other factors may, in different contexts be associated with resource use and cost variation. The analyses presented here could be expanded to define further levels of the hierarchy. For example, there is an economic rationale to suggest that patients in different wards within a centre, may face different incentives for discharge from hospital. Also, contextual factors may differ across health care centres within a country. In an international dataset with more centres, the approach presented here could be expanded to investigate which factors drive cost variation at a patient, centre, national or international level. Alternatively, the issue of cost variation could be explored in a national cost-effectiveness study, looking at variability amongst

different patients and health care providers. While the particular factors driving cost variation will differ according to the disease or the context, the methodological issues raised here would apply to any situation where the data may be regarded as hierarchical.

In this dataset, there was some residual variation in resource use and costs across the health care centres. This may reflect some of the differences highlighted in the literature review, such as technical, productive or scale efficiency or level of uptake of new technology. However, the intra-class correlation coefficient (ρ), which reports the proportion of the variability at the centre-level, was relatively low (0.16). This shows that most of the residual variation was at the patient level and suggests that unmeasured differences in patient characteristics (such as unmeasured case-mix, socioeconomic status, or morbidity) may be important in explaining overall variation. Whilst ρ is useful for understanding the degree to which variability exists at different levels, it should not be used as a basis for accepting or rejecting the use of the MLM approach. Indeed the relative levels of unexplained variation may be conditional on the number of centres versus number of patients included in the analysis. As is clear from this study, OLS methods can be severely misleading even when ρ is small.

Previous MLMs in health economics, in health services research and in other sectors such as education have assumed that the residuals are normally distributed (Carey 2000, Goldstein 1995). This assumption is unlikely to hold when the variable of interest is total cost per patient. Although generalised linear models (GLMs) have been recommended and used for analysing cost data (Manning and Mullahy 2001, Barber and Thompson 2004), they have not previously been evaluated in hierarchical models for comparing costs across health care settings. This study demonstrated their use in analysing hierarchical cost data. The GLMM random intercept model with the gamma distribution fitted the data better than the MLM random intercept model which assumed a normal distribution. This suggests that GLMMs are an attractive alternative for analysing hierarchical resource use and cost data.

Economic evaluations are often criticised for lacking generalisability (O'Brien 1997), particularly as costs may differ between locations (Hoch et al. 2002). Despite this, economic evaluations rarely examine why costs vary across settings and the few studies which have considered this issue have relied upon OLS analysis (Willke et al. 1998, Coyle and Drummond 1998). These results indicate that those studies which used OLS analyses to identify factors associated with resource use and cost differences, may have reached erroneous conclusions (Coyle and Drummond 1998). Other studies have used OLS to analyse variation in costs across centres (Willke et al. 1998). While this approach allows valid conclusions to be drawn about the magnitude of cost variation across centres, it does not consider *why* costs and cost-effectiveness vary across health care settings. Similarly, while tests for interactions can assess whether there is significant variation across centres (Cook et al. 2003), they are not suitable for assessing the reasons for cost variability. MLMs allow more correct inferences to be made about which factors are associated with cost variation. This can help decision-makers assess the applicability of results to their local setting. In this study the higher-level variable which explained most cost variation was the national level of spending on health care. This variable could therefore be used, alongside patient-level variables, to predict stroke costs for centres outside the study.

Previous multicentre costing studies have struggled to assess cost variation partly because unit costs have not been measured in each centre, and the costing methodology used has differed by location (Schulman et al. 1996). Although the case-study described was not a full cost-effectiveness analysis, it was chosen for this evaluation as it enabled cost variation to be carefully assessed. In particular, detailed resource use, unit costs and outcomes were collected in each centre using a consistent methodology. Thus any differences between the centres in observed cost, would be more likely to reflect real differences in resource use, and unit cost rather than differences in the methodology used. The numbers of patients and centres were reasonably similar to the numbers which could be recruited to an international economic evaluation, which made it a realistic setting for testing the usefulness of MLM versus OLS for assessing cost variation. MLMs could be applied to multicentre (national or international) economic evaluations, where the net-benefit is reported for

individual patients, and the incremental net-benefit of the new treatment is estimated using regression analysis (Manca et al. 2005, Hoch et al. 2002).

8.6 Conclusions

This chapter compared the use of MLMs with OLS regression models for identifying which factors were associated with resource use and cost variations across health care settings. By ignoring the hierarchical structure of the data the OLS models underestimated the standard errors associated with the higher-level variables and therefore overestimated their precision. MLMs correctly recognised the variance structure and found that none of the higher-level variables were associated with LOS. The only higher-level variable associated with total cost was the % GDP/health variable, which was a proxy for differences in health care infrastructure across the international settings concerned. The % GDP/health variable was still associated with total costs, when international differences in factor prices was adjusted for using the stroke specific PPP index. Patient factors were important in explaining variation in both LOS and total costs. However, the MLMs suggested that most of the variation left unexplained by the models was at a patient-level. The gamma model allowed for the skewed distribution of the data, and fitted the data best. The next chapter extends the comparison of MLMs and OLS regression analysis to estimate cost-effectiveness across different health care settings.

Table 8.5: Summary of hypotheses posed and tested for understanding why resource use, unit costs and total costs vary across settings.

Hypotheses from literature	Hypotheses from data	Hypotheses 'tested' in analyses
1. Production/ cost function		
Presence of care alternatives may explain different RU	Higher use of community or family support: shorter hospital stay	Substitution of community for hospital inputs
Differences in Technical inefficiency: differences in resource use	Differences in bed occupancy: differences in technical efficiency	Not tested, may be part of unexplained variation at centre level
Scale efficiency: Economies of scale in hospital production realised up to 300 beds	Provider units of similar size no hypotheses generated	Not tested
Differences in factor prices: differences in factor use	Wide variations in factor use and factor price	Variations in factor price not found to be association with factor use, or resource use differences
Larger providers may achieve economies of scale.	Provider units of similar size	Not tested
2. Patient factors		
More complex case-mix associated with higher resource use and/or unit costs.	Case-mix differs across centres	Case-mix found to be significantly associated with resource use and cost differences

3. Centre or national factors		
Presence of DRG system: lower RU.	Centres with DRG system: lower hospital RU	Presence of DRGs not associated with lower hospital LOS/ cost.
Use of patient copayments: lower RU	Centres with patient copayments: lower RU	Presence of copayments not significantly associated with LOS/ cost.
Labour market restrictions, may lead to variations in factor use and factor price	Factor prices and factor use varied widely between settings	Inflexibilities in labour market would appear important for understanding why factor price, and potentially factor use differences persist
Higher national spending on health care: improved level of health care infrastructure, better access to new technology	Higher national spending on health care, more resource use in individual centres	Centres with higher spending on health care, more resource use, could also reflect differences in productive inefficiency, and/or differences in quality of care
Better levels of health care infrastructure- better access to neurologists	Higher proportion using neurologists, higher RU	Centres with higher access to neurologists, higher RU
Better levels of health care infrastructure better access to rehabilitation hospitals	Better access to rehabilitation hospitals higher overall resource use	Centres with higher proportion using rehabilitation hospital, higher RU

RU: Resource use

4. Measurement issues		
Case-severity measures in aggregated production function studies inadequate, so observed differences inefficiency could reflect case-mix differences	Case-mix collected at disaggregated level anticipated to capture inter-patient differences	Using patient level case-mix variables proved helpful. Most outstanding differences still at patient level.
Random variation potentially important at patient or setting level	Wide variation within and across centres needs consideration.	Still high levels of unexplained variation even using appropriate technique/ disaggregated dataset
Aggregated measures fail to adjust for differences in quality of inputs or outputs	Disaggregated method would provide description of potential quality differences	Difficult to incorporate into analysis without data on outcomes/ more centres, suggestion though that centres with high GDP on health care, better quality should not be mistaken for inefficiency.
Disaggregated methods required to enhance methodological comparability	Variation in methods used to allocate overheads, may explain differences in unit costs	Tried to minimise methodological inconsistency, but some outstanding centre-level differences could reflect costing method
Choice of conversion factor requires consideration	Results could be sensitive to choice of conversion factor	Conclusions robust to choice of conversion factor

Chapter 9: Cost-effectiveness analysis

9.0 Introduction

The previous chapter suggested that considerable variation in health service costs can exist across health care settings, and that multilevel models (MLMs) can be more appropriate for identifying which factors are associated with cost variation across settings than ordinary least squares (OLS) regression models. This chapter compares the use of OLS models and MLMs for analysing multicentre cost-effectiveness data. These data may be hierarchical with patients clustered within centres, and centres clustered within countries. In this context, MLMs may be more appropriate than OLS models as their use may be more consistent with economic and statistical theory (see Chapters 4 and 5). This chapter assesses whether in practice the choice of statistical technique can have an important impact on the estimation of cost-effectiveness.

Previous studies have considered whether the relative cost-effectiveness of various interventions varies across health care centres located in different countries (Willke et al. 1998, Cook et al. 2003). Cook et al. (2003) found a lack of heterogeneity or variation across countries and concluded that pooling cost-effectiveness estimates across countries may therefore be appropriate. By contrast, Willke et al. (1998) found important differences in mean ICERs across the countries included in their study, and this limited the usefulness of just reporting an overall estimate of cost-effectiveness. However, both these studies lacked the statistical power needed to detect cross-national differences in cost-effectiveness, and neither study used MLMs to assess the variation in cost-effectiveness.

This chapter compares the use of OLS models with MLMs for analysing international cost-effectiveness data. As a suitable multicentre cost-effectiveness dataset is not available, the Biomed II dataset is extended to generate a multicentre cost-effectiveness dataset. It is recognised that by generating a dataset, it is not possible to identify factors that are associated with variation in cost-effectiveness. Instead, the aim of this chapter is to consider the potential impact of the choice of statistical technique on the results of a multicentre cost-effectiveness analysis. As with the cost analysis in chapter eight, the dataset covers 13 centres in 10 different countries. As there are insufficient centres in each country to have three levels in the MLM, two levels are used with patients nested within centres. The subsequent commentary discusses variability within and across centres. The context is multinational but the statistical principles could also be applied to cost-effectiveness data collected as part of a national multicentre study.

This chapter is divided into five main sections, the first provides an overview of the relevant statistical issues involved in estimating cost-effectiveness in this context, the second describes the development of the dataset, the third describes the statistical models used, the fourth compares the results of the OLS models and MLMs for assessing cost-effectiveness, and the final section discusses and interprets the results.

9.1 Relevant statistical issues in cost-effectiveness analysis (CEA)

The previous chapter compared the use of OLS models with MLMs for analysing which factors were associated with LOS and total costs. These techniques are now compared for the analysis of cost-effectiveness data, which raises further statistical issues. In CEA, recent developments in the methodological literature have seen incremental net benefits (INB) and cost-effectiveness acceptability curves (CEAC) become the preferred measures of cost-effectiveness (Fenwick et al. 2004, NICE 2004), and both these methods of presenting the results of CEA are used in this chapter. Net benefits are most commonly

estimated on the costs scale and are then termed net monetary benefits (NMB) and are defined by Hoch et al. (2002) as:

$$NMB_i = \lambda E_i - TC_i$$

where λ is the societal willingness to pay for a unit of health gain, E_i is the effectiveness of the intervention for individual *i*, TC_i is the total costs for individual *i*. When costs and outcomes are measured for each individual in an RCT, the NMB for each individual can therefore be calculated (Hoch et al. 2002). An OLS regression analysis can be used, with a dummy variable to estimate the effect of the intervention on net benefits-- the INB. The mean estimate of INB using this regression approach is equivalent to estimating the mean INB by estimating the difference between the mean NMB for the treatment group and the mean NMB for the control group in a standard net benefit analysis (Hoch et al. 2002).

9.11 Using total or aggregated OLS regression analysis to estimate cost-effectiveness

The main purpose of CEA is to provide the decision-maker with a relevant estimate of the cost-effectiveness of the intervention, i.e. one that can inform the resource allocation decision. A national decision-maker requires information that is relevant to the country concerned. However, the economic evaluation may be reliant on data collected from a multicentre RCT that recruits patients from several countries. The study may not have collected resource use data for each patient, or unit cost data in each health care setting. In this case, an aggregated regression analysis may be used to estimate an average measure of cost-effectiveness (e.g. the INB) across all the centres and countries in the study (see for example Johannesson et al. 1997). This approach has the advantage of maximising the power in the study to detect a positive INB, and produces a relatively precise estimate. The problem with this approach is that the overall estimate of cost-effectiveness may not be relevant to a national decision-maker. This approach ignores any variability in the mean or the distribution of the INB across the different centres or countries. This could lead to biased estimates of the INB, and the uncertainty that

surrounds it. Other techniques that can recognise the variability in the mean INB across different countries or centres therefore need considering.

9.12 Centre-specific OLS regression analysis

If resource use data for each patient and unit cost data for each centre are collected, then separate OLS analyses can be used to estimate INBs for each centre. While this approach provides specific unbiased estimates of the mean INB and its standard error, the analysis may lack power, and there may be insufficient cases in a particular centre or country to detect whether or not the mean INB is positive. It may also be regarded as statistically inefficient to base a decision on information collected solely from the setting most relevant to the decision-maker, without regard to the results in other centres, with potentially more cases. In reality, decision-makers are likely to be influenced by results in other contexts, which this model fails to acknowledge.

9.13 Fixed effects meta-analysis

To increase the statistical power needed to detect differences in NMB between the treatment and control groups in a multicentre RCT, it may be desirable to firstly estimate a mean INB for each centre, and then secondly to pool these results to estimate an overall INB using a fixed effects meta-analysis (Fleiss 1993). This increases the power of the analysis to detect differences between treatment and control groups. This method unlike the total OLS regression model recognises any differences in the distribution of the INB across the centres. However, both the total OLS model and the fixed effects meta-analysis assume that there is no variability in the mean INB across the centres, i.e. that there is no heterogeneity.[53]

---

[53] Even though using the separate OLS models does allow for different treatment effects in each centre, this is not strictly heterogeneity as the different centre effects are not drawn from a distribution.

9.14 MLMs

The assumption of no heterogeneity across the centres can be relaxed by moving to a MLM. A MLM can account for variability in the mean INB across centres by using the 'random slope model' described in detail in chapter five. The random slope model allows a regression coefficient representing the effect of the intervention to vary randomly across the study centres. The MLM can either assume a common patient-level variance across the treatment centres, or each centre can be allowed to have a different variance. The MLM with different variances can be estimated in one stage, or alternatively could be estimated in two stages using a random effects meta-analysis (Fleiss 1993, Thompson and Sharp 1999). This approach is the same as under the fixed effects meta-analysis, with centre specific estimates derived first (stage one), but these effects are then combined, allowing for variability in the INB within and across the centres (stage two). Random effects meta-analyses are most commonly used to pool effectiveness data from individual studies measuring treatment effects (Thompson and Sharp 1999), but the technique can be used to pool data from different centres (Localio et al. 2001). The random effects meta-analysis or two stage model, is similar to the one-stage MLM with different variances in each centre described above (Higgins et al. 2001).

Shrinkage estimates may be used in MLMs to improve the statistical efficiency of the individual centre's estimates. The use of shrinkage estimators aims to combine the advantages from using the overall mean estimate (additional power and using all information) and the centre-specific estimate (relevance). While shrinkage estimates may be appealing for the reasons outlined, they assume that the data are exchangeable (Spiegelhalter et al. 2000) (see Chapter 5 for a further discussion). In this context, this assumes that the INB for all the centres concerned are drawn from the same distribution i.e. there are no systematic differences in the INB across the centres. However, the evidence from the theoretical review and the cost analysis disputes this assumption (see Chapters 4 and 8). To address this, MLMs may use covariates to adjust for systematic differences across the centres, before using shrinkage estimates to give centre-specific

estimates. Again, this may be done in two ways, either as part of a one-stage MLM or using a meta-regression analysis (two stages). In a meta-regression analysis, the dependant variable in this context is the INB and centre or national-level factors can be included as independent variables to adjust for systematic differences across the centres.

Section 9.3 formally defines these alternative models for analysing cost-effectiveness data, but firstly the methodology and data used to generate the cost-effectiveness dataset are presented.

9.2 Development of the multicentre cost-effectiveness dataset

The original dataset recorded the costs and outcomes of routine stroke care for 1757 cases in 13 centres in 10 countries (see chapter 6). The dataset did not include an appropriate intervention for conducting a cost-effectiveness analysis. Instead, in this chapter the original data from the observational study are used to generate an international cost-effectiveness analysis. A hypothetical 'treatment' is defined by making several assumptions based on the results from the cost analysis (Chapter 8) and using the relevant literature on the effectiveness of interventions for stroke patients (Stroke Unit Trialists' Collaboration 2001). These assumptions are used to create a scenario that could arise when an economic evaluation is conducted alongside a multicentre RCT. The main features of this scenario are that the effectiveness of the 'intervention' is assumed to be similar across the centres, and it is assumed the intervention costs are the same for each patient and centre. In Chapters 6-8 the morbidity costs of stroke care were found to differ widely across the health care centres. It is therefore assumed that the impact of the intervention on morbidity costs varies across patients and centres.

The main methodological standpoints are the same as in the observational study (see Chapters 6 and 7). Once again, a hospital and community health service perspective is taken to cost measurement. The cost-effectiveness analysis includes cost and outcomes up to three-months post-stroke, as this is the time-horizon over which these data were recorded. As information was collected on survival but not on health-related quality-of life (HRQoL), the outcome measure used in the CEA is life-years gained.

9.21 Definition of the 'control' and 'treatment' groups and assumptions about the treatment's effectiveness.

The cost-effectiveness dataset is based on the original data from the 1757 patients included in the observational study (Chapter 6). It is assumed that 50% of these patients (n=878) have the costs and outcomes observed in the original empirical study. These patients therefore represent the 'control' group or routine practice in each centre. The remaining patients (n=879) are assumed to have the intervention. This hypothetical treatment is defined broadly as a new pharmaceutical intervention for stroke that reduces mortality and morbidity. It is assumed that the treatment group has a reduction in all cause mortality at three months of 34%, i.e. that the odds ratio of death for treatment versus control is 0.66. The odds ratio is similar to that observed from comparing post-stroke mortality following care on a specialised stroke unit versus care on a general medical ward (Stroke Unit Trialists' Collaboration 2001). It is also assumed that the intervention reduces morbidity. However, in the absence of a suitable measure of morbidity, this aspect of the intervention's impact is not captured by the effectiveness measure, but it is included in the cost analysis (see section 9.23).

9.22 Sampling from the observational data to achieve the assumed reduction in mortality.

The original observational dataset with costs and outcomes recorded for 1757 cases is divided into two sub-samples, a 'treatment' group and a 'control' group. To achieve the lower mortality rate in the treatment group, more of the cases in the observational study

who are alive at three months are allocated to the treatment than to the control group. Similarly, a higher proportion of the cases that died are allocated to the control group. The numbers of cases according to vital status who need to be allocated to each group, to achieve the odds ratio of 0.66 in favour of treatment are given in Table 9.1.

Table 9.1: 3-month mortality (N=1757) following treatment and control. N (%)

	Dead	Alive
Control	264 (30)	614 (70)
Treatment	195 (22)	684 (78)
Total	459 (26)	1298 (74)

Odds ratio (95% CI) for 3-month mortality for treatment versus control: 0.66(0.53 to 0.82).

The required numbers of cases are sampled for each group at random, from the cases in the observational dataset. The random sampling is over the whole cohort, with no stratification by centre. Each case could potentially be allocated to either treatment or control. The only constraints on the random sampling are that it has to achieve an odds ratio in favour of the treatment group of 0.66, and there have to be 879 patients in the treatment group, and 878 patients in the control group. This produces a single dataset with the required odds ratio for mortality, and equal numbers of patients in the treatment and control groups. For both the treatment and control groups, the mortality data are those observed in the observational study.

Repeating this random sampling with the same constraints leads to different combinations of patients being allocated to the two groups, and would lead to different eventual estimates of cost-effectiveness. A concern is that the ensuing statistical analysis could be based on samples that have outlying observations and are 'unrepresentative' of the likely combinations of patients in these two groups. The random sampling is therefore repeated

100 times to produce different 100 datasets. The dataset chosen for the analysis will be the one that has the mean measure of cost-effectiveness (see Section 9.24).

9.24 Costing Assumptions

The intervention itself is assumed to have a cost of $3,000 across all the patients and centres. It is recognised that making this assumption of uniform cost will lead to an underestimate of cost variability compared to an actual intervention. However, it is also assumed that the intervention reduces morbidity costs. For the 'control' group the CEA assumes that costs are the same as in the observational study but for each case in the treatment group, the baseline morbidity costs are reduced by 40%[54]. The rationale for this is that a successful intervention for stroke may result in less disability in addition to reducing mortality (Samsa et al. 1999), so assuming a reduction in morbidity costs would seem plausible. Previous international cost-effectiveness studies in other disease areas have suggested that those countries with higher absolute costs may experience a greater reduction in morbidity costs from an effective intervention, than those countries with lower baseline costs (Hull et al. 1981, Menzin et al. 1996). As the morbidity costs in the observational study varied across settings assuming that the intervention leads to a proportionate reduction in morbidity costs will lead to differences in the incremental costs of the intervention across the centres.

9.24 Calculation of life years, net-benefits and incremental analyses

The number of life years for the decedents is calculated by subtracting the date of stroke from the date of death, to give the number of days survived, and dividing this by 365. As the data are censored at three months post-stroke, patients alive at 90 days are not assumed to live any longer; these cases are therefore assumed to live for 90/365=0.25 life years. Similarly, no morbidity costs are included after three months post-stroke. By ignoring any impact the intervention has after three months, the CEA will underestimate

[54] This 40% reduction in morbidity costs is applied to decedents as well as survivors to simplify the analysis. However, it is recognised that this reduction in morbidity costs is less likely for patients who died.

the cost-effectiveness of the intervention. Net monetary benefits (NMB) are calculated for each individual i, by valuing their life years (LY) by the societal willingness to pay for a life year (λ), and subtracting from this, the individual's total costs of care (TC):

$$NMB_i = \lambda LY_i - TC_i$$

where λ is the societal willingness to pay for a life year, LY_i and TC_i are the number of life years and total costs for individual i. λ is initially set at \$60,000 per life-year gained (LYG) across all the centres, though this value is varied in subsequent sensitivity analyses.

OLS regression analyses are initially used to estimate the incremental effect of treatment on life years, total costs and NMB. The regression analyses each have a dummy variable for treatment group as an explanatory variable, and life years, total costs and NMB respectively as dependent variables. Using these regression models, the incremental effectiveness, incremental costs and incremental net benefits (INB) of the new treatment are estimated for each of the 100 costs and effects datasets. The mean INB varied across these different datasets, as each dataset had treatment and control groups with different mixes of patients. The dataset with the mean INB over the 100 datasets is therefore chosen for all the ensuing analyses. This choice was made to avoid basing the subsequent statistical analyses on a dataset with outlying observations.

9.25 Description of the cost-effectiveness dataset.

The cost-effectiveness dataset generated had a reasonably even distribution of treatment and control cases across the 13 centres, and could reflect the distribution of cases in a non-randomised treatment comparison (Table 9.2).

Table 9.2: Numbers (%) of treatment and control cases in each centre in the cost-effectiveness dataset.

Centre	N	Control	Treatment
Portugal	108	53 (49)	55 (51)
Spain	49	19 (39)	30 (61)
Italy	136	68 (50)	68 (50)
France	132	73 (55)	59 (45)
Denmark	297	160 (54)	137 (46)
Finland 1	95	52 (55)	43 (45)
Finland 2	71	32 (45)	39 (55)
Finland 3	43	20 (47)	23 (53)
UK	108	51 (47)	57 (53)
Poland	149	74 (50)	75 (50)
Lithuania 1	80	43 (54)	37 (46)
Lithuania 2	237	104 (44)	133 (56)
Latvia	252	129 (51)	123 (49)
All cases	**1757**	**878 (50)**	**879 (50)**

As the patients were selected for the treatment and control groups based on mortality, the case-mix for the treatment group was slightly less severe than for the control group (Table 9.3). For example, 39% of patients in the treatment group were incontinent at hospital admission compared to 42% of patients in the control group.

Table 9.3: Baseline Characteristics in treatment and control groups.

	Control (N=878)	Treatment (N=879)	All cases (N=1757)
Unconscious	212 (25%)	152 (17%)	364 (21%)
Incontinent	371 (42%)	345 (39%)	716 (41%)
Paralysis	706 (80%)	200 (80%)	351 (20%)

The treatment was associated with an overall gain in life years at a small but non-significant additional cost. The mean INB when λ was \$60,000 per LYG was therefore positive and statistically significant at the 95% level (Table 9.4).

Table 9.4: Life-years, total costs and INB, over 3-months post-stroke.

	Control Mean (sd)	Treatment Mean (sd)	Difference (tr-cont) Mean (95% CI)
Life years	0.19(0.10)	0.20(0.09)	0.016 (0.007 to 0.024)
Total costs (\$)	4,362(5,427)	4,733(3,268)	371 (-48 to 789)
ICER (\$/life year)			23,187
Overall INB (\$)			582 (23 to 1142)

(λ=\$60,000 per LYG)

To summarise, a cost-effectiveness dataset was generated, based on the observational stroke dataset. The important features of the cost-effectiveness dataset were that it assumed that the intervention's effectiveness was similar across the centres, and that the intervention costs were the same. However, it was assumed that the intervention led to a 40% reduction in morbidity costs, so that the impact of the intervention differs across individuals and centres. The overall mean INB estimated using OLS regression analysis

was positive and statistically significant. The next section defines the different OLS models and MLMs for estimating the relative cost-effectiveness of the intervention.

9.3 Statistical models used in the CEA

Each of the five models used in the CEA are defined below:

9.31 Model 1: Aggregated or total OLS regression model

In the most basic model, OLS regression can be used to estimate a single mean INB across all individuals and centres:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i ; \quad \varepsilon_i \sim Normal(0,\sigma^2) \text{ (model 1)} \quad i=1....1757 \text{ individuals}$$

where y_i is the NMB for the ith patient, x_i is an explanatory variable for treatment group, with the slope coefficient β_1, and β_0 is the intercept term. β_0 is a fixed effect that represents the NMB in the control group and β_1 represents the INB from treatment. This model is sometimes known as an aggregated or a total regression model (Kreft and De Leeuw 1998). This model assumes that the INB is the same across all individuals and centres. The model therefore assumes that there is a common variance across the centres and no variability in the mean INB across the centres.

9.32 Model 2: Centre-specific OLS models and fixed effects meta-analysis

Using separate OLS models for each centre allows different mean INB and variances to be estimated for each centre (stage 1).

Thus for each *j*th centre each model provides an average NMB for the control group: β_{0j}, an average effect of treatment (the INB): β_{1j}, and a centre-specific measure of the variance surrounding the residual: σ_j^2.

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_i ; \quad \varepsilon_i \sim Normal(0, \sigma_j^2) \quad \text{(model 2)}$$

To estimate an overall effect a second stage may be used in the model, with the centre-specific estimates pooled across the centres using a fixed effects meta-analysis (stage 2). In this context a fixed effects meta-analysis assumes that the true INB in each centre is the same. To estimate the overall mean INB the means for each centre ($\hat{\beta}_{1j}$) are weighted (w_j) according to the inverse of their variance (V) to give the overall mean estimate- β_1^F

$$\beta_1^F = \frac{\sum w_j \hat{\beta}_{1j}}{\sum w_j} \text{ where } w_j = 1/V(\hat{\beta}_{1j})$$

Those centres with a smaller variance surrounding their mean INB, are therefore given a higher weighting in calculating the pooled or overall mean INB. Therefore compared to the overall estimate from using the aggregated OLS, the results from the fixed effects meta-analysis move the overall mean towards those centres with smaller variances[55].

