Empirical Essays on the Cost Efficiency and Economic Regulation of Hospitals in the National Health Service in England

John Anthony Stephen Buckell

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Institute for Transport Studies and Academic Unit of Health Economics

August 2015

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 5 is based on Buckell, J., Smith, A., Longo, R. & Holland, D. 2015. Efficiency, heterogeneity and cost function analysis: empirical evidence from pathology services in the National Health Service in England. *Applied Economics,* 47 (31): 3311-3331. My contribution was to design the analysis with the oversight of Drs Smith and Longo. Between us, the policy commentary and efficiency context was developed. Dr Smith and I jointly developed the modelling approach, statistical testing procedure, the prediction exercise, and the MFP measure. I conducted all modelling exercises and the literature review. I wrote the majority of the text. I am the corresponding author.

Chapter 6 is based on Smith, A. S. J., Buckell, J., Wheat, P. & Longo, R. 2015. Hierarchical performance and unobservable heterogeneity in health: A dual-level efficiency approach applied to NHS pathology in England. *In:* Greene, W., Sickles, R. C., Khalaf, L., Veall, M. & Voia, M. (eds.) *Productivity and Efficiency Analysis.* New York: Springer. (accepted for publication 16/04/2015). My contribution was in the design of the analysis, and to provide the policy setting. Drs Smith, Wheat and I jointly developed the modelling and statistical testing strategy. Dr Smith and I developed the modelling implications discussion. I conducted all modelling exercises and the literature review. I wrote the majority of the text. I am the corresponding author.

Acknowledgements

This research has been carried out by a team which has included Dr. Andrew S. J. Smith, Dr Roberta Longo and Professor Claire Hulme. My own contributions, fully and explicitly indicated in the thesis, have been made clear on the previous pages with respect to the joint papers that have been modified appropriately for inclusion in this thesis.

This work was funded by an *'innovations in quantitative methods'* scholarship, provided by the University of Leeds.

I would like to express my enormous gratitude to those who have supported this work and without whom I would not have been able to produce this thesis. Firstly, to the late Professor Rick Jones, who was instrumental in providing data for this work and was a huge supporter of it. May he rest in peace. To colleagues and co-authors in AUHE and ITS who have been wholly supportive and a pleasure to work with. A special mention goes to Dr Phill Wheat, who co-authored one of the studies in this thesis.

To my supervisors, and friends. To Professor Claire Hulme, who was hugely supportive from day one, but who also stepped in without hesitation and went out of her way to make time for both me and my work. To Dr Roberta Longo, who gave me the confidence to believe in myself, and gave me the guidance I needed, especially at the beginning of my studies. Finally, to my lead supervisor and mentor, Dr Andrew Smith, whose contribution to my development would require a thesis of its own. Thank you for the time, effort and patience you have invested in me, I look forward to working with you in the future.

On a personal level, there are a number of people to thank. First, to have such great friends, who were always interested and encouraging. To proof readers, Rebecca Fry and Alec Peace, in particular to Alec who will whine *ad infinitum* if not specifically mentioned.

Lastly, I thank my family. To those who couldn't be here, grannies, grandpas and Uncle Tony. To my Stepdad, Dave and the Perkins 'gang'. To my wonderful aunts and uncles, Nick, Lindsay, Kate and Lizzie, who really are an inspiration. To my older brother and best friend, Chris, thanks for being there always. To my late father, Steve, who I miss every day. Last, and most importantly, to my mother, Judy, a wonderful woman and a true inspiration. This is for you, Mum.

Abstract

Rising global healthcare expenditures, the fallout from the global financial crisis and a commitment to improving patient outcomes have increased pressure on the budget of the National Health Service (NHS) in England to unprecedented levels. Therefore, ensuring services are delivered efficiently is key both politically and economically.

In the context of the NHS, the large share of spending in secondary care means that this area is well analysed in the literature. However, the scale of the savings needed requires that both (a) more research is needed to identify further possible gains; and (b) the potential for improvement that has been identified by these studies is captured. To these ends, there are two specific aims of this thesis. The first is to examine the regulation of NHS hospital efficiency. Drawing from health care and other sectors of the economy, a number of lessons for regulators to promote hospital efficiency in the NHS and beyond are proposed. The second is to look to areas of hospital activity for which empirical evidence on efficiency is limited to identify further available gains.

Many studies in the UK and beyond have sought to measure efficiency in health: the so-called "supply" of efficiency analysis is booming. However, despite their potential, the use of these studies has been limited in the NHS. In response to this, this thesis seeks to answer some of the methodological and practical issues raised around efficiency measurement and its application to the setting of NHS hospital efficiency targets. How these findings are useful more widely to health care systems around the world is also discussed.

Chapter 1 introduces the thesis, policy background and objectives of the work. Chapter 2 gives an economic history of the NHS and observe trends in global health care costs. Chapter 3 details the economic tools that are used to gather our empirical evidence. Chapters 4, 5 and 6 report the thesis's empirical work. Finally, chapter 7 concludes by drawing together findings and assessing the extent to which the aims set out have been achieved and comments on appropriate further research.

Table of Contents

List of Figures

List of Tables

List of Boxes

List of Appendices

Glossary

AC ..... Average Cost
BC92 ..... Battesse and Coelli (1992) Stochastic Frontier Model
C&C ..... Choice and Competition
CAA ..... Civil Aviation Authority
CCG ..... Clinical Commissioning Group
CD ..... Cobb-Douglas
CDF ..... Cumulative Density Function
CMA ..... Competition and Markets Authority
COLS ..... Corrected Ordinary Least Squares
CQC ..... Care Quality Commission
DEA ..... Data Envelopment Analysis
DfT ..... Department for Transport
DLSF ..... Dual-Level Stochastic Frontier
DRG ..... Diagnosis Related Group
GDP ..... Gross Domestic Product
GLS ..... Generalised Least Squares
H&T ..... Hierarchy and Targets
HES ..... Hospital Episode Statistics
HRG ..... Healthcare Resource Group
HSCIC ..... Health and Social Care Information Centrre
LFS ..... Labour Force Survey
LOS ..... Length of Stay
LR ..... Likelihood Ratio
LSDV ..... Least Squares Dummy Variable
MC ..... Marginal Cost
MFP ..... Multi-Factor Productivity
ML ..... Maximum Likelihood
NAO ..... National Audit Office
NHS ..... National Health Service
NTPS ..... National Tariff Payment System
OECD ..... Organisation for Economic Co-Operation and Development
OFCOM ..... The Office of Communications
OFGEM ..... The Office of Gas and Electricity Markets

OFWAT ..... The Water Services Regulation Authority
OLS ..... Ordinary Least Squares
ONS ..... Office for National Statistics
ORR ..... Office of Rail Regulation
P&L ..... Pitt and Lee (1981) Stochastic Frontier Model
PAF ..... Performance Assessment Framework
PbR ..... Payment by Results
PCR ..... Price-Cap Regulation
PDF ..... Probability Density Function
PROM ..... Patient Reported Outcome Measure
QIPP ..... Quality, Innovation, Productivity and Prevention
RCI ..... Reference Cost Index
REM ..... Random Effects Model
RTT ..... Referral To Treatment
RTS ..... Returns to Scale
SFA ..... Stochastic Frontier Analysis
SHA ..... Strategic Health Authority
SUR ..... Seemingly Unrelated Regression
T&A ..... Trust and Altruism
TFP ..... Total Factor Productivity
TPR ..... Transparent Public Ranking
TRE ..... True Random Effects
UK ..... United Kingdom
WHO ..... World Health Organisation

1. Introduction

1.1 Thesis Motivation

"Thesis, n.

Forms: Pl. theses /ˈθiːsiːz/ .
Etymology: < Greek θέσις putting, placing; a proposition, affirmation, etc., < root θε- of τιθέναι to put, place.

4. A proposition laid down or stated, esp. as a theme to be discussed and proved, or to be maintained against attack (in Logic sometimes as distinct from HYPOTHESIS n. 2, in Rhetoric from ANTITHESIS n. 3a); a statement, assertion, tenet."

Oxford English Dictionary Online (2015)

The subject of this thesis is the efficiency of hospitals in the National Health Service (NHS) in England.

The thesis – or proposition – posited here is that improvements can be made in both the regulation and the measurement of NHS hospital efficiency.

This thesis describes the economic history and context into which the empirical work is set. It makes use of previous studies in health care and other sectors to provide direction for the regulation of efficiency in the NHS hospital market. It sets out the methodologies used to measure inefficiency; and some health-based issues in conducting efficiency analysis. It provides, by way of solutions, some advances in efficiency measurement. It provides empirical insights into where inefficiency resides within NHS hospitals - specifically, pathology laboratories within hospitals.

1.2 Economic and Policy Landscape

A long run issue for many governments is the amount of money that is spent on health care services. In England, the NHS is a publicly funded, largely publicly operated and publicly

regulated health care system. As such, the government's focus on expenditure is particularly sharp. The British government spent around £140bn on the NHS in 2013/14 (HM Treasury, 2015). This is, in terms of the proportion of GDP, around 9%, which roughly accords with comparable systems around the world (OECD, 2014). This proportion has been steadily growing over time across all observed health care systems (Chernew and Newhouse, 2012).

Expenditure has been central in a litany of policies over the lifetime of the NHS, where we observe substantial increases over time. There appear to be a set of economic circumstances which explain this rise. Moreover, these reasons suggest rising expenditures are set to continue into the future. It is therefore of high importance to policy makers.

In the short run, in the fallout from the economic crisis, there are substantial pressures on the NHS budget. In response to this, the Nicholson Challenge set out targets for efficiency savings of £20bn by 2015 in the UK National Health Service (NHS) (Health Select Committee, 2010). However, financial pressure is expected to extend beyond 2015: the NHS will face a funding gap of £30bn by 2020 (NHS England, 2013). The NHS's *five year forward view* proposes that around £22bn of this is to be derived from efficiency savings (NHS England, 2014a). Thus, ensuring efficiency in all areas of health care is key.

Although there is a wide literature assessing efficiency in the NHS, new research is required since further gains are needed to meet the spending challenge. It has been argued that 'easy' efficiency savings have now been made across the NHS (National Audit Office, 2012). Further, surveys of NHS finance directors reveal growing scepticism about whether the Nicholson Challenge will be met at all (Appleby et al., 2013). Indeed, some hospital trusts are currently struggling to break even (Murray et al., 2014).

A recent report detailed possible savings of £5bn p.a. based on staffing and pharmacy in hospitals (Department of Health, 2015); this is short of the £22bn required by some margin. Therefore, identifying additional potential efficiency improvements and encouraging them to be captured may ease budgetary pressure.

The focus of this thesis is therefore efficiency in the NHS. To this end, a natural starting point is the question of whether the NHS is efficient. Although this question seems straightforward, beguilingly so, it has proved an area of contention for economists in the NHS setting and in

health care systems more widely, with little uptake of efficiency studies amongst policy makers (Hollingsworth and Street, 2006).

There is a body of literature of both academic and non-academic studies (e.g. think tanks such as the King's Fund, see Appleby et al., 2013) that has sought to measure inefficiency in the NHS. These may be at the macro or micro level. Typically, efficiency in the academic literature is measured by stochastic frontiers (SFs), data envelopment analysis (DEA) or multivariate, multilevel modelling (MVML), and using indicator analysis (such as spending per head, Davis et al., 2014) in the non-academic literature.

At the macro level, the NHS itself is the unit of analysis, and is thus compared to other national health care services across the world. In Spinks and Hollingsworth (2009), the UK compared unfavourably (in terms of efficiency) amongst its OECD peers. However, the authors note that theoretical issues limit the interpretation of DEA results. Elsewhere, Smith and Street (2006) argue against the use of SFs at the macro level on theoretical grounds. Greene (2010) takes the view that using microeconomic tools at the macroeconomic level may be inappropriate. Practically, the usefulness of macro efficiency studies is somewhat restricted in the context of the current financial challenge because these studies do not indicate where specific savings can be made within the NHS.

At the micro level, hospital studies dominate the national and international literature (Hollingsworth et al., 1999; Jacobs et al., 2006; Hollingsworth, 2003; 2008). Within NHS services, expenditure on hospitals is dominant: in 2013/14, the NHS in England spent £58.3bn of public money on 244 providers of hospital services, representing around 55% of total NHS expenditure, and this proportion is growing over time (Department of Health, 2014). At the same time the wealth of data available means that this is an area already well analysed in the more recent NHS-based literature (Farrar et al., 2009; Laudicella et al., 2010; Cooper et al., 2012; Gutacker et al., 2013a; Siciliani et al., 2013; Daidone and Street, 2013). There is work in other areas of service delivery, primary care services for example (Szczepura, 1993; Giuffrida and Gravelle, 2001), however, because the outputs of these services are difficult to define and to measure, eliciting meaningful efficiency estimates is challenging (Rosenman and Friesner, 2004; Lester and Roland, 2009; Amado and Santos, 2009; Murrillo-Zamorano and Petraglia, 2011; Longo et al., 2012). Perhaps it is unsurprising, then, that Hollingsworth (2008) finds no recent NHS primary care efficiency studies. The

story is similar for other micro level services such as intermediate care. Given these issues, for the purposes of this thesis, our focus is on efficiency amongst NHS hospitals.

Recently, following the introduction of the Health and Social Care Act (2012), the task of managing hospital efficiency has passed from the Department of Health to Monitor, the economic regulator of NHS hospitals that have achieved Foundation Status[1]. Since Monitor has assumed the role, it has begun to develop an approach to setting efficiency targets (known as the 'efficiency factor') based on econometric benchmarking (Deloitte, 2014b). This is in keeping with the aims of central government who have identified benchmarking as key to making efficiency savings in the public sector (HM Treasury, 2015). With this it aims to encourage hospitals to meet their efficiency targets and contribute to the top-level policy goal of plugging the oncoming funding gap. Our empirical work is designed around aiding this process. We therefore set out the following objectives.

1.3 Thesis Objectives

(i) To inform the process of setting efficiency targets for NHS hospitals by Monitor, by reviewing germane literature and conducting efficiency analysis and setting out empirical issues to which we are able to provide solutions;

(ii) To provide new economic evidence for an area of NHS hospital activity for which empirical evidence is scant: pathology laboratories. This is, in turn, to feed into the top-level policy goal of making efficiency savings across the NHS; and

(iii) To advance the measurement of efficiency in health markets and beyond.

1.4 Structure of this Thesis and Empirical Chapters

The rest of this thesis is set out as follows. Chapter 2 provides an economic history of the NHS in England where specific focus is given to NHS expenditure over time. This sets the

[1] Foundation status of a NHS trust (a trust is a hospital or small group of hospitals) means that it operates under an independent, not-for-profit regime, allowing it financial autonomy which it does not have without having foundation status (Marini et al., 2008). Trusts apply for foundation status, which is granted by the regulator, Monitor, if the trust has satisfied the regulator of its financial competence. Foundation status has not been awarded to all NHS trusts.

economic context for this thesis and argues the case for the importance of efficiency – and therefore its measurement.

Chapter 3 goes on to set out our methodological approach to efficiency measurement. We provide definitions of various concepts of, and relating to, efficiency. We justify of our approach, first by arguing in favour of frontier analysis, then our reasoning for adopting an econometric approach. We give an exposition of the cost function which we adopt in our empirical work, in both theoretical and empirical terms. We then set out, conceptually, our method for measuring inefficiency: the stochastic frontier model. We continue to describe three aspects of stochastic frontier methodology which are (a) of interest and importance in the health context; and (b) are the focus of empirical work in this thesis. These are efficiency change over time, unobserved heterogeneity and multi-level organisational structures.

The main contribution of this thesis is across the three following chapters. In general, it is to contribute to the academic field of efficiency measurement and regulation in health markets, whilst providing evidence to enable regulators and policy makers to answer the economic challenge that faces NHS hospitals. We take two approaches in pursuing these ends. First, we examine the issue of regulating the performance of NHS hospitals. Second, we go on to measure efficiency in the NHS hospitals; we report our empirical work in two studies of pathology laboratories. We provide details of each chapter below.

Chapter 4: National Health Service Performance Management, Price Regulation and Efficiency Measurement

This chapter is an analysis comprising several aspects pertaining to NHS hospital efficiency. We first review a number of performance management regimes that have been applied to NHS hospitals to try to drive out performance improvements. We focus on the general issues that arise as regards what is effective, or otherwise, when setting targets. We draw out lessons for Monitor for use when applying efficiency targets. Next, we review the hospital pricing mechanism for NHS hospitals, and suggest alterations that may encourage efficiency. Lastly, we review efficiency measurement and economic regulation in health care markets and other regulated industries in Britain. We make use of this to set out methodological challenges to efficiency measurement in health markets, and go on to propose solutions.

Chapter 5: Efficiency Over Time, Economies of Scale, Multi-Factor Productivity and Mergers in NHS Pathology

In our first empirical study, we answer several policy-based questions regarding pathology services in that we, for the first time, provide insights in to the extent of inefficiency; how efficiency changes over time; how costs vary with a number of exogenous factors; economies of scale properties in pathology production; an account of overall multi-factor productivity in pathology services; and a simulation exercise to determine the cost implications of laboratories merging. These features have been of interest to policy makers for a considerable length of time, as reflected in a number of prior studies. We are able to populate this policy debate with empirical evidence.

Next, this study, being the first approach to efficiency analysis in NHS pathology, fulfils our research agenda's goal of finding new areas of hospital services for efficiency gains.

In addition, our study provides some methodological advances to the measurement of efficiency in health markets. We are the first to adopt an econometric approach to efficiency measurement in pathology services; the first to adopt a flexible model allowing for individual laboratories to have specific paths for efficiency change over time (Cuesta, 2000); and the first to use a cost function to simulate laboratory mergers.

Chapter 6: Dual-level Inefficiency and Unobserved Heterogeneity in NHS Pathology

In our second empirical study, we adopt a multi-level model that allows us to examine the organisational structure of pathology laboratories, and the location of inefficiency therein (Smith and Wheat, 2012). We use these estimates to derive an overall inefficiency measure per upper tier unit.

We find that there are components of inefficiency at both of the levels examined. We find that inefficiency resides mostly at the lower, laboratory, level; whilst a small amount is found at the upper, strategic health authority (SHA)[2], level. As with the previous application of this model, not only do we find that there is inefficiency at both levels, but that failure to account for the structure of the organisation may lead to the underestimation of overall inefficiency (Smith and Wheat, 2012).

[2] we note that these have, subsequent to policy reforms in 2012, been abolished. However, our data collected at the time that SHAs were in place and our analysis therefore uses this structure

Methodologically, there are several novel aspects. This study is the first application of the dual-level stochastic frontier (DLSF) model in health markets. In addition, ours is the first study to measure inefficiency at two vertically distinct organisational levels in health markets.

The central contribution of this study, however, is the extension of the DLSF to account for unobserved heterogeneity between providers. We extend the model by taking advantage of methodological developments in the literature to augment the DLSF with statistical controls for unobserved heterogeneity (Farsi et al., 2005a; Kumbhakar et al., 2014). We use a set of statistical tests and adopt a measure which can account for different forms of unobserved heterogeneity. We demonstrate that it is important to do so, which is a key finding of this study.

This study is thus important not only for studies in health markets but for inefficiency measurement across all sectors of the economy.

Finally, chapter 7 concludes by bringing together the three studies. The major contributions of this thesis are in the review of incentive structures for NHS hospitals; the development of econometric techniques to measure areas of hospital activity that have not been studied previously to identify potential efficiency savings; novel application of econometric methods in health markets; the development of existing methods for health markets and for use beyond; development of appropriate testing strategies for the identification of various forms of unobserved heterogeneity; and the identification of vertically separate inefficiency in health. We discuss the extent to which the studies have answered the research objectives and go on to suggest useful areas for future work.

We now move to chapter 2 to discuss the economic history of the NHS.

2. National Health Service Structure, Expenditure and Productivity

Following our introduction in the previous chapter, this chapter examines government spending and economic issues related to health care services, with a particular focus on the NHS. We examine NHS productivity and draw out its importance for the provision of health care.

We begin by discussing health care system models to define the NHS and its position amongst its peers. We use this for reference in subsequent discussion. We then the progression of NHS expenditure over its lifetime, before moving to economic reasons for changes in health care expenditure. Finally, we move to NHS productivity and discuss how NHS productivity has changed in recent years. We conclude the chapter by suggesting that natural rises in health care expenditure over time implies that productivity and efficiency are crucial for policy makers, both in the NHS context and beyond.

Subsequent chapters go on to define efficiency and set out how it may be measured (chapter 3). Next, a chapter, 4, is devoted to efficiency in the NHS context: policy context, setting targets for efficiency and health issues pertaining to measuring efficiency. Following from this are the two empirical chapters, 5 & 6, which measure efficiency in NHS hospitals. Chapter 7 provides a synopsis and overall conclusions.

The rest of this chapter is set out as follows. Section 2.1 considers health care system models to make clear the NHS's position amongst its peers. In section 2.2 moves to the progression of NHS expenditure over its lifetime. Section 2.3 goes on to discuss economic drivers of health care costs and expenditure. Section 2.4 examines NHS productivity and concludes.

2.1 Health Insurance Models

To begin, consideration is given to the basic features of the NHS by way of comparison to other possible health care system structures. There are a number of ways in which a health system can be characterised, based on the regulation, financing and provision of health care. Bohm et al. (2013) set out ten basic types of system based on these features (of course, many more are possible, but these are deemed implausible; for example, a privately funded system with public provision). These are summarised in table 2.1.

Health System Type	Regulation	Financing	Provision	Example(s)
National Health Service	State	State	State	UK, Spain, Portugal, Scandinavia, Denmark, Iceland
Non-profit National Health System	State	State	Societal	
National Health Insurance System	State	State	Private	Australia, New Zealand, Canada, Italy, Ireland
Etatist social Health System	State	Societal	Societal	
Etatist Social Health Insurance	State	Societal	Private	France, Belgium, Poland, Israel, Japan, Korea
Etatist Private Health System	State	Private	Private	
Social Health System	Societal	Societal	Societal	
Social Health Insurance System	Societal	Societal	Private	
Corporatist Private Health System	Societal	Private	Private	Austria, Germany, Luxembourg, Switzerland
Private Health System	Private	Private	Private	USA

Table 2.1: Types of Health System

From a consumer's perspective, from Rothenburg (1951) and Nagendran (2010), there are four basic types of medical insurance. These are private medicine – fee for service (i.e. no insurance); private medicine – voluntary sickness insurance; private medicine – compulsory sickness insurance; and socialised medicine. There is some overlap between health care system and insurance schedule, but the two are different fundamentally.

The simplest way in which to conceptualise the various systems is to consider a continuum. On the far left is the National Health System, adopted in the UK and elsewhere (table 2.1), characterised by state regulation, state financing and state provision. The NHS model (also referred to as the Beveridge model, after Sir William Beveridge) is placed on the far left (or close thereto). On the far right is private medicine – fee for service characterised by complete private sector provision for all services and no state intervention. The remaining insurance models lie along the continuum, with the compulsory insurance model to the left of the voluntary insurance model. The Etatist Social Health Insurance model (sometimes referred to as the Bismarck model, after Otto Van Bismarck's late 19th century welfarist reforms) rests close to the compulsory insurance model, but has elements of socialised medicine. Its defining features include the stringent regulation of insurance (often but not necessarily sold on a not for profit basis), claims paid without being challenged, no exclusion for pre-existing conditions, prices fixed by the state and private primary and secondary care outlets. The system in the USA was close to a compulsory insurance model (the 1934 American Blue Cross and Blue Shield models), although the recent health care bill[3] looks to have shifted the system towards the Beveridge and Bismarck models. Indeed, more is spent on public health in the USA than in the UK: Medicare and Medicaid programmes (in 2013, Medicare cost $586bn, Medicare cost $449bn[4] - both of which are in excess of the approximate £140bn spent on the NHS). The Canadian National Health Insurance (NHI) model is close to the compulsory health insurance model. In less developed nations, e.g. Senegal, health care is almost always uninsured private medicine. These are demonstrated below.

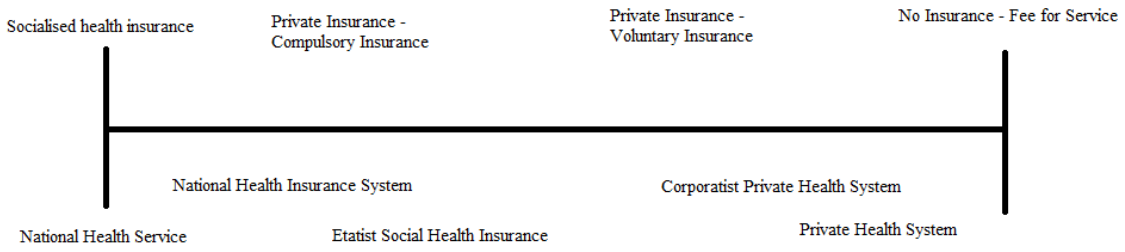

Figure 2.1: Health Systems and Insurance Models

The position of the various models along the continuum across time is subject to change. Remembering that all health systems were at one stage on the far right, there looks to be a

[3] See http://www.whitehouse.gov/health-care-meeting/reform-means-you for details
[4] See http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet.html

movement to the left, implying systems move to the left as development occurs. However, this is a discussion for another set of ends. The point here is that this continuum gives a way in which to observe whether reforms over the course of the NHS have changed it fundamentally, from its far left origin.