9.33 Model three: MLM with random slopes but common within-centre variance.

A MLM can account for variability in the INB across centres by using a 'random slope model'. The random slope model allows the regression coefficient representing the treatment effect or mean INB β_1, to vary randomly across the study centres. This requires that an additional error term, u_j is included, which represents a different slope or treatment effect for each centre (see Chapter 5 for a detailed explanation). The mean INB for each centre is therefore $\beta_1 + u_j$ where β_1 is the overall mean INB and u_j the mean

[55] Note, in this context the difference with the overall estimate may be a subtle one, as the aggregated estimate will partly reflect the numbers of cases in each centre, and those centres with more cases may also be the centres with smaller variances.

difference in the INB for the individual *j*th centre. u_j is assumed to be normally distributed with a mean of 0 and variance τ^2. Just as with the random intercepts model (chapter 8) interest is in the size of the between-centre variance or level of heterogeneity, τ^2.

$$y_{ij} = \beta_{0j} + (\beta_1 + u_j)\, x_{ij} + \varepsilon_{ij}\,; \quad \varepsilon_{ij} \sim Normal(0,\sigma^2)\; u_j \sim Normal(0,\tau^2) \text{ (model 3)}$$

This MLM assumes that the mean INB for each centre is sampled from a normal distribution and that this distribution represents the INB across all possible centres. Each centre is assumed again to have a different fixed intercept β_{0j}, but a common variance σ^2. Shrinkage estimates can be used to shrink the centre-specific estimates in towards the overall mean (see Chapter 5).

9.34 Model four: MLM with random slopes and different variances in each centre, estimated by either one-stage MLM, or two-stage MLM (random effects meta-analysis).

MLMs do not necessarily have to assume a common patient-level variance across the treatment centres as in Model 3, and instead each centre can be allowed to have a different variance. This can be estimated in the one-stage model below where σ_j^2 is the variance for the *j*th centre.

$$y_{ij} = \beta_{0j} + (\beta_1 + u_j)\, x_{ij} + \varepsilon_{ij}\,; \quad \varepsilon_{ij} \sim Normal(0,\sigma_j^{\,2})\; u_j \sim Normal(0,\tau^2) \text{ (model 4)}$$

Alternatively, the same model can be estimated in two stages. In the first stage, the mean INBs (and their standard errors) are estimated for each centre. In the second stage, the mean estimates from each centre, are weighted to take into account the level of within centre, and between-centre variability (DerSimonian and Laird 1986) and then combined using the formula below to give an estimate of the overall random effect:

$$\beta_1^R = \frac{\sum w_j \hat{\beta}_{1j}}{\sum w_j}$$

where β_1^R is the overall random effects estimate, $\hat{\beta}_{1j}$ is the effect in the *j*th centre and w_j is the weight given to the individual centre's INB. The weighting for each centre- w_j is given by the inverse of the within centre and between-centre variances:

$$w_j = \frac{1}{\sigma_j^2 + \tau^2}$$

This two-stage model is similar to the one-stage MLM described above (Higgins et al. 2001). Shrinkage estimates are again used to provide estimates of the centre-specific mean INBs and their standard errors.

9.35 Model five: MLM with different variances within and across centres with national-level covariates, estimated in either one stage or in two stages (meta regression).

Model four can be expanded to include additional covariates to identify which factors are associated with variability in the INB. It is no longer assumed that the "true" estimates of INB are drawn from the same distribution with no systematic differences across the centres. Instead, the covariates included can be used to adjust for systematic differences across the centres. The extent to which adding national-level covariates can reduce the between-centre variability and improve the fit of the model can be investigated. The choice of national-level variable for this analysis was driven by the literature review and by the empirical costing work, which suggested that the proportion of GDP spent on health care (% GDP/health) was associated with variability in costs. The % GDP/health

variable was included as a categorical variable (low, medium and high %GDP/health)[56]. This allows cost-effectiveness to be analysed and presented for centres that fall into different strata of GDP spent on health care. Two categories of % GDP/health were then included as dummy variables, excluding a third, as the reference case.

To investigate the effect of %GDP/health on the INB, an interaction with treatment was included for each of the %GDP/health dummy variables, as in the equation below:

$$y_{ij} = \beta_{0j} + (\beta_1 + u_j)\, x_{ij} + \beta_2\, x_{ij}\, z_{ij} + \beta_3\, x_{ij}\, w_{ij} + \varepsilon_{ij} \text{ ; (model 5)}$$

$$\varepsilon_{ij} \sim Normal(0, \sigma_j^2)\ u_j \sim Normal(0, \tau^2)$$

where $x_{ij}=1$ if treated, 0 if control

where $z_{ij}=1$ if medium % GDP/health, 0 otherwise

where $w_{ij}=1$ if high % GDP/health, 0 otherwise

y_i represents the NMB across all cases, β_{0j} the intercept term is the average NMB in the control group with the reference level of %GDP/health, that is low % GDP/health. β_1 represents the INB for the low %GDP/health group, and the INB is allowed to vary randomly across the centres. β_2 represents the additional effect of the medium % GDP/health group on INB, and β_3 represents the additional effect of high % GDP/health on INB.

[56] The % of GDP spent on health care was defined as low when <5.1%, medium between 5.1% and 5.5%, and high when % GDP/health >5.5. It is recognised that re-defining a continuous variable as a categorical variable is a simplification.

As with the previous MLM, a between-centre variance (u_j) and an individual variance term (ε_{ij}) that is allowed to differ across the centres, are included in the model. The variance τ^2 now represents the residual variability in the NMB across the centres after adjusting for the covariates.

This investigation of the effect of national-level covariates can also be done in two stages using a meta-regression model. Firstly, the mean and standard error of the INB are estimated in each centre, and secondly these estimates are used in a regression model with the mean INB in each centre as the dependant variable and the national-level variables as explanatory variables. Again, the two-stage meta-regression model should produce similar results to the one-stage MLM.

9.4 Estimation

OLS models and MLMs were fitted to estimate the effect of treatment on net monetary benefit- the INB. Treatment was defined as a dummy variable, with no treatment the reference category. The first OLS model was the aggregated or total regression model that estimated the effect of treatment over all cases and centres (model 1). The analysis was repeated fitting separate OLS regressions for each centre (model 2). The centre-specific INB and standard errors were pooled in a fixed effects meta-analysis to give an overall mean INB (model 2).

Two MLMs (models 3 and 4) were fitted with normal random slopes representing variability in the treatment effect, across the different centres. Both models were fitted with a dummy variable for each centre, to give a different, but fixed intercept in each centre. Model 3 was fitted with a common variance across the centres and model 4 was fitted with a different variance for each centre. Model 4 was also estimated as a two-stage

model by using the centre-specific estimates from model 2 in a random effects meta-analysis. Model 5 was an extension of Model 4 and included interaction terms to examine the differential effect of %GDP/health on the INB (Model 5). This analysis was again repeated as a two- stage model by using the estimates from model 2 in a meta-regression (Model 5). In addition, to reporting the overall mean INB for each MLM, the centre-specific estimates of INB and their standard errors were reported using shrinkage estimates (models 3-5).

The one-stage MLMs were estimated in MLwiN by restricted iterative generalised least squares (RIGLS), equivalent to restricted maximum likelihood (Rasbach et al. 2002), and using Markov chain Monte Carlo (MCMC) (Gilks et al. 1996). For each model residual deviances were reported which are equivalent to reporting -2*log likelihood and provide a measure of model fit.[57] Where one model was nested within another, the goodness of model fit was compared using a likelihood ratio test based on the difference in the residual deviance. The two-stage models were estimated by restricted maximum likelihood estimation in Stata 8.1 (Stata 2002). The DerSimonian and Laird test for heterogeneity was applied to the fixed and random effects meta-analyses (DerSimonian and Laird 1986).

Finally, models 1-5 were re-fitted for a range of ceiling ratios; λ was allowed to vary from \$0 to \$100,000 per life year gained (LYG). This enabled cost-effectiveness acceptability curves (CEACs) to be drawn that plot the probability that the intervention is cost-effective for different levels of the ceiling ratio (Fenwick et al. 2004). The probability that the intervention was cost-effective was estimated as 1-p/2, where p was the 2-sided value from the coefficient estimating the INB in each model (Hoch et al. 2002). This is a Bayesian interpretation of probability i.e the probability that the intervention is cost-effective given the data (Briggs 1999). CEACs were presented for

[57] These measures of model fit are reported in preference to R^2 values as these were not available for the MLMs.

models 1 and 4, and the fixed effects meta-analysis based on the results of model 2. The models are summarised in the table below (Table 9.5).

Table 9.5: Summary of the five models

	Model 1 Aggregated OLS	Model 2 Centre-specific OLS	Model 3 MLM, same variance	Model 4 MLM different variances	Model 5 MLM different variances
Model	OLS	OLS	MLM	MLM	MLM
Intercept	fixed	fixed	fixed	fixed	fixed
Variance	common	different	common	different	different
Slope	common	different/common[58]	random	random	random
Shrinkage	no	no	yes	yes	yes
Two-stage option	no	yes, fixed effects meta-analysis[59]	no	yes, random effects meta-analysis	yes, meta-regression
National-level covariates	no	no	no	no	yes

[58] Different slopes were specified for the centre-specific estimates (stage 1), but then a common slope was assumed for the fixed effects meta-analysis (stage 2).
[59] The centre-specific estimates from model 2 are also used in the random effects meta-analysis (model 4).

9.6 Results

This results section starts by considering whether variability in INB within and across centres is an issue in this dataset, by presenting the results from models 1 and 2. The overall results of the first four models are then summarised and discussed. The CEACs are presented. Following this the impact of using shrinkage estimates on the centre-specific estimates of INB are described. Finally, the results from model 5, that uses a national-level covariate to try and adjust for systematic differences across the centres, are presented.

9.61 Description of INB using OLS models (models 1 and 2).

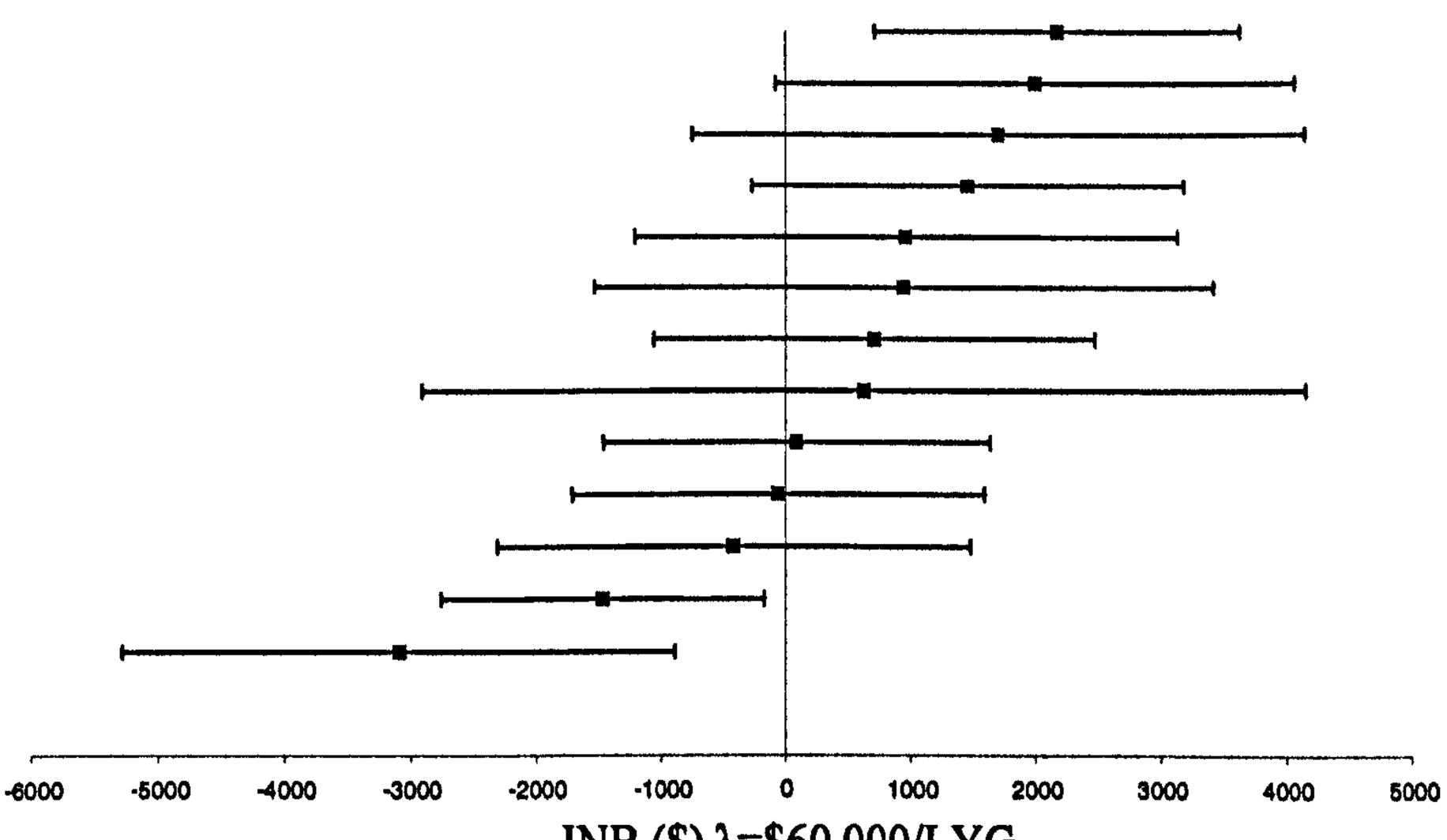
The results from the aggregated or total OLS model suggested that the overall INB was positive ($p<0.05$) (Table 9.6). However, the assumptions of no heterogeneity and common variance across all patients did not appear valid; heteroscedasticity was detected for the aggregated model ($p<0.001$ by the Cook-Weisberg test (Cook and Weisberg, 1982) and the distribution of the residuals differed widely across the centres (appendix 5). The histograms also suggest that in some centres the patient-level residuals did not follow a normal distribution (appendix 5). The results from the regression analyses for each centre (model 2) showed that both the mean INB and the standard errors differed across the centres (Table 9.6, Figure 9.1). For most of the centres the treatment was not associated with a significant gain or loss in NMB; the confidence intervals surrounding the centre-specific mean INB mostly overlapped zero (Figure 9.1).

Table 9.6: OLS models estimating the overall INB (model 1), and the centre-specific INB (model 2) [λ=$60,000/LYG]

Model	Centre	N	Estimate	(SE)
1	All cases	1757	582	(285)**
2	Portugal	108	959	(1109)
2	Spain	49	620	(1806)
2	Italy	136	-418	(966)
2	France	132	1457	(883)
2	Denmark	297	2177	(749)**
2	Finland1	43	2001	(1061)
2	Finland2	95	703	(902)
2	Finland3	71	941	(1264)
2	UK	108	1704	(1253)
2	Poland	108	85	(789)
2	Lithuania 1	80	-3089	(1123)**
2	Lithuania 2	237	-1468	(659)**
2	Latvia	252	-65	(842)

** p<0.05

Figure 9.1: Centre-specific INBs (95% CI) from model 2



9.62 Summary of overall estimates of INB for models 1-4

The overall results from the first four models are summarised below (Table 9.7). The results show that the overall INB from model 2, derived from the fixed effects meta-analysis, was smaller and less significant than the INB estimated by the total regression model (model 1). Model 1 gave most weight to those centres with the most cases, whereas the fixed effects meta-analysis (model 2) gave most weight to those centre estimates with the lowest variances. The centres with the lowest variances were mainly those centres with the lower mean INBs. The standard errors surrounding the overall estimates from both OLS models were relatively small compared to models 3 and 4 as they ignored the heterogeneity in the INB across the centres.

Table 9.7: Summary of the overall INB estimated by models 1-4

	Model 1 Aggregated OLS	Model 2 Centre-specific OLS	Model 3 MLM, same variance	Model 4 MLM different variances
Model	OLS	OLS	MLM	MLM
Intercept	fixed	fixed	fixed	fixed
Variance	common	different	common	different
Slope	common	different/common	random	random
Mean INB (SE)	582(285)	345(263)[60]	430(406)	425(403)
P value	0.04	0.19	0.29	0.29
τ^2			$11.6(11.0)*10^5$	$10.5(11.4)*10^5$
Goodness of fit Deviance (MCMC)	35,544	35,320	35,378	35,308

The MLMs (models 3 and 4) recognised the heterogeneity in the INB across the centres and therefore had much larger standard errors and p values for the overall mean INB, than either of the OLS models (Table 9.7). The overall mean INB for the MLMs was larger than estimated by the fixed effects meta-analysis (model 2), but smaller than from the aggregated OLS model. The MLMs estimates weighted each centre's estimate partly according to its own variance, so those centres with higher variance, mainly those with highest mean INBs, had their estimates down-weighted, so the overall mean INB was lower compared to model 1. However, the MLMs recognised that heterogeneity did exist across the centres, by weighting each centre's estimate according to the level of within and between-centre variance. The result was that when estimating the overall INB, the difference in the weighting given to each centre was smaller than for model 2. The down-

[60] The overall mean INB from model 2 is estimated using a fixed effects meta-analysis of the centre-specific results.

weighting of those centres with higher variances was therefore less extreme than for model 2 and therefore the effect sizes were larger.

Estimating the MLMs using either the one or two-stage approaches produced similar estimates of mean INB, and between-centre variability[61]. The estimates from the MLMs assumed that there were high levels of between-centre variance, and the DerSimonian and Laird test for heterogeneity suggested that there was significant heterogeneity that was ignored in the fixed effects meta-analysis, and OLS models (Test for heterogeneity Q=29.773, p=0.003). The comparisons of goodness of fit showed that the MLM that allowed for different variances in each centre (model 4) had lower levels of residual deviance and fitted the data best. A likelihood ratio test, comparing the residual deviance between models 2 and 4, showed that model 4 fitted the data significantly better (p=0.002)[62]. The OLS model (model 2) that allowed for different variances across the treatment centres had lower residual deviances and therefore fitted the data better than either the MLM or OLS models that assumed a common variance across the patients in different centres (models 3 and 1).

9.63 The impact of changing the threshold willingness to pay

The analyses were repeated for each model for different thresholds of the ceiling ratio λ. The CEAC plotted the probability the intervention was cost-effective against λ (Figure 9.2). The MLM that recognised that heterogeneity existed and allowed for different variances (model 4) was the model that made the most appropriate assumptions. Using the aggregate OLS (model 1) underestimated the standard error and overestimated the treatment effect; this model therefore overestimated the probability that the intervention was cost-effective at all realistic levels of λ.

---

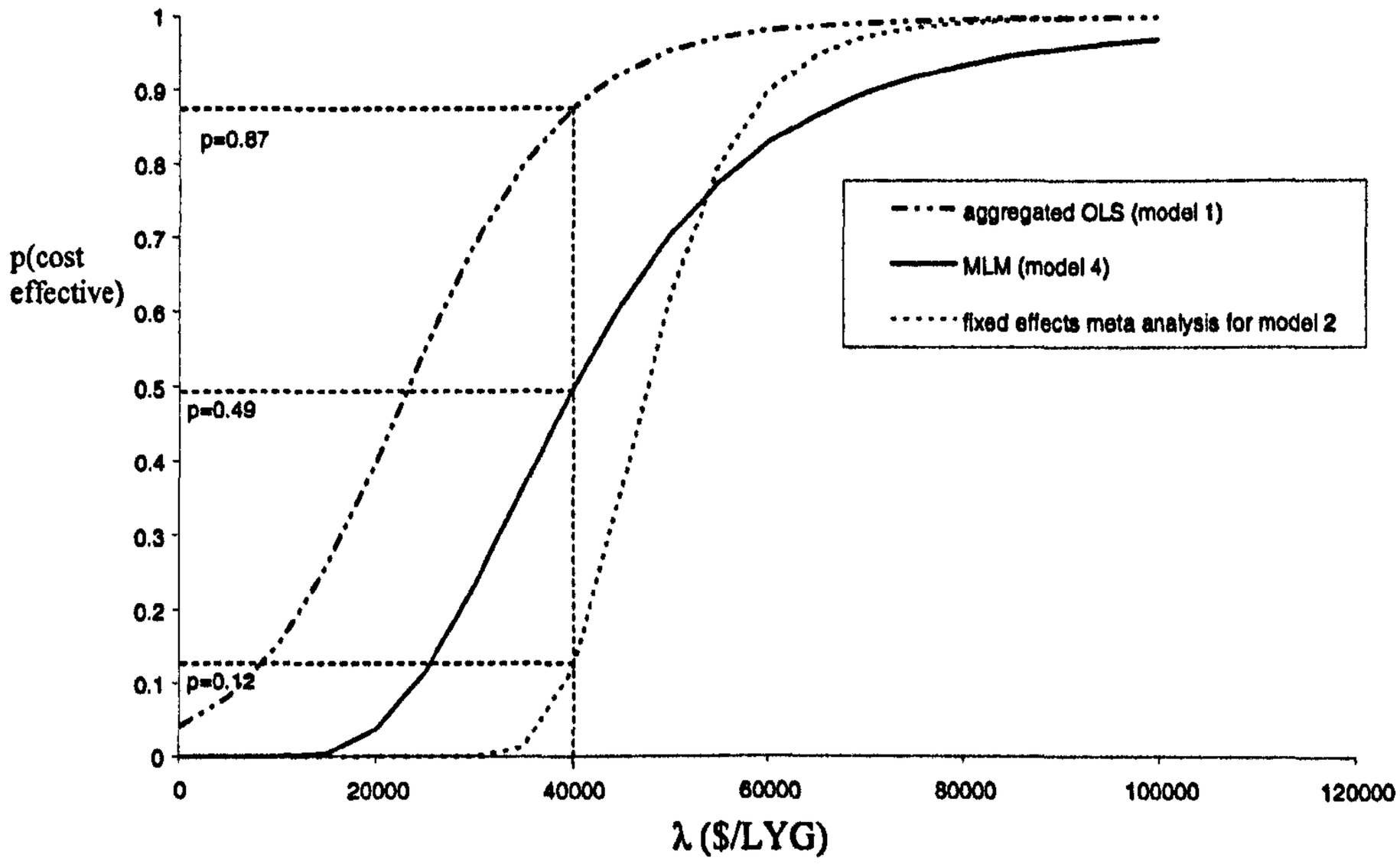
[61] The results presented were those from using the MCMC estimation in MLwin.
[62] The difference in residual deviance gives a chi-squared value of 18, which with 2 degrees of freedom gives a p value of 0.002.

Compared to the MLM (model 4) the fixed effects meta-analysis (model 2) underestimated the probability that the intervention was cost-effective at low levels of λ, and overestimated the probability that it was cost-effective when λ was greater than $60,000 per LYG.

The CEAC was steeper for the fixed effects estimate (model 2), and crossed the CEAC for the random effects estimate (model 4) at a λ of just under $60,000 per LYG. The fixed effects estimate was determined by the level of within-centre variance and this changed with λ; where λ was low the centres with the lowest variances tended to have lower mean INB, the estimates from these centres were given higher weighting in the fixed (model 2) than the random effects model (model 4). However, as λ increased the variances increased most in those centres with lower net-benefits, these centres estimates were down-weighted more in the fixed than the random effects estimate, and consequently the overall estimate in the fixed effects model increased relative to the random effects estimate. The random effects estimate was less dependent on the within-centre variances, and hence the estimates were less sensitive to λ. The more precise estimates for the fixed effects model (model 2) also caused the CEAC to be steeper than for the random effects model (model 4).

A general problem with the CEACs plotted using the models considered above is that they assume a general value of λ for each decision-context. As this study included centres from countries at different stages of development it is unlikely that the same value of λ applies in each context. CEACs allow decision-makers to consider the probability that the intervention is cost-effective at various values of λ. However, simultaneously changing the ceiling ratio in different decision contexts would make an overall CEAC very difficult to interpret. This issue is further considered in sections 9.65 and in the discussion (9.7).

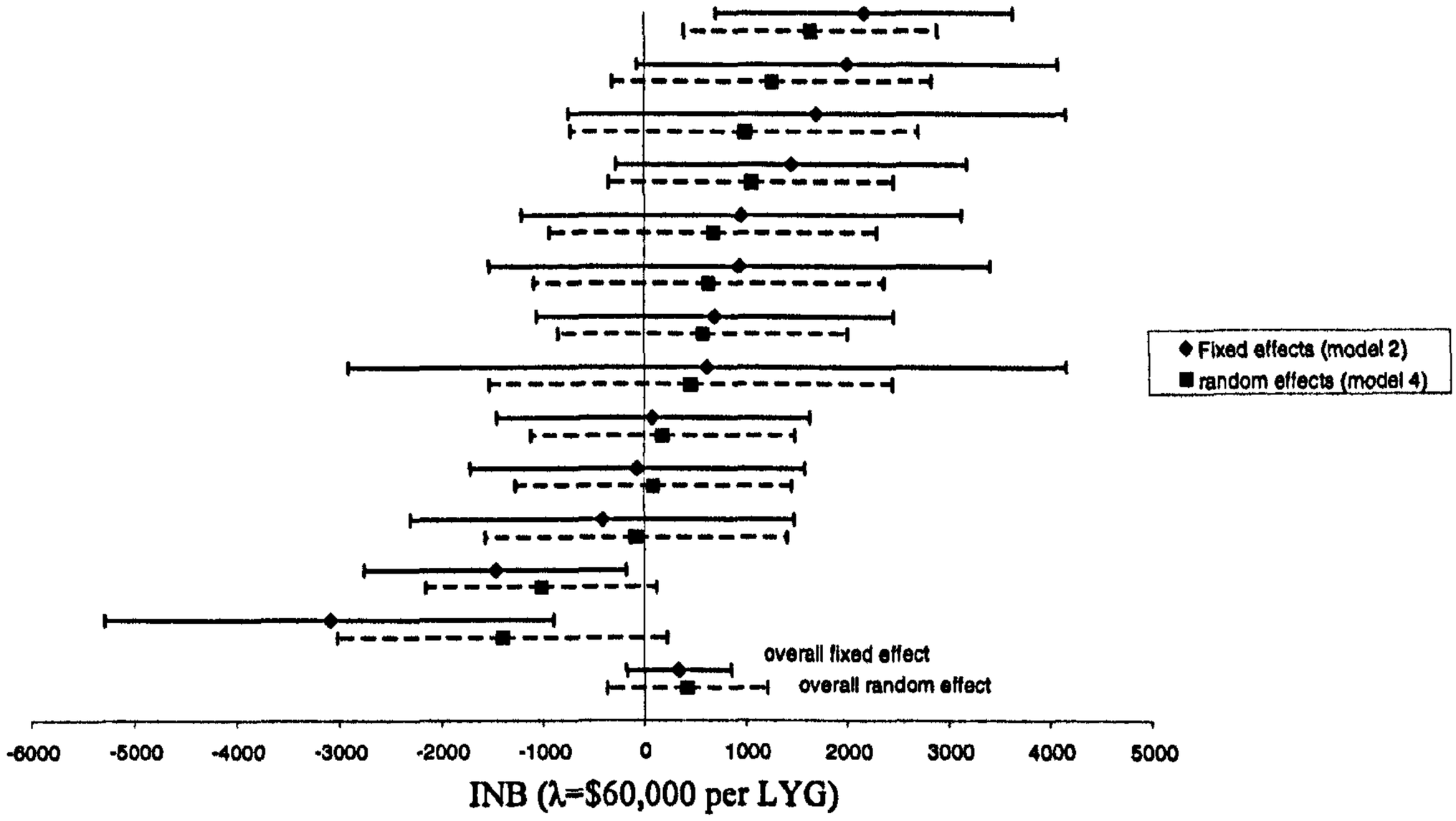
Figure 9.2: CEACs for aggregate OLS model (model 1) vs fixed effects estimate derived from the centre-specific OLS regressions (model 2) vs MLM with random treatment effect (model 4)



9.64 The effect of shrinkage

The mean INBs for each centre with 95% CI, are compared for models 2 and 4 to examine the effect of the shrinkage estimators used in model 4 (Figure 9.3). The results showed that, as anticipated the shrinkage estimators were closer to the overall mean, and the confidence intervals surrounding each estimator were narrower, indicating gains in statistical efficiency. Those centres estimates that changed most with shrinkage were those furthest from the overall mean and with the least precise results.