2.2 National Health Service Expenditure over Time

We begin this chapter by considering NHS expenditure in real terms and as a proportion of total government expenditure. These are presented in figure 2.5 below.

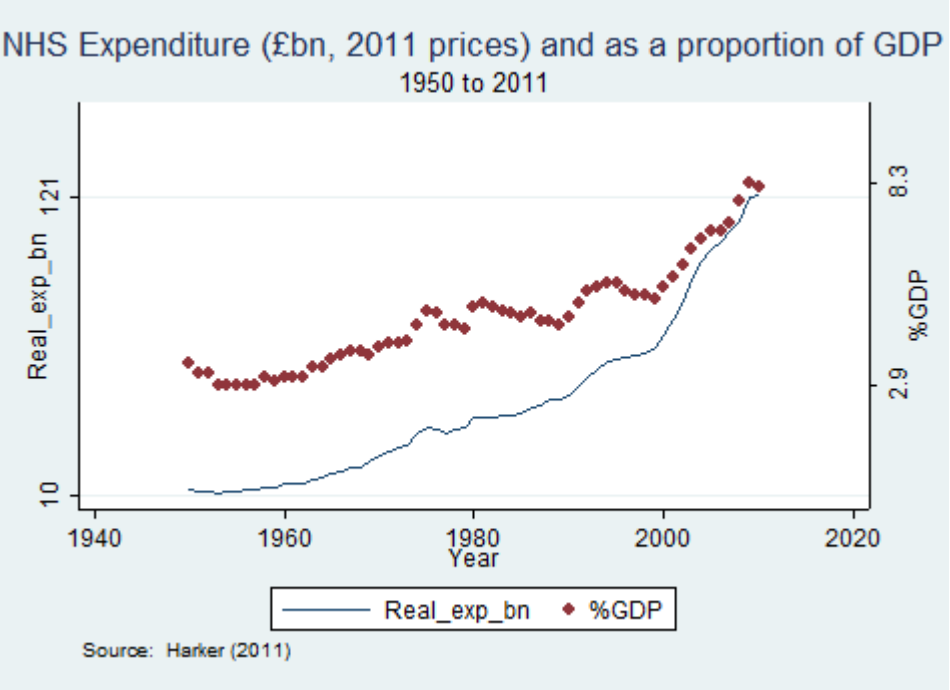

Figure 2.2: NHS Expenditure in real terms and as a proportion of GDP, 1950-2011. Source: Harker (2012)

Figure 2.2 shows the growth in NHS expenditure since its beginning. In real terms, spending has increased from its base of around £12bn in 1950 to its peak of around £121bn in 2011. In terms of GDP, the % share of NHS spending has increased from 3.5% in 1950 to 8.3% in 2011. Both series share a common trend which is that there is fairly uniform growth from 1950 until around 2000, at which point the trend increases sharply, with the increased trajectory holding until 2011. This is a substantial increase in spending, particularly in more recent years (coinciding with the Blair government, who mandated spending increases).

Given this rise in expenditure over time, a natural question to ask is whether other sectors in the economy have experienced similar trends in expenditure in terms of total GDP. Presented below are such trends.

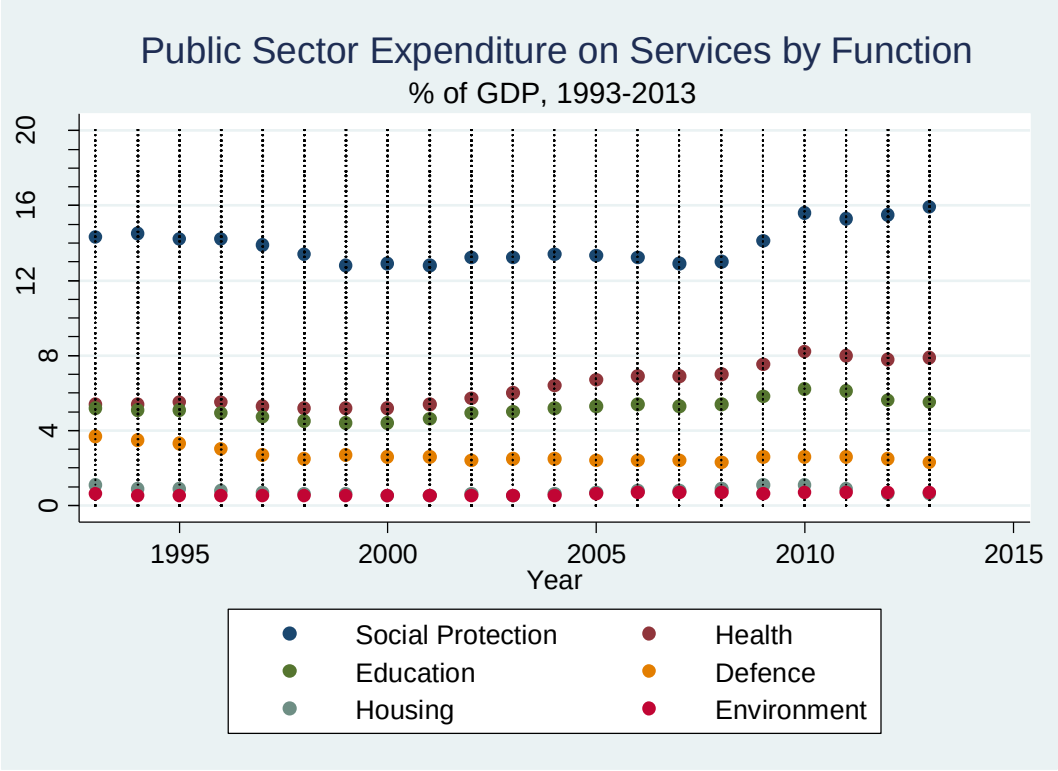

Figure 2.3: Government Expenditure by Sector as a % of GDP, 1993-2013. Source: HM Treasury (2014). NB – only a selection of sectors have been taken for ease of interpretation

Figure 2.3 shows the proportion of total GDP that is devoted to a number of sectors between 1993 and 2013. Five have been chosen for comparison, rather than all areas of spending, to make for ease of interpretation. Health represents government expenditure on publicly provided health, the provision of which may be either public or private (see table 2.1). Other sectors include social protection which is the governments largest outlay (health is second), education, defence, housing (and community amenities) and environmental protection.

As can be seen, health, education and social protection have had expenditure as a proportion of total GDP increased over the period. Health expenditure has outstripped increases in education spending over the period. The remainder of the sectors have either had fairly constant levels of expenditure, as in the case of environmental protection, or have had

expenditure reduced, as in the case of defence. Thus, not only has health expenditure grown as a proportion of GDP, it has grown relative to other sectors of the economy.

The next question to consider is whether these rises in expenditure are being driven by certain aspects within health care services or whether it is all aspects of health care that are contributing to the rising costs. To help answer this question, the following data are available.

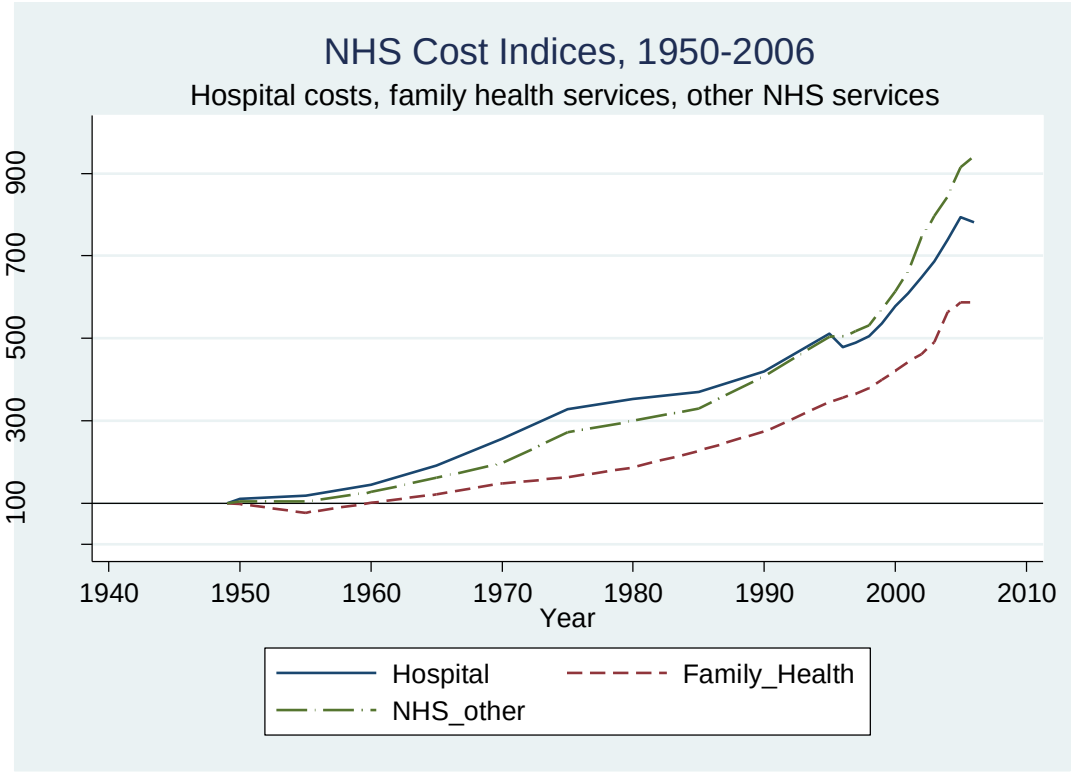

Figure 2.4: NHS Cost Indices, 1950-2006. Source: Hawe (2009)

Figure 2.4 shows cost indices for various services within the NHS over time. The *Hospital* index shows the cost index for hospitals. The *Family_Health* index refers to the costs of services that are considered to be for families, including primary care, optometry, dentistry and pharmacology. The *NHS_other* index refers to the costs of other services such as mental health, ambulances and special health authorities (e.g. National Blood Authority). These suggest that the costs of various subsets of NHS production are each increasing over time. The rates at which they increase vary: the costs of other services appear to have grown most over the period, followed closely by hospital costs. The costs of family health services seem to have lagged behind those of other services, reaching an index value of around 600, relative

to the 950 of other or the 800 of hospital costs. For Hospital and other services, the growth of the value of the indices appears to be reasonably constant from 1950 up until 2000, beyond which the trend appears to increase sharply. This corresponds to the pattern observed for overall spending, as in figure 2.2. For family health services, the trend appears to rise smoothly over the entire period.

In summary, health care expenditure has increased over time in both real terms and in terms of the proportion of GDP spend. The rise in GDP spend has occurred in many other countries – none of the reported countries' proportions of GDP spending reduced over the period. In addition, costs of various subsets of NHS activity have increased over time. In light of this, the natural question is as to why. To answer this, the following section explores reasons to explain these trends.

2.3 Health Care Expenditure Drivers

We first distinguish spending level and spending level growth. Spending level is simply the product of the quantity of outputs and their factor prices at a given equilibrium state. Spending level growth reflects a set of factors that cause this equilibrium state to shift over time; that is, those factors over and above the output alone that are causing health care costs to change over time. See Chernew and Newhouse (2012). Focus is given to spending level growth in this exposition.

Population trends include the rising population over time. Simply, more people require more health care services. This is especially true in the context of NHS where, under its universal service, there is non-excludability. Over the last 50 years, the UK has seen a population increase of around 18.7%, around 10 million in number, to 63 million (ONS, 2014). The population in the UK is set to rise by around 16% to 73 million by 2035 (ONS, 2012b).

The next population trend is the rise in long term conditions (LTCs). Rising incidence and/or prevalence across a number of conditions, for example diabetes (Diabetes UK, 2012), has increased pressure on health care services. This issue is linked to health behaviours, which were identified as a driver of service use, and correspondingly as a driver of costs, in the context of the NHS (Wanless, 2002). For example, BMI, indicative of health behaviours, was recently found to be influential on health care expenditures (Willeme and Dumont, 2015).

Further, multimorbidity is expected to rise in the coming years, which has significant cost implications (Department of Health, 2012). Multimorbid patients are resource intensive: it is estimated that the 30% of patients with one or more long term conditions account for 70% of health care spending in England (NHS England, 2013). Moreover, spending per case appears to have increased with time, largely due to increases in multimorbidities and as a result of technological advance (Chernew and Newhouse, 2012).

The next, and perhaps best documented, population trend is an ageing population. The median age in Britain is predicted to rise from just under 39 in 2010 to over 42 in 2035 (ONS, 2012c). Economists had for a long time believed that it was the age of patients themselves that was driving the association with increases in health care costs (Bos and Weizsacker, 1989). However, Zweifel et al. (1999) proposed the 'red herring hypothesis': that the driver of the observed expenditure rise was the patient's proximity to death, rather than their age. Age was therefore just a proxy for proximity to death. Later, Seshamati and Gray (2004) revisited this issue, finding that both age and proximity to death appeared to be driving increases in health care expenditure. This finding is corroborated by recent evidence, which proposes an interaction of the two factors as a further determinant of rising costs (Geue et al., 2014). Further evidence is in accordance but finds that whilst ageing did increase expenditure, the effect was meagre, by around a factor of four, relative to the effect of technological change (Dormont et al., 2006).

A corollary of average age is age structure, that is, proportion of the population across age strata. The belief is that given two populations of equal average age, the one with a higher share of elderly patients will be more costly. However, there is little empirical support for this (Baltagi and Moscone, 2010).

There is a growing recognition of frailty as a medical condition common amongst elderly populations (Clegg et al., 2013). It is inherently tied to both ageing and comorbidities, although is distinct from them (Fried et al., 2004). Frailty, both as an increased health risk itself and as an amalgam of several deleterious conditions, is likely to be financially burdensome (Clegg and Young, 2011).

Economic factors that have been suggested as driving health care expenditure growth include, firstly, Baumol's law[5] (Baumol, 1967; Baumol, 2012). The central idea is that, over time, the more labour-intensive industries will, ceteris paribus, become costly relative to the less labour-intensive industries, which are able to reduce production costs more rapidly through the adoption of technology. For example, the use of technology can be used to greater effect in reducing the unit cost of producing, say, an automobile than it can to reduce the unit cost of, say, an education, which requires a far higher share of labour input than its comparator. This is intuitively reasonable in the context of health markets, where there is a significant labour input. Indeed, labour cost shares are typically high in health markets (around 63% of total spend on hospitals, Department of Health (2015); as much as (circa) 80-90% in pathology production, see Department of Health (2008)). Much empirical testing shows that, whilst some results suggest that Baumol's cost disease pervades health markets (Hartwig, 2008; Hartwig, 2011; Bates and Santerre, 2013), other studies do not find evidence to support this idea (Gerdtham et al., 1992; Murthy and Ukpolo, 1994).

The next economic factor is technological change. Technological advance is thought to be a major driver of health care expenditure (Manning et al., 1987). Indeed, one study estimated 75% of expenditure growth was attributable to technological progress (Newhouse, 1992). This has become known as the 'Newhouse conjecture'. This issue has proven challenging for economists - finding a suitable proxy for technological change has been difficult in empirical applications (Baltagi and Moscone, 2010). If proxy measures are supported, then the Newhouse conjecture has been supported empirically. As proxies, Baker and Wheeler (1998) used the number of surgical procedures; Okunade and Murthy (2002) used R&D spending; and Gerdtham and Lothgren (2000) used the passage of time. A recent study made use of two novel indicators, namely the number of approved medical devices and pharmaceutical products (Willeme & Dumont, 2015). The study findings are consistent with prior literature, where technological progress accounted for, on average across OECD countries, around 43% of health expenditures between 1980 and 2009. Dormont et al. (2006) make use of changes in clinical practice (pharmaceutical expenditures) as a proxy for technological change. The authors find that technological change is substantially more effective on expenditure increases than population age.

[5] Elsewhere referred to as 'Baumol's cost disease'

Another issue around technology is the capability of clinicians: as technology advances, it is possible to treat conditions that had not been possible previously. This has been documented in the NHS, congenital heart disease services being a recent example (Glenwright et al., 2014).

Of course, the positive relationship may not hold true in all cases; some technological progress may help to lower costs in some settings. Much depends on the view – the Newhouse conjecture is a more macro, long-run effect. There may be different effects at different levels of aggregation or over different time horizons. For example, technological progress (as proxied by the passage of time) was thought to be explaining short run cost reduction in pathology services (Buckell et al., 2015). In some cases, medical intervention can be substituted by the use of drugs – termed 'drug cost offset' – evidence for which has been found in the literature (Willeme and Dumont, 2015).

Next is the effect of rising incomes. Economic theory predicts that demand for health care will increase with rising incomes, both at the micro and macro levels (Rice, 2003). Health care expenditure will consequently rise. This has not only been supported in many empirical applications where studies reveal that health care expenditure is highly correlated with income, income has been shown to explain a significant share of the variation in total health care spending (Morris et al., 2007). The current debate amongst economists is whether the income elasticity of demand is greater than or less than one, and thus whether health care is a 'necessary' (or 'normal') or 'luxury' good. Although many empirical studies have sought to answer this question, and a range of approaches employed, the evidence is mixed and the answer remains ambiguous (Lago-Penas et al., 2013).

There are also often political factors that have bearing on expenditure. The structure of health finance and provision may have implications for expenditure. Chernew and Newhouse (2012) do not find significant differences between insurance schemes in markets in the USA. Xu et al. (2011) found some evidence to suggest that social insurance is expensive relative to general taxation across OECD countries. Insurance models based on co-payments present an additional aspect to insurance. The central example of such a scheme was the Rand Health Insurance Experiment in the USA in the 1980s. Economic analysis has shown that the system of insurance can indeed have bearing on health expenditures (Manning et al., 1987).

However, the authors note that the effect is inferior to that of technological change. Nonetheless, this issue may be important with the growth in private insurance in the UK.

For public vs. private provision, Hollingsworth (2008), in an authoritative literature review of efficiency analyses, finds no clear consensus amongst empirical studies. If anything, the literature may point to public provision being preferable, but the answer remains unclear. Moreover, the optimum is likely to vary by region.

Reforms to health care systems are often expensive. The Health and Social Care Act (2012) was a major top-down reorganisation of the NHS. Not only is evidence around the savings dubious, the actual costs have exceeded projections and the upheaval in service provision itself has likely driven up costs (cf. NAO, 2013; Walshe, 2014). Political commitments can also drive spending: New Labour famously[6] matched NHS expenditure per capita to European levels during the 2000s (Smee, 2005; fig. 2.2). A further contributory factor can be major shocks, as Maynard and Ludbrook (1980, pp. 293) have described 1970s funding,

"…what you got last year, plus an allowance for growth, plus an allowance for scandals."

Recession, or rather the fallout of recession, can have implications for health care expenditure. For example, the 2008 global financial crisis increased pressure on public sector expenditure and so on the costs of the health care system in the UK. In response to this, the Nicholson Challenge set out targets for efficiency savings of £20bn by 2015 in NHS (Health Select Committee, 2010). There have been real terms freezes in expenditure during the last few years, and financial pressure is expected to extend beyond 2015, with a funding gap of £30bn expected by 2020 (Roberts et al., 2012; NHS, 2013).

In the long run, NHS expenditure has risen over time, both in absolute and real terms. Expenditure has grown as a proportion of GDP in Britain and across health care systems the world over. Expenditure has grown within all observed subsets of NHS activity. There are a number of reasons to explain this growth. Further, there are a number of reasons to expect that this growth will continue for the foreseeable future. As Baumol (2012) argues, this growth is not necessarily an issue; more an indication of gains made in other sectors.

[6] Tony Blair made the announcement on BBC 2's 'Breakfast with Frost'. This has since been dubbed 'the most expensive breakfast in history'.

In more recent times, there have been issues that have had implications for health expenditure, both positively and negatively. In some cases, political intervention is often associated with expenditure rises; whereas in the fallout of recessions, health budgets are more constrained.

The implication of these factors is that there is a delicate balance for budget holders in health markets. On the one hand, spending on health services must track the growth in costs. If spending reductions, or even spending held in line with inflation, services will, assuming constant productivity, be deprived. On the other hand, financial pressure on budgets often means budget holders must be cautious not to overspend. Therefore, increasing productivity, and its economic counterpart, efficiency, will allow policy makers to minimise over-spending. Efficiency and productivity will have a central role in enabling governments to maintain levels of service and levels of quality under increasingly stringent budgets. We therefore move to discuss these concepts in the context of the NHS.

2.4 NHS Efficiency and Productivity

In the previous section, the growth – and likely continuation thereof – of NHS expenditure was discussed. Governments ought not therefore, where possible, to reduce or freeze health budgets, as services – and/or service quality – are risked when (real terms) budgets are wilted by health cost inflation. In response to this, the government has two basic options, namely to increase spending to match rising costs, or to improve efficiency and/or productivity so that the same level of output (in terms of both volume and quality) can be achieved whilst costs rise. We therefore consider these concepts in turn to examine this issue, but also to take a general view as to how the NHS is performing.

The global level of NHS efficiency, that is considering the entire NHS as the unit of analysis, has been sought across a number of studies using a variety of methods. Three examples of these are presented in table 2.2 below. Here, SFA – stochastic frontier analysis and DEA – data envelopment analysis; see chapter 3 for definition. Indicator measures are just that, including, for example, mortality rates or spending per patient.

Study	Year	Method	Conclusion
Tandon et al.	2000	SFA	0.88-0.93; close to the top of rankings
Spinks and Hollingsworth	2009	DEA	0.96-0.99; close to the bottom of rankings
Commonwealth Fund	2014	Indicator Analysis	NHS most efficient amongst peers

Table 2.2: Macro NHS Efficiency Studies.

On the basis of this evidence, it would appear that the NHS is highly efficient: across a number of studies which use a variety of methods, the estimates of efficiency are close to 1. Even in the case of the DEA estimates where the ranking is low, the estimates indicate there is almost no inefficiency. However, a number of issues arise when attempting to gauge efficiency at this aggregate level.

First, the estimates and rankings are sensitive to estimation. Greene (2004), using the same data as Tandon et al. (2000), showed that both efficiency estimates and rankings are sensitive to model specification. Spinks and Hollingsworth (2009) outline theoretical issues in DEA estimates when methods are applied at the aggregate level which inhibit the validity of estimates. More broadly, Greene (2010) suggests that the application of microeconomic tools at the macroeconomic level may be inappropriate. Second, the results are somewhat conflicting. The DEA and SFA results both suggest that the NHS is highly efficient. However, the DEA results suggest the NHS is more efficient than the SFA results, yet the NHS ranks amongst the lowest according to DEA whilst it ranks amongst the highest in the SFA ranks. This conflict casts doubt over the legitimacy of these estimates. Thirdly, there are methodological issues. The indicator analysis uses the expenditure on health as a % of GDP as one of its measures of efficiency (Davis et al., 2014, pp. 23). These rather crude metrics neglect a number of aspects of expenditure, for example quality. Moreover, there are perverse incentives in adopting this approach: in order to rank more highly, expenditure on health should simply be reduced, which is undesirable.

Overall, whilst studies indicate that the NHS is highly efficient, there are technical issues for which the credibility of these estimates is questionable. We therefore look to measures of productivity for more reliable evidence as to how the NHS is performing.

Productivity[7], like efficiency, is critical for expenditure, as highlighted by the Office for Budget Responsibility's health spending projections (OBR, 2014). If productivity rises by 2.2% p.a. then spending is expected to be just over 8% of total GDP by 2063/64. If productivity rises, as historically, by around 1% p.a., then spending is expected to reach more the 20% of GDP by 2063/64.

NHS productivity has been analysed directly in a number of ways. One way is at the macro level, such that an overall NHS productivity index is derived (Bojke et al., 2015; ONS, 2015). Another approach is at the regional level, assessing the inputs and outputs of the NHS by region (Bojke et al., 2013). There are also studies aimed at a more disaggregate level, hospitals being a recent example (Castelli et al., 2015).

We consider two indices of overall NHS productivity to consider how the general level of NHS performance of the NHS has changed in recent years. Comparison of these indices is useful in identifying some issues that arise when constructing measures of productivity. Figure 2.5 below shows both the ONS's (ONS, 2015) and University of York's (Bojke et al., 2015; referred to as UoY hereafter) indices of NHS productivity.