Figure 9.3: The effect of shrinkage: centre-specific and overall mean INB (95% CI) for fixed effects (model 2) and random effects (model 4)



9.65 Inclusion of national-level covariates- model 5

The use of the shrinkage estimator assumes that those centres that diverge most from the random effects estimator are those that are the least precise. All the heterogeneity across the centres is therefore assumed to be random, and not due to systematic differences across the centres. However, the literature review and previous cost analyses suggested that certain factors, in particular the level of GDP spent on health care, are associated with differences in cost and potentially variation in INB across the centres.

The results from model 5 showed that including the treatment by % GDP/health interaction terms improved the fit of the model (Table 9.8). The interaction terms were statistically significant and their inclusion reduced the level of unexplained variance at a

centre-level (the residual deviance is lower for model 5 than model 4). A likelihood ratio test showed that the model fit improved with the inclusion of the interaction terms ($p = 0.04$)[63]. The mean INB for the reference group, those centres spending a low proportion of GDP on health care, was negative whereas those centres spending a medium or high level of GDP on health care had higher, positive, mean INBs.

Table 9.8: Results from meta-regression analysis (model 5) estimating the effect of % GDP/health on INB

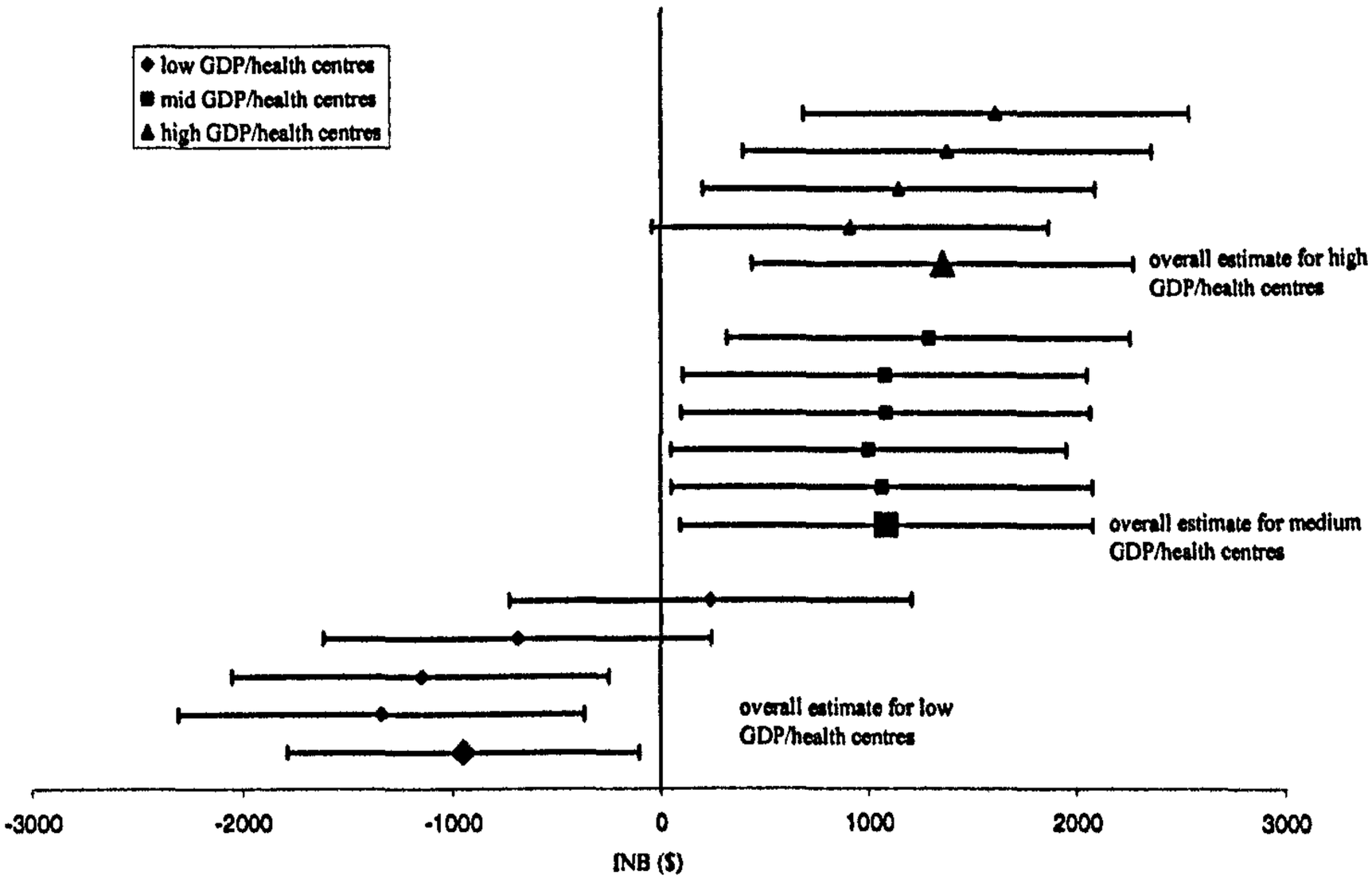
Coefficient	Estimate (SE)
Treatment	-949(430)
Treatment-covariate interactions	
Low GDP/Health*treatment	reference
Mid GDP/Health*treatment	2025(699)**
High GDP/Health*treatment	2280(652)**
τ^2	$1.21(3.6)*10^5$
Deviance (MCMC)	35,302

** p<0.05

The figure below (Figure 9.4) used the estimates from model 5 to plot the estimated INB for each stratum of % GDP/health, alongside the centre-specific shrunken estimates of INB (model 4). The results show that the overall INB was significantly less than zero for the centres with low %GDP/health, and significantly greater than zero for those centres with medium or high levels of GDP spent on health care (Figure 9.4).

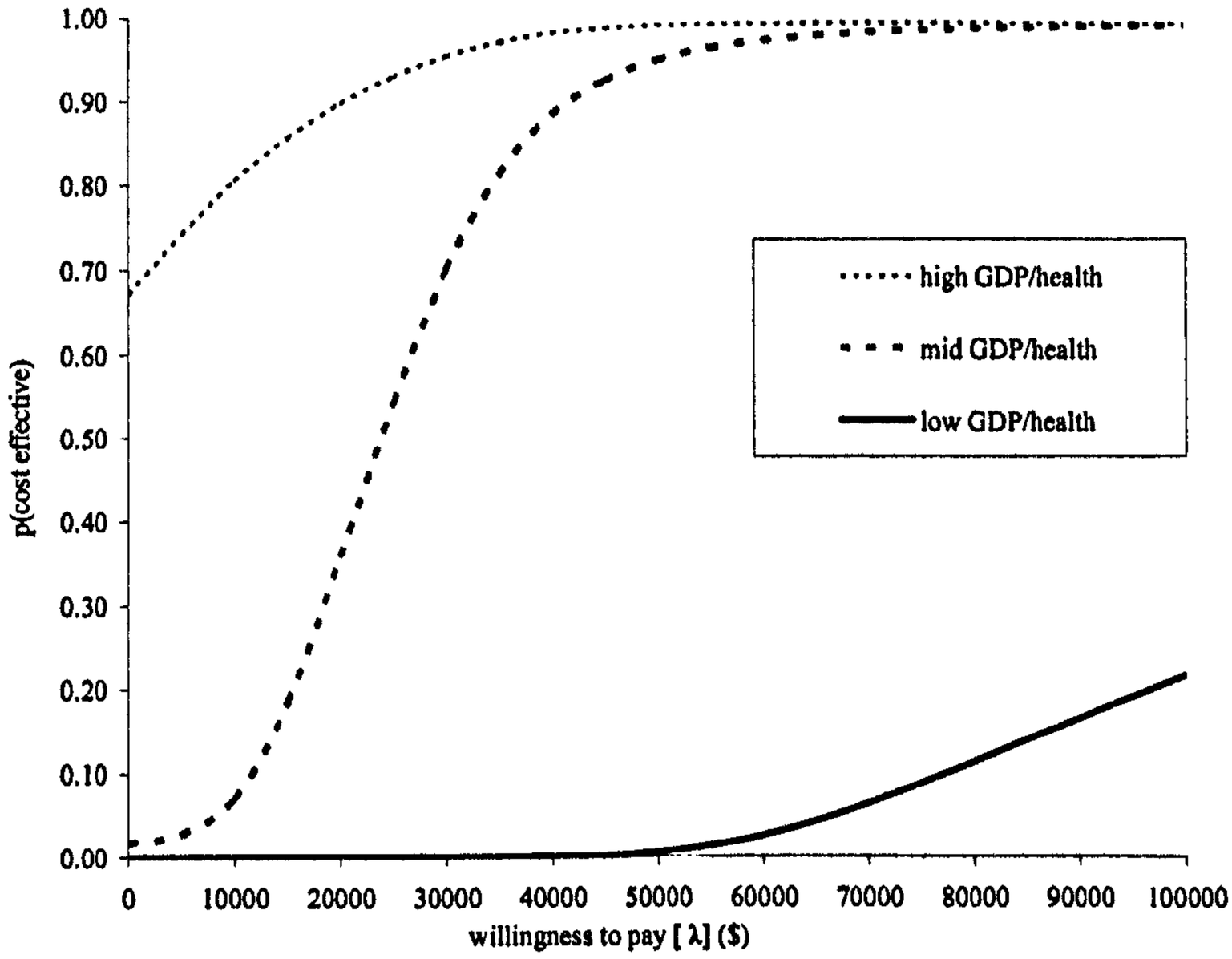
[63] Compared to model 4, model 5 had a reduction in the residual deviance giving a chi-squared of 6, with 2 additional degrees of freedom.

Figure 9.4: Estimates of mean INB (95% CI) for each centre and overall estimates of mean INB (95% CI) for centres in countries spending low, medium and high % of GDP on health care (from model 5).



Model 5 was re-estimated for different levels of the ceiling ratio λ, and the results used to plot CEACs according to different strata of GDP spending on health care. The results show that for low GDP/health centres, the intervention was not cost-effective at different levels of λ (Figure 9.5); for high GDP/health centres the intervention was cost saving and the intervention was cost-effective even if λ was zero. For the mid GDP/health centres, the probability that the intervention was cost-effective was above 0.5 when λ exceeded $25,000 per LYG.

Figure 9.5: CEACs based on estimates of INB from model 5, with adjustment for different levels of GDP/Health



9.7 Discussion

The aim of this chapter was to examine whether using MLMs rather than OLS models to analyse multicentre cost-effectiveness data could lead to different results. The results showed that using an aggregated OLS model to report an overall measure of cost-effectiveness underestimated the uncertainty that surrounded the results. Using separate OLS models for each centre suffered from a lack of statistical power and pooling the estimates from these models in a fixed effects meta-analysis, ignored heterogeneity across the centres and overestimated the precision of the results. The MLMs, allowed for heterogeneity across the different centres and provided more accurate estimates of the uncertainty surrounding the results. The cost-effectiveness dataset used to compare the techniques was generated and further research using multicentre studies is required to compare MLMs with OLS regression analyses. The results presented here do though raise certain issues that need careful consideration when analysing multicentre cost-effectiveness data.

The situation highlighted was where an intervention has the same cost and similar effectiveness across a range of centres, but has a differential impact on morbidity costs across centres. OLS regression analysis has recently been recommended for assessing the INB of one treatment alternative compared to another (Hoch et al. 2002), and could be applied to the situation described. However, using OLS regression analyses to estimate a single pooled mean INB does not appear appropriate in this context. This approach ignores differences in the mean and variance of the INB that may exist across centres. In this case study, compared to the MLM the aggregated OLS analysis overestimated the overall mean INB. It has been recently recommended that decision-makers choose the interventions that are on average relatively cost-effective (Claxton et al. 2002, Claxton 1999), i.e. where the *mean* INB is positive. If OLS analyses fail to accurately estimate the mean INB, basing decision-making on these estimates could lead to the misallocation of

health care resources. There may be other circumstances where using OLS analysis to estimate mean INB may underestimate the cost-effectiveness of an intervention. For example, if the mean INB is lower in those centres with more patients, then the overall OLS regression analysis may underestimate the mean INB.

While ignoring heterogeneity has an ambiguous impact on the mean estimate of cost-effectiveness the impact on the standard error surrounding the measure of cost-effectiveness is more generalisable. By ignoring heterogeneity, aggregated OLS analysis or fixed effects estimates based on centre-specific OLS analyses, underestimate the standard error surrounding the INB and overestimate the precision of the estimate. This would provide health care decision-makers with a false sense of certainty surrounding the overall result. The greater the uncertainty surrounding the mean estimate of cost-effectiveness, the higher the value of acquiring more information (Claxton 1999). So OLS analyses would overestimate the precision of the overall estimate of cost-effectiveness, and may understate the value of acquiring further information for assessing cost-effectiveness.

Rather than using pooled OLS results, a decision-maker could use the separate OLS regressions for each centre, and base their decision on the most relevant result for their own setting. The main problem with using separate OLS analyses is that there is a lack of statistical power to detect whether differences in NMB exist between the interventions, and high levels of sampling variation therefore surround the results. In this case study only 3 out of the 13 centres had INBs that were statistically significant at the 5% level. Other international cost-effectiveness studies have concluded that there was a lack of statistical power to detect differences in cost-effectiveness according to the country concerned (Willke et al. 1998, Cook et al. 2003). A decision-maker in this situation is likely to consider the results from other centres, or the overall estimate, when deciding

which intervention is relatively cost-effective. This approach does not formally recognise the transfer of information from other centres, and does not use all the available data.

The MLMs recognised the heterogeneity that existed across the centres, and adjusted the overall mean INB and standard error accordingly. The result was that compared to the overall estimates from the OLS analyses, the MLMs made a more accurate estimate of the mean INB and the sampling variation surrounding the result. The comparisons of model fit showed that the MLMs that recognised the differences in the mean and the variance of the INB across the centres, fitted the data best.

In addition to providing an overall estimate of cost-effectiveness that recognises heterogeneity, MLMs can also provide shrunken estimates of centre-specific estimates and their standard errors. Those centres with outlying or imprecise estimates can borrow strength from those centres with more precise estimates. The centre-specific estimates are moved in towards the overall mean and are more precisely estimated compared to those based on the observed data. However, the use of shrinkage estimators highlights a potential problem with using MLMs in this context, in that the estimates from the individual centres are regarded as exchangeable: i.e. they are drawn from the same distribution (Spiegelhalter et al. 2000). In the context of multinational data the assumption that there are no systematic differences across the centres is contradicted by insights from economic theory summarised in the literature review (Chapter 4), and by the previous empirical results (Chapter 8). In particular, differences exist in budget constraints, factor prices and production processes across these health care settings; *a priori* reasoning would not appear to justify the assumption of exchangeability across all the health care centres.

To allow for these systematic differences across the centres additional covariates can be included in the MLM. The literature review identified several variables potentially useful for characterising differences across centres (Chapter 4), and one of these-the % of GDP spent on health care (%GDP/health) was chosen for this analysis based on the results of the cost analysis (Chapter 8). Compared to the basic MLM the inclusion of these covariates reduced the level of unexplained variability at a centre-level and improved the model fit. Presenting the overall results of the MLM for different strata of GDP spent on health care would therefore seem more appropriate, and the results for individual centres were reasonably comparable to the random effects estimate for the particular stratum of GDP spent on health care (%GDP/health). This model provides results that have more precise estimates than the centre-specific estimates, but does not make such strong exchangeability assumptions as using the overall random effects estimates or shrunken estimates without covariate adjustment.

In multinational studies decision-makers in countries at different stages of development may face very different levels of λ. In principle, a CEAC allows a decision-maker to see whether the results of a CEA are robust to different values of λ. However, although decision-maker in different countries can specify alternative values of λ, incorporating these different values into a single CEAC, would make the overall CEAC very difficult to interpret. Instead, using the %GDP/health by treatment interaction in the regression models, allowed different CEACs to be drawn for different stratums of GDP/health. Using this approach assumed that the data, including the value of λ were exchangeable within, but not across different stratums of GDP/health.

It cannot be concluded that the % of GDP spent on health care is *the* most appropriate variable for stratifying centres based on these results. The cost-effectiveness dataset was constructed making the assumption that the intervention reduced morbidity costs by 40%, and these morbidity costs were associated with the % of GDP spent on health care

(chapter 8). Hence, the association of the variable %GDP/health with INB was contrived through the design of this study, and would only apply to other cost-effectiveness studies bearing similar characteristics. Nevertheless, the approach illustrates how to relax the exchangeability assumption, and the potential importance of using *a priori* reasoning to stratify centres in a CEA. It also illustrates how MLMs can include covariates to try and reduce the level of unexplained heterogeneity across the centres.

In other contexts, there may be strong *a priori* reasons for including different centre-level variables in a MLM for example, the presence of a DRG system, or the reimbursement method for physicians could in other circumstances be important variables. The choice of variables should be determined by *a priori* reasoning that uses theoretical principles when considering the likely variability in costs and cost-effectiveness for the settings and health care technologies in question. Theoretical reasoning would appear to provide stronger basis for choosing variables than simply identifying those variables that are statistically significant, which may just be an artefact of the dataset concerned.

The analysis considered two alternative techniques for including covariates in MLMs. The first required the inclusion of covariate interactions with treatment. The second used a two- stage approach, with the centre-specific mean estimates regressed against the level of GDP spent on health care, in a meta-regression. Both approaches produced the same result, that in this example there was a differential INB according to the level of GDP/health. The two-stage meta-regression approach was easier to implement than the one-stage approach. However, the advantage of the one-stage approach is that it can be easily expanded to include patient-level covariates. This could be potentially useful when adjusting for baseline differences between treatment and control groups. For example, in this case study more patients were in coma at admission in the control group. The one-stage approach could easily be extended to include a treatment by coma interaction term to adjust for these differences.

MLMs have traditionally been used in other areas, e.g. education and the meta-analysis of clinical trials, but their use is relatively recent in health economics (Rice and Jones 1997). When developing a MLM to examine multinational cost-effectiveness data, this case study suggested it was important to avoid the assumption of common variance across the treatment centres, an assumption commonly made by MLMs in education (Goldstein 1995) and recently in health economics (Carey 2000). In this example, the inter-patient variability clearly differed across the centres. Allowing the variance to differ across the centres improved the fit of the MLM, and should be considered during the analysis of multicentre cost-effectiveness data.

One limitation of the MLMs developed in this chapter is that they assume the residuals are normally distributed. Whereas the histograms for the centre-level residuals showed that this assumption may be realistic (see appendix 5), this assumption appeared less tenable for the patient-level residuals, which at least in some centres, were clearly non-normally distributed (appendix 6). Instead, it was assumed that there were sufficient cases to invoke the central limit theorem and to presume that with the large numbers involved the normality assumption held. This has been shown to be reasonable for some cost datasets (Barber and Thompson 1998), and this assumption has also been made when analysing net benefits (Hoch et al. 2002, Manca et al. 2005). However, this work could be extended to investigate whether the hierarchical gamma models used in the cost analysis in Chapter 8, are also appropriate for analysing hierarchical cost-effectiveness data. Another alternative would be to investigate the use of the non-parametric bootstrap for analysing multicentre cost-effectiveness data.

This cost-effectiveness analysis demonstrated that when the intervention leads to differences in morbidity costs across the centres, then the choice of MLM versus OLS regression analysis could make an important difference to the inferences drawn. Obviously, in multinational economic evaluations there may be other circumstances when

the difference in the results produced by the techniques is much smaller, for example when both the incremental costs and effects are similar across the centres. In other circumstances, drug prices might also differ across centres, leading to greater heterogeneity[64]. This chapter demonstrates the use of appropriate techniques for dealing with the heterogeneity that may exist across centres in a multicentre cost-effectiveness analysis. Few studies in the cost-effectiveness literature have used statistical techniques to test and adjust for differences across centres, when analysing cost-effectiveness. In a recent review, Barbieri et al. (2005) demonstrated that important differences in mean cost-effectiveness could exist across health care settings. However, this study did not provide any statistical measures of uncertainty or any formal tests of heterogeneity. Similarly Willke et al. (1998) demonstrated that important differences in the mean estimate of cost-effectiveness can exist across countries, however, the study lacked the power to detect whether these differences were statistical significant. Cook et al. (2003) used statistical techniques to report differences in the cost-effectiveness of simvastatin across several Scandinavian countries, and based on tests for heterogeneity concluded that there were not significant differences in the INB across the studies concerned. However, the tests for heterogeneity were underpowered, and they did not provide a strong basis for assuming that the INB did not vary across the countries concerned and that the overall INB was the appropriate measure of cost-effectiveness upon which to base decision-making.

9.8 Conclusions

To conclude, this chapter has shown that ignoring the hierarchical nature of multicentre cost-effectiveness data can lead to an inappropriate assessment of cost-effectiveness. The case study illustrated that by acknowledging the heterogeneity that exists across centres, MLMs can provide a more appropriate estimate of cost-effectiveness. MLMs also need to consider the systematic differences that exist across centres, using covariates suggested

[64] Countries may also differ in the outcome gains from an intervention, or the value of those outcome gains. Explicit consideration of these issues is outside the scope of this investigation which focuses on the impact of cost variation across settings on estimates of cost-effectiveness.

by economic theory. This chapter illustrated the issues raised when using these techniques for analysing a cost-effectiveness dataset.

Chapter 10: Discussion

10.0 Introduction

Economic evaluations are used in setting priorities in health care. For example, NICE use CEA to recommend which health care interventions the NHS in England and Wales should provide (NICE 2004). The use of CEA in decision-making has the potential to improve the efficiency of resource allocation in health care. However, this will only happen if these studies use sound methods, in particular if the costs used represent opportunity costs for the decision-context (Graves et al. 2002). Measuring opportunity costs is problematic and requires that costs reflect productive efficiency (Donaldson et al. 2002). However, the observed costs of health care may vary across settings, for example hospitals in different geographical locations (Willke et al. 1998). It is unclear whether such variations reflect variations in productive efficiency, differences in contextual factors such as factor prices, or inconsistencies in costing methods. Unless studies disentangle the reasons for observed cost variation across health care settings CEA may continue to use costs that diverge from opportunity costs.

The aim of this thesis is to assess why costs vary across health care settings, and the implications for the methodologies used in economic evaluation. The specific objectives stated in the introduction are:

1. To assess how economic evaluations currently consider cost variation across settings.
2. To generate hypotheses for why costs may vary across health care settings.
3. To identify which factors are associated with variation in resource use and cost using MLMs and OLS regression models.
4. To compare the use of OLS regression models to MLMs for analysing multicentre cost-effectiveness data.

The first section of this chapter considers the main findings from both the literature review and the empirical investigation in relation to each of these objectives. The second section reviews important methodological themes from the thesis alongside emerging issues in the literature. The third section discusses the limitations of the approach taken.

10.1 Main findings from the thesis

10.11 How does economic evaluation consider cost variation across settings?

The literature review highlights that economic evaluations commonly ignore cost variation across settings at both the design and analysis stages. Health care providers in different settings may face different levels of factor prices, use various combinations of factor inputs and differ in the observed costs of producing health care programmes (Chapters 2 and 3). The review suggests that economic evaluations make fundamental decisions at the design stage that limit the consideration of these and other reasons for cost variation. For example, if studies only measure costs in one health care setting, it is unclear whether the costs measured are those of efficient production and therefore represent opportunity costs for the decision context concerned. Unless the costs used represent opportunity costs, the evaluations may make inaccurate estimates of the relative costs and cost-effectiveness of different health care interventions. The use of inappropriate cost estimates can hinder moves to improve the allocative efficiency of resource allocation.

Recent multinational economic evaluations observed wide cost variations across health care settings, and found that ignoring this variation led to inaccurate estimates of cost-effectiveness (Chapter 3). However, these studies suffered from serious limitations. They did not use economic theory to pose hypotheses for why costs may vary. The studies failed to address measurement issues that pervade the comparison of costs across health care settings, and the analytical methods did not recognise the hierarchical structure of these cost data.

Studies are required that assess reasons why costs vary across health care settings, and identify circumstances where costs may depart from opportunity costs. An assessment of cost variation across settings can provide guidance on a number of methodological issues. For example, these studies can guide the numbers and characteristics of health care settings that an economic evaluation requires to help ensure that the costs collected represent opportunity costs for the decision context concerned.

10.12 To generate hypotheses for why costs may vary across health care settings.

a) Hypotheses generated from the literature

The literature review generated hypotheses for why costs observed may vary across health care settings (Chapter 4). This review provided a conceptual framework for the ensuing empirical investigation. The production and cost function literature highlighted that firms could face different factor prices and therefore choose alternative input combinations whilst still minimising costs. However, the firm's ability to cost-minimise may depend on the health care context. For example, the level of labour input available to the health care firm may vary across countries because of differences in the national labour markets for health care professionals. Health care firms in certain countries may be more constrained in choosing the level of labour inputs required to maximise productive efficiency. Differences in patient factors in particular case-mix are also important in explaining cost variation across health care settings (Chapter 4).

The literature review highlighted the problems involved in testing these hypotheses. The methodological guidelines for economic evaluation have not carefully considered the measurement issues that arise when comparing or using cost data collected from different health care settings (Chapter 2). For example, the guidelines are not prescriptive about the appropriate level of aggregation to use when measuring resource use or unit costs. However, where studies have taken an aggregated approach to cost measurement, it is unclear whether the same items are included in the unit costs or resource use measured in each health care setting (Chapter 3). An empirical investigation of cost variation across

health care settings needs to use a consistent method of cost measurement in different contexts. The review suggests that to reduce methodological inconsistencies across study settings a disaggregated approach to cost measurement is preferable. In addition, economic evaluations that considered cost variation across settings did not include sufficient cases or centres to identify systematic reasons for cost variation across health care settings (Chapter 3). A study comparing costs across settings should consider whether there are sufficient patients and settings to identify systematic cost differences.

The review of the applied cost function literature emphasises the importance of adjusting for differences in case-mix, when comparing costs across health care settings (Chapter 4). Studies that estimated differences in efficiency across health care settings required data from many health care units. These studies therefore relied on administrative costing datasets and used the DRG system of case-mix classification, to adjust for case-mix differences across health care providers. Using these highly aggregated measures of case-mix meant that observed cost differences across health care settings could reflect unmeasured differences in the patient case-mix. Studies of cost variation across settings therefore need to use appropriate datasets that collect sufficiently detailed measures of patient case-mix to explore the role of differences amongst patients in explaining cost variation.

b) Hypotheses from the dataset in the empirical investigation

The case study used for the empirical investigation was a multinational stroke dataset comprising case-mix, resource use, and cost data on 1300 stroke admissions from 13 centres in 10 European countries (Chapter 6). Detailed information on patient factors and resource use were measured for each patient for three months post-stroke. Information on unit costs and the characteristics of each centre and country were also collected (Chapters 6 and 7). These data were used together with the findings from the literature review (Chapter 4) to pose hypotheses for cost variations.

The resource use data illustrated that there were important differences in the way stroke care was provided across the study settings (Chapter 6). While each centre had an acute

hospital providing inpatient care, different alternatives to hospital care existed across the centres and the range of care alternatives available may be associated with differences in mean LOS across the centres. The access to and use of specialist technologies also varied across centres. In some settings, there was better access to specialist neurological services, or inpatient rehabilitation services. It was hypothesised that access to these more specialised technologies, led to a higher use of tests and investigations and higher health service costs. A national-level variable, the proportion of GDP spent on health care (% GDP/health), was proposed for inclusion in the subsequent regression analyses. This variable represented differences in the level of health care infrastructure across the countries concerned. There were also variations in patient factors such as case-mix across the centres and it was hypothesised that these differences may partly explain the observed variations in resource use and costs across the centres.

The analysis of the unit cost data (Chapter 7) suggested that there were wide differences in factor prices across the centres, particular between those located within Eastern compared to Western Europe. These differences may arise because of inflexibilities in the European labour markets for health care professionals. There were also differences in the use of factor inputs that did not appear to reflect differences in relative factor prices.

10.13. To identify which factors are associated with variability in resource use and cost using MLMs and OLS regression models.

The review of economic concepts and statistical techniques highlighted some of the difficulties in identifying factors associated with cost variation (Chapters 4 and 5). Economic theory suggests that factors associated with cost variation operate at different levels. Statistical theory states that where data are hierarchical, using OLS regression analysis and ignoring the clustered nature of observations could lead to incorrect inferences about the reasons for resource use and cost variation. While the theoretical literature suggests that the choice of technique for analysing cost data *might* matter, the results from the empirical investigation provided an example of where the choice of technique *did* matter (Chapter 8). Using OLS regression analysis led to incorrect

inferences, in particular the significance of each of the higher-level variables was overstated. Once MLMs rather than OLS regression models were used to assess the reasons for cost variation, the only higher-level variable associated with cost variation was the % GDP/health.

The % GDP/health variable may capture the effect of several factors that the data and the literature review suggested were likely to be associated with resource use, but which were not formally included in the regression equations. Insights from the literature review suggested that the % GDP/health variable was likely to represent differences across the centres in the quality of inputs, access to technology, and the national level of health care infrastructure. Those countries spending a high proportion of their GDP on health care are more likely to adopt new technologies, and ensure health professionals are highly trained. The descriptive data clearly suggested that in the Eastern and Southern European centres where the national levels of spending on health care were lower; there was less access to neurologists, organised stroke care, inpatient rehabilitation facilities and physiotherapists. It would therefore appear that there were different 'models of stroke care' and that these models of care were associated with different costs. The variable % GDP/health therefore appeared to differentiate between these different models of care, and summarised the effect they had on costs.

The MLMs confirmed that several other factors suggested by the literature review and data as likely to be associated with cost variation, were still statistically significant when an appropriate technique was used for assessing cost variation. More complex patient case-mix was associated with higher total cost, and the presence of care alternatives was associated with shorter LOS. The MLMs were therefore useful in identifying factors associated with systematic cost variations across the health care settings in the study.

10.14. Assessing the use of MLM compared to OLS for analysing multicentre cost-effectiveness data.

The observational dataset used for the main aspect of the empirical investigation was extended to generate a multicentre cost-effectiveness dataset, comparing a new intervention for stroke patients to existing practice in each centre (Chapter 9). The analysis examined the implications of using a pooled OLS estimate to assess incremental cost-effectiveness compared to a MLM that allowed the relative cost-effectiveness of the intervention to vary across the health care centres. The results showed that the pooled OLS analysis provided an inaccurate, over precise estimate of relative cost-effectiveness compared to the MLMs. The MLMs recognised the heterogeneity that existed across the health care centres and gave a more accurate estimate of the mean cost-effectiveness of the intervention and the associated uncertainty.

10.2 Central methodological themes emerging from the thesis

The central methodological themes from the thesis fall under three main headings: the use of explicit *a priori reasoning* to identify reasons for cost variation, the use of MLM in health economics, and the methodological implications for economic evaluation.

10.21 The use of *a priori* reasoning to identify reasons for cost variation

This thesis used relevant strands from both microeconomic and statistical theory. The thesis found that the theoretical strands from these two disciplines were complementary to one another when identifying reasons for cost variation. The review of microeconomic theory found *a priori* reasons why costs may vary across health care settings (Chapter 4). Theory also suggests that systematic reasons for cost variation operated at different levels, for example at a patient, centre or national level. The cost data may therefore be clustered, for example within health care centres. Statistical theory states that where data are clustered using OLS regression analyses that assume observations are independent is inappropriate. MLMs recognise the data hierarchy and are more suitable for testing the hypotheses suggested by microeconomic theory in this context (Chapter 5).

Insights from production and cost function theory were used in the empirical investigation to choose the variables for inclusion in the cost functions. The cost functions specified cannot be regarded as the analogue of well-behaved production functions (Breyer 1987). There were insufficient patients or centres, to estimate isoquants and to test the extent to which each health care centre was exhibiting cost-minimising behaviour. However, the choice of independent variables was still made with recourse to economic theory, rather than resorting to an 'anything goes' approach to cost function specification (Wagstaff 1989a). The variables included patient characteristics, the availability of substitutes for hospital care, incentives for the hospital to cost-minimise, and a measure for the level of health care infrastructure (% GDP spent on health care) in each country (Chapter 8).

Insights from microeconomic theory were used when interpreting the results of the cost functions presented in the empirical investigation. The unexplained variations across the centres could relate to a number of factors excluded from the cost functions. These unexplained variations could reflect residual variations in patient factors, factor prices or differences in technical or productive inefficiency, as well as random variation across the centres (Chapter 8).

The use of economic theory highlighted some of the measurement problems that arise when comparing unit costs across centres, particularly when those centres are drawn from economies at different stages of development. The use of the GDP Purchasing Power Parity indices to convert local currencies into US dollars was unlikely to reflect the relative opportunity cost of each input in each of the centres concerned. It must also be acknowledged that there was variability in the quality of inputs and outputs across the study centres. While the % GDP/health variable may have captured some of these effects, these factors may also have explained the residual differences found to exist at a centre-level (Chapter 7).

10.22 Issues arising when using MLMs in health economics

MLMs have been used for some time in the statistical literature and have been widely used in research and policy-making in education (Goldstein 1996). Despite their conceptual advantages, prior to the commencement of this thesis MLMs had rarely been used in health economics (Rice and Jones 1997). The empirical analysis in this thesis demonstrated the use and appropriateness of MLMs for analysing international cost and cost-effectiveness data. Recent work by Manca et al. (2005) showed that similar advantages exist for using MLMs compared to OLS regression models when analysing cost-effectiveness data collected in multicentre trials in a single country.

When applying the standard MLMs traditionally used in the education and HSR literature to the analysis of cost and cost-effectiveness variation, two key issues emerged that warrant further discussion: the skewed nature of the cost variable, and the assumption of exchangeability. MLMs commonly assume that the residuals from the regression equation are normally distributed. However, cost data are often highly skewed, and methods that assume either the raw data, or the residual from a regression equation are normally distributed, may not be appropriate. GLMs have been recommended for analysing non-hierarchical cost data as they can allow for the skewed distribution of the data whilst still reporting the arithmetic mean costs required by decision-makers (Barber and Thompson 2004). This thesis used GLMMs as they allow for the hierarchical structure of the cost data *and* a skewed distribution function can be chosen. Within the group of GLMMs the gamma model chosen allowed for the distribution of the cost data to vary by centre (Chapter 8). This model would seem particularly useful for analysing hierarchical cost data. An alternative model that also allows for the skewed nature of cost data is the non-parametric bootstrap. This model has been applied to clustered data in an educational context (Carpenter et al. 2003), but its use in analysing hierarchical cost data has yet to be tested.

MLMs also rest on the assumption that data are exchangeable, that is they are drawn effectively at random, from the same distribution. This assumption has been made by another study that used shrinkage estimates in MLMs (Manca et al. 2005), to move

centre-specific estimates of incremental cost-effectiveness in towards the overall, random effects mean estimate. These models regard the variability between centres as random rather than attributable to systematic differences. However, in this thesis both the theoretical review (Chapter 4) and the empirical investigation (Chapter 8) suggested there were systematic cost variations across the centres included in the study. This thesis avoided categorising cost variation across settings as 'random' and instead identified systematic reasons for cost variation. Before assuming that the cost-effectiveness data are exchangeable, the analysis included covariates to allow for systematic variations in costs across the centres (Chapter 9). As Spiegelhalter et al. (2000) point out:

"...where there are known reasons to suspect specific units are systematically different, then these reasons need to be modelled.." (Spiegelhalter et al. 2000, p21).

This thesis therefore argues in favour of using covariates to adjust for systematic differences in costs and cost-effectiveness across health care settings, before making the assumption of exchangeability. After making these adjustments, it is more likely that the data can be assumed exchangeable across the health care settings, i.e. drawn from the same distribution.

10.23 Methodological implications for economic evaluation

The literature review found gaps in the advice given on how CEA should be designed, analysed and presented. The empirical investigation has provided some potential solutions to the methodological issues raised relating to cost variation across settings. These solutions are described below and listed in Table 10.1.

a) Insights for the design of economic evaluations

There is little guidance available for those designing CEA alongside multicentre RCTs on the number of settings that should be included in the costing study, the characteristics of the study centre or the level of aggregation that should be used for cost measurement (Chapter 2). A further concern is whether it is necessary to measure costs in all centres or countries. Multinational studies have measured costs in a single country and generalised the results to other countries or measured country-specific unit costs, but combined these

with trial wide measures of resource use (Chapter 3). More recently, studies have measured country-specific resource use and unit costs (Willke et al. 1998). The guidelines for economic evaluation fail to offer advice on which study design is most appropriate.

This thesis found that there was potentially wide variation in both resource use and unit costs across study centres (Chapters 6 and 7). In a multinational study, it may not be appropriate to simply collect costs in one setting or to pool cost results across settings. Also while it may be easier for analysts to measure unit costs, but not resource use in each centre, the findings in this study, highlight the importance of measuring resource use in a range of international health care setting.

Basing estimates of total cost on resource use estimated from an atypical sub-sample of the countries recruited to a multinational study may generate inaccurate and misleading results. Taking this approach ignores systematic differences across the settings that may partly reflect differences in productive inefficiency. Basing cost estimates on an 'average' measure of resource use or unit cost is unlikely to represent opportunity costs in any particular decision context. Analysts should consider estimating resource use and unit costs in more centres than is currently the norm in economic evaluations. The use of MLMs reveals the need to have data available from enough centres to enable the effects of higher-level variables to be estimated with adequate precision. One preferred strategy is therefore to undertake costing sub-studies on a random selection of patients from all centres, rather than on all patients from only selected centres. Studies can then consider why any observed variations in costs exist across settings. This can allow analysts or decision-makers in different countries to identify which cost estimates are most likely to approximate opportunity costs in their decision-making context. Estimating the resource use and unit costs in many health care setting would help identify the opportunity costs of each health care programme in each decision context.

Once studies have identified factors that are associated with cost variation this can inform the design of future CEA. If the level of health care spending is found more generally to

be associated with total or incremental costs, then a multinational study could select centres likely to represent opportunity costs in countries with high, medium or low levels of spending on health care. Decision-makers in each of these settings could then base their decisions about relative cost-effectiveness on relevant cost data.

The methodological guidelines reviewed were not specific about the level of disaggregation that should be employed when estimating resource use or unit costs. In this study, a highly disaggregated approach was taken to resource use and unit cost measurement, particularly for labour costs. Factor inputs and factor prices were measured for each labour input, in each centre. Collecting these disaggregated data provided further insights into why costs varied across settings, in this case differences in wage rates across countries were found to be particularly important. This disaggregated approach meant that indices could be used to summarise differences in relative and absolute prices, and technology-specific purchasing power parity (PPP) indices were constructed. The use of technology specific PPPs showed that the main results were not sensitive to the choice of conversion factor. Other international costing studies have also taken a disaggregated approach to cost variation and constructed technology specific PPPs to test the robustness of their results to the choice of conversion factor (Hutton 2001, Wordsworth and Ludbrook 2005).

Although there were advantages associated with taking a disaggregated approach the opportunity costs of this use of research resources must be recognised. Taking a detailed approach limits the number of settings where the study is able to collect cost data. An extreme alternative would be to collect cost data across many different settings using highly aggregated, routinely collected data. The main problem with this approach is that methodological inconsistencies, in particular in the measurement of costs would limit the study's ability to detect systematic cost variations (Schulman et al. 1998). On balance, this thesis advises against using a highly aggregated approach certainly in a multinational context, where methodological differences in cost measurement across settings appear an important problem.

Table 10.1: Summary of the issues related to cost variation across settings that arise during the conduct of CEA and proposed solutions

Issue Raised	Proposed solution
Study Design	
• No centres for resource use or unit cost measurement?	• Multinational studies: measure resource use, unit costs in several countries.
• Choice of centres for resource use or unit cost measurement?	• Use higher-level factors for example % GDP on health care in international studies
• Appropriate level of aggregation in resource use or unit cost measurement?	• Further disaggregation into factor price and factor inputs, useful.
Study analysis	
• How to pool resource use, unit cost or total cost data across settings?	• Pooled data using OLS regression can lead to misleading results.
• How to analyse cost data to allow for differences in the mean and the distribution of costs across settings?	• MLMs provide accurate estimates of mean cost-effectiveness and associated precision.
• How to adjust the cost parameters in a model to estimate cost-effectiveness in different settings?	• Use relationships between patient and higher-level factors and total costs.
Presentation of study results	
• How to present context specific results in multicentre studies?	• Use shrinkage estimates after adjusting for systematic differences.
• How to use statistical methods to make results more representative of the decision context?	• Random effects estimate provides a potentially more generalisable estimate.
• How to consider the extent to which the results are generalisable?	• Centre and country-level covariates can be used to provide limits on generalisability/ assist with transfer of results.

b) Insights provided for the analysis of cost data

The review of the guidelines found that little attention has been given to how resource use and cost data should be analysed to take account of variation in costs and cost-effectiveness across settings (Chapter 2). The empirical investigation compared methods for identifying reasons for cost variation across settings. An important finding from this investigation was that when OLS regression analyses were used to identify factors associated with international cost variation, they overestimated the statistical significance of the centre and national-level variables (Chapter 8). Previous studies that used OLS regression analysis to compare costs across health care settings did not recognise the hierarchical nature of cost data (Coyle and Drummond 1998, Willke et al. 1998) and may have overestimated the statistical significance of centre-level variables associated with cost variation (Coyle and Drummond 1998). MLMs provided more accurate estimates of the effects of the centre and national-level variables and their standard errors. This thesis therefore recommends the use of MLMs for identifying reasons why costs may vary across health care settings (Chapter 8)

The problems associated with analysing skewed hierarchical cost data were examined. Previous studies suggested that GLMs are attractive for analysing non-hierarchical cost data (Manning and Mullahy 2001, Barber and Thompson 2004). This thesis suggests that GLMMs are attractive for analysing cost data that are both skewed and hierarchical (Chapter 8).

OLS regression models and MLMs were compared for analysing multicentre cost-effectiveness data in an international context. The results showed that pooling results across settings using OLS regression analysis may lead to inaccurate mean estimates of cost-effectiveness and an underestimation of uncertainty. Using OLS analyses to estimate cost-effectiveness separately for each setting, suffered from a lack of statistical power (Chapter 9). The problem of a lack of statistical power also pervades recent suggestions to use tests for heterogeneity in this context (Cook et al. 2003). MLMs would appear the more appropriate alternative for analysing hierarchical cost-effectiveness data. The

overall estimates from a random effects model provides an accurate estimate of the overall cost-effectiveness and the uncertainty that surrounds it.

c) Insights for the presentation of results

The presentation of study results should incorporate the different forms of uncertainty that surround the mean estimates of cost-effectiveness and much progress has been made on methods for accounting for sampling variation in economic evaluation (Briggs and Gray 1999, Fenwick et al. 2004). Recent developments in this area have acknowledged the heterogeneity that may exist across patient groups, and sub-group analysis is now recommended for CEA that are submitted to NICE in the UK (NICE 2004). However, less attention has been given to understanding variations across settings. The evidence from this and other recent studies suggests that there are systematic variations in costs across settings (Willke et al. 1998, Coyle and Drummond 1998, Manca et al. 2005) and this cost variation should be recognised alongside other forms of uncertainty in the presentation of results.

One way of highlighting this cost variation across settings would be to present setting specific cost-effectiveness results, however, there may be insufficient power in the study to do this, and this is an inefficient use of available information. The results from the empirical investigation suggested that using the centre-specific shrinkage estimates from the MLMs may be an appropriate way of presenting variation across settings (Chapter 9). However, a fundamental assumption that has not been challenged by other studies using MLMs (Manca et al. 2005) is that the data are exchangeable across the settings concerned, i.e. that they are drawn from the same distribution. This assumption was not plausible for this case study, as both the theoretical review and empirical analysis of cost data suggested that there were *a priori* reasons for expecting systematic variations in incremental cost-effectiveness across the study settings. This thesis therefore recommends that before making the exchangeability assumption, covariates (in this case % GDP/health) are used to adjust for systematic variations across health care settings. The use of covariates can highlight contexts where adopting a new intervention may not improve productive efficiency. This can help ensure that subsequent attempts to compare

the cost-effectiveness of interventions across different disease areas do not include inefficient alternatives (Chapter 9).

If there are systematic cost variations across settings, this has implications for the generalisability of a study's results. In the economic evaluation literature statements about the generalisability of results across health care settings tend to be made without recourse to economic theory or evidence (Mason 1997). The mean estimate of cost-effectiveness is usually assumed to apply to all health care settings within the study setting, and to generalise to health care settings not included in the study. The results from this study suggested some limits to the generalisability of cost and cost-effectiveness results across health care settings.

Covariates can be used to provide boundaries for the generalisability of cost-effectiveness results. These boundaries may help decision-makers in countries not represented by centres in the study to assess whether the results of the CEA apply to their decision context. The cost-effectiveness results presented in this thesis would be inaccurate if the results for those countries with high or medium levels of spending on health care were transferred to those countries with relatively low levels of spending on health care (Chapter 9). The costs in the study context would be highly unlikely to represent opportunity costs in the decision context. Instead, the random effects mean estimates for each stratum of GDP spent on health care (high, medium, low) could be applied to other centres that fall within the same stratum of GDP spending on health care. The use of covariates therefore provides a 'rule of thumb' for deciding which result is most appropriate for a centre outside the study[65].

Rather than arguing that the % of GDP spent on health care is *the variable* to use when generalising results in other multinational studies, this thesis presents a methodology for

[65] Clearly, care still has to be taken as there may be countries that meet the criteria for low or high levels of GDP spending on health care, that have very different characteristics to the countries participating in this study. It may be helpful to use the range of health care spending observed amongst countries included in the study to place upper and lower boundaries on the contexts that the results from such a study could be applied to. In addition, other data on covariates suggested in the literature review as being associated with costs, may be compared between centres included in the study and those in the new decision context to assess whether the result can be regarded as transferable.

considering cost variation that could be more generally applied. In particular, using MLMs to identify the covariates that are associated with cost variation, could further understanding of cost variation across health care settings. The literature review provided a list of covariates potentially associated with cost variation, and in other contexts these covariates, may provide an appropriate basis for deciding when the results of an economic evaluation can be generalised to different decision contexts.

The focus of this study was on improving the methods of cost and cost-effectiveness estimation where patient-level data are available. However, the literature review also highlighted that cost variation across settings is important in the context of model-based economic evaluations (Chapter 3). Little guidance exists on how models should consider cost variation across countries although models have been suggested as a panacea for problems of generalisability (Drummond and Davies 1991). Models have been used to 'adapt' costs collected in one setting to another (Drummond and Davies 1991). Often there is little basis for the way costs have been adapted, beyond the use of expert opinions' on the likely differences in resource use. Instead the covariates identified in this thesis could be used in a cost-effectiveness model to transfer the results of a cost analysis to a different country. This thesis argues that even if a cost-effectiveness model is used to transport costs across settings, the transfer of costs should be based on theoretical and empirical insights, and not according to commentators' speculations.

10.23 Methodological implications for the broader cost function literature

In their forward to the *Handbook of Health Economics* Culyer and Newhouse (2000) highlight the importance of importing ideas from other disciplines to health economics, but also of exporting findings from health economics to inform other disciplines. While the main contribution of this thesis has involved transferring ideas from microeconomics and statistics into economic evaluation, there are also findings from this thesis that can be exported to the more general literature on cost functions.

The cost function literature has been dominated by studies estimating variations in technical, productive and scale inefficiency across health care settings (Cowing and Holtmann 1983, Wagstaff 1989b, Jacobs 2000, Hollingsworth 2003). These studies have concentrated on cost variability at the level of the health care setting; patient-level cost variation has been largely ignored. This is true of studies using non-frontier estimation techniques (such as OLS models), but also of studies using frontier estimation techniques (such as deterministic OLS, DEA, and SFA). These studies have relied on routinely collected, highly aggregated datasets (Chapter 5). Disaggregated datasets with sufficient observations at both a patient and centre level have not been available.

The results presented suggest that differences in patient characteristics are important in explaining resource use and cost differences across health care settings. The use of MLMs found that unexplained variation at a patient-level was much larger than at a centre-level. Carey (2000) used a MLM to assess patient and centre-level cost variability amongst health care providers in the US, and came to similar conclusions. Ignoring the measurement issues posed by inter-patient variation therefore appears to be an important omission. Previous studies could have made inaccurate inferences about the effect of contextual factors such as incentives to cost-minimise on levels of inefficiency (Conrad and Strauss 1983, Cowing and Holtmann 1983).

When assessing reasons for inefficiency further consideration should be given to capturing variation across patients. As more disaggregated cost datasets become available in health care, it would seem desirable to extend existing techniques for measuring efficiency to allow for variability at different levels of the data hierarchy. Both SFA and DEA have been recently applied to panel datasets that allows for variability over time to be examined (Jones 2000), and this may in itself reduce the impact of ignoring unexplained patient variation. However, based on the results of this thesis, extending the use of DEA and SFA to include patient-level variation would appear highly desirable.

10.3 Limitations

The case study used in the thesis was an observational costing study covering 13 centres in 10 different countries. The study did not include sufficient patients or centres to formally test some of the hypotheses emerging from the literature review. For example, it was not possible to examine whether differences in technical or productive efficiency explained cost variation across settings. To estimate efficiency requires the use of a frontier estimation method, rather than one based on a measure of central tendency. Frontier estimation techniques tend to require the use of datasets with many observations at a centre level, which are only available from routine highly aggregated datasets. The use of a disaggregated dataset was able to assess some of the measurement issues that pervade the analysis of aggregated datasets, and should be regarded as complementary to more aggregated analyses.

The use of disaggregated data allows the thesis to tackle some but not all of the measurement issues involved in comparing costs. The cost analysis did not allow for differences across the centres in patients' outcomes. Production function theory highlights that if the inputs used in the production function of health care change, then outcomes may also vary. Previous work undertaken as part of the empirical study found that survival differed across the centres after adjusting for patient factors (Grieve et al. 2001a). The problem with incorporating survival into the cost function was that it led to difficulties in the interpretation in the cost function. Differences in cost across the centres could lead to differences in survival or vice versa. The pragmatic solution taken to this problem was that the main analysis in the thesis only included those patients who survived for at least three months post-stroke (Chapters 6-8). Any conclusions regarding the reasons for cost variation only therefore apply to this group of patients.

The cost-effectiveness analysis (Chapter 9) did include both survivors and decedents. This analysis considered the use of MLMs for analysing hierarchical cost-effectiveness data. The scope of the analysis was limited to considering the impact of cost variation across settings on the relative cost-effectiveness of a new technology. However, the

relative effectiveness of health care interventions may also vary across health care settings. Indeed, even if there is no differential effect in clinical effectiveness according to setting, there may still be important differences in the value of a given effect across different cultures. The estimates presented in this thesis therefore do not capture the full implications of variability between health care settings on the relative cost-effectiveness of different health care interventions.

The empirical investigation used cost data for a particular disease -- stroke care in a certain context -- care for hospital-admitted stroke patients in different European centres. Previous research found wide variations in the way hospital-admitted stroke patients were managed across different European centres (Beech et al. 1996, Wolfe et al. 1999). The patient factors that this thesis found are associated with resource use and total cost variation are particular to this disease area and context and should not be regarded as generalisable to other disease areas or geographical locations. While the country-level variable, the % of GDP spent on health care, might be associated with international cost variation in other contexts, further studies are required to identify factors associated with cost variation more generally. By contrast the conceptual framework, based on using microeconomic theory to generate hypotheses and testing these hypotheses using MLMs would appear transferable to other disease areas and contexts.

A central argument in this thesis is that to understand reasons for cost variation across settings, it is necessary to use a consistent costing methodology in each health care setting. In general, the disaggregated approach taken did enable measurement problems to be identified rather than concealed. However, some of the residual variations at both patient and centre-level may reflect problems in applying consistent methods in each international setting. In particular, the allocation of overheads in each setting relied on the methods used in each finance department, and this may partly explain some of the observed variations in unit costs. A further problem was that none of the conversion factors used allowed the costs reported in a common currency to represent the opportunity costs of those resources in each health care setting.

The thesis used MLMs to distinguish between cost variations at a centre as opposed to a patient-level. The choice of two levels in the models was a pragmatic one, based on the numbers of patients, and centres included in the study. Ideally, the analysis would have included separate levels for centres and countries. However, there were insufficient centres within each country to allow for this third level in the hierarchy. In studies including more centres in each country, the analytical methods presented here could be extended to analyse variation at more than two levels. More complex MLMs could be estimated that recognise the clustering of observations within individuals, individuals within particular care areas e.g. wards, wards within centres, centres within regions, and regions within countries.

Chapter 11: Conclusions

11.0 Contribution of this thesis

The overall contribution of this thesis was to raise awareness about the way costs are currently measured and analysed in economic evaluation. The thesis identified particular problems with the way CEAs currently consider cost variation across health care settings. The thesis highlighted that existing approaches to cost variation across health care settings are inconsistent with economic and statistical theory, and can lead to inaccurate estimates of costs and cost-effectiveness. Basing decision-making on inaccurate assessments of cost-effectiveness can move the allocation of health care resources away from the target of allocative efficiency.

CEAs that recognise cost variation across health care settings can provide a stronger basis for decision-making. This thesis exposed the methodological problems associated with previous attempts to examine cost variation across settings. The major contribution of the thesis was to develop a more appropriate methodology for identifying reasons for cost variation across settings and the implications for economic evaluations. This approach had three important methodological strands, each of which contributed to existing literature in this area.

Firstly, microeconomic theory was used to pose hypotheses for the reasons for cost variation. Previous studies have tended to disregard theory when analysing cost variation across settings. This thesis demonstrated the use of microeconomic theory for identifying factors associated with cost variation across settings. For example it highlighted that costs may vary because of differences in patient factors or centre-level factors such as factor prices.

Secondly, the literature review found that previous studies took a highly aggregated approach to cost measurement. This approach disregarded measurement issues that arose when comparing costs across settings. An important contribution of this thesis was to present a disaggregated approach to cost measurement that could disentangle the reasons for cost variation across international health care settings. Price and volume indices originally developed on time series data, were applied to a different context: to examine the role of price and volume differences in explaining cost variation across health care settings. The empirical investigation used a disaggregated multinational stroke dataset to investigate whether the *a priori reasons* suggested by theory did explain systematic cost variations across settings. The investigation demonstrated that there were wide differences in factor prices across the centres, particular between those located within Eastern compared to Western Europe. The thesis illustrated that taking a more disaggregated approach offered clear insights into the reasons for cost variation across international health care settings.

Thirdly, the thesis demonstrated the use of statistical techniques that recognised the hierarchical structure of international cost and cost-effectiveness data. The use of MLMs in health economics was previously recommended. However, recent cost and cost-effectiveness analyses have not used MLMs and have tended to rely on OLS regression analyses that disregard the inherently hierarchical structure of these data. Statistical theory states that where data are hierarchical, using OLS regression analysis that ignores the clustered nature of observations can lead to incorrect inferences. While the theoretical literature suggested that the choice of technique for analysing cost data *might* matter, the results from the empirical investigation provided an example of where the choice of technique *did* matter. Using OLS regression analysis led to incorrect inferences, in particular the significance of each of the higher-level variables in explaining cost variation was overstated. Once MLMs rather than OLS regression models were used to assess the reasons for cost variation, the only higher-level variable associated with cost variation was the proportion of GDP spent on health care. The empirical investigation also demonstrated that in an evaluative context, OLS regression analysis can lead to inaccurate estimates of the mean incremental cost-effectiveness, and the associated

uncertainty. The thesis used a generated international cost-effectiveness dataset to demonstrate that MLMs were also appropriate for analysing between-centre differences in cost-effectiveness. A further contribution of the thesis was to highlight that the assumptions made when using MLMs in other sectors such as education or health services research, may not be appropriate for analysing cost or cost-effectiveness data. In particular, MLMs typically assume that the residuals are normally distributed, an implausible assumption when analysing skewed cost data. The use of generalised linear mixed models (GLMM) was found to be more appropriate, and proved attractive for analysing cost data that were both skewed and hierarchical. The thesis also highlighted that when analysing multinational cost-effectiveness data it was important to challenge the assumption that the data were exchangeable. The methodology presented emphasised the use of national-level covariates to adjust for systematic differences across health care settings when estimating cost-effectiveness.