[7] For the purposes of discussion here, we refer to Total Factor Productivity (TFP); for a definition see chapter 3. This is in keeping with the nomenclature adopted in the measures examined here.

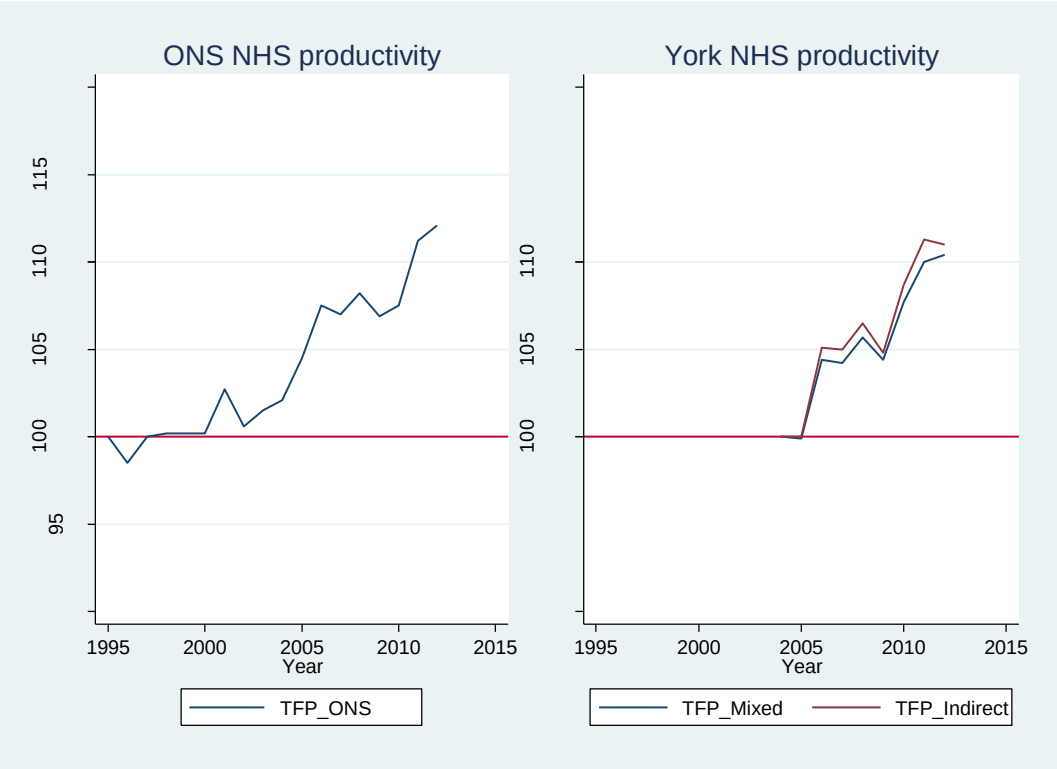

Figure 2.5: ONS Productivity Indices, ONS and University of York, 1995-2012. NB – the University of York use two indices which vary by input definition ("Mixed" uses staff numbers and expenditure; "Indirect" uses expenditure only). (TFP = Total Factor Productivity). Sources: ONS (2015), Bojke et al. (2015)

As shown above, the overall productivity of the NHS appears to have increased over time. In both cases, the level of growth is similar, from each index's origin at 100, to around 110 by 2012. However, the time series for both indices varies. The ONS index commences in 1995, whereas the UoY indices commence in 2004. This would suggest that the gradient in growth in the UoY indices is higher on average. As can be seen, however, this is not the case since the gradient of the ONS index is rather flat until around 2003. Indeed, the pattern of both indices is rather close in the years 2004-2013, in terms of both direction and magnitude. To examine this issue in greater detail, Figure 2.6 below shows the annual change of each index in percentage terms.

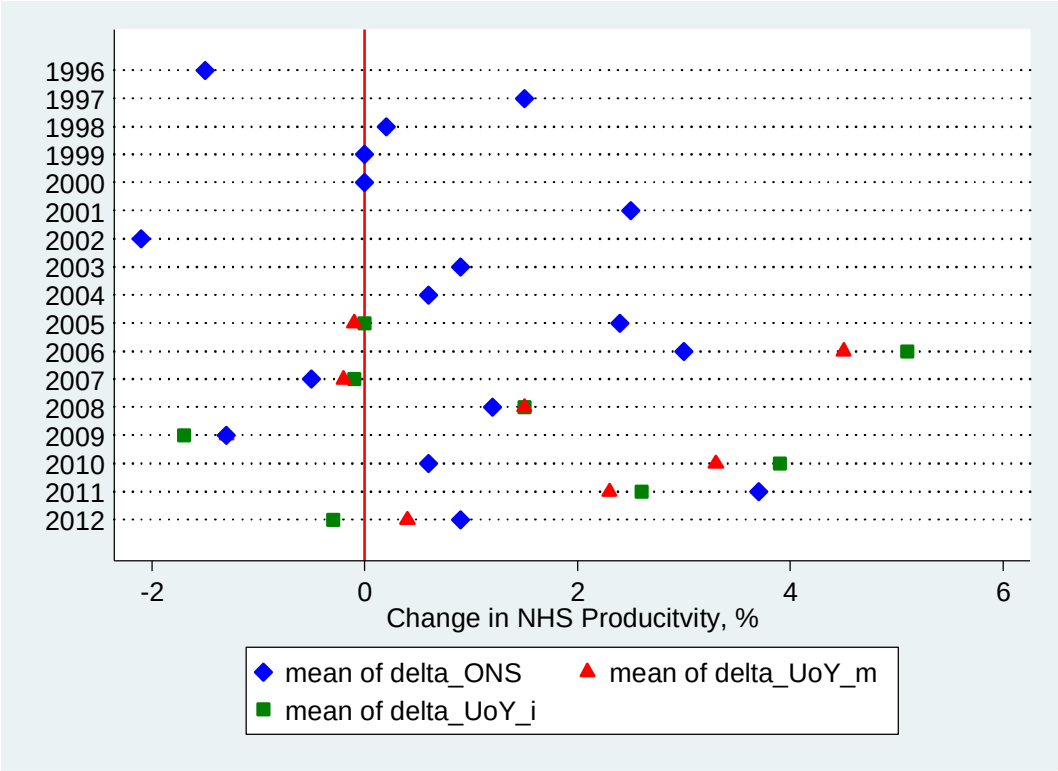

Figure 2.6: Annual NHS Productivity Change, ONS and UoY Productivity Indices, 1995-2012. Sources: ONS (2015), Bojke et al. (2015)

Figure 2.6 above shows that in some years productivity change was positive, where values are to the right of the line (at zero), and in other years where growth was negative and values are to the left of the line. That the majority of the values are to the right of the line, and their distances are generally further from it, reflects that, on average, according to these indices, NHS productivity has grown. The average change across the indices, reflecting average productivity growth in percentage terms, is 0.71, 1.30 and 1.38 for the ONS, UoY Mixed and UoY Indirect, respectively.

In many years, all three indices' values are in the same direction. In 2008 they are all positive and in 2009 they are all negative. In other years, this was not the case, as in 2012. In most cases, the estimate of productivity change is similar between indices, as in 2007. However, in other cases, the indices' values diverge, as in 2005. In these years, and more broadly, a question arises as to which is most reliable and thus how each index is constructed. To examine this issue, table 2.3 below compares the indices.

Index	Data Sources	Index Components		Years	Methods
		Inputs	Outputs		
ONS	DH; RC; WG; SG; DHSSPS; HSCIC (outpatients only); GOC; GDC; SD; Imputed; PCA	Labour; Goods and Services; Capital	Hospital and Community Health Services; Family Health Services; GP Drugs; Non-NHS	1995/96-2012/13	Chain-linked Laspeyres Indices
CHE York	DH; RC; HES; Qresearch; GPPS; PSSRU; QOF; PCAS; HSCIC	Labour; Goods and Services; Capital	Hospital Activity; Inpatient and Community Mental Health; Community Care; Primary Care; Accident & Emergency; Other	2004/05-2012/13	Chain-linked Laspeyres Indices; "mixed" (using staff numbers and expenditure for inputs) and "indirect" (using expenditure for inputs)

Table 2.3: Composition of NHS Productivity Indices[8]. Sources: ONS (2015) and Bojke et al. (2015).

Table 2.3 highlights some difference and similarities in the methods that have been applied to analyse NHS productivity. In terms of input categories and index methods, the indices are similar. However, in terms of data, output categories and years covered, the two methods diverge. Overall, whilst there is some divergence in some years, there is, considering the differences in construction, noticeable concordance between the two measures in other years in terms of both direction and magnitude (figure 2.5; figure 2.6): the two measures broadly agree.

In principle, productivity is central to the aims of NHS resource management. However, there are two theoretical issues which hinder its usefulness, both of which have been observed when looking to measure NHS productivity directly.

First, productivity, being a ratio of inputs to output, is unbounded. This means that the measure itself can neither define nor reach its own limit. The implication for using productivity measures in the NHS context is that targets can be continually reset with no regard to their potential limit, the 'goalposts are kept shifting'. In this regard, some have

[8] Glossary: DH – Department of Health, RC – Reference Costs, WG – Welsh Government, SG – Scottish Government, DHSSPS – Department of Health, Social Services and Public Safety Northern Ireland, HSCIC – Health and Social Care Information Centre, GOC – General Ophthalmic Council, GDC – General Dental Council, SD – Survey Data, PCA – Prescription Cost Analysis, GPPS – GP Patient Survey, PSSRU – Personal & Social Services Research Unit; QOF – Quality and Outcomes Framework, PCAS – Prescription Cost Analysis Service.

questioned whether the current push for productivity gains is sustainable (Appleby, 2012). One way to overcome this issue is to use productivity to benchmark, as per, for example, health regions in the NHS (Bojke et al., 2013).

The second issue is that the way in which these indices are constructed can often have a significant bearing on their magnitude. This is highlighted in the context of ONS's NHS productivity measure (figure 2.5), which shifted from suggesting a decline in NHS productivity to an increase following a revision of the index (Black, 2013). One way to overcome this issue is to measure various components of productivity in order to avoid having to aggregate various metrics. This is the starting point of our analysis.

Productivity comprises a number of aspects. Productivity can be improved through gains in efficiency, gains in the scale or scope of operation or through technological change. It is possible to measure these aspects of productivity separately in empirical settings (chapter 3 which follows defines and describes how to measure each of these features). Moreover, it is possible to combine them to arrive back at an overall measure of productivity.

For policy, efficiency is particularly useful. Here, the basic question is around how the government raises expenditure so as to cope with the natural rise in health costs, whilst concurrently not overspending (particularly during times of financial pressure). This would be a straightforward exercise if all the factors that cause health costs to rise are both known and perfectly observable, which they are not. Thus, defining a 'natural' rise in health costs is doubtless an impossible task. However, one clear way to satisfy these constraints is to maximise efficiency (or, conversely, to minimise inefficiency). Put differently, if budget holders know that services are efficient, then they also know that any cost increases are the result of natural economic factors rather than overspending.

Overall, efficiency and productivity are crucial tools for policy makers seeking to spend as much as necessary but as little as possible on the provision of health care, both in the NHS and for health care systems around the world. Therefore, in the following chapter, we proceed to define efficiency, and set our methods to measuring efficiency and productivity. We are then prepared to approach formally efficiency in health, to which we turn in the following three empirical chapters.

3. Measuring Efficiency

3.1 Introduction

In the previous chapter, we argued that, driven by natural rises in a number of factors over time, rising expenditure on healthcare was justifiable. To manage this growth, governments should seek to maximise efficiency in the delivery of health care services. To achieve this goal, efficiency must first be measured. We make use of econometric methods in our empirical applications in subsequent chapters, we therefore discuss the econometric approach to efficiency analysis here. We set out our justification for doing so by way of comparison to rival methods.

Once we have established the methods, we move to the first research chapter, 4, which reviews the literature on measuring performance in NHS hospitals; discusses the regulation of efficiency amongst NHS hospitals; and sets out the methodological landscape for measuring efficiency in health markets and other regulated sectors. From this, we set our research agenda which we go on to fulfil in two empirical chapters, 5 and 6. In all three chapters we develop the discussion of empirical efficiency analyses in health markets; we set the basis for this discussion in the remainder of this chapter.

In this chapter, we first introduce the concept of efficiency and discuss its theoretical underpinnings. We give an exposition of the economic approach to efficiency analysis based on the cost function, which is the framework that is employed in later chapters and also commonly in the literature. We then proceed to the econometric development of the economic models, making the case for our use of econometric techniques. We describe these methods – based on the stochastic frontier (SF) model - in detail, in particular paying attention to three aspects of importance in the health context: time-varying inefficiency, unobserved heterogeneity and multi-level organisational structures. We further review the measurement of efficiency in health care, paying particular attention to econometric approaches. Finally, we describe an overall measure of performance, Total Factor Productivity, which makes use of the various components of the models described during the chapter by bringing them together into a single measure of overall performance.

This chapter draws from a number of sources. We describe methodological aspects of use in this thesis, but leave a great deal for the purpose of brevity. For a fuller exposition of the theoretical aspects of the cost and production function, see Chambers (1988). For the link

between the production economics and the measurement of efficiency and productivity, see Coelli et al. (2005) and Fried et al. (2008). For the application of theory and methods in health, see Feldstein (1968), Jacobs et al. (2006), Morris et al. (2007) and Hollingsworth and Peacock (2008). Coverage of econometric techniques is provided in Gujarati (2003), Baltagi (2008) and Greene (2012c). For the estimation of econometric efficiency analysis, see Kumbhakar and Lovell (2000) and Greene (2012b). Finally, for the application of modelling techniques using contemporary software packages (e.g. LIMDEP, STATA), see Greene (2012b) and Kumbhakar et al. (2015).

3.1.1 Defining Efficiency

Koopmans (1951, pp. 460) defines efficiency as,

"An attainable set of commodity flows [or attainable point in the commodity space]…is called efficient if there is no other attainable set of commodity flows in which all flows are at least as large as the corresponding flows in the original set, while at least one is actually larger."

Debreu (1951) and Shephard (1953) provided graphical representations of efficiency through radial distances of producers from a frontier, both in an output-expanding direction (Debreu) and an input contracting direction (Shephard).

Farrell (1957) made, simultaneously, a number of significant steps. First, he defined cost efficiency as distinct from productive efficiency (and in doing so paved the way for the development of its analogue, revenue efficiency). Cost efficiency embeds input prices and asserts a behavioural assumption, cost minimisation, on the analysis of inputs and outputs in the production process. This leads to the second of Farrell's developments, which is the recognition that cost inefficiency is the product of two components, namely technical and allocative efficiencies. In this thesis, cost efficiency is measured. We therefore pay attention to its definition below.

Technical efficiency is the extent to which more resources are used in the production process than are absolutely necessary. Allocative efficiency is the extent to which suboptimal

combinations of inputs are used to produce a given level of output. In other words, there are usually many ways in which two (or more) inputs can produce a single desired output. It is unlikely that the two inputs are the same price to the producer. Then, the optimal combination is that which uses the least of the more expensive input whilst maintaining the production of the desired output. Any deviation from this optimum is the allocative inefficiency. These concepts are shown below in figure 3.1.

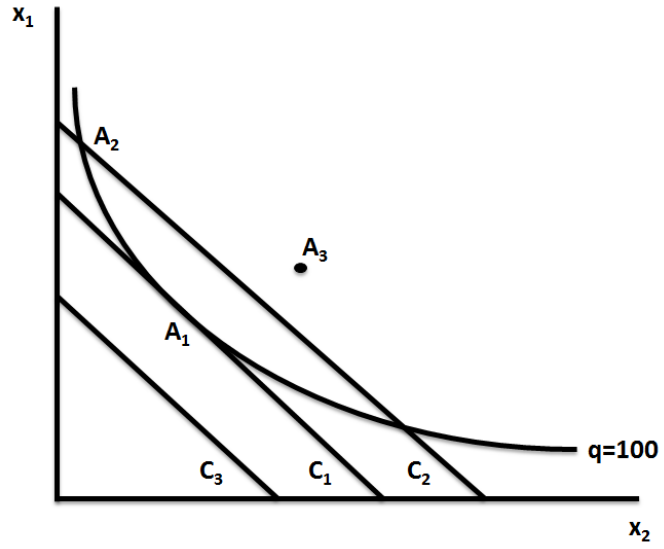

Figure 3.1: Allocative and Technical Efficiency

Figure 3.1 shows a firm's production of a given output, q. The graph shows a locus of production at q=100 representing the minimum combination of inputs, x_1 and x_2, that can produce 100q; the gradient represents the marginal rate of technical substitution between inputs. The parallel lines are isocost curves, C_1 and C_2 that represent uniform costs of production along them and increasing cost with distance from the origin: $C_2 > C_1$. The gradient reflects the ratio of input prices.

Assuming an output of q=100, the producer seeks to minimise its costs of production. The producer is technically efficient at any point along the locus of production, that is, it is not possible to produce 100q without using at least each amount of input along the curve. Any point above the curve for which production is 100q is technically inefficient, e.g. A_3, since it uses more inputs than are necessary to produce 100q.

Points A_1 and A_2 are both technically efficiency. However, at A_1 the cost of production is lower. This is due to the use of a combination of inputs which, for the same output, have a lower cost. This is, relative to A_2, allocatively efficient.

The optimal point of production is where the locus of production is tangential to the isocost curve; that is, it is not possible to produce 100q at lower cost. Here, technical and allocative efficiency are jointly achieved. A corollary of this is that allocative efficiency implies technical efficiency, but not the reverse. That is, technical efficiency is a necessary condition for cost minimisation and allocative efficiency is a sufficient condition for cost minimisation. Isocost line C_3 is unattainable for 100q.

Overall, cost efficiency is defined at the sum of technical and allocative efficiencies.

Finally, Farrell was the pioneer of empirical application, using linear programming techniques in agriculture. This seminal work inspired the development of two broad empirical techniques for frontier analysis, data envelopment analysis (DEA) and stochastic frontier analysis (SFA). In this thesis, we make use of SFA as our device for the measurement of efficiency. We justify our use of this method in the subsections that follow.

We proceed to the justification of frontier techniques on theoretical grounds, before presenting the cost function, which forms the basis of our approach. We then move to the econometric cost function, then to the stochastic frontier model and finally to total factor productivity.

3.2 The Theoretical Case for Frontier Techniques

As noted in previous chapters, we are interesting in NHS efficiency. Specifically we are interested in hospitals within the NHS. Despite a litany of theoretical groundwork, empirical analyses and sequential methodological advances, there exists, at present, no singly accepted framework for assessing hospital efficiency (cf. Hollingsworth, 2008; Hussey et al., 2009; Mutter et al., 2011). There is significant potential for frontier-type efficiency measures in health markets (Lovell, 2006; Mutter et al., 2011). We argue in favour of frontier techniques. The theoretical case for frontier-based analysis is set out as follows.

The implication of attempting to measure efficiency at all is that all economic agents are not completely efficient. A widely accepted theory of inefficiency is Leibenstein (1966), who labels organisational inefficiency X-efficiency. Leibenstein's contention is twofold: at the firm level, that as a result of information dissemination, motivation, difficulty in monitoring all staff and agency issues, any sizeable organisation is likely to be – at least to some degree - inefficient; and at the individual level, that human behaviour is composed of two parts, one of rationality (maximising their utility) and another of non-rational behaviour (suboptimal performance), which may lead to an inefficient level of individual or firm performance.

Other authors have proposed theories for both firm and individual behaviour (Table 3.1 – and this list is by no means exhaustive) which may lead to some suboptimal level of individual or (by extension) firm performance. These theories are thought to share an ontological core – inefficiency - and are thus interchangeable with Leibenstein's X-efficiency. Put differently, underlying these theories is the basic notion that for one reason or another, be it X-efficiency, the Peter Principle, weak identification, bounded rationality, etc., organisations may not be, always and everywhere, fully efficient.

Two further important developments in the literature are, firstly, Leibenstein & Maital (1992), who posit that frontier techniques are perhaps the best way in which to measure X-efficiency and secondly, Rice (2003), who suggests X-efficiency is valid in health markets. Rosko and Mutter (2011) is a recent example of an X-efficiency based health care frontier efficiency analysis. These together suggest that frontier methods are appropriate for determining the quantity of interest, namely inefficiency. Further, that frontier methods are derived from the foundations of empirical efficiency analysis (Cooper and Lovell, 2011) [9], and have been applied frequently in health care, give us confidence in this approach (Hollingsworth et al., 1999; Hollingsworth, 2003, 2008; Mutter et al., 2011).

Finally, Shleifer's (1985) theory of yardstick competition operationalises measurement of relative efficiency in a franchised monopoly market. This represents another health based use of frontier-type analyses and has been applied recently in health efficiency analysis (Olsen and Street, 2008).

[9] See http://www.terry.uga.edu/~knox/courses/READINGLIST8820I.pdf for a fuller literature survey of 'raw methods'.

These together constitute a robust theoretical case for frontier techniques in health care. Recently, economists in the NHS setting appear to have discarded frontier methods, but maintain the use of cost function-based approaches. Frontiers are elsewhere seen as the foremost hospital efficiency analysis tool and are seen to have great potential in health (cf. Lovell, 2006; Mutter et al., 2011). Then, the natural question is as to why these methods are out of favour with NHS-based economists. Some answers are to be found in the empirical setting. Indeed, we seek to answer some of these concerns in our empirical work. We return to this issue in due course.

Year	Author	Publication	Firm/Individual	Technical/Behavioural	Synopsis
1935	**Hicks**	Econometrica	Individual	Behavioural	Once a monopolist has obtained a monopolistic position, s/he is unlikely to continue to maximise profits; Monopolists enjoy a quiet life
1938	**Skinner**	The Behaviour of Organisms	Individual	Behavioural	Operant conditioning
1955	**Simon**	Quarterly Journal of Economics	Individual	Behavioural	People are 'satisficers' - they do enough; Bounded rationality
1962	**Alchian & Kessel**	Aspects of Labour Economics	Firm	Technical	Maximising profits is a 'cardinal sin'
1962	**Averch & Johnson**	American Economic Review	Firm	Technical	The A-J Effect
1964	**Williamson**	The Economics of Discretionary Behaviour: Managerial Objectives in a Theory of a Firm	Both	Behavioural	Maximising profit is but one of several managerial objectives
1965	**Alchian**	Il Politico	Firm	Technical	Defrayed ownership leads to diminished managerial monitoring and thus control
1966	**Leibenstein**	American Economic Review	Firm	Behavioural	X-inefficiency at firm level due to information, agency issues, monitoring, decentralisation of Command
1969	**Peter & Hull**	The Peter Principle	Individual	Technical	The Peter Principle – individuals rise to reach their level of incompetence
1971	**Niskanen**	Bureaucracy and Representative Government	Individual	Technical	Public mangers maximise their budgets regardless of inefficiency
1971	**Evans**	Canadian Economic Journal	Both	Technical	No reason for doctors to be efficient; hospitals do not cost minimise/profit maximise in the neo-classical sense
1974	**de Alessi**	Public Choice	Individual	Technical	Public managers have bias towards capital-intensive budgets
1976	**Lindsay**	Journal of the Political Economy	Individual	Technical	Public managers seek visible inputs
1976	**Stigler**	American Economic Review	Both	Technical	X-inefficiency is a myth; individuals maximise utility in different ways
1977	**Harris**	Bell Journal of Economics	Firm	Technical	Hospitals are two organisations each with differing objectives
1982	**Bailey & Freidlaender**	Journal of Economic literature	Firm	Technical	Hospitals are scarcely fully occupied always and everywhere
1988	**Hansmann**	Journal of Law, Economics and Organisation	Firm	Technical	Firms have more complex classification than simply public or private; problems arise via hierarchy, coordination, incomplete contracts, monitoring, agency costs
2003	**Rice**	The Economics of Health Reconsidered	Firm	Technical	X-efficiency applicable in health
2006	**Smith & Street**	The Elgar Companion to Health Economics	Firm	Technical	Principal-agent relationship at all levels
2012	**Oliver**	Journal of Health Politics, Policy & Law	Individual	Behavioural	Behavioural economics applied to the health sector (NHS)

Table 3.1: Theories of Inefficiency

3.3 The Cost Function

Frontier approaches to efficiency measurement are based on cost and production functions. In this section, we outline our tool of choice, the cost function, which is the basic economic model from which our econometric counterparts in later chapters are derived. Because we do not use the economic counterpart of the cost function, the production function, it is presented only tangentially, where necessary, in our discussion. The cost function is preferred to a production function as it embodies a richer economic problem (that is both allocative and technical inefficiencies; production function-based frontiers permit measurement of technical efficiency only) and allows multiple outputs to be included simultaneously (cf. Jacobs et al., 2006; Eakin, 2008). Attention is given to the conceptual features and properties of the cost function, and health-specific considerations. In our econometric work, we test these properties to validate empirical models. We pay attention only to aspects of theory relevant to this thesis; for comprehensive coverage of production theory, econometrics and health applications thereof, see Chambers (1988), Greene (2012c), Hollingsworth and Peacock (2008), respectively.