In summary, the major contribution of this thesis was to raise concerns about the way costs are measured and analysed in CEA. This thesis extended previous studies by demonstrating the use of a method rooted in economic and statistical theory for identifying reasons for systematic variation across health care settings. This method can be applied to multicentre cost and cost-effectiveness analysis. It is proposed that by recognising variation across health care settings, studies can provide more accurate estimates of cost and cost-effectiveness and the associate levels of uncertainty, and therefore provide a stronger basis for health care decision-making.

11.1 Areas for further research

In addition to making contributions to knowledge in this area, this thesis identified areas that are worthy of further research.

The scope of this thesis was limited to the assessment of cost variation across settings. It is plausible that if resource use varies across health care settings, then there may also be differences in outcomes. Ideally, an economic evaluation would identify factors

associated with variation across settings in outcomes, costs and incremental cost-effectiveness. To assess variations in these parameters, studies need to assess cost-effectiveness across sufficient health care settings.

While the focus of the empirical contribution was on observational cost data collected in a multinational context, the underlying methodological concerns raised by ignoring cost variation across health care settings may apply more generally, for example to national multicentre CEA. MLMs would also appear useful for assessing variation in this context. Indeed, Manca et al. (2005) have suggested that MLMs are appropriate for assessing variation in cost-effectiveness across different health care centres within a country. Further work, across more disease areas and contexts, is required to improve understanding of variations in costs, outcomes and cost-effectiveness using the methodologies presented in this thesis. This research could examine for example whether costs and cost-effectiveness vary if the interventions' costs are measured in a setting treating a high volume as opposed to a low volume of patients.

This thesis highlighted that CEAs often fail to consider the choice of centres for cost measurement. The empirical investigation suggested that in an international context there may be wide variations across centres in resource use, unit costs and total costs. Hence, in this context the choice of centre for costing purposes could have an important impact on the results. Further research is needed to establish *why* costs in particular centres may diverge from opportunity costs.

Finally, there is scope for considering further the measurement issues that arise when when estimating productive efficiency. It appears that practical constraints rather than conceptual reasoning have led the cost function literature to avoid considering inter-patient variability and inconsistency of cost measurement across settings. As economic evaluations continue to collect data alongside multicentre studies it may become more feasible to use techniques for measuring efficiency on more disaggregated datasets. Also, if the quality of routine datasets improve and have more consistent cost estimates and

collect more detailed data on patient case-mix, it may become worthwhile using these datasets to identify reasons for cost variation.

11.2 Policy implications.

The use of economic evaluation in policy-making can make the allocation of scarce health care resources more efficient. However, if economic evaluations provide results that are inaccurate or irrelevant to the decision-maker concerned, then their use may lead to further allocative inefficiency. This thesis argues that ignoring cost variation across health care settings leads to inaccurate results. Also, without information on why costs vary across settings, a decision-maker may be unable to assess how the results from a study can be applied to their particular context. Policy-makers may therefore ignore the results of economic evaluations or inappropriately apply them to their local context.

In several countries, decisions on the use of health care technologies are based, at least in part on the use of CEA. In England and Wales NICE now makes recommendations to the NHS on the use of health care technologies. It is clear from the methodological guidelines NICE issues that such decisions should be based on soundly conducted studies with particular attention given to the uncertainty that surrounds the studies' results (NICE 2004). Yet the guidelines do not give advice on how uncertainty that arises from cost variation across settings should be recognised or dealt with. The results in this thesis suggested that there can be systematic reasons why costs vary across health care settings. The thesis proposes methods for measuring, analysing and interpreting this variation. In particular it suggests that policy-makers should take great care before transferring results from different countries, for use in their own jurisdiction. Decision-makers should be aware that multinational economic evaluations that ignore cost variation across settings may produce inaccurate estimates and may not lead to appropriate health policy decisions. The use of covariates to stratify the results of cost-effectiveness analysis may allow decision-makers in countries not represented by a study to assess the relevance of their results to the local context.

This thesis illustrates a methodological approach that can encourage economic evaluations to provide results that are more relevant to policy-making. It highlights the potential importance of considering cost variation when making decisions about health care priorities. It may also help researchers to think about placing boundaries on the generalisability of their results. Just as developments such as the cost-effectiveness acceptability curve have been useful for communicating sampling uncertainty to policy-makers, so the use of MLMs, may help demonstrate where there are *systematic* variations in costs and cost-effectiveness across settings.

In summary, an important aim of health policy makers is to improve the efficiency of resource allocation. To move towards this goal policy-makers require economic evaluations that use appropriate methodologies. This thesis provides an approach that can identify reasons why costs vary systematically across health care settings. Using this approach can improve the methods used in costing studies. The thesis argues that if future economic evaluations use this approach this can improve the conduct of cost and cost-effectiveness analyses so that these studies provide a sounder basis for health policy-making.

References

Adam T, Evans D, Murray C (2003). Econometric estimation of country-specific hospital costs. *Cost Effectiveness and Resource Allocation* 1(3): 1-31.

Adams-Dudley R, Harrell FE, Smith LR, Mark DB, Califf RM et al. (1993). Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. *J Clin Epidemiol* 46(3): 261-271.

Afford C (2004). Personal communication

Aigner DJ, Chu SF (1968). On estimating the industry production function. *American Economic Review* 58: 226-239.

Aigner DJ, Lovell CAK, Schmidt P (1977). Formulation and estimation of stochastic production frontier models. *J Econom* 6: 21-37.

Aletras VH (1999). A comparison of hospital scale effects in short-run and long-run cost functions. *Health Econ* 8: 521-530.

Antonazzo E, Scott A, Skatun D, Elliott RF (2003). The labour market for nursing: a review of the labour supply literature. *Health Econ* 12: 465-478.

Audit Commission and Department of Health (1999). *NHS Trust Profiles Handbook-1997/8*. Audit Commission: London.

Baladi JF, Coyle D, Faienza B, Jacobs P, Lalonde A et al. (1996). *Guidelines for the Costing Process: the Canadian Experience*. Paper Presented to the Health Economists' Study Group, Brunel University.

Barber JA, Thompson SG (1998). Analysis and interpretation of cost data in randomised controlled trials: review of published studies. *BMJ* 317: 1195-1200.

Barber JA, Thompson SG (2000). Analysis of cost data in randomised trials: an application of the non-parametric bootstrap. *Stat Med* 19: 3219-3236.

Barber JA, Thompson SG (2004). Multiple regression of cost data: use of generalised linear models. *J Health Serv Res Policy* 9(4): 197-204.

Barbieri M, Drummond M, Willke R, Chancellor J, Jolain B, Towse A. Are the results of economic evaluation generalizable? Evidence from studies of pharmaceuticals in Western Europe? *Value in Health* 2005 (in press).

Barnum H, Kutzin J (1993). *Public hospitals in developing countries: resource use, cost, financing*. The Johns Hopkins University Press, Baltimore.

Battese GE, Coelli TJ (1995). A model for technical inefficiency effects in a stochastic frontier production for panel data. *Empirical Economics* 20: 325-332.

Beech R, Ratcliffe M, Tilling K, Wolfe CDA (1996). Hospital Services for stroke care: a European perspective. *Stroke* 27:1958-1964.

Bennett CL, George SL, Vose JM, Nemunaitis JJ, Armitage JL et al. (1995). Granulocyte-macrophage colony-stimulating factor as adjunct therapy in relapsed lymphoid malignancy: implications for economic analyses of phase III clinical trials. *Stem Cells* 13: 414-420.

Berger K, Fischer T, Szucs TD (1998). Cost-effectiveness analysis of paclitaxel and cisplatin versus cyclophosphamide and cisplatin as first-line therapy in advanced ovarian cancer. A European perspective. *Eur J Cancer* 34: 1894-901.

Berman P (1986). Cost analysis as a management tool for improving the efficiency of primary care: Some examples from Java. *Int J Health Plan Manage* 1: 275-288.

Bernt E (2000). International comparisons of pharmaceutical prices: what do we know, and what does it mean? *J Health Econ* 19: 283-287.

Berry RE (1973). On grouping hospitals for economic analysis. *Inquiry X* 5-12.

Bilodeau D, Cremieux P-Y, Oullette, P (2000). Hospital cost function in a non-market health care system. *The Review of Economics and Statistics* 82(3): 489-498.

Birch S, Gafni A (1992). Cost-effectiveness/utility analyses: do current decision rules lead us to where we want to be? *J Health Econ* 11: 279-296.

Birch S, Gafni A (1993). Changing the problem to fit the solution: Johannesson and Weintesin's (mis) application of economics to real world problems. *J Health Econ* 12: 469-476.

Birch S, Gafni A (2002). On being *NICE* in the UK; guidelines for technology appraisal for the NHS in England and Wales. *Health Econ* 11: 185-191.

Bliss S (1999). Spent Force. *Health Services Journal.*

Blundell R, Windmeijer F (1997). Cluster effects and simultaneity in multilevel models. *Health Econ* 6: 439-443.

Bosanquet N, Franks P (1998). Stroke care reducing the burden of disease. Stroke Association, London.

Boskin MJ, Dulberger E, Gordon R, Griliches Z, Jorgenson D (1996). Towards a more accurate measure of the cost of living. Final Report to the US Senate Finance committee, Washington DC, USA.

Bradford WD, Kleit AN, Krousel-Wood A, Re RN (2001). Stochastic frontier estimation of cost models within the hospital. *The Review of Economics and Statistics* 83(2): 302-309.

Brazier J, Deverill M, Green C, Harper R, Booth A (1999). A review of the use of health status measures in economic evaluation. *Health Technol Assess* 3(9): 1-176.

Breyer F (1987). The specification of a hospital cost function. A comment on the recent literature. *J Health Econ* 6: 147-157.

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

Briggs AH (2000). Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics* 17(5): 479-500.

Briggs A, Clark T, Wolstenholme J, Clarke P (2003). Missing... presumed at random: cost-analysis of incomplete data. *Health Econ* 12: 377-92.

Briggs A, Gray A (1998). The distribution of health care costs and their statistical analysis for economic evaluation. *J Health Serv Res Policy* 3(4): 233-245.

Briggs A, Nixon RM, Dixon S, Thompson SG (2005). Parametric modelling of cost data: some simulation evidence. *Health Econ* (in press).

Briggs AH, Gray AM (1999). Handling uncertainty when performing economic evaluation of health care interventions. *Health Technol Assess* 3(2): 1-131.

Brown LC, Epstein D, Manca A, Beard JD, Powell JT et al (2004). The UK Endovascular Aneurism Report (EVAR) trials: Design, methodology and progress. *Eur J Endovasc Surg* 27: 372-381.

Brown RE, Hutton J, Nuijten M (2001). Can unit costs be compared across western European countries? *Value in Health* 4(2): 48-48.

Brouwer W, Rutten F, Koopmanschap M (2001). Costing in economic evaluations. In *Economic Evaluation in Health Care: Merging theory with practice.* Drummond M, McGuire A (eds). Oxford University Press, Oxford.

Bryan S, Brown J (1998). Extrapolation of cost-effectiveness information to local settings. *J Health Serv Res Policy* 3(2): 108-112.

Burgess JF, Christiansen CL, Michalak SE, Morris CN (2000). Medical profiling: improving standards and risk adjustments using hierarchical models. *J Health Econ* 19: 291-309.

Butler JR, Furnival, CM, Hart RF (1995). Estimating treatment cost functions for progressive diseases: A multiproduct approach with an application to breast cancer. *J Health Econ* 14: 361-385.

Byford S, Palmer S (1998). Common errors and controversies in pharmacoeconomic analyses. *Pharmacoeconomics* 13(6): 659-666.

Cairns J, van der Pol M (1997). Saving future lives: a comparison of three discounting models. *Health Econ* 6: 341-350.

Call ST, Holahan WL (1983). *Microeconomics* (second edition). Wadsworth Inc, California.

Canadian Coordinating Office for the Health Technology Assessment of Pharmaceuticals (CCOHTA) (1997). *Guidelines for economic evaluation of pharmaceuticals* (2nd ed). CCOHTA, Ottawa, Canada.

Carey K (2000). A multi-level modelling approach to analysis of patient costs under managed care. *Health Econ* 9: 435-446.

Carpenter JR, Goldstein H, Rasbach J (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Appl Statis* 52, Part 4: 431-443.

Carr-Hill R, Currie L, Dixon P (2003). *Scoping exercise skillmix in secondary care.* Final Report to the NHS SDO R&D Programme, NHS SDO R&D Programme, London.

Carr-Hill RA, Dixon P, Griffiths M, Higgins M, McCaughan D et al. (1995). The impact of nursing grade on the quality and outcome of nursing care. *Health Econ* 4: 57-72.

Casciano J, Doyle J, Casciano R, Kopp Z, Marchant N et al. (2001). The cost-effectiveness of doxazosin for the treatment of hypertension in type II diabetic patients in the UK and Italy. *Int J Clin Pract* 55: 84-92.

Caves DW, Christensen LR, Tretheway MW (1980). Flexible cost functions for multiproduct firms. *Review of Economics and Statistics* 62: 477-481.

Cerniauskas G, Murauskiene L (2002). *Health care systems in transition: Lithuania.* European Observatory on Health Care Systems, Copenhagen.

Claxton K (1999). The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ* 18(3): 341-364.

Claxton K, Sculpher M, Drummond M (2002). A rational framework for decision making by the National Institute For Clinical Excellence (NICE). *Lancet* 360: 711-715.

Coast J, Richards SH, Peters TJ, Gunnell DJ, Darlow MA et al. (1998). Hospital at home or acute hospital care? A cost minimisation analysis. *BMJ* 316: 1802-1806.

Commonwealth Department of Human Services and Health (1995). *Guidelines for the pharmaceutical industry on preparation of submissions to the pharmaceutical benefits advisory committee.* Australian Government Publishing service, Canberra.

Conrad RF, Strauss RP (1983). A multiple-output multiple-input model of the hospital industry in North Carolina. *Applied Economics* 15: 341-352.

Cook JR, Drummond M, Glick H, Heyse JF (2003). Assessing the appropriateness of combining economic data from multinational clinical trials. *Stat Med* 22(12): 1955-1976.

Cook RD, Weisberg S (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

Cowing TG, Holtmannn AG (1983). Multiproduct short-run hospital cost functions: Empirical evidence and policy implications from cross-section data. *Southern Economic Journal* 49: 637-653.

Coyle D, Drummond MF (2001). Analyzing differences in the costs of treatment across centres within economic evaluations. *Int J of Technol Assess Health Care* 17: 155-163.

Coyle D, Davies L, Drummond MF (1998). Trials and tribulations: Emerging issues in designing economic evaluations alongside clinical trials. *Int J of Technol Assess Health Care* 14(1): 135-144.

Cromwell J, Mitchell JB (1986). Physician-induced demand for surgery. *J Health Econ* 5:293-313.

Culyer A (1990). The normative economics of health care finance and provision. In *Providing Health Care*. McGuire A, Fenn P, Mayhew K (eds). Oxford University Press, Oxford.

Culyer AJ, Drummond MF (1978). Financing medical education: interrelationships between medical school and teaching hospital expenditure. In *Economic Aspects of Health Services*. AJ Culyer, KG Wright (eds). Martin Robertson, Oxford.

Culyer AJ, Newhouse J. (2000). State and scope of health economics. In *Handbook of Health economics* volume 1A. Culyer AJ, Newhouse JP (eds). Elsevier, Amsterdam.

Culyer AJ, Wiseman J, Drummond MF, West P (1978). What accounts for the higher costs of teaching hospitals? *Social and Economic Administration* 12: 20-30.

Danzon P, Chao LW (2000). Cross-national price differences for pharmaceuticals: how large, and why? *J Health Econ* 19: 159-195.

Danzon PM, Kim JD (1998). International price comparisons. *Pharmacoeconomics* 14 Suppl 1: 115-128.

Davenport RJ, Dennis MS, Warlow CP (1996). Effect of correcting outcome data for case-mix: an example from stroke medicine. *BMJ* 312: 1503-1505.

Davey Smith G, Barley M, Blane D (1990). The Black report on socioeconomic inequalities in health 10 years on. *BMJ* 310: 373-377.

Dawson D (1994). *Costs and prices in the internal market: markets vs the NHS Management Executive guidelines*. Discussion paper 115, CHE, University of York, York.

Dawson D, Street A (1998). *Reference costs and the pursuit of efficiency in the NHS.* Discussion Paper 161, CHE, University of York, York.

de Pouvourville G, Tasch RF (1993). The economic consequences of NSAID-induced gastrointestinal damage. *Br J Med Econ* 2: 93-102.

Deaton A (2002). Policy implications of the gradient of health and wealth. *Health Affairs* 21(2): 13-30.

Department of Health (2002). *Reference costs 2002: National Schedule of Reference costs*. Department of Health, Leeds.

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

Dervaux B, Ferrier G, Leleu H, Valdmanis V (2004). Comparing French And US Hospital Technologies: A Directional Input Distance Function Approach. *Applied Economics* 36: 1065-1081.

Diehr P, Cain K, Connell F, Volinn E (1990). What is too much variation: The null hypothesis in small-area analysis. *Health Serv Res* 24: 741-771.

Diewert WE (1982). Duality approaches to microeconomic theory. In *Handbook of Mathematical economics, vol 2.* Arrow KJ, Intriligator MD (eds). Academic Press, New York.

Diewert WE (1999). Axiomatic and economic approaches to international comparisons. In *International and interarea comparisons of income, output and prices.* Heston A, Lipsey RE (eds). The University of Chicago Press, Chicago.

Doll JF, Orazem F (1973). *Production economics: theory with applications.* Grid Inc, Columbus, Ohio.

Donaldson C, Currie G, Mitton C (2002). Cost effectiveness analysis in health care: contraindications. *BMJ* 325: 891-4.

Donaldson C, Gerard K (1993). *Economics of health care financing: the visible hand.* The Macmillan Press Ltd, Basingstoke.

Dor A (1994). Non-minimum cost functions and the stochastic frontier: on applications to health care providers. *J Health Econ* 13: 329-334.

Doyle JJ, Casciano J, Arikian S, Tarride JE, Gonzalez MA, Casciano R (2001). A multinational evaluation of acute major depressive disorder (MDD): a comparison of cost-effectiveness between venlafaxine, SSRIs and TCAs. *Value Health 4(1)*: 16-31.

Drummond MF, Bloom BS, Carrin G, Hillman A, Hutchings CH et al. (1991). Issues in the cross-national assessment of health technology. *Int J Technol Assess Health Care* 8(4): 671-682.

Drummond MF, Davies L (1991). Economic analysis alongside clinical trials: revisiting the methodological issues. *Int J Technol Assess Health Care* 7(4): 561-573.

Drummond M, Torrance G, Mason J (1993). Cost-effectiveness league tables: more harm than good? *Soc Sci Med* 37(1): 33-40.

Drummond MF, Jefferson TO (1996). Guidelines for authors and peer reviewers of economic submissions to the BMJ. *BMJ* 313: 275-283.

Drummond MF, O'Brien B, Stoddart GL, Torrance GW (1997a). *Methods for the Economic Evaluation of Health Care Programmes.* Oxford University Press, Oxford.

Drummond M, Cooke J, Walley T (1997b). Economic evaluation under managed competition: evidence from the UK. *Soc Sci Med* 45(4): 583-595.

Drummond M, Sculpher M (2005). Common methodological flaws in economic evaluations. *Medical Care* (in press).

Eccles M, Mason J (2001). How to develop cost-conscious guidelines. *Health Technol Assess* 5(16): 1-69.

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

Elbourne D, Dezateux C, Clarke A, Gray A, King A et al (2002). Ultrasonography in the diagnosis and management of developmental hip dysplasia (UK Hip Trial): clinical and economic results of a multicentre randomised controlled trial. *Lancet* 360: 2009-2017.

Elliott B (2003). Labour markets in the NHS: an agenda for research. *Health Econ* 12: 797-801.

Elliott R, Buxton M (1998). *Valid approximations in individual patient costing on intensive care.* Paper presented at the Health Economics Study Group Meeting, Galway, July, 1998.

Elliott RF, Scott A, Skåtun D, Farrar S, Napper M (2003). *The impact of local labour market factors on the organisation and delivery of health services.* Final Report to the NHS SDO R&D Programme, NHS SDO R&D Programme, London.

Ellis RP, McGuire TG (1986). Provider behaviour under prospective reimbursement: cost sharing and supply. *J Health Econ* 5: 129-251.

Ensor T (2004). Informal payments for health care in transition economies. *Soc Sci Med* 58: 237-246.

Ensor T, Witter S (2001). Health economics in low income countries: adapting to the reality of the unofficial economy. *Health Policy* 57: 1-13.

Escarce JJ (1993). Would eliminating differences in physician practice style reduce geographic variations in cataract surgery rates? *Medical Care* 12: 1106-1118.

Evans RG (1971). Behavioural cost functions for hospitals. *Canadian Journal of Economics* 5: 198-215.

Evans, RG (1990). The dog in the night-time. Medical practice variations and health policy. In *The challenge of medical practice variations.* Anderson TF and Mooney GH (eds). Macmillan, London.

Farrell MJ (1957). The measurement of productive efficiency. *J Royal Stat Soc* 120, series A Part III: 253-281.

Feldstein MS (1967). *Economic analysis for health service efficiency: Econometric studies of the British National Health Service.* North Holland Publishing Company, Amsterdam.

Fenwick E, O'Brien BJ, Briggs A (2004). Cost-effectiveness acceptability curves – facts, fallacies and frequently asked questions. *Health Econ* 13: 405-415.

Fliess JL (1986). Analysis of the data from multiclinic trials. *Control Clin Trials* 7: 267-275.

Fleiss JL (1993). The statistical basis of meta-analysis. *Stat Methods Med Res* 2(2): 121-145.

Folland S, Goodman A, Stano M (1997). *The economics of health and health care.* Prentice Hall, New Jersey.

Folland S, Stano M (1990). Small area variations: A critical review of propositions, Methods and evidence. *Medical Care Review* 47: 419-465.

Folland ST, Hofler RA (2001). How reliable are hospital efficiency estimates? Exploiting the dual to homothetic production. *Health Econ* 10: 683-698.

Forbes JF, Dennis MS (1995). *Costs and Health Outcomes of stroke patients: a prospective study.* Final Project report to the Chief Scientists' Office, Scottish and Home Health Department. Edinburgh.

Fuchs VR, Hahn JS (1990). How does Canada do it? A comparison of expenditures for physicians' services in the United States and Canada. *N Eng J Med* 323: 884-890.

Gafni A, Birch S, Mehrez A (1993). Economics, health and health economics: HYEs versus QALYs. *J Health Econ* 11: 325-339.

Gelman A, Carlin JB, Stern HS, Rubin DR (1998). *Bayesian Data Analysis*. Chapman and Hall, London.

Gerard K, Smoker L, Seymour J (1999). Raising the quality of cost-utility analyses: lessons learnt and still to learn. *Health Policy* 46: 219-238.

Gerdtham UG, Sogaard J, Andersson F, Jonsson B (1992). An econometric analysis of health care expenditure: a cross-section study of the OECD countries. *J Health Econ* 11: 63-84.

Gerschenkron A (1951). *A dollar index of Soviet machinery output, 1927-8 to 1937.* Report-197. Rand, Santa Monica, Poland.

Gilks WR, Richardson S, Spiegelhalter DJ (1996). *Markov chain Monte Carlo in practice.* Chapman and Hall, London.

Glick H, Cook J (2003). Estimating cost-effectiveness in multinational clinical trials. In *Statistical Methods for cost-effectiveness research.* Briggs AH (ed). Office of Health Economics, London.

Glick HA, Orzol SM, Tooley JF, Polsky D, Mauskopf JO (2002). Design and analysis of unit cost estimation studies: How many hospital diagnoses? How many countries? *Health Econ* 12(7): 517-527.

Glick H, Willke R, Polsky D, Llama T, Alves WM et al. (1998). Economic analysis of tirilazad mesylate for aneurysmal subarachnoid hemorrhagge. *Int J Technol Assess Health Care* 14(1): 145-160.

Goddard M, Hutton J (1991). Economic evaluation of trends in cancer therapy: Marginal or average costs. *Int J Technol Assess Health Care* 7: 594-603.

Gold MR, Siegel JE, Russell LB, Weinstein MC (eds) (1996). *Cost-Effectiveness in Health and Medicine.* Oxford University Press, NewYork.

Goldstein H (1992). Statistical information and the measurement of education outcomes. *J R Statist Soc A* 155: 313-315.

Goldstein H (1995). *Multilevel statistical models.* Edward Arnold, London.

Goldstein H, Leyland A (2001). Further topics in multilevel modelling. In *Multilevel Modelling of Health Statistics.* Leyland AH, Goldstein H (eds). John Wiley and Sons, Chichester, England.

Goldstein H, Rasbach J, Yang M, Woodhouse G, Pan H et al (1993). A multilevel analysis of school examination results. *Oxford Review of Education* 19: 425-433.

Goldstein HR, Spiegelhalter DJ (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Statist Soc* 159, Part 3, 385-443.

Goree R, Gafni A, Hannah M, Myhr T, Blackhouse G (1999). Hospital selection for unit cost estimates in multi-centre economic evaluations. Does the choice of hospital make a difference? *Pharmacoeconomics* 15(6): 561-672.

Gravelle H, Rees R (1982). *Microeconomics* (second edition). Longman, New York.

Graves N, Walker D, Raine R, Hutchings A, Roberts JA (2002). Cost data for individual patients included in clinical studies: no amount of statistical analysis can compensate for inadequate costing methods. *Health Econ* 11: 735-739.

Gray AM, Marshall M, Lockwood A, Morris J (1997). Problems in conducting economic evaluations alongside clinical trials. *British Journal of Psychiatry* 170: 47-52.

Gray A, McGuire A, Stuart P (1986). *Factor input in NHS hospitals*. Discussion paper 02/86, Health Economics Research Unit, University of Aberdeen.

Greene WH (1980). Maximum likelihood estimation of econometric frontier functions. *J Econom* 13: 27-56.

Grieve R, Porsdal V, Hutton J, Wolfe C (2000). A comparison of the cost-effectiveness of stroke care provided in London and Copenhagen. *Int J Technol Assess Health Care.* 16: 684-95.

Grieve R, Hutton J, Bhalla A, Rastenytë D, Ryglewicz D et al. (2001a). A comparison of the costs and survival of hospital-admitted stroke patients across Europe. *Stroke* 32: 1684-1691.

Grieve R, Dundas R, Beech R, Wolfe CDA (2001b). The development and use of a method to compare the costs of acute stroke across Europe. *Age and Ageing* 30: 67-72.

Grieve R, Nixon R, Thompson SG, Normand C (2005). Using multilevel models for assessing the variability of multinational resource use and cost data. *Health Econ* (in press).

Gujarati DN (1988). *Basic Econometrics*. 2nd edition McGraw-Hill, Singapore.

Halliday RG, Darba J (2003). Cost data assessment in multinational economic evaluations: some theory and review of published studies. *Appl Health Econ Health Policy* 2(3): 149-55.

Hantrais L, Letablier M-T (1996). *Families and Family Policies in Europe.* Longman, London.

Hatona S (1976). Experience from a multicentre stroke register, a preliminary report. *Bull World Health Organ* 54: 541-553.

Healey A, Mirandola M, Amaddeo F, Bonizzato P, Tansella M (2000). Using health production functions to evaluate treatment effectiveness: an application to a community mental health service. *Health Econ* 9: 373-383.

Heaney DC, Shorvon SD, Sander JW, Boon P, Komarek V et al. (2000). Cost minimization analysis of antiepileptic drugs in newly diagnosed epilepsy in 12 European countries. *Epilepsia* 41 Suppl 5: S37-S44.

Heathfield DF, Wibe S (1981). *An introduction to cost and production functions.* MacMillan Education, Basingstoke.

Henderson RA, Brown R (1999). The costs of routine eptifibatide use in acute coronary syndromes in Western Europe: an economic substudy of the PURSUIT trial. *Eur Heart J* (Suppl N): N35-N41.