3.3.1 The Economic Cost Function

In theory, firms are assumed to seek to minimise costs according to output(s) and input (or factor) prices. Therefore, a cost function represents the minimum attainable cost for a firm in a fixed period of time for a given combination of outputs and input prices. These are assumed exogenous to the firm; thus the mix of inputs is sought which minimises costs. Then, the problem is typically defined mathematically as one of minimisation,

$$C(y, w) = \min_{x \geq 0}(w'x : x \in V(y)) \tag{3.1}$$

Where C are costs and $y = (y_1, y_2, \dots, y_n)'$ is a vector of outputs. Similarly, w and x are vectors of input prices and inputs, respectively. $w'x$is the inner product of the vectors w and x[10]. $V(y)$ is the feasible set of outputs. A feasible set denotes the range of outputs that are attainable to the firm; not all outputs are attainable (there may be minimum levels of output, indivisible units of output, etc.).

The cost function is said to embody a set of regularity conditions, meaning that the following properties are upheld:

[10] that is, $w'x = w_1.x_1 + w_2.x_2 + \cdots + w_n.x_n$

(i) Non-negativity: Costs can never be negative. Mathematically,

$$c(y,w) > 0 \; \forall \; y > 0, x > 0 \tag{3.2}$$

(ii) No fixed costs: Zero output is costless[11]. Mathematically,

$$c(0,w) = 0 \tag{3.3}$$

(iii) Non-decreasing in w: When input prices are increased, costs do not decrease. Mathematically,

$$if \; w^0 \geq w^1 then \; c(y,w^0) \geq c(y,w^1) \tag{3.4}$$

(iv) Non-decreasing in y: When outputs are increased, costs do not decrease. Mathematically,

$$if \; y^0 \geq y^1 then \; c(y^0,w) \geq c(y^1,w) if \; w^0 \geq w^1 then \; c(y,w^0) \geq c(y,w^1) \tag{3.5}$$

(v) Positive linear homogeneity: multiplying all input prices by an amount *k* will result in a *k*-fold increase in costs. Mathematically,

$$c(y,kw) = kc(y,w) \; \forall \; k > 0 \tag{3.6}$$

(vi) Concavity in w: This property is derived directly from the *fundamental inequality of cost minimisation* (Chambers, 1988 pp.53). Mathematically,

$$c(y,\theta w^0 + (1-\theta)w^1) \geq \theta c(y,w^0) + (1-\theta)c(y,w^1) \tag{3.7}$$

For a fuller discussion of these properties, and their proofs, the reader is referred to Chambers (1988). The fundamental uses for these properties in empirical settings are discussed by Coelli et al., 2005 (pp. 24). One important feature is that these properties allow validation of

[11] In some circumstances, there may be short run costs incurred for zero production, e.g. start-up costs before production begins.

econometric cost functions. We make use of the cost function as the basis of assessing producer inefficiency. For the purposes of the empirical work in this thesis, it is therefore important that the cost function is justifiable on economic grounds.

3.3.2 Economies of Scale and Scope

Economies of scale properties are often a quantity of interest to both researchers and policy makers. Economies of scale properties in production can readily be assessed via the cost function. Economies of scale are defined as the proportional change in costs that corresponds to a change in the level of output. By taking natural logarithms, proportional changes are observed, then scale economies can be assessed by differentiating the cost function with respect to output(s),

$$\varepsilon_{cy} = \left[\sum_{n=1}^{N} \frac{\partial lnc}{\partial lny_n}\right]^{-1} \tag{3.8}$$

Where there are up to *n* outputs[12]. If $\varepsilon_c > 1$, economies of scale exist; $\varepsilon_c = 1$ denotes the optimal scale of production; $\varepsilon_c < 1$ reflects that the firm is operating under diseconomies of scale. We make use of this measure in our empirical work in subsequent chapters.

When $n > 1$, a further quantity of interest is economies of scope; that is, the extent to which costs vary under joint production. Economies of scope can be measured empirically,

$$S = \left[\sum_{n=1}^{N} \frac{c(y_n, w)}{c(y, w)}\right] - 1 \tag{3.9}$$

Where S is the global (i.e. across all outputs) economies of scope (for product-specific measures refer to Coelli et al. 2005, pp.30). This represents the proportional change in costs if all outputs are produced separately. If $S < 0$, then production should be separate; where $S > 0$, produce jointly. In our empirical chapters, $n = 1$, we are therefore unable to compute a measure of economies of scope. We retain its exposition for completeness.

[12]Which reduces to $\varepsilon_c = \left[\frac{\partial lnc}{\partial lny}\right]^{-1}$ when n=1

3.3.3 Additional Features of the Economic Cost Function

The basic cost function defined above relates firms' costs of production to levels of output and input prices, considered to be exogenous to the firm. That is, the basic cost function takes the form,

$$c = f(y, w) \tag{3.10}$$

Where c are costs, y is (are) output(s) and w are input prices. In reality, there are a number of other exogenous characteristics of the production environment in which firms operate. This point is of particular relevance in the field of health. It is therefore important to augment the basic function with these features.

The first of these features is time. Incorporating time into the cost function can accommodate exogenous shifts in the production environment that firms face over time. Time trends are interpreted as technological change (Kumbhakar and Lovell, 2000); indeed, they are typically used to compute total factor productivity (TFP) indexes from econometric efficiency models (Coelli et al., 2005). In empirical settings, this can be operationalised in a number of ways, for example by inserting a time trend or time period dummy variables into the cost function. The cost function then becomes,

$$c = f(y, w, t) \tag{3.11}$$

Next, there may be exogenous heterogeneity present in the production environment - that has bearing on firms' costs – that can be captured in the cost function. Empirically, capturing this observable heterogeneity is conducted using what are termed environmental variables (Coelli et al., 2005).

To the extent that hospitals offer a range of services and specialisations, it is unlikely that two are the same. Indeed, hospitals are commonly in various stages of investment cycles, under differing ownership regimes, providing varying levels/types of teaching, and to varying extents are part of service networks, inter alia (Mutter et al., 2011). Unless these features are controlled for, assigning common cost or production functions is questionable. Some features can be readily incorporated into efficiency analysis, ownership status for example (Tiemann et al., 2012). For other sources of heterogeneity, data are continually refined and developed for various aspects of service heterogeneity in health. For instance, the way in which healthcare diagnoses and procedures are coded - by ICD or OPCS coding – are subject to

regular updates to reflect developments in practice (WHO, 2004; HSCIC, 2013). These should be, where possible, incorporated into the cost function.

In addition, there is heterogeneity at the patient level. Patient-level heterogeneity is a clear issue when characterising cost functions (Iezzoni, 2009). Daidone and Street (2013) used patient-level data to control for patient-level heterogeneity in the costs of specialised care in the NHS, in part to make judgements on performance.

However, even in the case that highly granular data are to hand, there are likely many differences that remain unobserved, the age of hospital buildings or their physical layout, for example. This implies controlling for unobservable heterogeneity is critical. We return to this issue in subsection 3.5.4.4.

The incorporation of environmental variables (observable heterogeneity), denoted z, leads to,

$$c = f(y, w, t, z) \tag{3.12}$$

Lastly, it is important to characterise the quality of the services provided in the hospital setting (Sloan, 2000). Capturing service quality in health efficiency analyses is a challenging task owing to it being unobservable directly and complexity. There are a number of ways to proxy its measurement, a number of health studies that have incorporated the measurement of quality in their efficiency analyses. We return to discuss this issue in detail in section 4.5.

Irrespective of the specific measure of quality, the amendment to the general cost function yields,

$$c = f(y, w, t, z, q) \tag{3.13}$$

With this, we complete the definition of the economic cost function in hospital markets. It is then possible to define and test a cost function empirically as the basis for our analysis of efficiency. We return to the specific realisations of these general features both later in this chapter, and in our empirical applications.

Of course, it is necessary to make the link between the definition of the cost function and the measurement of efficiency using frontier-based techniques. As a first step in this process, we move to the discussion of the econometric cost function, which is the basis of our empirical endeavours in later chapters.

One central issue to highlight in passing is that of unobserved heterogeneity. In the case that data for the features of the cost function are lacking, the features are measured imperfectly or that the properties of the cost function are breached, there may be inaccuracy in the cost

function itself. Thus, when making use of frontier techniques, which are based on distances from the cost function, it is of importance to make allowances for any unobserved influences on costs. We return to this issue in detail in subsequent sections in this chapter and in chapters, 4, 5 and 6.

3.4 The Econometric Cost Function

In this section, we develop the econometric counterpart to the economic cost function defined above. We begin by setting out the cross-sectional (or pooled panel) model. Next, we discuss panel models, that is, models based on several cross-sections observed over time. Next, functional form is considered. Estimation and testing are discussed throughout. These models represent the basis from which our efficiency analysis tools are developed. We therefore proceed to develop the efficiency analysis tool – the stochastic frontier model.

3.4.1 Cross-sectional Econometric Cost Functions

In cross-sectional settings, firms are observed only once, and observations are assumed to be independent. The cross-sectional cost function takes the form,

$$c_i = \alpha_0 + \beta_1 y_i + \beta_2 w_i + \beta_3 z_i + \beta_4 q_i + \varepsilon_i \qquad (3.14)$$

Where c_i is the cost of firm i and α_0 is a constant term. y_i is a kx1 vector of outputs for firm i, w_i, z_i and q_i are kx1 vectors of input prices, environmental variables and quality, respectively. The β terms are parameters to be estimated. $\varepsilon_i$is the error term (also referred to as the residual or the disturbance) which captures any variation in costs that are not captured by the regressors (Gujarati, 2003).

The betas are the first derivative of cost with respect to each variable, so, for example in the case of output, $\frac{\partial c}{\partial y} = \beta_1$. This allows the researcher to, ceteris paribus, estimate the relationship of cost and each variable.

The model is assumed to embody Gaussian assumptions (for detail, see Gujarati, 2003 pp.66-76). Estimation proceeds typically via Ordinary Least Squares (OLS).

3.4.2 Econometric Cost Functions with Panel Data

We now consider the case where the cross-section of firms is observed repeatedly over several time periods. When such data exists, it is termed panel data. One way in which to

proceed is to ignore the structure of the data, assuming the observations are independent. This is called pooling, and the treatment is as section 3.4.1 above. However, there are many advantages to using panel data that are of use in our empirical work. First, there is likely to be information about firms held in the structure of the data, which can be exploited. Specifically, repeated observations of the firm mean that a firm-specific effect can be observed. Second, panel data are more informative: there is more variability in the data, a greater number of degrees of freedom and parameter estimates are more efficient from panel data models. Importantly, collinearity is less of an issue than in the cross sectional equivalent models. Third, the dynamics of data can be studied using panel data. Lastly, issues around aggregation across firms may be reduced with panel data. See Baltagi (2008) for details.

The panel data cost function is, with the reintroduction of time into the cost function, of the general form,

$$c_{it} = \alpha_i + \beta_1 y_{it} + \beta_2 w_{it} + \beta_3 z_{it} + \beta_4 q_{it} + \beta_5 t + \varepsilon_{it} \qquad (3.15)$$

Where c_{it} are the costs of the i^{th} firm in time period *t*. y_{it} is a kx1 vector of outputs for firm *i* in time period, *t*; w_{it}, z_{it} and q_{it} are kx1 vectors of input prices, environmental variables and quality, respectively. The βterms are parameters to be estimated. ε_{it} is the residual. α_i is the firm effect, that is, the capture of all factors that are unobserved, firm-specific and time-invariant.

Estimating the model can be done in two ways, namely fixed or random effects. For fixed effects, the α_i are fixed parameters - simply firm dummy variables; estimation proceeds via Least Squares Dummy Variable (LSDV) regression with the constant term removed. In this setting, any correlation between the regressors and firm effects is captured in the effect. Estimates of beta are *within (group)* estimates and are unbiased.

For random effects, estimation proceeds via either Generalised Least Squares (GLS) or maximum likelihood (ML). In this setting, the firm effects are assumed uncorrelated with the regressors; this can be duly tested. Estimates of beta are thus *within* and *between (group)* and are (potentially) biased, if there is correlation between regressors and firm effects. Advantages of the random effects approach is that the beta estimates are, relative to the fixed effects, efficient. In addition, the random effects model can incorporate time-invariant variables, which is not possible for fixed effects estimation.

The choice between fixed and random effects can be made in a number of ways. The researcher may have their own preference depending on their own circumstances. For example, if there are important binary variables, then a random effects approach is preferable because the fixed effects approach cannot accommodate these variables. Statistically, a Hausman test (Hausman, 1978) based on the strength of the correlation between the firm effects and the regressors, can indicate whether fixed or random effects are preferred. However, this should only be an indicator; there should not be an overreliance placed on this test (Baltagi, 2008). An alternative approach is the Wu test proposed by Greene (2012b) which makes use of group mean variables. In a similar spirit, it is possible to retrieve within beta estimates in random effects estimation using group mean variables, which capture the correlation between the regressors and the firm effects (cf. Mundlak, 1987; Baltagi, 2006). This allows estimation of unbiased and consistent beta parameters. However, this is at the expense of degrees of freedom, and so the approach may be of limited use in smaller samples.

3.4.3 Data Transformation and Functional Form

In preceding sections, the variables are in their raw form. In empirical settings, researchers typically use transformations of the data when estimating models. This is for a number of economic and technical reasons. We begin with the commonly used Cobb-Douglas functional form (Nerlove, 1963).

A Cobb-Douglas cost function, which is said to be the dual[13] of the Cobb-Douglas production function, is imposed by taking the natural logarithms of the dependent and independent variables, such that, in the case of the (cross-sectional) cost function,

$$\ln(c) = \alpha + \sum_{n=1}^{N} \beta_n \ln(x_n) + \varepsilon \tag{3.16}$$

Where $x_n = (y, w, z, q)$. The Cobb-Douglas functional form has a number of appealing features. First, it imposes concavity (and thus is in agreement with property (vi) section 3.3.1) of the economic cost function. Second, it implies that (as per its name), the estimates of beta can be directly interpreted as cost elasticities. Third, it allays (to some extent) heteroscedasticity concerns (Jacobs et al., 2006), which is in keeping with the Gaussian assumptions of the econometric model.

[13] That is, from the cost function, it is possible to work back to the production function via the transformation function, and vice versa (see Coelli et al., 2005, chapter 2).

Next, another commonly used functional form, is the transcendental logarithmic or translog (Christensen and Greene, 1976). A translog is a generalisation of the Cobb-Douglas, with the addition of squared and interaction terms for all variables. Thus for *n* variables, there are approximately $n(n+1)/2$ parameters. The translog takes the form,

$$\ln(c) = \alpha + \sum_{n=1}^{N} \beta_n \ln(x_n) + \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \beta_{mn} \ln(x_n) \ln(x_m) + \varepsilon \tag{3.17}$$

Where $x_n = (y, w, z, q)$ and $\ln(x_n)\ln(x_m)$ denotes interactions between variables. A translog has some appealing empirical and economic features: its flexible nature means it provides a second-order differential approximation to any unknown function $f(.)$ (Kumbhakar and Hjalmarsson, 1995); it does not impose restrictions on substitution possibilities; and allows economies of scale to vary with output levels. This is likely to provide a better empirical approximation of the unknown cost function than the Cobb-Douglas. The price is the addition of variables, which may affect the precision of estimates. The translog has the useful feature that it is possible to mean-scale the regressors in order to interpret the first order terms as elasticities. See Appendix A for derivation of this result.

For a complete approximation of an unknown function, the Fourier functional form has been proposed (Gallant, 1981). The Fourier cost function comprises squared terms and linear combinations of the sine and cosine of the variables, thus,

$$\begin{aligned} \ln(c) = \alpha + \sum_{n=1}^{N} \beta_n \ln(x_n) + \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \beta_{mn} \ln(x_n) \ln(x_m) \\ + \sum_{n=1}^{N} \left[\beta_m \cos(\ln(x_n)) + \beta_p \sin(\ln(x_n)) \right] + \varepsilon \end{aligned} \tag{3.18}$$

Where $x_n = (y, w, z, q)$. Calculus shows that a Fourier series can exactly represent any underlying function, $f(.)$ (Mitchell & Onvural, 1996). To achieve its exact representation, the addition of (potentially infinite) higher order sine and cosine terms are required. Thus estimation may problematic in small samples (Mitchell & Onvural, 1996). Therefore, a trade-off may be required between the number of parameters and the desired fit.

As shown above, the Fourier form nests the translog. The Fourier form – an exact representation of $f(.)$ - may be preferable to the translog, which is an approximation. Indeed, a translog may have difficulty capturing the true underlying cost function when the size of firms varies significantly (Feng & Serletis, 2009). The translog, in turn, nests the Cobb-

Douglas, which again is more restrictive. Therefore, using this sequence of functional forms, the researcher is equipped with the full range of flexibility when seeking to characterise the underlying cost function. Given that these forms are nested, it is readily possible to test down to arrive at a preferred specification. In empirical applications, the Fourier functional form is not widely used; the translog is used most often.

Other common functional forms include the linear, quadratic, normalised quadratic and the generalised Leontief (Coelli et al., 2005 pp. 211).

In passing, we note some alternative methods for transforming variables, including the Box-Cox (of which a special case is the Cobb-Douglas as being a logarithmic transform) and the Inverse Hyperbolic Sine transformation (Burbidge et al., 1988). However, we do not make use of these in this thesis, in keeping with the wider literature in making use of the translog.

Finally, we discuss the imposition of positive linear homogeneity in the cost function (assumption (v) section 3.3.1). Consider a cross-sectional, Cobb-Douglas cost function with a single output and two input prices variables,

$$\ln(c_i) = \alpha_0 + \beta_1 \ln(y_i) + \beta_2 \ln(wl_i) + \beta_3 \ln(wk_i) + \varepsilon_i \quad (3.19)$$

Here, wl_i are labour input prices and $wk_i$are capital input prices. For linear homogeneity of degree one in input prices, we require $\sum_w \beta_w = \beta_2 + \beta_3 = 1$[14]. Imposing this is possible in two ways. First, the restriction can be imposed for estimation (this is done straightforwardly in any modern software package). Alternatively, it is possible to normalise costs and input prices variables by one of the input price variables and substitute in terms (see Kumbhakar et al., 2015, pp103-104 for derivation). The choice of input price with which to normalise is irrelevant. That is,

$$\ln\left(\frac{c_i}{wk_i}\right) = \alpha_0 + \beta_1 \ln(y_i) + \beta_2 ln\left(\frac{wl_i}{wk_i}\right) + \varepsilon_i \quad (3.20)$$

3.4.4 Summary: Cost Functions

In sections 3.3 and 3.4, we have defined and described the cost function, both in terms of its economic features and its econometric representation and estimation. As noted, the cost function is the tool from which our efficiency analysis is derived. Further, we make use of features of the cost function in our empirical work, underlining the importance of its

[14] Which is extended to include $\sum_w \beta_{wk} = 0 \; \forall \; k$ and $\sum_w \beta_{wy} = 0 \; \forall \; y$ in the translog setting. *k* are other input prices; y are outputs.

exposition. Specifically, we look to validate our empirical models by testing correspondence with the economic properties of cost functions. Next, we make use of the ability to measure both scale properties and technical change in our empirical application. We make use of and test a number of functional forms, as detailed here.

3.5 The Stochastic Frontier Model

We have introduced the concept of efficiency, argued in favour of frontier approaches for measuring efficiency and presented the cost function, both economically and econometrically. We now move to the final methodological stage, which is to define our efficiency measurement method of choice: the stochastic frontier (SF).

We present the model conceptually and go on to justify its use over mathematical programming alternatives. We then present the stochastic frontier in its simplest form, the model's assumptions and estimation. In the next subsection we discuss the retrieval of firm-specific inefficiency predictions. Next, we consider extensions of the SF for panel data and additional features for capturing efficiency change over time and unobserved heterogeneity. Finally, we consider SF models for inefficiency measurement at vertically separate organisational levels.

In the section that follows, we round off the methodological discussion with an overview of the features defined in this section and introduce the concept of Total Factor Productivity. We demonstrate how to measure the change in Total Factor Productivity based on cost frontiers. This is important insofar as it allows an overall account of performance in the sample; we go on to estimate such a quantity in chapter 5 of this thesis.

The SF model, in essence, uses the cost function as the efficiency frontier faced by firms in the market. The frontier assumes the shape of the cost function. The frontier represents, in the case of the cost frontier, the minimum attainable cost for a firm, given its levels of outputs and input prices (and other features defined by the cost function, see section 3.3.3). Deviations of firms from the frontier are considered to be, in part, due to inefficiency. The distance to the frontier represents the magnitude of the inefficiency. In addition, the SF model allows for the removal of random statistical noise in the data from inefficiency estimates[15].

[15] It is from this feature that the model derives its name; without the treatment of noise, the frontier is deterministic

Random noise can encompass a number of features. Typically, researchers suggest the noise comprises random shocks to production, including untoward events such as strikes, unusual weather, force majeure, etc. In addition, this component can account for measurement error and approximation error (from the choice of functional form) (Coelli et al., 2005). The defining feature of the SF model is its ability to remove these factors so that they do not distort the underlying metric of interest – inefficiency. We discuss how this is achieved in following subsections.

Then, the overall observed deviation from the frontier is considered to comprise both inefficiency and random shocks to firms' production that impinge on their costs. Figure 3.2 shows the conceptual features of the SF model.

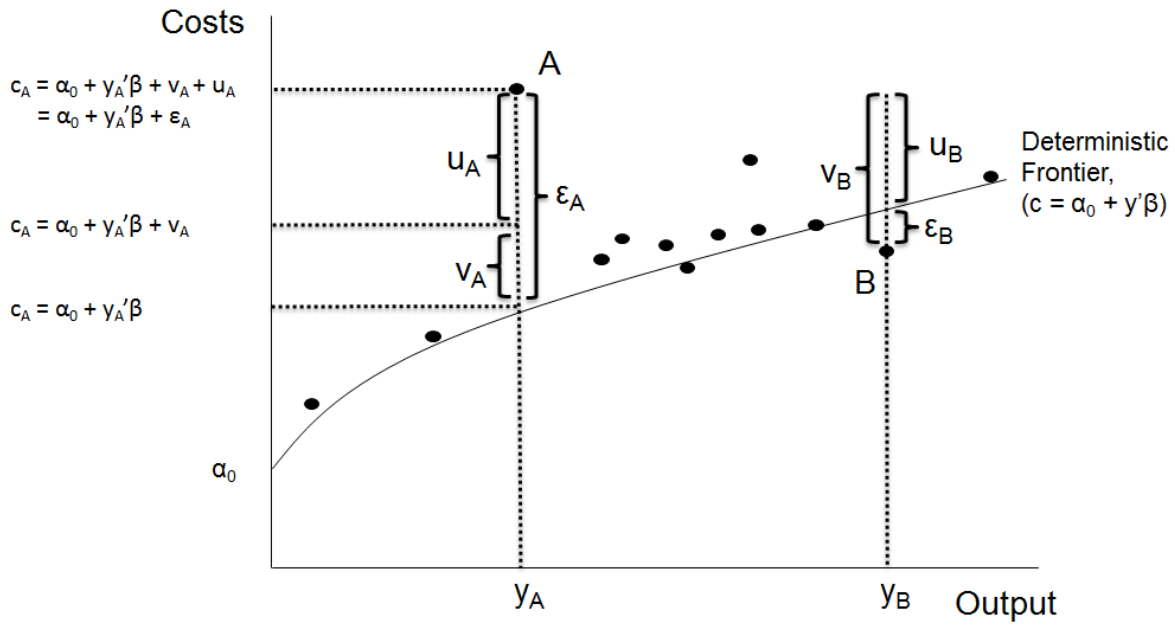

Figure 3.2: The Stochastic Cost Frontier Model

Fig 3.2 shows the basic features of the SF model. The gradient of the frontier is defined according to the cost function and is the deterministic element of the model. In this case, costs, c, are defined as a function of outputs, y (i.e. $c = \alpha_0 + y'\beta$).

We consider observation A, representing firm A. Here, as shown, firm A has output, y_A, at which the conditional costs for firm A – shown via the cost frontier - are $c_A = \alpha_0 + y_A'\beta$. The observed value is in fact point A, which is higher than the expected (conditional) costs for firm A, given its level of output. Then, the firm-specific observed deviation, ε_A, comprises

both the firm's inefficiency, u_A, and random statistical noise, v_A. In this case, both noise and inefficiency have a positive influence on costs.