Higgins JPT, Whitehead A, Turner RM, Omar RZ, Thompson SG (2001). Meta-analysis of continuous outcome data from individual patients. *Stat Med* 20: 2219-2241.

Hill RJ (1999). International comparisons using spanning trees. In *International and interarea comparisons of income, output and prices.* Heston A, Lipsey RE (eds). The University of Chicago Press, Chicago.

Hoch JS, Briggs, AH, Willan AR (2002). Something old, something new, something borrowed, something BLUE: A framework for the marriage of econometrics and cost-effectiveness analysis. *Health Econ* 11: 415-430.

Hollingsworth B (2003). Non-parametric and parametric applications measuring efficiency in health care. *Health Care Manag Sci* 6(4): 203-218.

Hollingsworth B, Parkin D (1998). *Developing efficiency measures for use in the NHS.* A report to the NHS Executive Northern and Yorkshire R&D Directorate, Health Economics Group, University of Newcastle.

Holmes DR, Califf RM, Van de Werf F, Berger PB, Bates ER et al. (1997) Differences in countries' use of resources and clinical outcome for patients with cardiogenic shock after myocardial infarction: results from the GUSTO trial. *Lancet* 349: 75-78.

Hull RD, Hirsh J, Sackett DL, Stoddart GL (1981). Cost-effectiveness of clinical diagnosis, venography and non-invasive testing in patients with symptomatic deep-vein thrombosis. *New Eng J Med* 304: 1561-1567.

Hurst J (1977). *Saving hospital expenditure by reducing inpatient stay.* Government Economic service occasional paper 14. HMSO, London.

Hurst JW (1991). Reforming health care in seven European nations. *Health Affairs* 10(3): 7-21.

Huttin C, de Pouvourville G (2001). The impact of teaching and research on hospital costs. An empirical study in the French context. *Eur J Public Health* 2(2): 47-53.

Hutton G (2001). *Can the costs of the world health organization antenatal care programme be predicted in developing countries?* PhD thesis. University of London, London.

Hutton J (1994). Economic evaluation of health care: a half-way technology. *Health Econ* 3: 1-4.

Hutton J, Maynard A (2000). A nice challenge for Health Economics. *Health Econ* 9: 89-93.

Iezzoni LI (1997a) *Risk adjustment for measuring healthcare outcomes.* (second edition) Iezzoni LI (ed). Health Administration Press, Chicago.

Iezzoni LI (1997b). Data sources and implications: Information from medical records and patients. In *Risk adjustment for measuring healthcare outcomes.* (second edition) Iezzoni LI (ed). Health Administration Press, Chicago.

Iezzoni LI, Schwarz M, Ash AS, Mackiernan YD (1996). Does severity explain differences in hospital length of stay for pneumonia patients? *J Health Serv Res Policy* 1(2): 65-76.

International Labour Organisation (ILO) (2002). *Health Care in Central and Eastern Europe: Reform, Privatization and employment in four countries.* A draft report to the International Office Infocus Programme on Socio-Economic Security and Public services International. ILO, Geneva.

Jacobs P, Baladi J (1996). Biases in cost measurement for economic evaluation studies in health care. *Health Econ* 5: 525-529.

Jacobs R (2000). Alternative methods to examine hospital efficiency: Data Envelopment Analysis and Stochastic Frontier Analysis. Discussion Paper 177, CHE, York.

Jacobs R, Dawson D (2003). Hospital efficiency targets. *Health Econ* 12(8): 669-684.

Jansen RB, Capri S, Nuijten MJC, Burrell A, Marini MG et al. (1997). Economic evaluation of meloxicam (7.5 mg) versus sustained release diclofenac (100 mg) treatment for osteoarthritis: A cross-national assessment for France, Italy and the UK. *Br J Med Econ* 11: 1-2.

Jansen R, Redekop WK, Rutten FF (2001). Cost effectiveness of continuous terbinafine compared with intermittent itraconazole in the treatment of dermatophyte toenail onychomycosis: an analysis of based on results from the L.I.ON. study. Lamisil versus Itraconazole in Onychomycosis. *Pharmacoeconomics* 19: 401-410.

Jefferson T, Mugford M, Gray A, Demicheli V (1996). An exercise on the feasibility of carrying out secondary economic analyses. *Health Econ* 5: 155-165.

Jensen GA, Morrisey MA (1986). The role of physicians in hospital production. *Review of Economics and Statistics* 68: 432-442.

Jian Z, Jing-Jin Y, Rong-Zhen Z, Xing-Lu Z, Jun Z et al. (1998). Costs of polio immunization days in Cinga: Implications for mass immunization strategies. *Int J Health Plan Manage* 13: 5-25.

Johannesson M, Jonsson B, Kjekshus J, Olsson AG, Pedersen TR et al. (1997). Cost effectiveness of simvastatin treatment to lower cholesterol levels in patients with coronary heart disease. Scandinavian Simvastatin Survival Study Group. *N Engl J Med* 303: 332-6.

Johannesson M, Weinstein MC (1993). On the decision rules of cost-effectiveness analysis. *J Health Econ* 12: 459-467.

Johnston K, Buxton MJ, Jones DR, Fitzpatrick R (1999). Assessing the costs of healthcare technologies in clinical trials. *Health Technol Assess* 3(6): 1-76.

Johnston K, Gerard K (2001). Assessing efficiency in the UK breast screening programme: does size of screening unit make a difference? *Health Policy* 56: 21-32.

Johnston K, Gerard K, Brown J (1998). Generalizing costs from trials. Analyzing centre selection bias in a breast screening trial. *Int J Technol Assess Health Care* 14(3): 494-504.

Jones AM (2000). Health econometrics. In *Handbook of Health Economics, Vol 1.* Culyer AJ, Newhouse JP (eds). Elsevier, Amsterdam.

Jonsson B, Weinstein MC (1997). Economic evaluation alongside multinational clinical trials. Study consideration for GUSTO IIb. *Int J Technol Assess Health Care* 13(1): 49-58.

Karaskevica J, Tragakes E (2001). *Health care systems in transition: Latvia.* European Observatory on Health Care Systems, Copenhagen.

Karski JB, Koronkiewicz A, Healey J (2002). *Health care systems in transition: Poland.* European Observatory on Health Care Systems, Copenhagen.

Kanavos P, Mossialos E (1999). International comparisons of health care expenditures: what we know and what we do not know. *J Health Serv Res Policy* 4:2 122.

Knapp M, Beecham J (1993). Reduced list costings: examination of an informed short cut in mental health research. *Health Econ* 2: 313-322.

Koop GJ, Osiewalski J, Steel MFJ (1997). Bayesian efficiency analysis through individual effects: hospital cost frontiers. *J Econom* 76: 77-105.

Koopmanschap MR, Touw KCR, Rutten FFH (2001). Analysis of costs and cost-effectiveness in multinational trials. *Health Policy* 58: 175-186.

Krasnik A, Groene Wegen P, Pedersen PA, Scholler P, Mooney G et al. (1990). Changing remuneration systems: effects activity in general practice. *BMJ* 300: 1698-1701.

Kreft I, de Leeuw J (1998). *Introducing multilevel modelling*. Sage, London.

Lave J, Lave LB (1970). Hospital cost function. *American Economic Review* 60: 379-395.

Lavers RJ, Whynes DK (1978). A production function analysis of English maternity hospitals. *Socioeconomic Planning Sciences* 12: 85-93.

Lee ML, Wallace RL (1973). Problems in estimating multiproduct cost function: an application to hospitals. *Western Economic Journal* 11: 350-363.

Leyland AH, Goldstein H (eds) (2001). *Multilevel modelling of health statistics*. Wiley, Chichester.

Lindsay CM (1980). *National health issues: the British experience*. Roche Laboratoire, Nutley.

Linna M (1998). Measuring the hospital cost efficiency with panel data models. *Health Econ* 7: 215-427.

Linna M, Hakkinen U, Linnakko E (1998). An econometric study of costs of teaching and research in Finnish hospitals. *Health Econ* 7: 291-305.

Lipscomb J, Anukiewicz M, Parmigiani G, Hasselblad V, Samsa G, Matchar DB (1998). Predicting the cost of illness: a comparison of alternative models applied to stroke. *Med Decis Making* 18: S39-S56.

Localio RA, Berlin JA, Ten Have TR, Kimmel SE (2001). Adjustments for centre in multicenter studies: an overview. *Ann Intern Med* 135: 112-113.

Longworth L, Young T, Ratcliffe J, Bryan S, Buxton M (2001). *Economic evaluation of the Transplantation Programme in England and Wales: An assessment of the costs of liver transplantation.* Unpublished Report to the Department of Health.

Lopez H, Li LZ, Balan DA, Willke RJ, Rittenhouse BE et al (2003). Hospital resource use and cost of treatment with linezolid versus teicoplanin for treatment of serious gram-positive bacterial infections among hospitalized patients from South America and Mexico: results from a multicenter trial. *Clin Ther* 25: 1846-71.

Lorenzoni R, Pagano D, Mazzotta G Rosen SD, Fattore G, et al. (1998). Pitfalls in the economic evaluation of thrombolysis in myocardial infarction. The impact of national differences in the cost of thrombolytics and of differences in the efficacy across patient subgroups. *Eur Heart J* 19: 1518-1524.

Luce B, Manning W, Siegel JE, Lipscomb J (1996). Estimating costs in cost-effectiveness analysis. In *Cost-effectiveness in health and medicine*. Gold MR, Siegel JE, Russell LB, Weinstein MC (eds.) Oxford University Press, New York.

Manca A, Rice N, Sculpher MJ, Briggs AH (2005). Assessing generalisability by location in trial-based cost-effectiveness analysis: the use of multilevel models. *Health Econ* (in press).

Maniadakis N, Thanassoulis E (2000). Assessing productivity changes in UK hospitals reflecting technology and input prices. *Applied Economics* 32: 1575-1589.

Manning W, Fryback D, Weinstein M (1996). Reflecting uncertainty in cost-effectiveness analysis. In *Cost-effectiveness in health and medicine*. Gold MR, Siegel JE, Russell LB, Weinstein MC (eds). Oxford University Press, New York.

Manning WG, Mullahy J (2001). Estimating log models: to transform or not to transform? *J Health Econ* 20: 461-494.

Mark DB, Flatky MA, Califf RM, Naylor CD, Lee KL et al. (1995). Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *New Eng J Med* 332: 1418-1424.

Marshall E, Spiegelhalter DJ (2001). Institutional Performance. In *Multilevel Modelling of Health Statistics.* Leyland AH, Goldstein H (eds). John Wiley and Sons, Chichester, England.

Mason J (1997). The generalisability of pharmacoeconomic studies. *Pharmacoeconomics* 11(6): 503-524.

Maynard A, Kanavos P (2000). Health Economics: an evolving paradigm. *Health Econ* 9: 183-190.

McClellan M, Kessler D (1999). A global analysis of technological change in health care: the case of heart attacks. The TECH Investigators. *Health Affairs* 18(3): 250-255.

McClellan M, Newhouse JP (1997). The marginal cost-effectiveness of a medical technology. A panel instrumental-variables approach. *J Econ* 77: 39-64.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. Chapman and Hall, London.

McGuire A (2001). Theoretical concepts in the economic evaluation of health care. In *Economic Evaluation in Health Care: Merging theory with practice.* Drummond M, McGuire A (eds). Oxford University Press, Oxford.

McKevitt C, Beech R, Pound P, Rudd AG, Wolfe CDA (2000). Putting stroke outcomes into context: assessment of variations in process of care. *Eur J Public Health* 10: 120-126.

McPake B, Kumaranayake L, Normand C (2002). *Health Economics- an international perspective.* Routledge, London.

McPherson K, Wennberg J, Hovind OB, Clifford P (1982). Small area variations in the use of common surgical procedures: An international comparison of New England, England and Norway. *New Engl J Med* 307: 1310-1313.

Medical Research Council (MRC) Laparoscopic Groin Hernia Trial Group (2001). Cost-utility analysis of open versus laparoscopic groin hernia repair: results from a multicentre randomized clinical trial. *British Journal of Surgery* 88: 653-661.

Menzin J, Oster G, Davies L, Drummond MF, Greiner W et al. (1996). A multinational economic evaluation of rhDNase in the treatment of cystic fibrosis. *Int J Technol Assess Health Care* 12: 52-61.

Miller RH, Luft HS (1997). Does managed care lead to better or worse quality of care? *Health Affairs* 16(5): 7-25.

Mitchell JB, Ballard DJ, Whisnant JP, Ammering CJ, Samsa G et al (1996). What role do neurologists play in determining the costs and outcomes of stroke patients? *Stroke* 27: 1937-1943.

Morey RC, Dittman DA (1996). Cost pass-through reimbursement to hospitals and their impacts on operating efficiencies. *Annals of Operations Research* 18: 435-44.

Murray CJ, Evans DB, Acharya A Baltussen R (2000). Development of WHO guidelines on generalized cost-effectiveness analysis. *Health Econ* 9: 235-251.

National Institute for Clinical Excellence (NICE) (2004). *Guide to the methods of Technology Appraisal.* NICE, London.

Netten A, Curtis L (2002). *Unit costs of health and social care 2002.* Personal Social Services Research Unit, University of Kent, Canterbury.

Netten A, Dennett J (1996). *Unit costs of health and social care.* Personal Social Services Research Unit, University of Kent, Kent.

Newhouse J (1994). Frontier estimation: How useful a tool for health economics? *J Health Econ* 13:317-322.

NHS R&D Programme (1998). *The stability of cost-effectiveness analyses.* NHS Health Technology Assessment Programme. SGHT no. 98/22.

Nixon R, Thompson SG (2005). Incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Econ* (in press)

O'Brien BJ 1997. A tale of two (or more) cities: geographic transferability of pharmacoeconomic data. *Am J Manag Care* 3: S33-S39.

O'Donnell O, Propper C (1991). Equity and the distribution of National Health Service resources. *J Health Econ* 10: 1-10.

OECD (2000). *Main economic indicators* 2000. OECD, Paris.

OECD (2001). *OECD Health Data 2001, A comparative analysis of 30 countries.* OECD, Paris.

Oostenbrink J, Koopmanschap MA, Rutten FFH (2000). *Costing manual for economic evaluation in health care* (in Dutch). Health Insurance Council, Amstelveen, Holland.

Palmer S, Raftery J (1999). Economic Notes: opportunity cost. *BMJ* 318: 1551-1552.

Pang F (2002). Design, analysis and presentation of multinational studies: the need for guidance. Pharmacoeconomics 20(2): 75-90.

Parente ST (1989). *Measuring the substitution of outpatient for inpatient services.* Masters thesis. University of Rochester, Rochester.

Phelps CE (1995). Perspectives in health economics. *Health Econ* 4: 335-353.

Phelps CE, Mooney C (1993). Variations in medical practice use, causes and consequences. In *Competitive approaches to health care reform*. Arnoud RJ, Rich RF, White WD (eds). The Urban Institute Press, Washington, DC, 139-178.

Phelps CE, Parente ST (1990). Priority setting in medical technology and medical practice assessment. *Med Care* 28: 703-723.

Porsdal V, Boysen G (1997). Cost-of illness studies of stroke. *Cerebrovascular Dis* 7: 258-63.

Pritchard C (2004). *Developments in economic evaluation in health care: a review of HEED*. OHE, London.

Raikou M, Briggs A, Gray A, McGuire A (2000). Centre-specific or average unit costs in multi-centre studies? Some theory and simulation. *Health Econ* 9: 191-198.

Rasbach J, Browne W, Goldstein H, Yang M, Plewis I et al. (2002). *A user's guide to MLwiN: version 2.1c.* Institute of Education, University of London, London.

Reed SD, Friedman JY, Gnanasakthy A, Schulman KA (2003). Comparison of hospital costing methods in an economic evaluation of a multinational clinical trial *Int J Technol Assess Health Care* 19(2): 396-406.

Reed SD, Friedman JY, Velazquez EJ, Gnanasakthy A, Califf RM, Schulman KA (2004). Multinational economic evaluation of valsartan in patients with chronic heart failure: results from the Valstartan Heart Failure Trial (Val-HeFT). Am Heart J 148(1): 122-8.

Register C, Bruning ER (1987). Profit, incentives and technical efficiency in the production of hospital care. *Southern Economic Journal* 53(9): 899-914.

Reinhardt UE (1970). *An Economic Analysis of Physicians' Practices.* PhD thesis, Yale University, Yale.

Ricci S, Celani M, LaRosa F, Vitali R, Duca E, Ferraguzzi R et al (1991). SEPIVAC: a community-based study of stroke incidence in Umbria, Italy. *J Neurol Neurosurg Psychiatry* 54: 695-698.

Rice N (2001). Binomial regression. In *Multilevel Modelling of Health Statistics.* Leyland AH, Goldstein H. (eds). John Wiley and Sons, Chichester, England.

Rice N, Jones A (1997). Multilevel models and health economics. *Health Econ* 6: 561-575.

Rice N, Leyland A (1996). Multilevel models: application to health data. *J Health Serv Res Policy* 1(3): 154-164.

Richardson G, Maynard A (1995). *Fewer doctors? More nurses? A review of the knowledge base of doctor-nurse substitution.* Discussion paper 132, Centre for Health Economics, University of York, York.

Roos N, Roos L (1982). Surgical rate variations Do they reflect the health or socioeconomic characteristics of the population? *Med Care* 20: 945-958.

Rosen S (1986). The theory of equalising differences. In *Handbook of Labor Economics.* Vol 11. Ashenfelter O, Layard PRG (eds). North Holland, Amsterdam.

Rosko M (1999). Impact of internal and external environmental pressures on hospital inefficiency. *Health Care Mgt Science* 2: 64-78.

Rosko MD (2001). Cost efficiency of US hospitals: a stochastic frontier approach. *Health Econ* 10: 539-551.

Rutten-van Molken PMH, van Doorslaer EKA, Till MD (1998). Cost-effectiveness analysis of Formoterol versus Salmeterol in patients with asthma. *Pharmacoeconomics* 14(6): 671-684.

Samsa G, Reutter RA, Parmigiani G, Ancukiewicz M, Abramse P et al. (1999). Performing cost-effectiveness analysis by integrating randomised trial data with a comprehensive decision model: application treatment of acute ischemic stroke. *J Clin Epidem* 52(3): 259-271.

Scandinavian Simvastatin Survival Study (4S) (1994). Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Study (4S). *Lancet* 244: 1383-1389.

Schmidt P, Lovell CAN (1979). Estimating technical and allocative efficiency relative to stochastic production and cost frontiers. *J Econom* 9: 343-366.

Schreyer P, Koechlin F (2002). Purchasing power parities- measurement and uses. *Statistics Brief* 3: 1-8.

Schulman KA, Burke J, Drummond M, Davies L, Carlsson P et al. (1998). Resource costing for multinational neurologic clinical trials: methods and results. *Health Econ* 7: 629-638.

Schulman K, Buxton, Glick H, Sculpher M, Guzman G, et al. (1996). Results of the economic evaluation of the first study. A Multinational Prospective economic Evaluation. *Int J Technol Assess Health Care* 12(4): 698-713.

Scott A, Shiell A (1997). Do fee descriptors influence treatment choices in general practice? A multilevel discrete choice model. *J Health Econ* 13: 323-342.

Sculpher MJ, Bryan S, Dwyer N, Hutton J, Stirrat GM. (1993). An economic evaluation of transcervical endometrial resection versus abdominal hysterectomy for the treatment of menorrhagia. *Br J Obstet Gynaecol* 100(3):244-52.

Sculpher MJ, Drummond MF, Buxton M (1997). The iterative use of economic evaluation as part of the process of health technology assessment. *J Health Serv Res Policy* 2: 26-30.

Seshamani M, Gray AM (2004). A longitudinal study of the effects of age and time to death on hospital costs. *J Health Econ* 23: 217-235.

Sheiman I (1994). Forming the system of health insurance in the Russian Federation. *Soc Sci Med* 39(10): 1425-1432.

Smith PC (1995). Performance indicators, control in the public sector. In *Management Control: Theories, Issues and Practices*. Berry AJ, Broadbent J, Otley D (eds). Macmillan, Basingstoke, 163-178.

Sloan F (2000). Not-for-profit ownership and hospital behaviour. In *Handbook of Health Economics, Vol 1*. Culyer AJ, Newhouse JP (eds). Elsevier, Amsterdam.

Söderlund N, Gray A, Milne R, Raftery J (1996). Case mix measurement in English hospitals: an evaluation of five methods for predicting resource use. J Health Serv Res Policy 1(1): 10-19.

Söderlund N, Milne R, Gray A, Raftery J (1995). Differences in hospital case-mix, and the relationship between case-mix and hospital costs. *J Public Health Med* 17(1): 25-32.

Späth HM, Carrere M-O, 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: 161-77.

Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR (2000). Bayesian methods in health technology assessment: a review. *Health Technol Assess* 4(38): 1-130.

Stalhammar NO, Carlsson J, Peacock R Muller-Lissner S, Bigard MA et al (1999). Cost effectiveness of omeprazole and ranitidine in intermittent treatment of symptomatic gastro-oesophageal reflux disease. *Pharmacoeconomics* 16: 483-97.

Stata v 8.1 (2002) Stata Corp 4905 Lakeway Drive, College station, Texas. 77485 USA. http://www.stata.com.

Steele R, Gray AM (1982). Statistical cost analysis: the hospital case. *Applied Economics* 14: 491-502.

Street A (2003). How much confidence should we place in efficiency estimates? *Health Econ* 12: 895-908.

Stroke Unit Trialists' Collaboration (1999). Organised inpatient (stroke unit) care for stroke (Cochrane Review). In The Cochrane Library, Issue 1. Update Software, Oxford.

Stroke Unit Trialists' Collaboration (2001). Organised inpatient (stroke unit) care for stroke. The Cochrane Database of Systematic Reviews. In the Cochrane Library, Issue 3. Update Software, Oxford.

Suri R, Metcalfe C, Lees B, Grieve R, Normand C et al. (2001). RhDNase in children with cystic fibrosis: a randomised trial. *Lancet* 358: 1316-1321.

Tan-Torres Edejar, Baltussen R, Adam T, Hutubessy R, Acharya A et al. (2003). *WHO guide to cost-effectiveness*. WHO, Geneva.

Thompson R, Witter S (2000). Informal payments in transitional economies: Implications for health sector reform. *Int J Health Plan Manage* 15: 169-187.

Thompson SG (1993). Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Statistical Methods in Medical Research* 2: 173-192.

Thompson SG, Barber JA (2000). How should cost data in pragmatic randomised trials be analysed? *BMJ* 320: 1197-2000.

Thompson SG, Sharp SJ (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* 18: 2693-2708.

Torrance G, Siegel J, Luce B (1996). Framing and designing the cost-effectiveness analysis. In *Cost-effectiveness in health and medicine*. Gold MR, Siegel JE, Russell LB, Weinstein MC (eds). Oxford University Press, New York.

Tsuchiya A, Williams A (2001). Welfare economics and economic evaluation. In *Economic Evaluation in Health Care: Merging theory with practice*. Drummond M, McGuire A (eds). Oxford University Press, Oxford.

Tuomilehto J, Sarti C, Narva EV, Salmi K, Sivenius J et al (1992). The FINMONICA stroke register: community-based stroke incidence in Finland, 1983-1985. *Am J Epidemiol* 135: 1259-1270.

UK Small Aneurysm Trial Participants (1998). Health service costs and quality of life for early elective surgery or ultrasonographic surveillance for small abdominal aortic aneurysms. *Lancet* 352: 1656-60.

US General Accounting Office (GAO) (1992). *Prescription Drugs: Companies typically charge more in the United States than in Canada*. GAO/HRD 92-110.

Valdmanis V (1990). Ownership and technical efficiency of hospitals. *Medical Care* 28(6): 552-561.

Valdmanis V, Walker D, Fox-Rushby J (2003). Are vaccination sites in Bangladesh scale efficient? *Int J Technol Assess Health Care*. 19(4): 692-697.

Van Andel (1997). *Health financing and pharmaceutical policy reform in the countries of central and eastern Europe*. OHE briefing No 35. OHE, London.

Van Ark B, Monnikhof E, Timmer M (1999). Prices, quantities, and productivity in industry: a study of transition economies in a comparative perspective. In *International and interarea comparisons of income, output and prices*. Heston A, Lipsey RE (eds). The University of Chicago Press, Chicago.

Vita MG (1990). Exploring hospital production relationships with flexible functional forms. *J Health Econ* 9:1-22.

Vitaliano DF (1987). On the estimation of hospital cost functions. *J Health Econ* 6: 305-318.

Vitaliano D, Toren M (1994). Cost and efficiency in nursing homes: a stochastic frontier approach. *J Health Econ* 13: 281-300.

Voss GBWE, Hasman A, Rutten F, de Zwann C, Carpay JJ (1994). Explaining cost variations in DRGs 'Acute Myocardial Infarction' by severity of illness. *Health Policy* 28: 37-50.

Wade DT (1994). Stroke (acute cerebrovascular disease). In *Health care needs assessment: the epidemiologically based needs assessment reviews*. Vol. 1. Stevens A, Raftery J (eds). Radcliffe Medical Press, Oxford.

Wagstaff A (1989a). Econometric studies in health economics. A survey of the British literature. *J Health Econ* 8: 1-51.

Wagstaff A (1989b). Estimating efficiency in the hospital sector: a comparison of three statistical cost frontier models. *Applied Economics* 21: 659-672.

Walker A, Major K, Young D, Brown A (1997). *Economic costs in the NHS: A useful insight or just bad accountancy?* Paper presented to Health Economists' Study Group, University of Liverpool, Liverpool.

Weatherly H, Drummond M, Smith D (2002). Using evidence in the development of local health policies. *Int J Technol Assess Health Care* 18(4): 771-781.

Wennberg, JE (1984). Dealing with Medical Practice variations: A proposal for action. *Health Affairs* 3: 6-32.

WHO (1979). *Expanded programme on immunization: Costing guidelines.* EPI/GEN/79/5.

Whynes DK, Walker AR (1995). On approximations in treatment costing. *Health Econ* 4: 31-39.

Willke R, Glick HA, Polsky D, Schulman K (1998). Estimating country-specific cost-effectiveness from multinational clinical trials. *Health Econ* 7: 481-493.

Wolfe CDA, Tilling K, Beech R, Rudd AG (1999). Variation in case fatality and dependency from stroke in Western and Central Europe. *Stroke* 30: 350-356.

Wolstenholme J (2001). *Counting the costs of cancer care: breast cervical and lung cancer in Trent.* PhD thesis. University of Nottingham, Nottingham.

Wordsworth S, Ludbrook A (2005). Comparing costing results in across country economic evaluations: the use of technology specific purchasing power parities. *Health Econ* (in press).

World Bank (2000). *World Bank Indicators 2000.* 4th edition. World Bank, Washington DC.

Wouters A (1993). The cost and efficiency of public and private health care facilities in Ogun state, Nigeria. *Health Econ* 2: 31-42.

Zuckerman SJ, Hadley J, Iezzoni L. (1994). Measuring hospital efficiency with frontier cost functions. *J Health Econ* 13: 255-280.

Appendices

Appendix 1: Biomed 2: resource use and costing questionnaire

This questionnaire lists the resource and cost data that we need to collect during our fieldwork visits to centres. We have also indicated who is likely to supply these data based on our experiences in United Kingdom hospitals. Please modify the contact list if you think that it is more appropriate for us to obtain the data from an alternative source.

The focus of the costing exercise will be the main study hospital so most of our visit will concentrate on getting information from cost centres at this hospital. However, where a substantial amount of care is provided at more than one hospital site (as in Dijon where patients are transferred to a rehabilitation unit), we will need to visit these other hospitals to collect the information required.

The questionnaires will be completed at the meetings we will have during our visit. We are distributing copies of the questionnaire now, to help centre co-ordinators plan our visits, and to give staff adequate warning of the type of data we need.

(1) Background information

Contact: the centre co-ordinator

- A confirmation of the wards where the major stroke care provision is.
- An update on data collection at the centre
- A discussion of concerning more detailed study in this centre (see section 8)

(2)Cost category: Nursing costs

Contact: Nurse Managers at the wards providing the majority of care for stroke patients.