In some cases, firms are observed as below the deterministic frontier, as in observation B where the observed deviation from the frontier, ε_B, is negative. This is the result of a noise component, v_B, which is negative and greater than the firm's (positive) inefficiency, u_B. Here, firm B has output, y_B, at which the cost frontier is $c_B = \alpha_0 + y_B'\beta$.

We discuss our preference for the stochastic frontier model next, before considering the econometric estimation of the SF model.

3.5.1 Stochastic Frontier Analysis versus Data Envelopment Analysis

It would be possible to measure efficiency using data envelopment analysis (DEA), an approach to efficiency measurement based on mathematical programming[16]. Here, the frontier is a perimeter around the extreme points of the data (the minimum in the case of a cost frontier); measures of inefficiency are based on distances to this frontier. This is shown in figure 3.3 below.

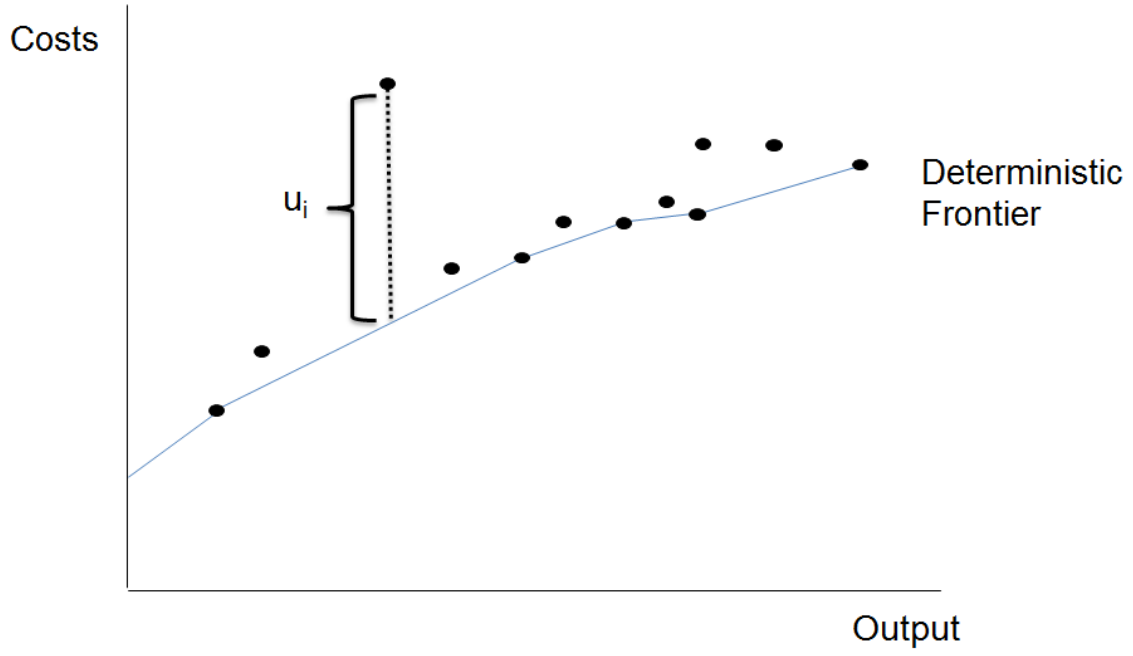

Figure 3.3: Data Envelopment Analysis

[16] See Fried et al. (2008) for methods and applications of DEA

We make use of SF models exclusively - we do not make use of DEA in this thesis. The SF approach is preferred for a number of reasons, which can be broadly categorised as methodological, cost function-based advantages and the incorporation of uncertainty.

Methodological issues include, firstly, that the residual is composed in DEA measures – there is no distillation of noise from the measure of inefficiency (fig. 3.3). This means that the inefficiency measure is necessarily biased by noise, unless noise is assumed away. This may distort the expected duality between costs and production functions (Greene, 2008). Importantly, total inefficiency is overestimated, which is undesirable in the policy context. Of course, in the SF world, an assumption must be made around the distribution of inefficiency to allow estimation, which some have argued is arbitrary (Newhouse, 1994). However, there is evidence to suggest that efficiency predictions and corresponding ranks are correlated between distributional assumptions, both in the general setting (Coelli et al., 2005) and specifically in the NHS context (Jacobs et al., 2006, pp. 68). Moreover, in some cases, distributional assumptions are preferable to distribution-free approaches based on panel data (Kim and Schmidt, 2000).

Next, that the DEA frontier is constructed on the extreme data points in the sample implies that inefficiency estimates are susceptible to extreme outliers, meaning that, again, overall inefficiency is overestimated. Further, the method is unable to account for measurement error and approximation error. This may introduce bias, both positively and negatively, which is unwanted.

DEA measures do not have the advantage of being able to analyse features of the cost function which is possible with SFs. Specifically, cost efficiency cannot be delineated into its technical efficiency and allocatively efficiency components, which is possible in the SF framework. Next, although it is possible to observe scale changes, it is not possible to examine scale properties in the depth that can be achieved in the SF model. Using a translog functional form, for example, allows the researcher insights into the economies of scale properties across the output range. This is not possible in the DEA framework. Further, the incorporation of other variables into the cost frontier allows for the measurement of the effect of certain features on costs, which is likely of high value to policy makers. For example, estimating the marginal cost of quality, or, as we do in our empirical work, estimating the effect on costs of providing teaching is key information sought by policy makers. A further advantage is that it is possible to check that the coefficients correspond to the economic

reality, or not as the case may be. This is helpful when looking to validate the model. This is not possible with DEA.

Lastly, the estimated cost function can be used for predictions, which can be useful in a number of ways. For example, in our empirical work in chapter 5, we are able to make predictions based on the estimated parameters of the cost function to simulate the effect on costs of organisational units merging. This was useful for policy makers who observed a trend of mergers, but lacked data, and thus evidence, on the effect on costs. Again, this would not be possible using DEA.

Lastly, the DEA framework does not allow the incorporation of uncertainty into the analysis in the same way that the SF framework does. Firstly, uncertainty can be reflected in confidence intervals around parameter estimates of cost function variables; and prediction intervals around point estimates of inefficiency (intervals are not confidence intervals for inefficiency predictions - this is an important distinction, see Wheat et al., 2014). There are bootstrapping techniques available for confidence intervals around DEA estimates to reflect the sampling uncertainty (Simar and Wilson, 2000). There are further methods to use a DEA approach and separate efficiency and noise, the Stochastic Non-smoothed Envelopment of Data (STONED) method (Kuosmanen and Kortelainen, 2012). However, these approaches do not allow the wealth of advantages relative to an econometric approach.

In addition, that SF modelling is based on econometrics means that there is a wealth of testing available for sensitivity in a number of forms (model specification, functional form, etc.), which there is not for DEA being based on mathematical programming.

Therefore, for these three general categories of reasons, we adopt an econometric approach in favour of a mathematical programming approach.

3.5.2 Cross-Sectional Stochastic Frontiers

We begin the discussion with the simplest form of econometric efficiency analysis: Corrected Ordinary Least Squares (COLS). This model makes use is the econometric cost function, equation (3.14), and subtracts the minimum error from each observed residual. This, in effect, shifts the cost function to the minimum observation in the sample. The measure of inefficiency is then the distance of each observation from the frontier. Thus,

$$\hat{u}_i = \varepsilon_i - \min_i\{\varepsilon_i\} \tag{3.21}$$

This model is not strictly a stochastic frontier, given that there is no decomposition of noise and inefficiency; COLS is, in essence, the econometric equivalent of DEA. The practical appeal of COLS is its simplicity – it can be run via OLS, and does not impose the distributional assumptions of the stochastic frontier model. Accordingly, it is typically used in settings where the stochastic frontier model is unidentified[17]. Technically, COLS allows the possibility of fully efficient firms, which is not the case in the stochastic frontier model, to which we now turn.

The stochastic frontier was proposed simultaneously by Aigner et al. (1977) and Meeusen and van den Broeck (1977). It takes the form (a Cobb-Douglas functional form is assumed),

$$\ln(c_i) = \alpha_0 + \ln(x_i)'\beta + \varepsilon_i \tag{3.22}$$

$$\varepsilon_i = u_i + v_i \tag{3.23}$$

Where $c_i$are the costs of firm i, α_0 is a constant term, x is a vector of cost function variables, $x = (y, w, z, q)$, and β are corresponding parameters to be estimated. The observed error, ε_i, comprises both random noise, v_i, and the metric of interest, the firm-specific inefficiency, u_i as per eqn. (3.22).

The model is based on a number of assumptions. First, the two components of the error term are assumed orthogonal to each other and to the regressors. In addition, there a number of assumptions made about the noise and inefficiency components. Noise is assumed to have zero mean, to be homoscedastic and uncorrelated,

$$E(v_i) = 0 \tag{3.24}$$

$$E(v_i^2) = \sigma_v^2 \tag{3.25}$$

$$E(v_i v_j) = 0 \; \forall \; i \neq j \tag{3.26}$$

The inefficiency component is assumed homoscedastic and uncorrelated,

$$E(u_i^2) = constant \tag{3.27}$$

$$E(u_i u_j) = 0 \; \forall \; i \neq j \tag{3.28}$$

The inefficiency component is not assumed to have zero mean since some inefficiency is assumed.

[17] This is often the case in regulatory settings where the sample size is small; see chapter 4.5

These assumptions are problematic for estimation by OLS. Parameters can be estimated consistently, however, the inefficiency causes upward bias to the intercept, meaning that via OLS it is not possible to estimate cost efficiency. However, by making suitable distributional assumptions regarding both components of the error term, it is possible to estimate the model via maximum likelihood. Therefore, the following assumptions are made,

$$v_i \sim iidN(0, \sigma_v^2) \tag{3.29}$$

$$u_i \sim iidN^+(0, \sigma_u^2) \tag{3.30}$$

With these at hand, Aigner et al. (1977) proposed the log likelihood function for a production frontier, which can readily be adapted for a cost frontier,

$$lnL(c|\beta, \sigma, \lambda) = -\frac{I}{2}\ln\left(\frac{\pi\sigma^2}{2}\right) + \sum_{i=1}^{I} ln\Phi\left(-\frac{\varepsilon_i\lambda}{\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\varepsilon_i^2 \tag{3.31}$$

Where c are costs, I is the number of observations, $\sigma^2 = \sigma_u^2 + \sigma_v^2$ and $\lambda^2 = \sigma_u^2/\sigma_v^2$. When $\lambda^2 = 0$, the variance is due solely to random noise – there is no inefficiency[18]. The disturbance is defined as $\varepsilon_i = v_i + u_i = c - \alpha_0 - x'\beta$. Finally, $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal variable evaluated at x.

Estimation of the parameters cannot be achieved analytically since the first derivatives are highly non-linear, meaning that they have no analytical solution. Instead, an iterative maximisation procedure can be conducted, by using some starting values (say those derived from OLS) and altering their magnitude to maximise the likelihood function. See Greene (2012c) for maximum likelihood estimation techniques.

It is possible to use alternative assumptions of the distribution of inefficiency. Alternative distributions include exponential and gamma (Kumbhakar and Lovell, 2000). Using different distributional assumptions will have bearing on the predictions of inefficiency. Given that the choice is somewhat arbitrary, this poses issues for the validity of estimates. Helpfully, estimates and their rankings are often robust to the assumptions imposed (Coelli et al., 2005; Jacobs et al., 2006). However, independent of the choice of distribution, all firms are, by construction, to some degree inefficient[19].

Truncation can be introduced to allow for the centre of probability mass to leave zero. This parameter can be tested. The Rayleigh distribution allows for similar flexibility without the

[18] A test on this parameter has been suggested as a test for the presence of inefficiency, however, researchers typically use a likelihood ratio based test of the SF model versus ordinary least squares (Coelli et al., 2005 pp. 258).

[19] This is due to the fact that the probability of drawing any value from the distribution exactly is zero, so although zero is a possible value, the probability that zero is drawn from the distribution is exactly zero (Rho and Schmidt, 2015). Recent models have been developed to overcome this issue (Kumbhakar et al., 2013; Rho and Schmidt, 2015)

need to estimate an additional parameter (Hajargasht, 2014). It is further possible to introduce variables into the mean and variance of the inefficiency. This allows insights into factors which affect inefficiency. However, there is little consensus as to whether variables should be included in the cost function or in the mean of the inefficiency (Kumbhakar and Lovell, 2000). It is possible to introduce the variables in both parts of the model, however, this may induce endogeneity, which is known to distort inefficiency estimates in SF models (Mutter et al., 2013).

3.5.3 The Retrieval of Firm-Specific Inefficiency Predictions

The estimate of cost efficiency from the SF model above is,

$$\widehat{CE}_i = \exp(-\hat{u}_i) \tag{3.32}$$

Where $\widehat{CE}_i$ is the estimate of firm *i*'s cost efficiency and $\hat{u}_i$ is the component of inefficiency in the SF model. This is not observed directly. The SF model yields residuals, $\hat{\varepsilon}_i$, which are a composition of both statistical noise and inefficiency. Having estimated the model, therefore, an additional stage is required to compute the firm-specific inefficiency predictions.

By using the assumptions about the distributions of both noise and inefficiency, it is possible to derive the joint density of the composed error term (see Kumbhakar et al., 2015 pp. 319-322 for the derivation). From this, the conditional mean (or mode) can be taken as the point estimate of inefficiency, u_i. This work was pioneered by Jondrow et al. (1982). For the case of the half-normal distribution, the prediction of firm-specific inefficiency is[20],

$$E[u_i|\varepsilon_i] = \frac{\sigma\lambda}{(1+\lambda^2)}\left[\frac{\phi(\varepsilon_i\lambda/\sigma)}{\Phi(-\varepsilon_i\lambda/\sigma)} - \frac{\varepsilon_i\lambda}{\sigma}\right] \tag{3.33}$$

Where $\sigma, \lambda \; and \; \varepsilon_i$ are as before. Equally, $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal variable evaluated at x; $\phi(x)$ is the corresponding probability density function (PDF). u_i is the inefficiency component of the model which can be used to compute the prediction of firm cost efficiency, as above.

Once firm-specific predictions are obtained, the level of efficiency across the sample can be computed by extension,

$$\overline{CE} = \frac{\sum_{i=1}^{I} \widehat{CE}_i}{I} \tag{3.34}$$

[20] Alternatively, the minimum squared error predictor can be used to derive point estimates (Kumbhakar and Lovell, 200 pp. 104)

Where $\overline{CE}$ is the market (sample) average cost efficiency and $\widehat{CE}_i$ are firm-specific cost efficiencies.

3.5.4 Stochastic Frontier Models for Panel Data

The stochastic frontier model can be extended when panel data are available. Panel data sets are often much larger than their cross-sectional counterparts, allowing greater precision of the estimated model parameters. In addition, the richer information that can be derived from panel data models can be exploited in measuring firm inefficiency. Panel data models also allow characterisation of three conceptual features, namely temporal efficiency change, unobserved heterogeneity and multi-level organisational structures. We focus on these three aspects of SF models because they feature in our empirical work. They do so because these issues are of particular relevance in health markets and specifically in the NHS. We discuss these in turn. First, we present the general form of the panel data stochastic frontier,

$$\ln(c_{it}) = \alpha_0 + \ln(x_{it})'\,\beta + u_i + v_{it} \tag{3.35}$$

Where $c_{it}$are the costs of firm *i* in time period *t* and x_{it} are cost function variables. α_0 is the intercept and β is a vector of parameters to be estimated. u_i is the inefficiency of firm *i* which is time-invariant; and is u_{it} when inefficiency varies over time. v_{it} is random statistical noise.

Estimation proceeds in a variety of ways, depending on the model's features and specification. Greene (2012b) provides technical details for the estimation of all models detailed below.

3.5.4.1 Stochastic Frontier Models for Panel Data: Time-Invariant Inefficiency

Perhaps the simplest of the panel data stochastic frontiers are those that consider the firm-specific effect in panel data models as the firm's inefficiency (that is, the α_is in section 3.4.2). These models include a fixed effects model based on LSDV (Schmidt and Sickles, 1984) and random effects model based on GLS (Kumbhakar and Lovell, 2000) formulations. The following transformation is made on the firm effect to derive the measure of inefficiency,

$$\hat{u}_i = \alpha_i - \min_i\{\alpha_i\} \tag{3.36}$$

The advantage of the fixed effects formulation is that it does not impose the distributional assumptions of SFs more widely. The drawback of this approach is that the effect captures all firm-specific, time-invariant effects on costs which, although likely include inefficiency, may

also include a number of other features – anything that is not captured by the regressors. Further, this method neglects any variation in inefficiency over time.

In the random effects setting, the firm effects are assumed uncorrelated with the regressors. The central advantages of this approach are that estimation of the firm effects is more efficient than its fixed effects equivalent; and it allows for time-invariant regressors, which the fixed effects approach does not. Of course, the estimated of beta may be biased if the assumption of no correlation with the regressors is breached. As with its fixed effects counterpart, this method does not allow for temporal inefficiency change.

Finally, in the case where distributional assumptions are tenable and when there is little concern for correlation between inefficiency and noise, then a time-invariant SF model (estimated ML) is available (Pitt and Lee, 1981). The advantage is that in general, this method is more efficient than its alternatives, since it exploits the distributional information. The drawback is that assumptions need to be imposed for estimation. The equivalent to the Jondrow et al. (1982) prediction of firm inefficiency is,

$$
\begin{aligned}
&E\left[u_i|\varepsilon_{i,t=1}, \varepsilon_{i,t=2}, \dots, \varepsilon_{i,t=n}\right] \\
&= (1-\gamma_i).(-\bar{\varepsilon}_i) \\
&+ \sigma_u^2\gamma_i\left[\phi\frac{(1-\gamma_i).(-\bar{\varepsilon}_i)}{\sigma_u^2\gamma_i}\Big/\Phi\frac{(1-\gamma_i).(-\bar{\varepsilon}_i)}{\sigma_u^2\gamma_i}\right]
\end{aligned}
\tag{3.37}
$$

Where $\sigma, \lambda, \varepsilon_{it}, \phi \; and \; \Phi$ are as per section 3.5.3. In addition, $\gamma_i = \frac{1}{1+\lambda^2 T_i}$ and $\bar{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{it}$ (Greene, 2012b).

Empirical evidence suggests that, in general these three approaches produce similar results in terms of rank correlation (Kumbhakar and Lovell, 200 pp. 106-107 for discussion of studies which make comparisons). However, there is Monte Carlo evidence to suggest that, in cases where the technology that firms face is complex, the performance of these models deteriorates (Gong and Sickles, 1989). This point is particularly relevant in health markets which are characterised by vast heterogeneity. Therefore, we pay attention to this issue in our empirical work.

This general class of models is appropriate for short panels (that is, panels where the number of time periods is limited). As the length of the panel increases, the assumption of constant inefficiency becomes less plausible. Schmidt and Sickles (1984) suggest these models are appropriate for N>T. We turn to models that allow this assumption to be relaxed.

3.5.4.2 Stochastic Frontier Models for Panel Data: Time-Varying Inefficiency

The general panel data stochastic frontier with time varying inefficiency is,

$$\ln(c_{it}) = \alpha_0 + \ln(x_{it})'\beta + u_{it} + v_{it} \quad (3.38)$$

There are broadly two approaches to modelling the temporal evolution of firm inefficiency. These are, firstly, those that estimate a time-invariant component and apply to it a parameter estimated from changes over time. These models specify a deterministic relationship between time and inefficiency, indeed they have been termed time-dependent rather than time-varying (Greene, 2012b). These include the models of Kumbhakar (1990), Cornwell et al. (1990) and Battese and Coelli (1992) (amongst others, see Kumbhakar and Lovell, 2000 pp. 108-115 for details). The Battese and Coelli (1992) model specifies,

$$u_{it} = u_i . \exp\{-\eta(t - T)\} \quad (3.39)$$

Where u_i is the time-invariant, firm-specific inefficiency[21], *t* is the time period, *T* is the final year in the data and η is a parameter estimated from the data. Here, the sign on η determines whether firms are becoming more or less inefficient over time. Whether firms' inefficiency changes at all over time can be tested readily via a t-test on η[22]. Where $\eta = 0$, the model reduces to that of Pitt and Lee (1981). This model is referred to as BC92 hereafter.

This model has the advantage of being estimable by maximum likelihood, which is generally more efficiency that LSDV or GLS (as per alternative models such as Cornwell et al. (1990)) and has relatively few parameters relative to its alternatives.

In some cases, this formulation results in 'drop-off', which is when firms' inefficiency in some cases curiously drops off the frontier in their final year. The model can be amended to correct for this by using an estimated parameter, rather than T (Wheat and Smith, 2012). Thus,

$$u_{it} = u_i . \exp\{-\eta(t - \xi_i)\} \quad (3.34)$$

[21] As per the Pitt and Lee (1981) formulation from section 3.5.4.1
[22] $H_0: \eta = 0$

The model can be extended further in two important ways. First, as proposed by Battese and Coelli (1992), to allow (and test) for efficiency change to be non-monotonic, it is possible to specify a squared term as[23],

$$u_{it} = u_i . \exp\{-\eta_1(t-T) + (-)\eta_2(t-T)^2\} \quad (3.41)$$

Second, to address that the BC92 model imposes the rather restrictive assumptions that all firms' inefficiencies move in the same direction over time (including, by implication that rankings are constant), a generalisation can be made. The BC92 model has been extended by introducing individual time trends for each firm (Cuesta, 2000). The model with this feature has proved useful in the regulatory setting (Smith, 2012). The model takes the form,

$$u_{it} = u_i . \exp\{-\eta_i(t-T)\} \quad (3.42)$$

Where now there are individual time paths for firm-specific inefficiency, η_i. Of course, this is at the cost of having up to *i-1* additional parameters to estimate. Each can, accordingly, be tested. We utilise this model in our empirical work in subsequent chapters.

The second approach to capturing changes in inefficiency over time is one in which realisations of inefficiency are independent over time. Models in this category include the simple pooled panel model, the Battese and Coelli (1995) model for the truncated normal panel data SF model and the 'true' models proposed by Greene (2005). We present the 'true' models in subsequent subsections, and so do not present them here for brevity.

It is thought that the independence of inefficiency over time allows a more realistic reflection of firms' performance than an invariant component with a systematic time trend applied to it. This essentially implies that inefficiency in previous time periods has no bearing on inefficiency in current periods, which is questionable. Indeed, in regulatory settings, a model in which there is no link between firms' efficiency year-on-year is unattractive. On the other hand, whilst a smooth pattern of temporal evolution may be realistic, that same evolution of inefficiency over time may neglect any volatility in inefficiency change, which again may be questionable. Ultimately, the researcher makes choices as to their preferred assumptions about how inefficiency evolves over time. In our empirical work, our solution is to test a number of specifications with differing assumptions.

[23] The non-monotonicity derives from the squared term which allows a point of inflection in the time path of inefficiency. It would be possible to add higher order terms to allow more than a single point of inflection. This may be of use when T is large and the direction of inefficiency changes multiple times. Higher order terms can readily be tested.

Finally, we note that, in contrast to time-invariant inefficiency models that assume away time-varying inefficiency, this class of model cannot isolate any firm-specific time-invariant inefficiency. We therefore move to models that are able to capture both forms of inefficiency.

3.5.4.3 Stochastic Frontier Models for Panel Data: Short and Long Run Inefficiency

In both of the preceding approaches, make somewhat restrictive assumptions about the nature of inefficiency: either that it is entirely time-invariant or that it is entirely time-varying. Of course, in reality, there may be a component of both in the firm's total inefficiency. Models have been developed that allow these assumptions to be relaxed. They do so by allowing estimation of both time-invariant and time-varying inefficiency. The time-invariant component of inefficiency is referred to as 'long run inefficiency' and the time-varying inefficiency is known as 'short run inefficiency'. These models are of the general form,

$$\ln(c_{it}) = \alpha_0 + \ln(x_{it})'\beta + \gamma_i + u_{it} + v_{it} \qquad (3.43)$$

Where $\gamma_i$represents the firm-specific, time-invariant inefficiency and u_{it} is the firm-specific, time-varying inefficiency. This model can be estimated in a number of ways. Kumbhakar and Hjalmarsson (1995) take a multi-stage approach.

Despite the conceptual appeal of being able to capture short and long run inefficiency, these models, like the Schmidt and Sickles (1984), rely on the regressors to capture all heterogeneity. In the case of hospital production, this is highly implausible. Therefore, we extend our review of SF models to consider models that seek to capture unobserved heterogeneity.

3.5.4.4 Stochastic Frontier Models for Panel Data: Unobserved Heterogeneity

The conceptual appeal of making an allowance for unobservable heterogeneity is to allay concerns about differences in production environments between providers, across a number of dimensions, which are not captured by a set of explanatory regressors. Significant developments in the recent literature have been made regarding methods to control for unobservable heterogeneity.

In the frontier literature, much attention has been given to this topic, and a number of methods have been developed to accommodate unobservable heterogeneity. Approaches based on restrictions to the cost or production function have been applied in models across a

number of sectors, including health. Simply adding time-invariant dummy variables to account for unobserved characteristics is perhaps the simplest approach – in Chapter 5 (Buckell et al. (2015)) regional dummies are used as one (of a range) control for unobserved heterogeneity. In this case, the SF model (3.35) is extended to,

$$\ln(c_{it}) = \alpha_0 + \ln(x_{it})'\beta + \psi z_i + u_{it} + v_{it} \quad (3.44)$$

Where z_i are time-invariant variables capturing unobserved heterogeneity. This approach relies on the there being such variables available and that these variables are appropriate, that is, there is economic reasoning for the unobserved heterogeneity being captured in this way. A test on ψ can be readily conducted.

Another approach to capturing unobserved heterogeneity is based on Mundlak (1978) and decomposes firm-specific effects using group mean variables. The central assumption in this framework is that unobserved heterogeneity is correlated with regressors, which is disentangled by the use of the group mean variables. In random effects models, these variables are inserted directly into the equation. This method has been applied in health markets to nursing homes (Farsi et al., 2005a). The SF model (3.35), if estimated by GLS is extended as follows,

$$\ln(c_{it}) = \alpha_0 + \ln(x_{it})'\beta_a + \ln(\bar{x}_i)'\beta_b + u_{it} + v_{it} \quad (3.45)$$

Where $\bar{x}_i$ are group mean variables. This approach can be readily tested via a joint test on the group mean parameters, β_b.

Next, there are a set of models that can account for unobservable heterogeneity making use of firm-specific effects. These include, firstly, the "true" models (Greene, 2005). These models make the assumption that the firm effect captures the unobserved heterogeneity; inefficiency is then captured by applying the standard SF decomposition to the residual. That is, unobserved heterogeneity is assumed time-invariant, and inefficiency is assumed to vary over time. Thus, if all inefficiency is time-varying (that is, the firm effect is considered to comprise unobserved factors only), then these models are able to correctly identify firm inefficiency. In cases where there is time-invariant inefficiency, total inefficiency is underestimated. These models assume the general form,

$$\ln(c_{it}) = \alpha_0 + \ln(x_{it})'\beta + \omega_i + u_{it} + v_{it} \quad (3.46)$$

Where the firm effect, ω_i is considered to be unobserved heterogeneity. These models can be based on either fixed effects or random effects. The model in this form has a clear parallel with model (3.42). Indeed, the difference here is in the interpretation of the firm effect.

For fixed effects, the 'true fixed effects' (TFE) model is tantamount to the classic panel data SF with firm dummies (Greene, 2012b). The model is estimated via maximum likelihood, and as such is at risk from the incidental parameters problem, especially when the T, the number of periods, is small (Greene, 2012c). Two approaches to deal with this issue have been proposed. Firstly, by model transformation, using either within-transformation or first differences (Wang and Ho, 2010). Second, an approach based on deviations from the mean (Chen et al., 2014).

For random effects, the 'true random effects' (TRE), estimation is slightly more involved than its fixed effect namesake. The procedure makes use of maximum simulated likelihood (Greene, 2005) and results in the segregation of three model components: time-invariant unobserved heterogeneity, time-varying inefficiency and random statistical noise. In addition, this model, being based on random effects, has the advantage that it can be augmented with Mundlak approach described above, as per (Farsi et al., 2005a; Filippini and Greene, 2015).

Next are a set of models that are an extension of the 'true' models to allow for the separation of firm-specific unobserved heterogeneity and time-invariant inefficiency. These have been referred to as 'four component' or 'generalised true random effects' models. These models take the general form,

$$\ln(c_{it}) = \alpha_0 + \ln(x_{it})'\beta + \omega_i + \gamma_i + u_{it} + v_{it} \tag{3.47}$$

Where the model components are as per previous models, with the additional component, γ_i, which is the firm-specific, time-invariant inefficiency. ω_i captures unobserved heterogeneity. Here, overall cost efficiency is computed as the product of the short and long run inefficiencies, such that,

$$OCE = LRCE * SRCE = \exp(-\gamma_i).\exp(-u_{it}) \tag{3.48}$$

Where OCE is overall cost efficiency, LRCE is long run cost efficiency and SRCE is short run cost efficiency. The model can be estimated in a variety of ways, including the multi-stage residual decomposition approach of Kumbhakar et al. (2014), a single-stage approach based on Bayesian estimation (Tsionas and Kumbhakar, 2014) and a single-stage approach based on maximum likelihood (Columbi et al., 2014; Filippini and Greene, 2015). As with

the true random effects models, this model, being based on random effects, has the advantage that it can be augmented with the Mundlak approach described above.

It is useful to point out that unobserved heterogeneity may arise in a number of forms. Using the Kumbhakar et al. (2014) approach is able to disentangle unobserved heterogeneity that is uncorrelated with the regressors. Using the Mundlak approach, it is possible to remove unobserved heterogeneity that is correlated with the regressors. This means that it is useful to combine the two approaches. Indeed, each can be tested separately. This is the approach we have adopted in chapter 6, which corresponds to Smith et al. (2015).

Finally, there are approaches to capturing unobserved heterogeneity through differences in the parameters of the cost function variables. These include, firstly, parameter heterogeneity by group, or class, known as the latent class stochastic frontier (LCSF) (Orea and Kumbhakar, 2004; Besstremyannaya, 2011). Here, it is assumed that the firms are members of a finite number of groups, or classes, unobserved to the researcher. Estimation allows these groups to be identified, and individual firms assigned to classes. The LCSF is of the general form,

$$\ln(c_{it}) = \alpha_0 + \ln(x_{it})' \beta_j + u_{it|j} + v_{it|j} \tag{3.49}$$

In this setting, subscripts *i* and *t* denote firm and time as before. Subscript *j* denotes the class. Here, j classes have their own group-specific parameter estimates, β_j, and corresponding inefficiency and noise terms, $u_{it|j}$ and $v_{it|j}$, respectively. The number of classes is typically determined by the fit of the data, but it is possible to specify a predetermined number of classes (Greene, 2012b). Assignment of firms to classes is conducted probabilistically post-estimation. Inefficiency is relative to the group's own frontier. Estimation proceeds via maximum likelihood.

A related but distinct approach is the random parameters stochastic frontier model (RPSF) (Greene, 2005). In this model, heterogeneity is captured by allowing a firm-specific parameter to be estimated for each firm. This is, in essence, a fully generalised case of the latent class SF (that is, in the case that the number of classes is equal to the number of firms). The general form of the RPSF,

$$\ln(c_{it}) = \alpha_0 + \ln(x_{it})' \beta_i + u_{it} + v_{it} \tag{3.50}$$

Where β_i reflects the firm-specific estimates of the parameters. Estimation proceeds via maximum simulated likelihood.

Although these models can incorporate unobserved heterogeneity, they have some drawbacks. Firstly, they require large data since there are a large number of parameters to be estimated. Indeed, when T is small, especially in the case of the RPSF, estimation may be problematic. Second, as with some comparator models, these models are unable to identify time-invariant inefficiency. Thirdly, estimation may be prohibitively slow. Here, Halton sequences provide a solution to accelerating the estimation (Greene, 2005). These methods are not commonly used in empirical applications.

Before we turn to the estimation of vertically separate inefficiency, we summarise the models presented above. Table 3.2 below shows the model features and empirical specifications of the models presented above (we suppress the dummy variable and Mundlak approaches for brevity; other than additional variables, their specification is as per other models).

	REM	P&L	BC92	CUESTA	TRE	GTRE	LCSF	RPSF
Firm-specific component, α_i	$iid(0,\sigma_\alpha^2)$	$iid(0,\sigma_\alpha^2)$	$iid(0,\sigma_\alpha^2)$	$iid(0,\sigma_\alpha^2)$	$N(0,\sigma_\omega^2)$	$N(0,\sigma_\omega^2)$	$iid(0,\sigma_\alpha^2)$	$iid(0,\sigma_\alpha^2)$
Random Error, ε_i	$iid(0,\sigma_\varepsilon^2)$	$\varepsilon_{it} = u_{it} + v_{it}$	$\varepsilon_{it} = u_{it} + v_{it}$	$\varepsilon_{it} = u_{it} + v_{it}$	$\varepsilon_{it} = \omega_i + u_{it} + v_{it}$	$\varepsilon_{it} = \gamma_i + \omega_i + u_{it} + v_{it}$	$\varepsilon_{it\|j} = u_{it\|j} + v_{it\|j}$	$\varepsilon_{it} = u_{it} + v_{it}$
		$u_{it} \sim \lvert N(0,\sigma_u^2)\rvert$	$u_{it} \sim \lvert N(0,\sigma_u^2)\rvert$	$u_{it} \sim \lvert N(0,\sigma_u^2)\rvert$	$u_{it} \sim \lvert N(0,\sigma_u^2)\rvert$	$u_{it} \sim \lvert N(0,\sigma_u^2)\rvert$ $\gamma_i \sim \lvert N(0,\sigma_\gamma^2)\rvert$	$u_{it\|j} \sim \lvert N(0,\sigma_{u\|j}^2)\rvert$	$u_{it} \sim \lvert N(0,\sigma_u^2)\rvert$
		$v_{it} \sim N(0,\sigma_v^2)$	$v_{it} \sim N(0,\sigma_v^2)$	$v_{it} \sim N(0,\sigma_v^2)$	$v_{it} \sim N(0,\sigma_v^2)$ $\omega_i \sim N(0,\sigma_\omega^2)$	$v_{it} \sim N(0,\sigma_v^2)$ $\omega_i \sim N(0,\sigma_\omega^2)$	$v_{it\|j} \sim N(0,\sigma_{v\|j}^2)$	$v_{it} \sim N(0,\sigma_v^2)$
Time-Invariant Inefficiency Component	$\hat{\alpha}_i - min\{\hat{\alpha}_i\}$	$E[u_i \lvert u_{it} + v_{it}]$	N/A	N/A	N/A	$E[\gamma_i \lvert \ln(c_{it})]$	N/A	N/A
Time-Varying Inefficiency Component	N/A	N/A	$E[u_{it} \lvert u_{it} + v_{it}]$	$E[u_{it} \lvert u_{it} + v_{it}]$	$E[u_{it} \lvert \alpha_i + \varepsilon_{it}]$	$E[u_{it} \lvert \ln(c_{it})]$	$E\left[u_{it\|j} \middle\lvert u_{it\|j} + v_{it\|j}\right]$	$E[u_{it} \lvert u_{it} + v_{it}]$
Temporal Inefficiency	N/A	N/A	$u_{it} = \exp[\eta(t-T)].u_i$	$u_{it} = \exp[\eta_i(t-T)].u_i$	$u_{i,t} \perp u_{i,t-1}\ \forall\ t$	$u_{i,t} \perp u_{i,t-1}\ \forall\ t$	$u_{i,t} \perp u_{i,t-1}\ \forall\ t$	$u_{i,t} \perp u_{i,t-1}\ \forall\ t$
Unobserved Heterogeneity	N/A	N/A	N/A	N/A	ω_i	ω_i	$\alpha_{0\|j}, \beta_j$	$\alpha_{0\|i}, \beta_i$

Table 3.2: Empirical Specifications and Features of Stochastic Frontier Models. REM – random effects model (Kumbhakar and Lovell, 2000); P&L – Pitt and Lee (1981); BC92 – Battesse and Coelli (1992); CUESTA – Cuesta (2000); TRE – True Random Effects (Greene, 2005); GTRE – Generalised True Random Effects (Filippini and Greene, 2015); LCSF – Latent Class SF (Orea and Kumbhakar, 2004); RPSF – Random Parameters SF (Greene, 2005)

3.5.4.5 Stochastic Frontier Models for Panel Data: Multi-Level Organisational Structures

Many organisations are typified by hierarchical organisational structures, where upper tier entities have some degree of control over entities lower down in the hierarchy. Due to this, it may be the case that the upper tier has an effect on the overall inefficiency of the organisation. If this issue is overlooked, there may be several implications for analysis of inefficiency, for example, that inefficiency beyond the control of lower tier units is incorrectly apportioned to them. Worse, there may be distortions to overall estimates. There have recently been models developed that are able to incorporate the organisational structure, and decompose inefficiency variation vertically.

There are two basic approaches that have been applied to account for this. We consider the cost frontier for the purposes of illustration. The first is the two tier stochastic frontier (2TSF) first proposed by Polacheck and Yoon (1987). In this model, the effect on cost at the lower tier is, as per standard SF models, considered to be positive. Conversely, the effect at the upper tier of the organisation is negative; that is, the upper tier is assumed to be reducing inefficiency, rather contributing to it. The second approach is the dual-level stochastic frontier (DLSF) proposed by Smith and Wheat (2012). In this model, the effects on costs at both organisational levels are positive, that is, both upper and lower tiers are assumed to be contributing to the total inefficiency. We make use of the DLSF in our empirical work, we therefore present this model.

The DLSF model is derived from panel data stochastic frontier models, with the exception that the structure of the panel is amended from firm and time to firm and sub-company, where the sub-company units are repeat observations of their respective firms. In this way, the structure of the organisation is embodied in the model. This allows the decomposition of inefficiency at the two organisational levels in the hierarchy.

The imposed form of inefficiency is well suited to the multi-level model. As discussed above, in traditional panels, having an overall inefficiency comprising a component of upper tier inefficiency that is time-invariant and a lower tier component that varies randomly over time may not accurately capture the natural temporal evolution of inefficiency. In contrast, imposing an upper tier-invariant component and a lower tier-varying component to the structure of inefficiency (that allows for independence between observations) befits the aim of characterising the organisational structure.

The DLSF takes the general form,

$$\ln(c_{is}) = \alpha_0 + \ln(x_{is})'\beta + \gamma_i + u_{is} + v_{is} \quad (3.51)$$

Where c_{is} are the costs of lower tier unit, *s*, nested within its upper tier unit, *i*. γ_i is the component of upper tier inefficiency and u_{is} is lower tier unit-specific inefficiency. v_{is} is random noise, x_{is} are cost function variables and β is a vector of parameters to be estimated.

Estimation can be carried out broadly in two ways. The first follows the multi-stage approach of Kumbhakar and Hjalmarsson (1995). Here, the first stage is an upper tier-stratified within or GLS regression (depending on fixed or random effects, respectively), followed by a second stage stochastic frontier applied to the residuals, stratified by the lower tier, once purged of the upper tier effect.

The second approach is based of the 'true' formulations of Greene (2005), though our interpretation of this effect is upper tier inefficiency, rather than unobserved heterogeneity. For a true fixed effects approach, required is the insertion of upper tier dummies directly into the lower tier-stratified stochastic frontier. This has the clear advantage of being a single stage procedure. Potential drawbacks include that it may be difficult to identify upper tier-specific effect (if the dummy variables are not significant) and, as with all true fixed effects models, may suffer from the incidental parameters problem. There have been solutions found to this problem in standard models, see subsection 3.5.3.4.

The true random effects formulation is available as an alternative and a feasible single-stage approach. Its disadvantage is that it may be difficult to estimate in small samples, which has been found in the literature (Farsi et al., 2007).

Once predictions of inefficiency at two vertically distinct levels are retrieved, it is necessary to compute and overall measure, as below,

$$\bar{u}_i = \gamma_i + \frac{\sum_{s=1}^{S} c_{is}.u_{is}}{\sum_{s=1}^{S} c_{is}} \quad (3.52)$$

Where $\bar{u}_i$ is the measure of overall inefficiency across unit *i*. This is the sum of the upper tier-specific inefficiency and the cost weighted lower tier-specific inefficiencies (Smith and Wheat, 2012).

This model has significant potential in health markets where organisations often have some hierarchical structure. However, the model in this form is of limited use due to the issue of unobserved heterogeneity. Smith and Wheat (2012) assume the unobserved heterogeneity

away, which is untenable in health markets. Therefore, in the empirical work, we extend this model in a number of ways to account not only for unobserved heterogeneity, but to allow for it to enter the model in a number of forms. We apply a suitable testing procedure to identify the presence and/or form of the unobserved heterogeneity. For details, see chapter 6.

3.5.5 Summary: Stochastic Frontier Models

In seeking to measure inefficiency in NHS hospitals, we have chosen to use the stochastic frontier model. We have presented a number of stochastic frontier models which are able to reflect a number of features. The literature on stochastic frontiers is vast, we could not hope to provide a complete coverage. For this, the reader is referred to Kumbhakar and Lovell (2000), Greene (2012b) and Parmeter and Kumbhakar (2014). In light of this, we have focussed our discussion methods to capture three important issues that relate to health markets. These are the temporal variation of inefficiency, unobserved heterogeneity and multi-level organisational structures. These features form the basis of the empirical work in this thesis and, therefore, the contribution of this thesis. In subsequent chapters, we discuss why these features in particular are of relevance. Before doing so, we round off our methodological discussion by combining elements from the preceding discussion into an overall measure of performance: total factor productivity. We then complete this chapter with an overall view of efficiency measurement in health with a particular focus on the NHS.

3.6 Total Factor Productivity

We noted in section 2.7 that productivity comprises several separate features. We have shown that it is possible, using the cost function (or derivatives of cost functions, notably the stochastic frontier), to measure these features, namely economies of scale, technical change and inefficiency. Whilst these are individually doubtless important to policy makers, it may be of use to take an overall account of performance over time, by combining these features. This is achieved by making use of productivity, or its full economic title, Total Factor Productivity (TFP) (or Multi-Factor Productivity (MFP), in the absence of a full set of components). Moreover, given the overarching goal of cost reduction, savings in scale and through technical change may be as important as efficiency.

It is possible to observe Total Factor Productivity change, between subsequent years, defined by Coelli et al. (2005) pp. 300-306,

$$TFP\ change = technical\ change \times efficiency\ change \times scale\ change \quad (3.53)$$

These concepts are demonstrated below in figure 3.4, with the concepts drawn from prior subsections.

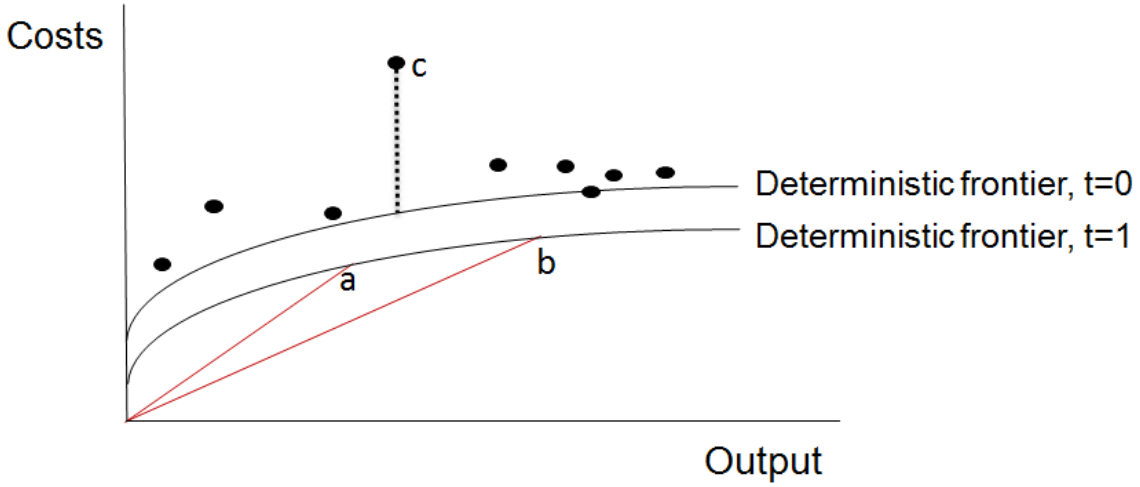

Figure 3.4: Total Factor Productivity Concepts

Technical change: technical change, also termed frontier shift, is the change in technology that firms face – the frontier itself - between time periods. On fig. 3.4, this is shown as the shift of the cost frontier from t=0 to t=1. In this case, technical change is positive, in the sense that costs are declining over time. This is called technical progress; the reverse is technical regress. Technical change is measured empirically by the time trend in the cost function (see section 3.3.3).

$$Technical\ change = exp\left\{\frac{1}{2}\left(\frac{\partial \ln(c_{i,t=0})}{\partial t = 0} + \frac{\partial \ln(c_{i,t=1})}{\partial t = 1}\right)\right\} \quad (3.54)$$

Efficiency change: whereas technical change concerns the frontier itself, efficiency change concerns the movement of firms either towards or away from the frontier. Then, efficiency change is the average change across firms between time periods. Point c on fig 3.4 shows that a firm is inefficient. If the position c is unchanged from t=0 to t=1, then the firm will be further from the frontier than in the previous period. If this is true across all firms, efficiency change will be negative. Of course, it is realistic to expect that some firms will gain whereas others will lose between periods. For this reason, models which allow this are preferred. Efficiency change is measured empirically as,

$$Efficiency\ change = \frac{\overline{CE}_{t=1}}{\overline{CE}_{t=0}} \tag{3.55}$$

Where $\overline{CE}$ is the sample average cost efficiency, see section 3.5.2.

Scale change: scale change pertains to how the (average) scale of production affects the average cost of production. On fig 3.4, rays from the origin reflect the average cost per unit of production (i.e. cost/output), thus the shallower the slope gradient, the lower the average cost. Therefore, at point b, the average cost per unit of output is lower than at point a. This is reflected in the shape of the frontier which shows increasing returns to scale. If, on average, firms increase their scale of production, scale change is positive in the sense that average costs are reduced. Scale change is calculated as,

$$Scale\ change = exp\left\{\frac{1}{2}\left(\varepsilon_{cy,t=0} + \varepsilon_{cy,t=1}\right)\right\} \tag{3.56}$$

Where ε_c is the elasticity of cost with respect to output, see section 3.3.2.

3.7 Efficiency Measurement in Health Care

In this section, we discuss the general health efficiency literature and, given that we have discussed the econometric approach based on econometric techniques, discuss specific applications of econometric techniques to NHS facilities. Whilst we adopt the stochastic frontier approach in our empirical work, we discuss other econometric approaches that have been applied to measure efficiency.

Hollingsworth and Street (2006) note that the supply side of the efficiency analysis market is booming. The supply of health care efficiency studies has grown substantially in recent years: between 1983 and 1987, the average annual number of health care efficiency studies was 1.6; whereas between 2002 and 2006 the annual average was 25 (Hollingsworth, 2008). A number of journal articles and text books review the application of efficiency analysis techniques to health care (Hollingsworth et al., 1999; Hollingsworth, 2003; Worthington, 2004; Jacobs et al., 2006; Hollingsworth, 2008; Hollingsworth and Peacock, 2008; Mutter et al., 2011).

Hollingsworth (2003; 2008) provides useful meta-analysis which is helpful in characterising the application of efficiency measurement to health care. As noted, the literature is large and growing. Non-parametric analysis dominates, with 80% of studies using DEA, Multiple

methods or Malmquist indices[24]. These are applied mainly to cross-sectional data sets, often with a low number of observations (<100). This may explain the low proportion of parametric studies, 18%, although the author notes that this proportion is growing. This growth may also be due to the development of methods to overcome some of the issues posed by users of efficiency analyses (cf. Hollingsworth, 2008; Mutter et al., 2011). This is one of the central themes of this thesis and we discuss these issues in detail in chapter 4.

In terms of areas of application, hospitals dominate, claiming over 50% of applications. The next two areas of application, namely nursing homes and physicians, account for under 10% of the applications each.

Hospital efficiency analyses are analysed by hospital type. Of these, public hospitals are the most efficient, with an average efficiency of 0.90. These are closely followed by Defence (military) hospitals, whose average efficiency is just under 0.90. For profit hospitals average around 0.86, whereas teaching hospitals average around 0.65 (although the sample size is small). Psychiatric hospitals appear to be least efficient, with an average of 0.60, but again numbers are low, with only two studies having been conducted.

Hospital applications are mainly in the USA, with approximately two thirds of studies here; European studies account for around a third. Of these, the average measured efficiency is 0.86 (0.84 for all hospitals), with a minimum of 0.72 (0.47). That is, European hospitals appear to be more efficient than their peers in the USA and around the world, according to the evidence presented.

Within these European studies are, of course, those conducted on NHS hospitals, which are of particular interest. Perhaps a slight limitation of this literature review is that is it conducted on studies until mid-2006. Therefore, table 3.3 below is a literature review table containing these studies and those in more recent years. We concentrate on econometric studies so as to be in keeping with the analysis in this thesis.