(e.g. general medicine, elderly care, neurology, intensive care unit etc)

- The total number of beds on the ward.

- The number of occupied bed days used per year and the current occupancy of the ward.

- The number of occupied days per year used by stroke patients.

- The nursing levels on each of these wards in terms of whole time equivalents by grade/ type of nurse (during night shift/ day shift).

- The role of each type/grade of nurse in the care of patients and the qualifications and experience that they need before they can be employed in that role.

(3) Cost category: Doctors' costs

Contact: Representatives from the main clinical specialities involved in the care of stroke patients.

- The total number of clinical staff available for inpatient and non-inpatient care (all patients) in terms of whole time equivalents by grade or clinical sessions by grade.

- The total number of beds that the inpatient staff cover (all patients).

- The total number of clinical staff available for the inpatient and non-inpatient care of stroke patients (measured in terms of whole time equivalents by grade).

- The role of each type/grade of clinician in the care of patients and the qualifications and experience that they need before they can be employed in that role.

- The proportion of time spent by clinicians on clinical duties (as opposed to lecturing, research etc).

(4)Cost category: Investigation costs

Contact: Heads of departments providing CT scans, MRI, angiography, doppler and Echocardiogram services

- The average cost of providing each of these scans
- For a CT scan the average staff time involved in conducting a scan and the type of equipment used.

(5) Cost category: Cost of therapy

Contact: Physiotherapists/ Occupational therapists

- The total number of therapists by grade/type available for inpatient and non-inpatient therapy (all patients).

- The number of beds covered by the therapists

- The number of hours that a therapist works per week.

- The role of each type/grade of therapist in the care of patients and the qualifications and experience that they need before they can be employed in that role.

- The options for rehabilitation for stroke patients post discharge

(6) Cost category: Overall cost per hospital day

Contact: A representative of the hospital finance department

- The average cost per month of each grade of the following:

(salary + any additional costs which employers have to pay) of all:

nurses

doctors

therapists

- The total inpatient and non-inpatient expenditure of the hospital over the most recent financial year.

- The total number of occupied bed-days in the whole hospital during the most recent financial year

The **total inpatient spending** in the department most relevant to stroke patients (e.g. neurology) on the following areas

- Drugs
- Consumables (e.g. surgical equipment, bandages etc)
- Food
- Transport
- Maintenance (repairs to machinery etc)
- Administration (secretarial costs, management costs both direct to the department or allocated from the central hospital)
- Cleaning
- Porters (people who move the patients from place to place)
- Energy (Gas, electricity, water)

(7) More detailed study

Contact: Centre co-ordinator.

- In order to find out more detailed information on the resource use associated with stroke care we would like if possible to look at the notes of patients with the project co-ordinator. The purpose of this study is to collect information on:

- Staff contacts
- Investigations
- Drugs

- To facilitate this, please could you have **10 sets of medical records available** for us to look at with you on the site visit for patients who are about to be discharged or have been discharged from the study hospital.

(8)Cost category: Other hospitals

If stroke patients who are included in the Biomed centre are likely to be treated as inpatients at local/ rehabilitation hospitals we will need visit these centre and repeat the exercise there.

(9)Cost category: Services in the community

Contact: Centre co ordinator and representative from community rehabilitation facilites.

- To discuss rehabilitation and other services which are available place in the community
- Possibly to pilot a 3 month questionnaire looking at resource use in the community

(10) General

In addition to the above, we would be grateful for any other data that are routinely available which describe services used by stroke patients or the costs of stroke in your hospital or locality.

Appendix 2: Unit costs ($/PPP)

	Port	Spain	Italy	France	Den	Fin 1	Fin 2	Fin 3	UK	Poland	Lith 1	Lith 2	Latvia
Acute day	263	226	333	172	234	157	227	112	194	112	50	48	23
CT scan	209	161	181	124	66	221	139	203	98	124	62	62	49
Doppler	377	99	97	124	88	102	226	102	77	99	114	114	4
Rehabilitation day	Na	Na	188	179	234	86	112	112	Na	111	62	62	32
Outpatient visit	23	23	21	27	75	22	30	29	38	43	4	4	4
Outpatient therapy	20	27	33	16	27	32	32	32	38	72	19	19	6
Nursing home day	Na	84	132	35	98	63	63	63	56	36	29	29	20
GP visit	18	21	20	19	16	10	10	10	7	25	14	14	2
District Nurse visit	24	26	32	27	32	26	26	26	33	18	4	4	8
Home carer visit	5	7	23	18	26	26	26	26	7	18	Na	Na	Na

NA: not applicable, Port: Portugal, Den: Denmark, Fin: Finland, Lith: Lithuania

Appendix 3: Paper by Grieve et al. (2005)

HEALTH ECONOMICS
Health Econ. 14: 185–196 (2005)
Published online 14 June 2004 in Wiley InterScience (www.interscience.wiley.com). DOI:10.1002/hec.916

Using multilevel models for assessing the variability of multinational resource use and cost data

Richard Grieve[a,*], Richard Nixon[b], Simon G. Thompson[b] and Charles Normand[a]
[a] *Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, UK*
[b] *MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK*

Summary

Multinational economic evaluations often calculate a single measure of cost-effectiveness using cost data pooled across several countries. To assess the validity of pooling international cost data the reasons for cost variation across countries need to be assessed. Previously, ordinary least-squares (OLS) regression models have been used to identify factors associated with variability in resource use and total costs. However, multilevel models (MLMs), which accommodate the hierarchical structure of the data, may be more appropriate. This paper compares these different techniques using a multinational dataset comprising case-mix, resource use and cost data on 1300 stroke admissions from 13 centres in 11 European countries. OLS and MLMs were used to estimate the effect of patient and centre-level covariates on the total length of hospital stay (LOS) and total cost. MLMs with normal and gamma distributions for the data within centres were compared. The results from the OLS model showed that both patient and centre-level covariates were associated with LOS and total cost. The estimates from the MLMs showed that none of the centre-level characteristics were associated with LOS, and the level of spending on health was the centre-level variable most highly associated with total cost. We conclude that using OLS models for assessing international variation can lead to incorrect inferences, and that MLMs are more appropriate for assessing why resource use and costs vary across centres. Copyright © 2004 John Wiley & Sons, Ltd.

Keywords multilevel modelling; multinational studies; costing methodology; economic evaluation

Introduction

An increasing number of economic evaluations are being conducted on a multinational basis [1–6]. For example, RCTs often recruit patients from several countries and calculate an average measure of the effectiveness of the new intervention across the study centres. Multinational economic evaluations have then used these trial data, combined with a pooled measure of the incremental cost, to calculate a single, global measure of cost-effectiveness [4,5]. This practice has been criticised as unit costs, resource use and patient outcomes may all vary across countries [1]. To understand the implications of this variability on the design, analysis and interpretation of multinational cost-effectiveness studies, further research is needed on why these parameters vary across health care settings.

A good starting point for assessing why cost-effectiveness varies across health care settings, is to examine why resource use varies, since resource use is associated with total costs and outcomes. Previous work has suggested that reasons for resource use variability may operate at different levels. For example Raikou *et al.* argue [7] that

*Correspondence to: Health Services Research Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. E-mail: richard.grieve@lshtm.ac.uk

Received 2 June 2003
Accepted 23 January 2004

relative price, which is usually measured at the level of the health care centre or country concerned, is likely to be important in determining the relative level of resource utilisation in different centres. The practice variations literature suggests that patient-level variables, in particular case-mix, are associated with resource use differences [8–10]. To investigate the accuracy (or inaccuracy) of pooling data across different health care settings, the extent to which these different factors are associated with variations in resource use needs to be investigated. The approach can then be extended to examine differences in cost and cost-effectiveness across settings.

The reasons for variability should be assessed using appropriate analytical methods. Studies in this area have generally used ordinary least-squares (OLS) models [1,11]. These models assume that observations across patients are independent and have a common variance. This assumption would seem unlikely to hold when using data from different centres, as patients' resource use (or costs) within a particular centre may be more similar than that in different centres. Furthermore, centre-level variables included in an OLS model are considered as if they were measured at a patient level, thus spuriously inflating the amount of information they supply. By contrast, multilevel models (MLMs) are able to incorporate the hierarchical structure of the data (that is, of patients within centres), and provide more appropriate estimates of patient and centre-level effects. MLMs have been recommended for use in health economics [12], but despite the obvious intuitive appeal of using MLMs to assess multinational resource use and cost data, they have not yet been used for this purpose.

The aim of this paper is to compare the use of MLMs and OLS models for assessing the reasons for international resource use and cost variations. These methods are illustrated using an observational costing study which included patient and centre-level data from 13 centres in 11 different European countries.

Methods

This study compares the use of OLS models and MLMs for estimating the extent to which factors are associated with length of hospital stay and total cost. The basic structure of MLMs and the rationale for their use in health care and health economics has been previously described [12–14], so only a summary is provided here. In our context of multinational data, patients (level-1 units) are nested within centres (level-2 units).

Models

An OLS regression model takes the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \text{Normal}(0, \sigma^2)$$

where y_i is the outcome variable for the ith patient, x_i is an explanatory variable, with associated slope coefficient β_1. β_0 is the intercept and ε_i, the error term which represents unexplained variability between patients, is assumed to be normally distributed with a mean of zero. The OLS model assumes that the variance of the error term is the same for all patients. Extra explanatory variables can be included. One representing a centre-level covariate (z_i say) necessarily takes the same value for all patients in a particular centre

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i, \quad \varepsilon_i \sim \text{Normal}(0, \sigma^2)$$

Moving to a MLM structure changes the way the unexplained variation, the random error term, is modelled. The most basic MLM, the random intercepts model, includes an additional term which represents the unexplained variation that exists between centres. Using subscripts i and j for the ith patient in jth centre, the model may be written as

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{Normal}(0, \sigma^2),$$
$$u_j \sim \text{Normal}(0, \tau^2)$$

where β_0 is a fixed quantity applying to all patients, u_j is a random variable with zero mean and constant variance (τ^2) which applies to the patients in centre j, and ε_{ij} is a random error term which represents the unexplained variation for patients within a centre. u_j indicates the random effect of centre on the outcome variable, over and above that explained by the set of explanatory variables. The intercept for the jth centre (previously given as β_0) is now given as a fixed component (β_0) plus a random component (u_j). The model can be developed by including additional explanatory variables at the level of the patient. The regression coefficients (such as β_1) can also be allowed to vary between centres, but we do not use such 'random slopes' models in this paper.

A centre-level explanatory variable, taking the value z_j for centre j, can be included as

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + u_j + \varepsilon_{ij},$$

$$\varepsilon_{ij} \sim \text{Normal}(0, \sigma^2), \quad u_j \sim \text{Normal}(0, \tau^2)$$

The variance τ^2 now represents the residual variability between centres not explained by the covariates [15]. Hence MLMs allow the effect of adding centre-level variables on the extent of unexplained variation between centres to be estimated. In addition, the degree of dependency between observations can be measured by the intra-class correlation coefficient (ρ) defined as $\rho = \tau^2/(\tau^2 + \sigma^2)$. This reflects the strength of 'nesting' within the data hierarchy.

The basic MLM above assumes that the errors terms are normally distributed. However, resource use and cost data are usually not normally distributed [16]. Generalised linear models (GLMs) models have been recommended for analysing cost data [Barber JA, Thompson SG. Multiple regression of cost data: use of generalised linear models. *J Health Serv Res Policy* 2004, in press], as a variety of non-normal distributions can be specified but, unlike data transformations in OLS regression, they make inferences about the mean cost directly. A number of alternative GLMs were considered for our analysis. A gamma distribution model with random mean (θ_{ij}), and different shape (ϕ_j) parameters for each centre, was chosen as it fitted the positively skewed LOS and cost data reasonably well:

$$y_{ij} \sim \text{Gamma}(\theta_{ij}, \phi_j), \quad \theta_{ij} = \beta_0 + \beta_1 x_{ij} + u_j,$$

$$u_j \sim \text{Normal}(0, \tau^2)$$

This 'generalised linear mixed model' (GLMM) is a type of multilevel model similar to model 2 in that it allows for random centre effects (u_j), but it accommodates positively skewed data by the use of a gamma rather than a normal distribution, allowing the gamma distributions to have different shapes in each centre. Again this can be extended to include centre-level covariates z_j:

$$y_{ij} \sim \text{Gamma}(\theta_{ij}, \phi_j),$$

$$\theta_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + u_j, \quad u_j \sim \text{Normal}(0, \tau^2)$$

Data

The case-study used to compare these methods was an observational study which compared the costs and outcomes of different ways of providing stroke care across 13 centres in 11 different European countries. The methodology and main results have been reported elsewhere [18,19]. Briefly, the study centres recruited patients prospectively who had suffered a first-ever stroke and attended each of the study centres over a 1-year period during 1996–1997. Data were collected for 3 months post-stroke.

Patient-level variables

Designated investigators at each centre completed standard forms to record patient-level variables [20]. Data were collected on patient characteristics (sex, age, pre-stroke living conditions); stroke severity measures (incontinence during the first week after stroke, dysphasia, paralysis at hospital admission) and stroke subtype (cerebral infarction, intra-cerebral haemorrhage, or unspecified stroke). The use of hospital and community services was recorded for three months post-stroke. The unit costs of each resource use item were collected in a consistent and disaggregated manner at each centre on site visits [18]. For example for labour inputs, the costs per hour of employing each grade of health care professional were measured, and multiplied by the level of labour inputs (hours per occupied bed-day) to give the labour costs per bed-day. Unit costs were initially reported in local currencies (1998 prices) and then converted into US dollars using the purchasing power parity (PPP) index to adjust for differences in the opportunity cost of resources across the economies concerned [21–23]. Finally, total 3-month costs ($/PPP) were calculated for each patient.

The main outcome measure for the original study was survival post-stroke, but for this analysis patients who had died during the 3 months post-stroke were excluded. This was because our focus was on comparing methodologies for assessing resource use and cost data, rather than trying to assess the role of outcome in understanding resource use differences. Some patient-level variables for the patients included in the analysis ($n = 1298$) are presented in Table 1.

Centre-level variables

Data on centre-variables potentially associated with resource use differences were obtained from

Table 1. Summary of patient-level variables

Centre	*N*	Mean (sd) *LOS* (days)	Mean (sd) total cost (US $/PPP)	Mean (sd) age (years)	% paralysed	% incontinent
Spain	32	7.8 (6.3)	2805 (2514)	72.3 (8.4)	84	28
Portugal	73	13.0 (14.2)	5405 (4353)	64.4 (11.5)	84	25
Latvia	145	18.7 (11.1)	550 (370)	62.7 (10.0)	82	12
Italy	109	20.2 (23.0)	4155 (2565)	74.2 (11.0)	81	40
Finland 1	37	19.1 (20.2)	6170 (5293)	72.9 (9.4)	65	19
France	105	24.2 (27.5)	6203 (5015)	70.9 (17.6)	69	18
Lithuania 1	62	28.0 (13.5)	1439 (919)	69.9 (11.1)	89	25
Poland	92	25.5 (15.9)	1659 (711)	69.6 (11.6)	95	20
Lithuania 2	186	26.5 (16.5)	4072 (2303)	66.9 (10.5)	68	16
Denmark	246	37.0 (33.7)	9311 (7983)	67.6 (13.9)	70	33
UK	73	39.3 (33.8)	6515 (3215)	71.3 (11.8)	67	26
Finland 2	81	36.5 (31.1)	9599 (7706)	70.1 (11.6)	74	35
Finland 3	57	37.4 (29.0)	9036 (5815)	80.8 (7.3)	60	42
Total	1298	27.3 (25.6)	5340 (5884)	69.3 (12.6)	77	26

reviewing the literature, and visiting the centres concerned. The variables used in the analysis are presented in Table 2, and defined as follows: the level of health spending was taken as the proportion of Gross Domestic Product (% GDP) spent on public health care; the reimbursement system for acute hospital costs was defined as being either a global budgeting or a diagnosis-related group (DRG) system; the level of patient co-payment for acute care was recorded for each centre, and defined as a categorical (Yes/No) variable.

International price indices were constructed to estimate differences in the relative price of health inputs across the centres. Labour input prices were used to construct the indices as labour costs make up the majority of total stroke care costs [24]. Laspeyres price indices were calculated which weight the relative price of each input by the reference centre's resource volumes [25]. The French centre was taken as the reference centre as it had the median level of unit costs. Paasche indices were also constructed which weight relative prices by each comparison centre's resource volumes [25]. The geometric mean of the Laspeyres and Paasche indices was calculated to give the Fisher price index [25]. The Fisher price index was used as the relative price variable in the analysis, and is presented in Table 2.

Estimation

An OLS regression model was fitted to estimate the effect of patient-level variables on the total length of hospital stay (LOS) (model 1). Variables were included if there was an *a priori* reason that they were likely to be associated with LOS, and were retained in the equation even if they did not reach conventional levels of significance. A series of diagnostic tests were performed for heteroscedasticity, multi-colinearity and correct functional form. The analysis was repeated with normal MLM random intercepts (model 2) and gamma GLMM random intercepts (model 3). The normal model was estimated by restricted iterative generalised least squares in MLwiN [30], equivalent to restricted maximum likelihood. The gamma models were fitted using Markov chain Monte Carlo methods in WinBUGS [31]. The LOS analyses were then repeated also including centre-level variables (models 4–6). Finally, each of the models with patient and centre-level variables were re-run with total cost as the dependent variable (models 7–9). For the total cost models price index was not considered as an independent variable as it is likely to be highly correlated with total cost. The goodness of fit of the different cost models was compared using log-likelihoods, and plots of deviance residuals [31].

Table 2. Summary of centre and country-level variables

	National Public Health Expenditure (% share of GDP)	Method of reimbursement for acute hospital	Patient co-payment for acute care?	Unit costs acute care ($/PPP)	Fisher Price index (relative to French centre)
Spain	5.4	Global budget	No	239	0.88
Portugal	5.2	DRG	No	263	1.00
Latvia	3.9	DRG	Yes	23	0.07
Italy	5.7	DRG	No	333	1.00
Finland 1	5.3	Global budget	No	157	0.75
France	7.3	DRG	Yes	172	1.00
Lithuania 1	5.1	DRG	No	48	0.16
Poland	4.7	Global budget	No	112	0.38
Lithuania 2	5.1	DRG	No	50	0.15
Denmark	6.8	Global budget	No	234	1.00
UK	5.6	Global budget	No	194	0.87
Finland 2	5.3	Global budget	No	227	1.19
Finland 3	5.3	Global budget	No	112	1.18

Sources: OECD Health Data 2001 [26], European Observatory on Health Care systems [27–29], Grieve *et al.* [18].
GDP, Gross Domestic Product: DRG, Diagnostic Related Group.

Results

(i) *LOS models with patient-level variables*. The results from the OLS analysis suggested that most of the variables considered were associated with LOS (Table 3, model 1). Compared to the OLS model, the normal random intercepts model (model 2) gave somewhat different coefficients and standard errors, and the estimated significance of the covariates changed accordingly. For example, both age and independence pre-stroke had smaller coefficients and were no longer significant, whereas paralysis had a larger estimated effect. The random intercepts model (model 2) fitted the data better as shown by the lower level of unexplained variation at a patient level (σ^2) and the higher log-likelihood statistic. In addition, model 2 provided an estimate of the level of unexplained variation which existed between the centres. The results showed that the majority of the unexplained variation was among patients, rather than among centres, and the intra-class correlation coefficient (ρ) was 0.16. MLMs with random slope parameters were also fitted to the data, but as the results did not alter appreciably from the random intercepts model they are not presented here.

Diagnostic testing for the OLS model suggested that severe multi-colinearity was unlikely to exist; for example, the pairwise correlation coefficients between the explanatory variables did not exceed 0.6. Heteroscedasticity was detected for the OLS model ($p<0.001$ by the Cook-Weisberg test [32]). It can be caused by using an incorrect functional form, and here the residuals were found to be non-normally distributed. The analysis was therefore repeated using a gamma model with intercept and slope parameter varying by centre (Table 3, model 3). The estimated effect sizes and their associated standard errors differed somewhat from both the two previous models. For example, the coefficient for age was now larger and significant at the 5% level. The log-likelihood statistic indicated that the gamma model fitted the data substantially better than either of the other two models.

(ii) *LOS models with patient and centre-level variables*. The models were re-fitted including centre-level variables (Table 4). The results showed that the estimated direction of effect for these variables was that anticipated by theory: proxies for the presence of incentives to discharge patients earlier – a DRG system, patient co-payments, or higher relative prices (Fisher price index) – were all associated with shorter LOS, whereas a higher proportion of GDP spent on health was associated with longer LOS. For the OLS analysis (model 4) each of these effects was statistically significant at the 5 or 10% level. However, the normal (model 5) and gamma random intercepts models (model 6) found that these centre-level variables were far

Table 3. OLS, MLM and GLMM models estimating the effect of patient-level variables on length of hospital stay (LOS): coefficient (SE)

	Model 1: OLS	Model 2: MLM Normal, with random intercept	Model 3: GLMM Gamma, with random intercept
Constant term	41.7 (5.5)**	29.5 (6.2)**	31.3 (4.8)**
Patient variables			
Age	−0.10 (0.06)*	−0.08 (0.05)	−0.13 (0.04)**
Independent pre-stroke	−6.04 (2.38)**	−3.22 (2.32)	−4.82 (2.22)**
Living alone	Reference	Reference	Reference
Living with others	−8.23 (1.55)**	−5.07 (1.55)**	−0.65 (1.13)
Living in nursing home	−10.85 (4.61)**	−8.49 (4.44)*	4.37 (4.31)
Incontinent	19.29 (1.55)**	18.81 (1.49)**	15.65 (1.63)**
Paralysed	7.34 (1.56)**	8.48 (1.50)**	2.45 (0.81)**
Ischaemic stroke	Reference	Reference	Reference
Haemorrhagic stroke	5.35 (2.13)**	5.80 (2.04)**	6.82 (1.65)**
Unknown stroke type	−2.27 (1.73)	−7.71 (2.45)**	−3.67 (1.13)**
Onset to admission <6 h	Reference	Reference	Reference
Onset to admission 6–24 h	−0.55 (1.58)	0.59 (1.52)	0.77 (0.89)
Onset to admission 1–7 days	−0.96 (1.80)	−0.70 (1.73)	1.22 (0.96)
Onset to admission >7 days	−1.33 (3.27)	−0.85 (3.17)	0.28 (1.50)
Onset to admission unknown	10.17 (2.64)**	8.05 (2.60)**	6.39 (2.85)**
Family support	−1.53 (1.39)	2.51 (1.49)*	2.46 (0.81)**
Community support	−4.06 (1.71)**	−5.76 (1.68)**	−0.20 (1.04)
Random effects			
σ^2 (within centres)	522	470	442
τ^2 (between centres)		87 (37)	85 (47)
Log-likelihood	−5896	−5845	−5267

**$p<0.05$.
*$p<0.10$.

from statistically significant. Although the coefficients for the centre-level variables were generally similar for the three models, the standard errors were much smaller for the OLS model. The OLS model does not take into account the hierarchical structure of the data, and thus severely overestimated the significance of these variables. The gamma model had the highest log-likelihood, and therefore fitted the data best.

The variance terms for both multilevel models showed that most of the unexplained variation was, again, at the level of the patient rather than the centre ($\rho = 0.18$ for the normal random intercepts model). The MLMs which included centre-level variables (models 5 and 6) had a similar extent of unexplained variation between centres, and a similar log-likelihood, compared to the models with just patient-level variables (models 2 and 3). This showed that, once the hierarchical nature of the data was recognised, these centre-level variables did not help explain the variability in LOS between centres.

Figure 1 shows the deviance residuals from models 5 and 6 in the form of normal plots. If a model is appropriate, deviance residuals should approximately follow a normal distribution and will lie along the line of identity shown in the plots. For the normal MLM (model 5), the residuals show considerable positive skewness, in keeping with the positive skewness of the raw LOS data. The residuals from the gamma GLMM (model 6) show a much better behaviour, indicating the greater appropriateness of this model for these data and confirming the improvement in fit shown by the log-likelihoods in Table 4.

(iii) *Total cost models with patient and centre-level variables.* The results for total costs (Table 5) showed that the OLS models again severely

Table 4. OLS, MLM and GLMM models estimating the effect of patient and centre-level variables on length of hospital stay (LOS): coefficient (SE)

	Model 4: OLS	Model 5: MLM Normal, with random intercept	Model 6: GLMM Gamma, with random intercept
Constant term	15.3 (7.8)*	4.7 (24.6)	12.6 (26.1)
Patient variables			
Age	−0.08 (0.05)	−0.07 (0.05)	−0.13 (0.04)**
Independent pre-stroke	−3.84 (2.37)	−3.13 (2.32)	−5.09 (2.19)**
Living alone	Reference	Reference	Reference
Living with others	−6.29 (1.59)**	−4.98 (1.56)**	−0.73 (1.11)
Living in nursing home	−7.03 (4.55)	−8.38 (4.44)*	4.25 (4.19)
Incontinent	19.28 (1.53)**	18.82 (1.49)**	15.57 (1.65)**
Paralysed	7.64 (1.54)**	8.52 (1.51)**	2.31 (0.84)**
Ischaemic stroke	Reference	Reference	Reference
Haemorrhagic stroke	6.03 (2.10)**	5.84 (2.04)**	6.84 (1.63)**
Unknown stroke type	−3.22 (2.17)	−7.82 (2.48)**	−3.74 (1.12)**
Onset to admission <6 h	Reference	Reference	Reference
Onset to admission 6–24 h	0.36 (1.56)	0.64 (1.52)	0.69 (0.88)
Onset to admission 1–7 days	−0.53 (1.78)	−0.68 (1.73)	1.16 (0.97)
Onset to admission >7 days	−2.65 (3.25)	−0.87 (3.17)	0.21 (1.53)
Onset to admission unknown	6.66 (2.64)**	7.90 (2.60)**	6.12 (2.83)**
Family support	1.72 (1.48)	2.67 (1.49)*	2.46 (0.81)**
Community support	−6.37 (1.73)**	−5.86 (1.69)**	−0.31 (1.04)
Centre variables			
% share GDP	5.71 (1.09)**	6.58 (5.13)	5.39 (5.53)
DRG system	−8.11 (1.77)**	−6.16 (7.03)	−4.96 (7.39)
Price index	−10.91 (2.58)**	−10.79 (10.26)	−9.60 (10.93)
Copayment	−3.82 (2.19)*	−4.92 (9.68)	−3.93 (10.30)
Random effects			
σ^2 (within centres)	502	470	369
τ^2 (between centres)		102 (42)	120 (97)
Log-likelihood	−5868	−5845	−5268

**$p<0.05$.
*$p<0.10$.

underestimated the standard errors of the centre-level variables, so that their significance was overstated. Model 8, which correctly recognises the hierarchical structure of the data, estimated that the only centre-level variable which was significantly associated with total cost was the % share of GDP spent on health care. Model 9 which incorporated the hierarchical structure of the data *and* used a more appropriate functional form which better fitted the data, again found that only % GDP was associated with total cost. Like the LOS models, the proportion of unexplained variability in total costs between patients was substantially higher that between centres ($\rho=0.16$).