Table 3.3 shows that studies vary considerably across a number of dimensions. Methods vary from those based on cost/production functions using OLS to more complex analyses including SFA and multi-level models to policy evaluation techniques such as difference-in-differences regression. The unit of analysis across studies also varies, from those that conduct analysis at the whole hospital level down to analysis at the level of the individual procedure (e.g. hip replacement). Perhaps the area in which the studies vary most is sample size, where

[24] A form of productivity index; see Coelli et al. (2005) for exposition

the range across the studies varies by a factor of close to two million – from 31 to 54 million observations.

In light of this variation, a question arises as to why we adopt a SFA framework as opposed to other available methods. First, we note the link between the cost function methods used and the SFA literature. In several studies, a hospital-specific effect from a cost function is used as the measure of hospital efficiency (some authors use the term performance) (Laudicella et al., 2010; Daidone and Street, 2013; Gutacker et al., 2013; Moran and Jacobs, 2015). In this sense, they are akin to the panel data SF models identified in chapter 3.5.4.1 that measure time-invariant inefficiency (Schmidt and Sickles, 1984; Kumbhakar and Lovell, 2000). Indeed, in other applications of these models in health care, the direct link has been expounded (Sorensen et al., 2009).

There are three main reasons for which we adopt the SFA approach. First is the policy debate. We began by identifying spending pressures and the need not only to find efficiency savings, but to go beyond this in quantifying them. Some of the methods available in table 3.3 do not, in the form presented, allow explicit valuation of available monetary savings, e.g. difference-in-differences regression. This point extends to the analysis of Moran and Jacobs (2015) who, although identify variation in provider performance, do not provide results in monetary terms. For our purposes, we prefer a method which does.

Second, the nature of the data does not allow the application of these methods. In our data, we have neither changes in policy across observed units nor patient level data, meaning that alternative methods such as difference-in-differences or the 2 stage least squares approach are unavailable here.

Third, The SFA framework allows us to develop methods for examining important aspects of hospital performance that we have identified in this chapter. We begin by extending the existing work on temporal variation of inefficiency by introducing new flexible models to NHS hospitals that have previously not been employed (chapter 5). This, in turn, allows us to develop an econometric measure of multi-factor productivity (the full form of which is total factor productivity, as defined in section 3.6 above). This is lacking in table 3.3. Next, we introduce a measure of multi-level inefficiency to NHS hospitals. Whilst previous studies have adopted multi-level approaches (e.g. Gutacker et al., 2013), their efficiency measures remain limited to a single level. We develop models for the analysis of multi-level efficiency (chapter 6). Lastly, whilst some studies have used controls for unobserved heterogeneity, we

adopt a range of methods and testing procedures to look more closely at this issue (chapters 5 and 6).

Year	Authors	Methods	Unit of Analysis	Sample size	Years	Observations	Findings
1967	Feldstein	Production function	Whole hospital	177	1961	177	more and less efficient units identified; efficiency itself not quantified explicitly
1987	Wagstaff	SFA	Maternity hospitals	193	1971/72	193	hospitals are all technically efficient; caveats issued
1995	Scott and Parkin	Cost Function	Whole hospital	76	1992/93	76	statistical issues inhibit interpretation; no results presented
2001	Harper, Hauck and Street	COLS	General Surgery Specialities	31	1998/99 - 1999/00	62	no efficiency predictions reported; rank correlations only which were highly correlated between models
2002	Street and Jacobs	COLS; SFA	Whole hospital	217	1999	217	Average inefficiency: 0.74 COLS; 0.90-0.92 for SFA.
2003	Street	COLS; SFA	Whole hospital	236	1999	236	Average inefficiency: 0.69 COLS; 0.87-0.90 for SFA.
2006	Ferrari	SFA (distance function approach)	Whole hospital	52	1991/92 - 1996/97	312	productivity gain average 3% p.a.; no time-varying inefficiency; no efficiency estimates reported
2006	Jacobs, Smith and Street	COLS; SFA	Whole hospital	185	1994/95 -1997-98	740	mean efficiency across range: COLS 0.69; SFA cross section 0.87-0.90; panel 0.61-0.92
2009	Farrar, Yi, Sutton, Chalkley, Sussex and Scott	Difference-in-differences	Whole hospital	297 hospitals; 53,954,201 patients	2001/02 - 2005/06	53,954,201	hospital costs reduced following introduction of PbR; no observable effect on quality

2010	Laudicella, Olsen and Street	2SLS cost function	Obstetrics	136 hospitals; 952,273 patients	2005/06	952,273	more and less efficient units identified; efficiency itself not quantified explicitly
2012	Cooper, Gibbons, Jones and McGuire	Difference-in-differences	Individual procedure	161 hospitals; 1,882,750 patients	2002/03 - 2010/11	1,882,750	competition appeared to induce efficiency improvements
2013	Daidone and Street	2SLS cost function	Specialised Services	163 hospitals; 12,154,599 patients	2008/09	12,154,599	some variation in hospital efficiency but not reported
2013	Gutacker, Bojke, Daidone, Devlin, Parkin and Street	Multi-level cost function	Individual procedure	147 hospitals; 194,570 patients	2009/10	194,570	for hip replacement, 95% of providers within range -£2740 and +£3690 (mean = £6335)
2013	Siciliani, Sivey and Street	OLS	Hip replacement	193 hospitals; 42,948 patients	2006/07	42,948	treatment centres can provide more efficient care than hospitals due to specialisation; private providers have lower LOS
2015	Moran and Jacobs	Ordered Probit; linear model	Mental Health	58 providers; 185,281 patients	2011/12 -2012-13	342,288	Performance based on outcomes; variation of around 11% in ordered probit and around 2% in linear model attributable to providers

Table 3.3: Econometric Studies of NHS Hospital Efficiency

3.8 Summary

In this chapter, we have established the meaning of efficiency. We have then argued that frontier techniques, derived from economic theory, are our favoured method for capturing efficiency. We have set out what we mean specifically by the term cost efficiency, and detailed our methodological approach to measuring it. Further, we have discussed the aspects of the cost function, which are of use empirically and for policy purposes. Finally, we have detailed a procedure to reflect overall change in the sample over time, based on production theory.

With these economic tools at hand, we are now ready to proceed to our application: NHS hospitals. The following chapter, the first study of this thesis, is an overture to our empirical work. In this chapter, we set out our interest in hospitals and hospital efficiency; and the policy context, specifically with regard to costs and expenditure. Our starting point is that the duty to set hospital efficiency targets for NHS hospitals has shifted from the Department of Health to the economic regulator, Monitor.

We consider the various policy regimes that have been used for performance improvement and review evidence on how NHS hospitals have responded to them. Next, we draw on the extensive experience of economic regulators in Britain in measuring the efficiency of firms in their respective markets. We combine this with a review of the measurement of efficiency in healthcare markets – in particular hospitals in the NHS. We conclude by setting our research agenda for measuring inefficiency in NHS hospitals, which we pursue in the following empirical chapters.

4. National Health Service Performance Management, Price Regulation and Efficiency Measurement

4.1 Introduction

As noted in earlier chapters, following the introduction of the Health and Social Care Act (2012), the task of managing hospital efficiency has passed from the Department of Health to Monitor[25], which is the economic regulator of NHS hospitals that have achieved Foundation Status[26]. Since Monitor has assumed the role, it has begun to develop a more transparent approach to setting efficiency targets (known as the 'efficiency factor') than the Department of Health previously, based on benchmarking e.g. by comparing expenditure between services across hospitals (Monitor, 2013a). This is in keeping with the aims of central government who have identified benchmarking as key to making efficiency savings in the public sector (HM Treasury, 2015). Benchmarking is well developed, being used by other economic regulators in England across a range of industries (Crew and Parker, 2006). Monitor is interested in developing a more rigorous approach using economic techniques (Deloitte, 2014a; Deloitte, 2014b). With this, it aims to encourage hospitals to meet their efficiency targets and contribute to the top-level policy goal of plugging the oncoming funding gap.

The aim of this chapter is threefold: to inform the setting of efficiency targets for hospitals by reviewing incentive schemes applied to NHS hospitals; to inform the setting of efficiency targets for hospitals by reviewing the regulation of prices in other sectors of the economy; and to inform the setting of efficiency targets for hospitals by reviewing the measurement of efficiency in health markets and other sectors. These together are aimed at the top level cost-based policy goals. We do so by bringing together literature from NHS performance measurement, health-based efficiency measurement and regulatory economics.

[25] The responsibility is held jointly by Monitor and NHS England. For brevity, we use monitor throughout to denote Monitor and NHS England.

[26] Foundation status of a NHS trust (a trust is a hospital or small group of hospitals) means that it operates under an independent, not-for-profit regime, allowing it financial autonomy which it does not have without having foundation status (Marini et al., 2008). Trusts apply for foundation status, which is granted by the regulator, Monitor, if the trust has satisfied the regulator of its financial competence. Foundation status has not been awarded to all NHS trusts.

In section 4.2 we set out the hospital reimbursement scheme, discuss hospital efficiency since 2006, set out central issues and how we go on to answer them. In section 4.3 we discuss performance management schemes applied to NHS services since 1991. In section 4.4 we discuss the theory and practice of economic regulation and its implications for Monitor. In section 4.5 we consider the measurement of inefficiency, in the contexts of both economic regulation in other sectors and health care markets. Section 4.6 brings together the preceding 3 to draw out lessons for Monitor and section 4.7 concludes.

4.2 Hospital Price Setting and Efficiency Targets in the National Health Service

For NHS hospitals in England, under the National Tariff Payment System (NTPS) (formerly Payment by Results (PbR) under Department of Health – we consider these as interchangeable for the purposes of this discussion) activity-based reimbursement scheme, a "national tariff" (i.e. price) is set for each service provided (known as Healthcare Resource Group, HRG[27]) based on the national average cost for that service. These are termed Reference Costs. NTPS has been in operation since April 2013; PbR previously operated from 2003/04 (Audit Commission, 2004). Setting prices at average cost is a form of yardstick competition used to mechanise productive efficiency (Shleifer, 1985).

Under NTPS, the national tariff for each service provided is adjusted annually according to two factors. The first, to reflect inflation and other rises in the costs of service provision, is known as 'cost uplift', which raises the service price. The second, to encourage efficiency gains, an 'efficiency factor' which reduces the service price. The sum of these two factors determines the net annual adjustment applied to the price of each HRG.

Therefore, under NTPS, there are two basic mechanisms to encourage efficiency improvements: HRG prices set at average costs and the efficiency factor of the national tariff. This is shown in equation 4.1 below,

$$HRG\ price_{i,t} = \overline{HRG\ cost}_{i,t-1} + cost\ uplift_t - efficiency\ factor_{\ t} \qquad (4.1)$$

[27] Analogous to Diagnosis Related Groups, or DRGs, used in Europe and the USA

That is, the reimbursement for a given HRG, *i*, in year, *t*, is the sum of the preceding year's average cost (across all hospitals)[28], the year-specific cost uplift and the year-specific efficiency factor. The combination of the cost uplift and the efficiency factor is termed the net deflator[29]. In this study, we focus on setting the efficiency factor in our methodological discussion, but consider broader efficiency improvement throughout.

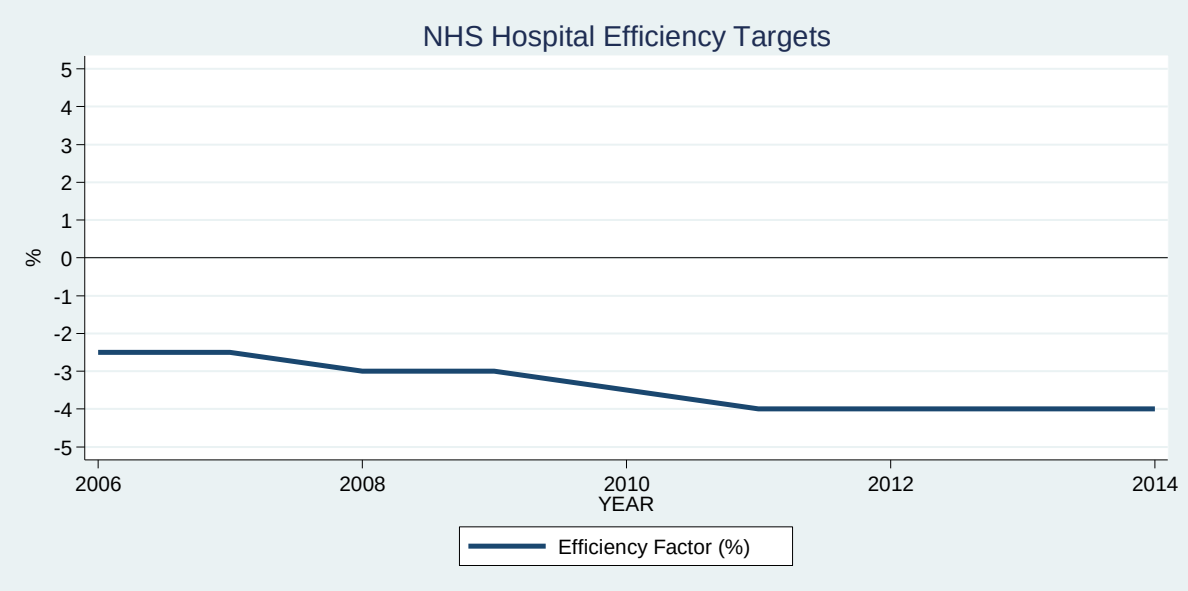

Figure 4.1: Efficiency Factor for NHS Hospitals, 2006/7 to 2014/15[30]. Sources: Department of Health Reference Costs Guidance and Monitor NTPS Guidance.

Figure 4.1 shows the efficiency factor for English NHS hospitals (Efficiency Factor)[31]. During this period, the Department of Health was responsible for setting the efficiency factor until 2014/15, when Monitor began to set the target. Hospitals have failed to reach these targets in recent years (Monitor, 2013b). Average targets were 3.8% between 2010 /11 and 2012/13, with savings, at best, 3.4% p.a. (Monitor, 2013b, pp.11). Therefore, despite the two mechanisms to encourage efficiency improvements, it does not appear that targets have been met. We seek to answer why this is the case.

[28] For ease of exposition; it may be the case that prior years' (t-2, t-3, etc.) are used, or that some average is used, depending on circumstances. For example, the 2014/15 HRG prices were based on 2010/11 reference costs because Monitor wished to maintain stability in its first setting of prices (Monitor, 2014).
[29] In practice, the net deflator is calculated as $(1 + cost\ uplift)(1 + efficiecny\ facotr) - 1$, but we retain equation (4) for ease of exposition
[30] For ease of reading, we have used just the first year of the financial year on the graph. 2008 refers to 2008/09, 2012 to 2012/13, etc.
[31] NB – the change in cost is compared to its target from the year at the start of the period, e.g. the 2008/09 target is compared to cost changes between 2008/09 and 2009/10

Historically, the efficiency factor has been decided in a number of ways, including funding disparities between NHS service commissioners' allocations and their expected expenditures (Deloitte, 2014a). Other targets were set as part of national government policy on efficiency gains in the public sector. For example, the 3% target in 2008/09 is derived from the NHS's settlement from central government's Comprehensive Spending Review (Department of Health, 2007). Thus the efficiency factors have been set based on central expenditure saving requirements by governments. This may be a reason contributing to hospitals' failure to meet targets. Another reason may be that these savings are unrealistic; NHS hospital managers suggested that around 1% annual savings were possible (Jacobs and Dawson, 2003). (We note that this survey is now over fifteen years old.) To these ends, Monitor is interested in a more data-driven approach to measuring efficiency and then setting the efficiency factor (Monitor, 2013).

Monitor's first year of NTPS was 2014/15 (Monitor, 2013c). Prices from the preceding year's tariff were kept for stability given that this was Monitor's first regulatory review period. The efficiency factor was 4%, drawn from indicator measures between hospitals, reports by consultants and past productivity gains (Deloitte, 2014b).

The current proposals for the 2015/16 NTPS have been referred to the competition and markets authority (CMA), given that 75% of providers had rejected the pricing proposal[32]. Although the efficiency factor has been reduced to 3.8%, the challenge remains. Monitor has this year adopted a more rigorous approach, making use of econometric benchmarking techniques and bottom-up hospital modelling (Deloitte, 2014b). This substantial opposition reflects objection to the methods used. We therefore look to motivate our empirical work by examining in detail the methods used in the determination of the efficiency factor.

Another important element to Monitor's role is controlling for quality, following from the Francis Report in which Monitor was thought to be too focussed on financial aspects of performance (Francis, 2013).

Thus, the central question for Monitor is how to ensure that these targets are appropriately set and then met. In this paper we bring together several strands of literature to inform two aspects of this question.

[32] Under NTPS, all providers are consulted on each year's pricing determination. If over 51% reject, the prices are referred to the CMA for arbitration.

(i) How to measure hospital efficiency whilst controlling for the quality of care; and

(ii) How to incentivise efficiency amongst NHS (i.e. publicly owned) hospitals

As a starting point, we review studies on the response of NHS hospitals[33] to various policy mechanisms that have been applied in the NHS's recent history. We describe the various policy regimes, and consider evidence on how hospitals have responded to them in order to inform Monitor's role of promoting efficiency amongst hospitals. We do not try to draw parallels with efficiency measurement; rather we look to observe the features of applying performance measures that are effective and those that are not. This is section 4.3.

We then turn to the theory of regulation and its implications for NTPS; see 4.4.

Next, we examine the practice of efficiency measurement in the joint contexts of health markets and across regulated industries. To the extent that there are some similarities between measuring efficiency in a cross-sectorial way, there are likely lessons for regulators in health markets. We further assess efficiency measurement in health and focus on methodological issues that have hindered the uptake of efficiency analyses amongst policy makers. We identify solutions to espoused issues based on recent developments in the literature. This is section 4.5.

Our main contribution is in the discussion: we seek to identify lessons for Monitor (and regulators in health more widely) for setting and enforcing efficiency targets for hospitals, both methodologically and practically, based on the recent developments in the literature. These lessons are drawn from the critical review of NHS performance management initiatives, from the regulation of other sectors that are subject to economic regulation; and from efficiency measurement in health markets and the wider literature. See section 4.6.

[33] Predominantly English hospitals, although in some cases English hospitals are compared to Scottish or Welsh hospitals; or indeed Scottish or Welsh hospitals in their own right. We are careful to note this throughout.

4.3 Performance Management in the National Health Service

We are concerned in this thesis with efficiency. Whilst we have identified the background to NHS hospital efficiency and its regulation in the previous section and we go on to its measurement in subsequent sections, it is important to consider other aspects of performance. There have been many policy schemes applied to NHS hospitals since the early 1990s. Insofar as hospitals have reacted to each of these schemes, there are likely general lessons to be drawn in observing hospitals' responses to them. We therefore examine literature that has assessed these schemes, looking for features of targets that are effective, and those that are less so. We then proceed to discussion of regulation, and its relevance to NTPS. We then consider measuring efficiency, both in the health and regulatory contexts.

A reason for missed efficiency targets may be hospitals' incentives. There have been a number of concurrent initiatives that successive governments have implemented to encourage performance improvements. These are broadly termed performance management (Smith et al. 2009). They have had varying degrees of success, and have been assigned varying levels of prioritisation (cf. Bevan, 2006; Oliver, 2009). Then, the consideration of the efficiency factor is more complicated than the efficiency factor; efficiency targets may have risen up/fallen down hospital mangers' agendas over time. In other words, given that there have been a number of other schemes that may have taken priority over the efficiency factor, managers' focus may have shifted from improving efficiency. In light of this, for enforcing the efficiency factor, there may be important hospital incentives found in by observing hospitals' reaction alternative policies, as well as potential pitfalls to be avoided. We therefore examine studies that have analysed these policy schemes.

Until the early 1990s, the belief was that system agents acted altruistically and that, to maximise performance, all that was needed was the proliferation of information – agents for whom performance was sub-optimal would aspire to improve. Agents are motivated via identification with the ideals and values of the system (Oliver, 2012). This system is referred to as 'trust and altruism' (T&A) (Bevan and Wilson, 2013). However, concerns arose around hospitals' efficiency towards the late 1980s, which undermined this belief, leading policy makers to reassess their approach (Maynard, 1991). Indeed, there are two major flaws with

T&A that act to erode incentives, namely, that rewards are maintained irrespective of performance and that failure is rewarded (Bevan, 2015).

Bevan and Wilson (2013) describe, as departures from T&A, three basic approaches to performance management that have, subsequent to the reforms in the early 1990s, been adopted by NHS policy makers. We review evidence to shed light on the above issues with a view to informing regulatory practice in the NHS context and beyond. Table 4.1 below gives an overview of the four basic regimes.

Scheme	Description	Economic Rationale	NHS examples
Trust & Altruism	Publicly-spirited individuals need only performance information as incentive	Identity (Oliver, 2012)	Until the 1991 reforms
Choice & Competition	Choice induces improved performance through competition	Invisible hand (Smith, 1776); contestable markets (Dranove, 2012).	Internal market 1991-1999; choose and book (2006); patients choose any provider (2008); gatekeeper-aided choice (2012)
Transparent Public Ranking	Publishing performance stimulates improvements amongst poor performers, "naming and shaming"	Loss aversion/Prospect Theory (Oliver 2012; Maynard, 2012; Bevan and Wilson, 2013)	Hospital star ratings 2000-2005; Surgeon league tables, 2013
Hierarchy and Targets	Setting targets with rewards for high performers and sanctions for missing targets, "targets and terror"	Loss aversion/Prospect Theory (Oliver 2012; Maynard, 2012; Bevan and Wilson, 2013); 'humans' and 'econs' (Bevan and Wilson, 2013)	Cancer waiting times; ambulance response times; accident and emergency 4-hour waiting times

Table 4.1: Policy Regimes for NHS Performance Management

A timeline of major policies is given below in table 4.2. For some schemes, e.g. waiting times policies, policies are aimed at specific services, meaning that reporting all of them in a table would be impossible. For a review of waiting times polices, see Smith and Sutton (2013). We discuss each of the approaches and review evidence in what follows.

Table 4.2: Overview of Policy Regimes Applied to NHS hospitals, 1991-2015. Sources: Oliver (2009); Cooper et al. (2010); Smith and Sutton (2013); Bevan and Wilson (2013). NB – this is meant to provide a general overview; the actual implementation of these regimes was more complex.

Notes: the NHS constitution set out a raft of waiting times polices, including those for ambulances, cancers, access to primary care, revascularisation and others, see Smith and Sutton (2013). RTT – referral to treatment, meaning the time taken from an initial GP consultation to seeing a specialist. C&C – choice and competition, TPR – transparent public ranking, H&T – hierarchy and targets, CCG – clinical commissioning group

4.3.1 Choice and Competition (C&C)

Choice and competition is based on the microeconomic doctrine of allowing market forces to engender efficiency. However, theoretically, the effects of competition will vary depending on the underlying market conditions (Draonve and Satterthwaite, 2000; Siciliani, 2005; Dawson et al., 2007). There needs to be sufficient competition in the marketplace for it to be preferable to monopoly (Brekke et al., 2008). Once providers begin to compete on quality and not prices, there does not appear to be a preferable theoretic market structure (Brekke et al., 2011). Allowing patient choice introduces contestability into hospital markets with, effectively, no sunk costs (hospitals already exist), implying that competitive conditions arise (Dranove, 2012).

C&C has been introduced in the NHS broadly in three waves (Propper, 2012). The first, during the 1990s, is referred to as the internal market. This introduced price competition through allocating funds to local providers[34] who, in turn, prospectively purchased annual bundles of services from hospital providers on behalf of the patients they served (Propper, 1996). The second was through a set of policy changes between 2002 and 2008. The thrust was to fix prices for reimbursement and allow patients both choice and information on providers. Thus, hospitals were to compete on quality rather than price. The third was implemented by the coalition government under the Health and Social Care Act (2012), which gave purchasing responsibility to GP consortia to procure hospital services on behalf of the patients they served. Between these periods, patient choice remained, but after 1997 when labour came to power, competition was actively discouraged by the government and fundholding was itself abolished in 1999 (Dusheiko et al., 2004; Propper et al., 2008).

A number of studies have reviewed these policy changes. Overall, there seem to be two emerging themes. First, under the internal market's price competition, whilst waiting times and prices seemed to be reduced, outcomes (in terms of mortality rates) declined (Dowling, 1997; Soderlund et al., 1997; Propper and Soderlund, 1998; Hamilton and Bramley-Harker, 1999; Propper et al., 2002; Dushieko et al., 2004; Propper et al., 2004; Ferrari, 2006; Propper et al., 2008). Second, under quality competition, outcomes, efficiency and waiting times seem to have improved (Dawson et al., 2007; Siciliani and Martin, 2007; Cooper et al., 2010; Bloom et al., 2010; Cooper et al., 2011; Cooper et al., 2012; Gaynor et al., 2013).