Discussion

This study compared the use of OLS and MLMs for assessing which factors were associated with international resource use and cost variation. The results showed that the OLS analysis severely overestimated the precision of the centre-level associations, and made incorrect inferences. MLMs were more appropriate for analysing the multinational data, and led to different results. In particular, whereas the OLS analyses found that all the centre-level variables considered were associated with resource use and total cost, the MLM analysis showed that, once the hierarchical nature of the data was recognised, none of these

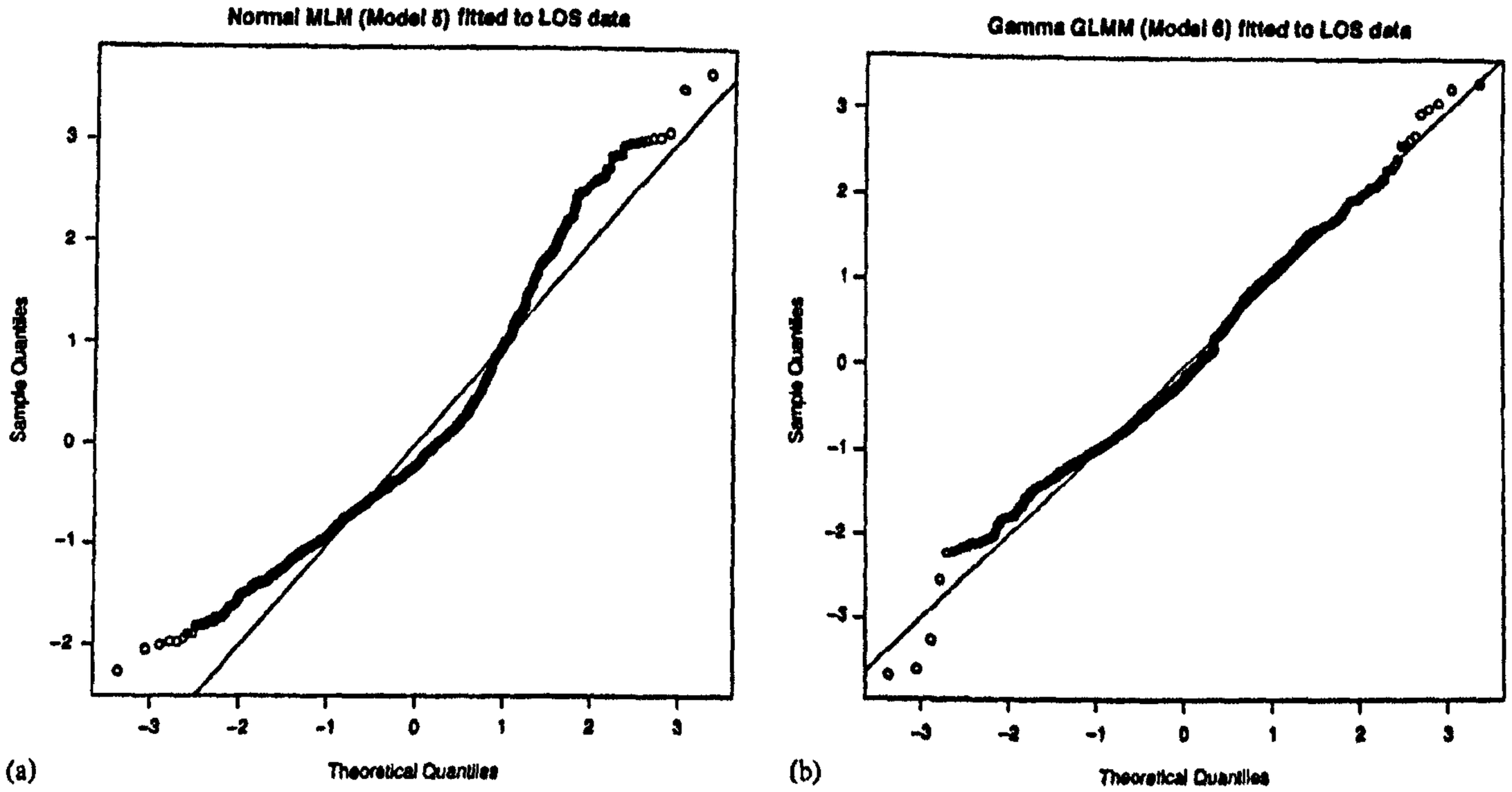


Figure 1. Quantiles of the standardised deviance residuals from length of hospital stay (LOS) plotted against the quantiles from a standard normal distribution

variables predicted resource use and only the level of health spending was associated with total cost.

The reason for this is as follows [15]. In an OLS regression of, say, cost on % GDP, all 1298 patient costs are considered. Thus the regression line is very precisely estimated (Figure 2a). In a MLM regression, the correct hierarchical structure of the data is recognised, and essentially the mean cost in each of the 13 centres is regressed on % GDP. The resulting regression line is imprecisely estimated (Figure 2b). In this example, the regression lines for the OLS and MLM analyses have a similar slope. However, the SE is much larger in the MLM since it involves a regression of 13 rather than 1298 points. OLS regression models also give incorrect results for the effects of patient-level variables because of differences between centres which are not accounted for in the analysis – a bias due to confounding. MLMs avoid this bias.

Economic evaluations are often criticised for lacking generalisability [33], particularly as costs may differ between locations [34]. Despite this, economic evaluations rarely examine why costs vary across settings and the few studies which have considered this issue have relied upon OLS analysis [1,11]. Our results indicate that those studies which used OLS analyses to identify factors associated with resource use and cost differences may have reached erroneous conclusions [11]. Other studies have used OLS to analyse variation in costs across centres [1]. Whilst this approach allows valid conclusions to be drawn about the magnitude of cost variation between centres, it does not consider *why* costs and cost-effectiveness vary across health care settings. Similarly, whilst tests for interactions can assess whether there is significant variation across centres [35], they are not suitable for assessing the reasons for cost variability. MLMs allow correct inferences to be made about which factors are associated with cost variation. This can help decision-makers assess the applicability of results to their local setting. In this study the centre-level variable which explained most cost variation was the national level of spending on health care. This variable could therefore be used, alongside patient-level variables, to predict stroke costs for centres outside the study. This illustrates how MLMs can be used to improve generalisability.

Using MLMs to assess cost variation could also assist in the design of multinational economic evaluations. Currently, multinational evaluations often only measure costs for a subsample of centres recruited to the overall study [1–6]. The choice of centres for the costing sub-study is usually based on pragmatic grounds, rather than

Table 5. OLS, MLM and GLMM models estimating the effect of patient and centre-level variables on total cost: coefficient (SE)

	Model 7: OLS	Model 8: MLM Normal, with random intercept	Model 9: GLMM Gamma, with random intercept
Constant	−4470 (1432)**	−6156 (4353)	−2034 (6529)
Patient variables			
Age	−10.0 (10.8)	−6.2 (10.8)	−7.4 (2.6)**
Independent pre-stroke	−916 (478)*	−831 (467)*	−3985 (3679)
Living alone	Reference	Reference	Reference
Living with others	−1344 (316)**	−1309 (309)**	−20 (86)
Living in nursing home	412 (917)	0 (892)	994 (602)
Incontinent	3657 (310)**	3561 (300)**	942 (176)**
Paralysed	1709 (309)**	1811 (302)**	56 (54)
Ischaemic stroke	Reference	Reference	Reference
Haemorrhagic stroke	985 (424)**	971 (411)**	182 (116)*
Unknown stroke type	−2140 (385)**	−1110 (497)**	−244 (69)**
Onset to admission <6 h	Reference	Reference	Reference
Onset to admission 6–24 h	−125 (315)	−141 (305)	−5 (62)
Onset to admission 1–7 days	−400 (358)	−500 (348)	39 (64)
Onset to admission >7 days	−740 (652)	−300 (638)	−54 (82)
Onset to admission unknown	1836 (531)**	2063 (523)**	419 (280)*
Centre variables			
% share GDP	2042 (155)**	2200 (755)**	2218 (946)**
DRG system	−1688 (354)**	−1475 (1259)	−1860 (1560)
Price index	NA	NA	NA
Copayment	−1722 (405)**	−1637 (1758)	−1477 (2153)
Random effects			
σ^2 (within centres)	20.5×10^6	19.1×10^6	18.3×10^6
τ^2 (between centres)		$3.6\ (1.5) \times 10^6$	$5.8\ (4.4) \times 10^6$
Log-likelihood	−12761	−12731	−11800

**$p<0.05$.
*$p<0.10$.
NA not applicable.

on any rational basis. Instead factors identified as being associated with total costs could be used to choose where best to measure costs. For example, if the level of health care spending is associated with total or incremental costs, then centres could be selected which were broadly representative of countries with high, medium or low levels of spending on health care. Decision-makers in each of these settings could then base their decisions about relative cost-effectiveness on relevant cost data. Perhaps more importantly, the use of MLMs reveals the need to have data available from enough centres to enable the effects of centre-level variables to be estimated with adequate precision. One preferred strategy is therefore to undertake costing sub-studies on a random selection of patients from all centres, rather than on all patients from only selected centres.

Apart from correctly estimating coefficients and standard errors for patient and centre-level variables, the MLM approaches provide additional information. In particular, the extent of residual variation is reported at both patient and centre level. For this dataset there was some residual variation in resource use and costs across the health care centres, which may reflect differences in technical or scale efficiency. However, the intra-class correlation coefficient (ρ), which reports the proportion of the variability at the centre-level, was relatively low (0.16). This shows that most of

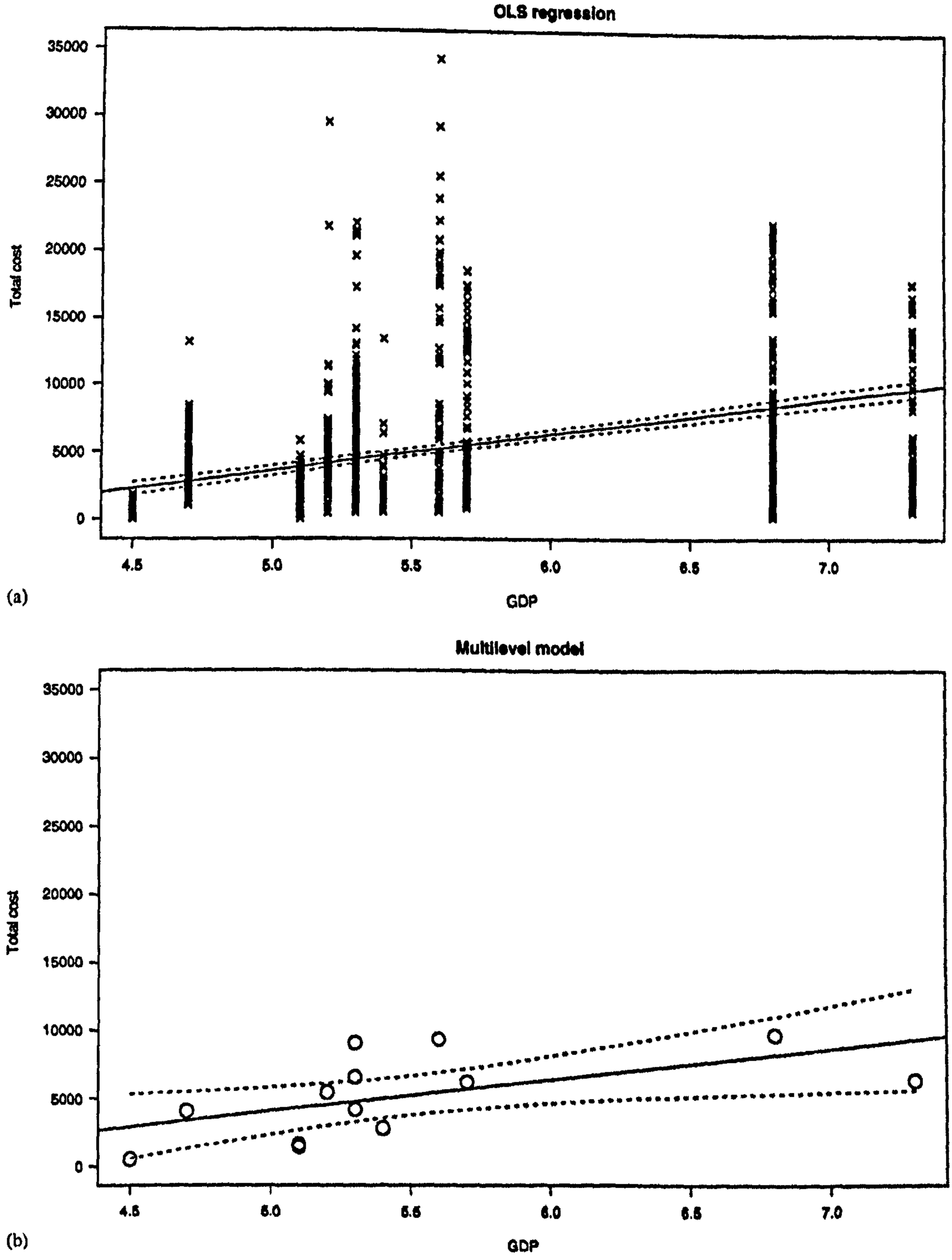


Figure 2. The relationship between cost and % share of GDP, based on (a) ordinary least-squares regression, (b) multilevel modelling; 95% confidence interval for the regression lines shown

the residual variation was at the patient level and suggests that unmeasured differences in patient characteristics (such as socioeconomic status) may be important in explaining overall variation. Whilst ρ is useful for understanding the degree to which variability exists at different levels, it should not be used as a basis for accepting or rejecting the use of the MLM approach. As is clear from the examples in this paper, OLS methods can be severely misleading even when ρ is small.

Previous MLMs in health economics have assumed that the residuals are normally distributed [14]. This assumption is unlikely to hold when the variable of interest is total cost per patient, or a measure of cost-effectiveness such as incremental net-benefit. Although generalised linear models (GLMs) have been recommended and used for analysing cost data [Barber, 2003, [17]], they have not previously been evaluated in hierarchical models for analysing cost or cost-effectiveness data. This study demonstrated their use in analysing hierarchical cost data. The GLMM random intercept model with the gamma distribution fitted the data better than the MLM random intercept model which assumed a normal distribution. This suggests that GLMMs are an attractive alternative for analysing multinational resource use and cost data.

Previous multicentre costing studies have struggled to assess cost variation partly because unit costs have not been measured in each centre, and the costing methodology used has differed by location [6]. Although the case-study described was not a full cost-effectiveness analysis, it was chosen for this evaluation as it enabled cost variation to be carefully assessed. In particular, detailed resource use, unit costs and outcomes were collected in each centre using a consistent methodology. The numbers of patients and centres were reasonably similar to the numbers which could be recruited to an international economic evaluation, which made it a realistic setting for testing the usefulness of MLM versus OLS for assessing cost variability. The technique could be applied to multicentre (national or international) economic evaluations, where the net-benefit is reported for individual patients, and the incremental net-benefit of the new treatment is estimated using regression analysis [35,36]. To fully test the use of MLMs for multicentre economic evaluations, future studies are needed which measure resource use and unit costs in more centres using consistent costing methodologies.

Acknowledgements

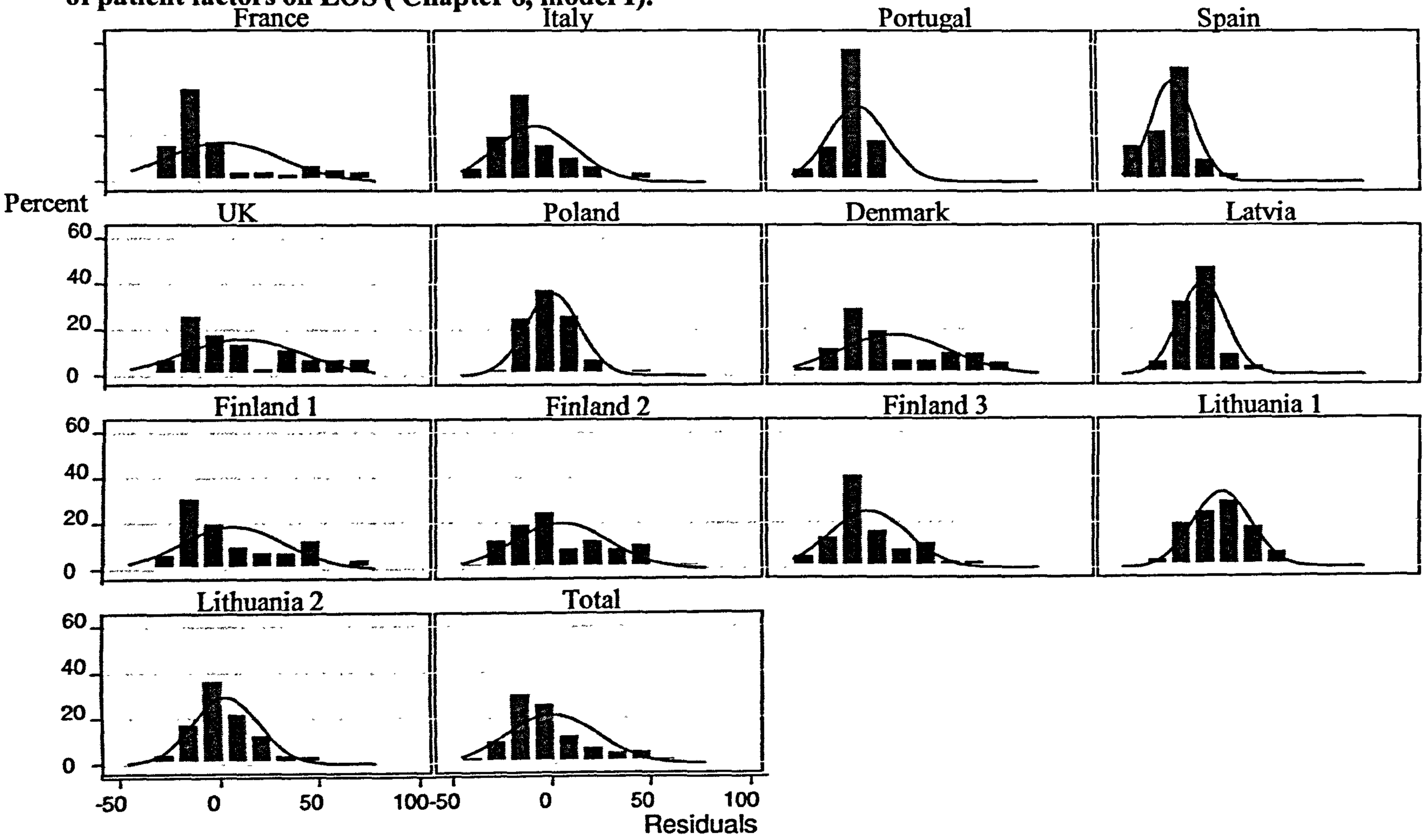
The authors are grateful to Dr Charles Wolfe and the participants of the Biomed II European Study of Stroke Care for access to the stroke data.

References

1. Willke R, Glick HA, Polsky D, Schulman K. Estimating country-specific cost-effectiveness from multinational clinical trials. *Health Econ* 1998; **7**: 481–493.
2. Rutten-van Molken PMH, van Doorslaer EKA, Till MD. Cost-effectiveness analysis of Formoterol versus Salmeterol in patients with asthma. *Pharmaco-Econ* 1998; **14**: 671–684.
3. Menzin J, Oster G, Davies L *et al.* A multinational economic evaluation of rhDNase in the treatment of cystic fibrosis. *Int J Technol Assess Health Care* 1996; **12**: 52–61.
4. Johannesson M, Jonsson B, Kjekshus J *et al.* Cost effectiveness of simvastatin treatment to lower cholesterol levels in patients with coronary heart disease. Scandinavian Simvastatin Survival Study Group. *N Engl J Med* 1997; **303**: 332–336.
5. Glick HA, Willke R, Polsky D *et al.* Economic analysis of tirilazad mesylate for aneurysmal subarachnoid hemorrhage. *Int J Technol Assess Health Care* 1998; **14**: 145–160.
6. Schulman K, Glick HA, Buxton M *et al.* The economic evaluation of the FIRST study: design of a prospective analysis alongside a multinational phase III clinical trial. *Controlled Clinical Trials* 1996; **17**: 304–315.
7. Raikou M, Briggs A, Gray A, McGuire A. Centre-specific or average unit costs in multi-centre studies? Some theory and simulation. *Health Econ* 2000; **9**: 191–198.
8. Wennberg JE. Dealing with Medical Practice variations: a proposal for action. *Health Affairs* 1984; **3**: 6–32.
9. McPherson K, Wennberg J, Hovind OB, Clifford P. Small area variations in the use of common surgical procedures: An international comparison of New England, England and Norway. *New Engl J Med* 1982; **307**: 1310–1313.
10. Phelps CE, Mooney C. Variations in medical practice use, causes and consequences, In *Competitive Approaches to Health Care Reform*, Arnoud RJ, Rich RF, White WD (eds). The Urban Institute Press: Washington, DC, 1993; 139–178.
11. Coyle D, Drummond MF. Analyzing differences in the costs of treatment across centres within economic evaluations. *Int J Technol Assess Health Care* 1998; **17**: 155–163.

12. Rice N, Jones A. Multilevel models and health economics. *Health Econ* 1997; **6**: 561–575.
13. Leyland AH, Goldstein H (eds). *Multi-level Modelling of Health Statistics*. Wiley: West Sussex, England, 2001.
14. Carey K. A multi-level modelling approach to analysis of patient costs under managed care. *Health Econ* 2000; **9**: 435–446.
15. Goldstein H. *Multilevel Statistical Models*. London: Edward Arnold, 1995.
16. Briggs AH, Gray AM. Handling uncertainty when performing economic evaluation of health care interventions. *Health Technol Assess* 1999; **3**(2).
17. Manning WG, Mullahy J. Estimating log models: to transform or not to transform? *J Health Econ* 2001; **20**: 461–494.
18. Grieve R, Dundas R, Beech R, Wolfe CDA. The development and use of a method to compare the costs of acute stroke across Europe. *Age Ageing* 2001; **30**: 67–72.
19. Grieve R, Hutton J, Bhalla A *et al.* A comparison of the costs and survival of hospital-admitted stroke patients across Europe. *Stroke* 2001; **32**: 1684–1691.
20. Wolfe CDA, Tilling K, Beech R, Rudd AG. Variation in case fatality and dependency from stroke in Western and Central Europe. *Stroke* 1999; **30**: 350–356.
21. World Bank. *World Bank Indicators 2000* (4th edn.), Washington: World Bank, 2000.
22. OECD. *Main Economic Indicators*. Paris: OECD, 1999.
23. Kavanos P, Mossialos E. International comparisons of health care expenditures: what we know and what we do not know. *J Health Serv Res Policy* 1999; **4**: 122–126.
24. Alexandrov AV, Smurawska LT, Bartle W, Oh P. Cost considerations in the pharmacological prevention and treatment of stroke. *PharmacoEconomics* 1997; **11**: 408–418.
25. Diewert WE. Axiomatic and economic approaches to international comparisons. In *International and Interarea Comparisons of Income, Output and Prices*, Heston A, Lipsey RE (eds). The University of Chicago Press: Chicago, London, 1999; 13–108.
26. OECD. *OECD Health Data 2001, A Comparative Analysis of 30 Countries*. OECD: Paris, 2001.
27. Cerniauskas G, Murauskiene L. *Health Care Systems in Transition: Lithuania*. European Observatory on Health Care Systems: Copenhagen, 2002.
28. Karaskevica J, Tragakes E. *Health Care Systems in Transition: Latvia*. European Observatory on Health Care Systems: Copenhagen, 2001.
29. Karski JB, Koronkiewicz A, Healey J. *Health Care Systems in Transition: Poland*. European Observatory on Health Care Systems: Copenhagen, 2002.
30. Rasbach J, Browne W, Goldstein H *et al. A User's Guide to MLwiN: Version 2.1c*. Institute of Education, University of London: London, 2002.
31. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. London; Chapman & Hall, 1996.
32. Cook RD, Weisberg S. *Residuals and Influence in Regression*. Chapman & Hall: New York, 1982.
33. Spath H-M, Carrere M-O, Fervers B, Philip T. 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* 1999; **49**: 161–177.
34. O'Brien BJ. A tale of two (or more) cities: geographic transferability of pharmacoeconomic data. *Am J Managed Care* 1997; **3**: S33–S39.
35. Cook JR, Drummond M, Glick H *et al.* Assessing the appropriateness of combining economic data from multinational clinical trials. *Stat Med* 2003; **12**: 1955–1976.
36. Hoch JS, Briggs, AH, Willan AR. Something old, something new, something borrowed, something BLUE: a framework for the marriage of econometrics and cost-effectiveness analysis. *Health Econ* 2002; **11**: 415–430.

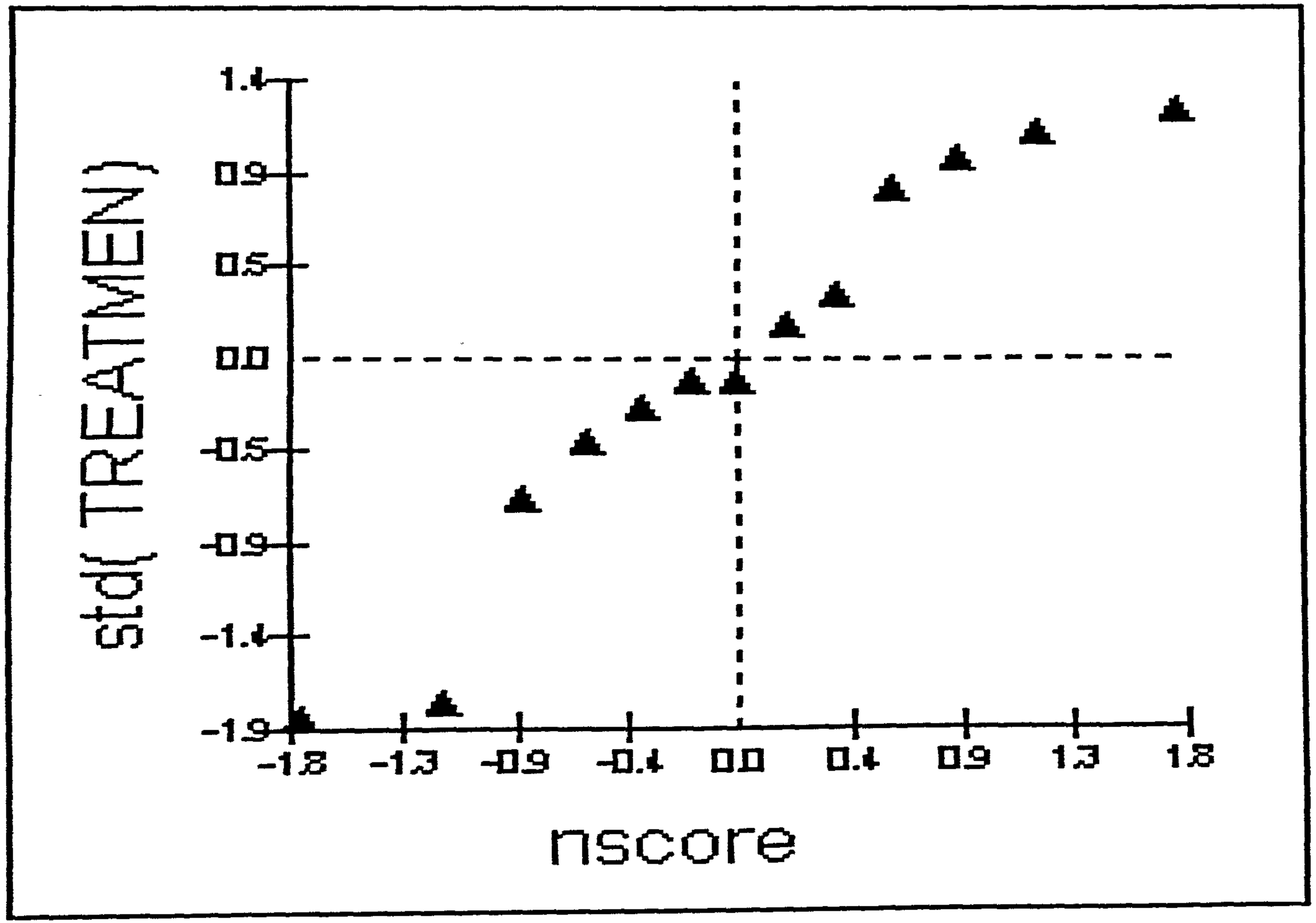
Appendix 4: Histogram of residuals by centre from OLS regression model estimating the effects of patient factors on LOS (Chapter 8, model 1).



369

Appendix 5: Plot of centre-level residuals against normal scores from MLM estimating the effect of treatment on net benefits (Chapter 9, model 4).

A straight line of 45 degrees indicates that the residuals are approximately normally distributed.



Appendix 6: Histograms of patient-level residuals from centre specific OLS regression models estimating the effect of treatment on net benefits