[34] either District Health Authorities (DHAs), who served larger patient populations (circa 315,000); or GP fund holder schemes (GPFH) who served smaller patient populations (circa 9000) and could select patients (usually by registers), i.e. could cream-skim.

Bevan and Skellern (2011) compare and contrast evidence on the two periods of provider competition. They attach a number of important caveats to the positive findings from studies of competition on quality. They note that metrics are focussed on single departments and therefore findings are difficult to generalise. In addition, they note that the potential for gaming[35] and for the diversion of resources from less visible areas of hospital activity may weaken the findings. This is reflected in the findings of Propper et al. (2006), who note that, although there is some evidence of improvement in outcomes due to competition, there is insufficient theoretical and empirical support to make its case conclusively. Thus the overall effect of competition and choice remains ambiguous.

4.3.2 Transparent Public Ranking (TPR)

TPR is a system under which entities' (which could be hospitals, departments or, as recently introduced, individual surgeons) performance is judged in some way and the entities are ranked. These ranks are then publicly disseminated. The intuition follows from the human condition of loss aversion, that is, humans naturally respond to loss more than gain. Being close to or bottom in rank carries a sense of shame and disappointment. Thus there is motivation in avoiding being ranked last. That said, there is also clear motivation derived from the prestige associated with being ranked highly. Thus, the system can be characterised as one in which victors are rewarded and failures are penalised: 'knights and knaves' (Bevan, 2010). Proponents of TPR argue its justification on this idea, but also that it is in the public interest to be transparent and that we are, collectively, better able to identify substandard or harmful practice when results are in the public domain (Iacobucci, 2012). Critics argue that TPR is inaccurate and may apportion blame incorrectly, that it instils anxiety amongst the workforce and necessarily casts entities as 'poor', which may be undue (Adab et al., 2002; Westaby, 2014).

Under New Labour, TPR was introduced in the form of hospital 'star ratings' in 2000 (Bevan, 2010). Hospitals were judged in three domains of performance, and awarded a star if each were met. An overall score along a scale of 0-3 stars was then awarded. Those for which 0 stars were awarded came under fire, the 12 hospitals with 0 rating in year one were labelled the 'dirty dozen' (Bevan and Wilson, 2013).

[35] An example of gaming is given in Bevan and Hood (2006). After the implementation of 8-minute ambulance response times targets for critical calls was implemented, data show a significant spike 8 minutes, implying that many 8 and 9 minute responses had been recoded as just under 8 minutes in order to meet the target.

The evidence suggests that the 'naming and shaming' of hospitals has been effective in stimulating performance improvement. The star ratings system was reported by NHS trusts themselves as helpful in transmitting priorities from central government, to help modernise practice amongst trusts and to expose substandard management (Mannion et al., 2005). However, the same study reported evidence of substantial gaming, unintended consequences (such as disincentives to invisible services, myopia, tunnel vision, bullying) and erosion of public trust; although these issues seemed to arise due to the measures used for analysis (mainly waiting lists and performance indicators, i.e. incomplete measures) rather than the TPR system itself. Besley et al. (2009) find that TPR successfully lowered waiting times for English hospitals relative to Welsh hospitals which maintained a T&A-type policy regime. Bevan (2010) reached a similar conclusion, but emphasised a number of conditions that were satisfied to enable success. These were that the ratings were given clear priority, were communicated clearly and applied with consistency (pp. 48). Crucially, relative to the Welsh system, there were sanctions applied for failure. Bevan and Wilson (2013) conclude that TPR is more effective than T&A. However, they note that there was evidence of gaming and that there were a number of policy models simultaneously in place.

4.3.3 Hierarchy and Targets (H&T)

Under H&T, a target is set for hospitals (e.g. four hour accident and emergency maximum waiting times). High performers are rewarded; those that fail to reach the targets face sanction. As with TPR, there is a loss aversion rationale here. Further, Bevan and Wilson (2013) propose a 'humans' and 'econs' argument[36] in that providers, characterised as 'econs' as opposed to 'humans', respond equally to potential gains as to potential losses. This implies hospitals should be motivated in seeking to avoid the sanctions and equally be motivated to reach the targets. In the NHS, the implementation of H&T in the 2000s was referred to as 'targets and terror'. Waiting times have been the primary focus of the application of H&T. As an example of current policy, missing the 18 week referral to treatment (RTT, see table 4.2) target results in a reduction of up to 5% of revenue, depending on the speciality, for the month of the breach (Smith and Sutton, 2013).

There are a number of studies which report the basic finding that, subsequent to the implementation of targets, performance against these targets improved (Alvarez-Rosete et al.,

[36] Thaler and Sunstein (2008)

2005; Bevan and Hood, 2006; Hauck and Street, 2007; Dimakou et al., 2009; Harrison and Appleby, 2009; Bevan and Hamblin, 2009; Propper et al., 2010; Marques et al., 2014). Further, evidence suggests that the removal of sanctions for failing to reach targets weakens performance (Smith and Sutton, 2013).

There have, however, been a number of criticisms aimed at these measures. Bevan and Hood (2006) and Bevan (2006) describe three assumptions that underpin H&T which have been breached in application to NHS performance management. These are that the target-setting authority is able to determine a system which prioritises what matters; that failure to meet features not built into performance measures are irrelevant (i.e. unintended consequences); and that the advantages of any outcomes offset any gaming activities. Indeed, Bevan and Hood (2006) find significant evidence of gaming. Bevan and Hamblin (2009) outline further issues with performance targets: There is an issue of selecting the correct or appropriate measures; the issue that indicators give an incomplete picture of the production process; there is information loss when aggregating; and that public targets risk workforce morale. Harrison and Appleby (2009) also note that targets were set against a backdrop of substantial funding increases, thus the attribution of success to targets in not entirely straightforward. Marquez et al. (2014) found substantial gaming. Further, they found no evidence of H&T-induced efficiency gains.

Contrary to much criticism, there have been some positive aspects of these policies. Harrison and Appleby (2009) suggest that the sustained pressure on hospitals to improve performance has led to the avoidance of 'quick-fix' solutions and the development of longer term management strategies to meet the required gains. Oliver (2009) takes a longer term view, showing that the maximum inpatient waiting time target has reduced from 15 months (2002/03) to 3 months (2008/09) and is projected to be cut to 2 weeks (2022/23). Propper et al. (2010) take the general view that the targets have achieved their objectives. Crucially, they did not find substantive evidence of gaming, nor did they find any decrease in the quality of care[37].

In this section, we have reviewed evidence on hospitals' responses to several performance management regimes. There is, to our knowledge, no literature that assesses how NHS hospitals have responded to specific efficiency targets that have been set in recent years. However, given the links between the measures we have reviewed in this chapter and hospital

[37] They did find evidence of waiting list manipulation, but did not class this as gaming.

efficiency, there are likely generalizable features of applying targets that can be used in the context of setting the efficiency factor. There are some key themes that emerge from the literature. Firstly, hospitals have responded well under the following conditions, as shown in box 4.1 below.

(i)	Performance analysis that is widely disseminated;
(ii)	Targets that are easily understood;
(iii)	Targets that are prioritised;
(iv)	Targets that are clearly communicated;
(v)	Targets applied consistently between providers;
(vi)	Sanctions are applied for failure;
(vii)	Underlying increases in funding;
(viii)	The sustained application of measures (avoiding “quick-fixes”);
(ix)	Specific (disaggregate) targets; and
(x)	Measures that have staff engagement.

Box 4.1: Features of Targets Associated with Favourable Responses

There have also been a number of issues identified with certain performance improvement schemes. These are summarised in box 4.2 below.

(i)	Gaming, e.g. misreporting;
(ii)	Flaws in the target itself[38];
(iii)	The potential for information loss when aggregating measures;
(iv)	Diversion of resources from less 'visible' services (for indicator measures);
(v)	Disincentives to fund/promote quality amongst less 'visible' services (for indicator measures);
(vi)	Potentially damaging to morale;
(vii)	Tunnel vision – focusing on an indicator only;
(viii)	Myopia – short-term responses instead of measured, long term strategies;
(ix)	Bullying;
(x)	Erosion of public trust; and
(xi)	Multiple concurrent polices mean ascribing improvement to a single policy may be challenging.

Box 4.2: Issues Encountered with Performance Management Schemes.

We return to discuss the implications for Monitor in section 4.6.

4.4 Price-Cap Regulation and the National Tariff Payment System

4.4.1 Idiosyncrasies in Health Care Markets

It is often suggested that health care markets differ from other markets. The underlying differences stem from uncertainty (Arrow, 1963). This uncertainty lies on both the demand and supply side, and surrounds the incidence of disease and the efficacy of treatment.

In respect of regulating, our focus is on the implications for firm behaviour. The thrust of Arrow's argument is that doctors do not behave as firms do in other markets. There is far less by way of competition; recommended courses of treatment should be based on need rather than profitability; and providers' goals are more diverse than pure profit maximisation (Morris et al., 2007). At a more macro level, Mooney (1992) identifies a range of goals of health care systems: technological innovation, equity (both horizontal and vertical), effectiveness, efficiency, professional status, patients' rights, clinicians' rights, communities'

[38] an 8 minute ambulance response time means that, in terms of the target, to reach a patient in 7 minutes who dies is better than to get to a patient in 9 minutes that lives.

preferences and medical ethics. For these reasons, it is argued some distinction should be made in the analysis of health care markets.

Morris et al. (2007) make the point that whilst these issues hold for health care markets, they may also hold true for others. In fact, many of the aforementioned list may be argued as pursuits for regulators in other sectors, technological advance being an obvious example. To this extent, there may be some parallels between health care markets and other regulated industries. However, the authors also observe that across many of these features, health is an extreme case, which is not necessarily true for all other industries.

We make this link in the following section when discussing the mechanics of pricing structures in regulated industries. Later in this chapter, we highlight these differences when discussing measuring efficiency.

When budgets are limited, the concept of scarcity implies that it is necessary to assess cost effectiveness of health care interventions as a means of allocating resources optimally (Olsen, 2009). Whilst this is not an issue specific to health care, an immediate issue arises at this juncture in health: for optimal allocation, there must be a measure of value. Individuals' concept of value is distorted by many features in health care, e.g. monopolistic powers of firms, poor information, agents' conflicts of interest, inter alia (McCabe et al., 2014). A consequence of this is that, in many analysis of cost effectiveness, value is quantified as 'health gains', typically assessed via use of the Quality-Adjusted Life-Year (QALY) or Disability-Adjusted Life-Year (DALY). These measures comprise both measures of (changes in) health status, assessed using instruments such as the EQ-5D[39], and the changes in length of life. Here, both the value (private or social) and the opportunity cost (gains of patients who do not receive treatment when another patient does) are expressed through the provider's willingness to pay (McCabe et al., 2014).

With these at hand, judgements can be made as to which treatments yield highest gains, can be used for health technology assessment, i.e. whether new treatments are preferable to existing technologies. Using QALYs and DALYs, a new technology's value can be assessed, or, as McCabe et al. (2014), pp.4, note,

"the value of the health displaced as the use of other technologies has to fall."

The concept of value in health care extends further to the measurement of output, which is directly relevant for the purposes of this thesis. Whilst in many regulated industries,

[39] The EuroQol 5-Dimension measure, which is a self-reported measure of health status across 5 dimensions

measuring output can be readily recorded, kilowatts of electricity, for example, this is not the case in health care. Whilst the volume of patients can be used as a health metric, there are aspects of the service that are neglected in so doing: health care is more than treating patients, the efficacy – or quality - of treatment (and thus its value) is critical. Many measures have been developed to embed service quality into output, or to measure it directly. We return to discuss this issue explicitly in following sections of this chapter.

Before doing so, we turn to the pricing mechanism employed under NTPS, and its relation to comparable systems in other sectors.

4.4.2 PCR and NTPS

Regulatory need derives from market failure in the form of natural monopoly (as in, for example, the electricity transmission, overseen by Ofgem) or oligopolistic competition, where large firms are in a position to (potentially) exploit market power (as in airports, regulated by the CAA). Market failures of this nature have been observed in health markets (Dranove, 2012). In utilities markets, these failures have led regulators to develop strategies to mitigate their negative effects (Viscusi et al., 2005). One of these strategies is the control of prices, termed price-cap regulation (PCR; see Train (1991) for an exposition). We consider some fundamental differences between health economics and other industries below, before turning to PCR and its application to health care markets.

Under NTPS (and PbR before it), there are two basic mechanisms to encourage efficiency improvements, namely average cost based HRG prices (enabling yardstick competition) and the efficiency factor of the national tariff (see equation 4.1). These are both considered "high-powered" efficiency-inducing mechanisms (Shleifer, 1985; Viscusi et al., 2005). However, in practice, PbR looks to have moved costs toward their average rather than to have applied downward pressure (Maynard, 2012).

Insofar as there are fixed prices and efficiency targets to incentivise efficiency improvements, PCR, as used by non-health regulators in Britain and internationally, has clear parallels with NTPS. PCR has been effective in reducing costs in regulated industries (Baldwin and Cave, 1999; Viscusi et al., 2005; Hauge and Sappington, 2010; Sappington and Weisman, 2010). Distinguishing features of PCR include the length of control period, termed regulatory lag, and the review date being set and non-negotiable. When these conditions are breached, regulatory lag is endogenous (either party can request a review), and, importantly, efficiency

incentives are significantly weakened; the system becomes comparable to Rate-of-Return (ROR) regulation (cf. Armstrong et al., 1994; Viscusi et al., 2005).

In general, setting the length of regulatory lag can be problematic for regulators, and there does not appear to be an optimal length (Liston, 1993). Regulators in non-health industries in Britain have commonly opted for a five year lag. In Monitor's case, the lag is a single year, meaning, theoretically, efficiency incentives are eroded. Therefore, lengthening the lag is likely to induce efficiency improvements.

Further advantages of longer control periods include that the regulator has ample time to consult with firms and the public to deliver comprehensive performance reviews over the price control periods[40]. It would also be possible to build specific policy objectives into the regulatory structure, such as the NHS's *five year forward view* (NHS England, 2014a). Indeed, efficiency targets could be aligned to any policy target with a simple calculation, by deriving the savings requirement in secondary care from the total (based on the proportion of total expenditure on secondary care, for example), and setting the efficiency factor over the period to align with the requirement.

One further consideration is that NHS England or the Department of Health may have minimum (or specific) service requirements for its budget. Here, Monitor may draw on the experience of the Office of Rail Regulation (ORR), which has developed a tripartite price-cap based model of regulation that incorporates policy directly, as follows.

In this setting there are three bodies: the ORR (regulator), Department for Transport (DfT; the government) and Network Rail (owner/operator of rail infrastructure; public body). Here, the DfT sets out its requirements (e.g. improvement in trains running on time) and the amount of money it is willing to pay (has at its disposal). Then, Network Rail responds with a counter offer. Lastly, ORR defines a specification at an agreed cost by reconciling the two sides' submissions – this is termed the 'final determination'. At this point, prices are effectively fixed and Network Rail can increase profits through cost reductions over the period. This has been effective in both reducing costs and improving standards, e.g. operating costs have fallen 40% since 2004 (ORR, 2013). This system would be implementable in the health setting, with Department of Health specifications answered by trust estimations and central arbitration by Monitor. This system is important in the context of the NHS as follows.

[40] see Parker et al., 2006, pp. 124-133, for a history of benchmarking for price control across gas, electricity, water and telecommunications industries in Britain.

Historically, under public ownership, firms often lack adequate funding for capital investment. Privatisation coupled with independent economic regulation is a means to rectify this issue. In the case of rail, the issue is further complicated by the need for additional subsidy. This mechanism, in turn, ensures Network Rail has enough by way of funding for necessary levels of service and quality – assuming it is efficient. The key here is the regulator, as arbiter, prevents the political tussle between parties with competing interests. This has potential in health, particularly at present when trusts' budgets are highly pressured whilst Monitor's NTPS has been rejected and referred to the CMA. So, if trusts are struggling for funding, Monitor can call on the government to increase levels of funding. Equally, if Monitor's decree is that hospitals are inefficient, it is difficult for them to seek additional funding.

Economic theory predicts disincentives to quality under fixed prices (Spence, 1975)[41]. To counter this effect, quality standards must be upheld (Laffont and Tirole, 2000). This may be difficult due to measurement, ascribing costs/benefits or ascribing responsibility for quality (Sappington and Wiseman, 2010)[42].

In the context of the NHS, the Care Quality Commission (CQC) is responsible for controlling service quality. However, this was the case in the period 2005-2008 during which failures in service quality occurred (Francis, 2013). Therefore, Monitor need to ensure quality is maintained. Quality can be incorporated directly into efficiency analysis (see following section for details) as one solution. Another solution would be to have off-model quality control, perhaps with targets and sanctions for target failure. Quality requirements can readily be embedded into the tripartite approach (or derivative system) the ORR has implemented, where quality specifications are built in to HRG price determinations.

Overall, an adjustment to the current pricing mechanism under NTPS which draws on PCR theory and practice is likely to have significant benefits to the status quo, both in terms of incentives and the pursuit of policy goals.

4.5 Efficiency Measurement for National Health Service Price-Capping

[41] For balance, we note that this is also true of ROR. This is because the firm bears all costs of quality improvement, but has to share the rewards of the improvements (Sappington and Weisman, 2010)
[42] See section 5 for a discussion of quality in health

The final objective of this chapter is to examine the measurement of efficiency in health markets and regulated industries. We elaborate on the basic tools described in chapter 3 by identifying health-based issues that have arisen in the academic literature. These have been raised in the NHS context and beyond. We further complement this discussion with reference to wider regulatory and methodological literature, where appropriate.

Efficiency measures are commonly, but not always, frontier-based, following Farrell (1957). The use of frontier-type techniques seems to have gained primacy amongst academics in health markets (Hollingsworth et al., 1999; Hollingsworth, 2003, 2008; Hollingsworth and Peacock, 2009; Rosko and Mutter, 2011; Mutter et al., 2011). There is significant potential for frontier-type efficiency measures in health (Lovell, 2006; Mutter et al., 2011). However, for a number of reasons, policy makers have, hitherto, found the results of these studies of limited use (Hollingsworth and Street, 2006; Hollingsworth, 2008; 2012). There have been developments in the literature – notably in econometrics - to these issues which may mean these techniques are of use to managers and policy makers, and specifically to Monitor.

We further justify this framework by making reference to the 'best practice' criteria for judging regulators' approaches to benchmarking developed in electricity distribution (Haney and Pollitt, 2009; 2011). We extend these criteria for health by including the health-based issues (and/or nuanced issues from regulated industries) that have been raised specifically in health markets. Table 4.3 shows the issues that have been encountered by analysts in health and non-health settings. We use this as the basis for the following discussion.

For 2015/16 tariff, the efficiency factor was initially set at 3.8% which is based partly on benchmarking of hospitals (Deloitte, 2014b). Following a rejection from 75% of providers, there has been a referral to the CMA. Therefore, a question arises as to the methods used to calculate the efficiency factor. These are presented in Deloitte (2014b). In our discussion, we set out the methods used for benchmarking and contrast them with our review in this section. From this, we suggest how benchmarking could be improved. We then proceed to demonstrate a number of these improvements in our empirical chapters that follow.

4.5.1 Large Data

A low number of observations in data sets has been prohibitive for some regulators wishing to make use of a full set of analytical methods. In the case of NHS hospitals, sample size issues are unlikely to pose problems given that there are vast data collections available to the

regulator, namely Hospital Episode Statistics (HES) and Reference Costs (RC). In regulation, sample sizes are typically very small; see Appendix B in which all have fewer than 100 observations save for an international comparison of 560. In contrast, Monitor used a sample of 750 observations (Deloitte, 2014b). If patient level data were used, there are 18.2 million admitted patient records in HES in 2013/14 alone (HSCIC, 2015). Moreover, these data can be mapped to national datasets such collections held by the Office for National Statistics (ONS). Thus in terms of size and potential information, Monitor are well equipped.

4.5.2 Data Quality

The quality of data may present a number of issues. Data may be collected in an inconsistent manner in both cross-sectional and time-series dimensions. This issue can stem from a number of factors, including, inter alia, firms' interpretation of data guidance, changing definitions over time, gaming, allocation of capital costs, firms' ability to learn how to record data (and regulators' engagement with them), major shocks to the firm (e.g. board changes), the value placed in the data by the firm and the level of aggregation. These are germane issues in the health context: Updates to RC mean that comparisons over time are difficult – accounting regulation is set to fundamentally change costing from procedure-based to patient level costing (Mason et al., 2011); gaming is well known in health (Bevan and Hamblin, 2009); and there is evidence to suggest variation in clinical coding[43] (Joy et al., 2008; Pett & Clark, 2012).

There have been initiatives to encourage higher quality data recording such as Ofgem's Information Quality Incentive (IQI) which rewards firms for high-quality data reporting (Ofgem, 2011). In addition, Ofgem use a 'fast-track' scheme whereby firms that produce business plans that are deemed of sufficient quality are agreed without further scrutiny (Ofgem, 2011).

A major issue for regulators is the consistency of data over time. For example, OFWAT, although had access to panel data in the late 1990s, used only cross-sectional analyses (Parker et al., 2006). A solution to time series data inconsistency is to use a hierarchical approach (e.g., Smith and Wheat, 2012; Gutacker et al., 2013; Smith et al., 2015), which is based on panel data methods, but does not require data collected over time. If the data is hierarchical in structure, then observations of units in the lower tier can be used as repeat observations of the

[43] The general problem of up-coding, i.e. artificially recording a higher severity of a patient's condition, is well known in health (Newhouse, 1996)

unit at the upper tier, thus a panel structure is contained in the data. Here, the benefits of panel data can be realised without the need to collect data over a period of years. Therefore, concerns about data consistency over time evaporate. This approach has also been referred to as the dual-level efficiency model (Smith and Wheat, 2012; Smith et al., 2015).

Issue	Health	Non-health	Solutions available
Large Data	Jacobs et al., 2006	ORR, 2008; Ofgem, 2011;Haney & Pollitt, 2009, 2011	Collect large data; dual-level efficiency; International comparison
Data Quality	Scott & Parkin, 1995; Joy et al., 2008; Pett & Clark, 2012	Ofgem, 2011; Haney & Pollitt, 2009, 2011	Dual-level efficiency; Data quality incentives
Allocating capital costs	Drummond et al., 2005; Dranove, 2012; Buckell et al., 2015	Parker et al., 2006; CEPA, 2014	OPEX/CAPEX modelling; Smoothing; Estimation of capital costs
Engagement with industry	Hollingsworth, 2008; Smith, 2015	CEPA, 2014	Consultation with end users
Range of methods	Jacobs et al., 2006	Haney & Pollitt, 2009, 2011	Triangulation; Model selection; Cross-checking
Panel data and temporal efficiency	Jacobs et al., 2006; Hollingsworth, 2008	Weyman-Jones et al., 2006; Haney & Pollitt, 2009, 2011	Dual-level efficiency; Cross-sectional analysis; Model temporal change
Heterogeneity: Organisational	Dormont & Milcent, 2004; Hollingsworth, 2008	Arocena et al. (2012); Haney & Pollitt, 2009, 2011	Disaggregate level of analysis; Hierarchical modelling; Data to capture heterogeneity
Heterogeneity: Patient level	Dormont & Milcent, 2004; Iezzoni, 2009		Data on patient characteristics: age, gender, ethnicity, deprivation, etc.
Heterogeneity: Quality/Outcomes	Hollingsworth, 2008; Smith & Street, 2013	Parker et al., 2006; Haney & Pollitt, 2009, 2011	Data on outcomes, waiting times, readmissions, cleanliness, etc.
Unobservable Heterogeneity	Greene, 2004; Farsi et al, 2005a		Statistical approaches for unobservable heterogeneity
Uncertainty and sensitivity	Newhouse, 1994; Street, 2003; Hollingsworth, 2008	Weyman-Jones et al., 2006; Haney & Pollitt, 2009, 2011	Statistical testing/specification; Efficiency prediction by interval; Distribution-free approaches

Table 4.3: Issues for Measuring Efficiency in Health and for Economic Regulators

4.5.3 Allocating Capital Costs

The allocation of capital costs is problematic for regulators. Ideally, capital costs are contained in firm cost data and regressions run on total expenditure, i.e. TOTEX modelling. This has been carried out by several regulators (Parker et al., 2006; NERA, 2008; CEPA, 2014). In many cases, however, TOTEX modelling is inappropriate. Firstly, capital expenditure is often significant and sporadic: 'lumpy'. This implies that there may be undue cost variation that is cast as inefficiency in TOTEX models for firms that invest heavily (Rossi and Ruzzier, 2000). An additional issue is how shared capital costs are allocated between services, which occurs in health markets (Drummond et al., 2005; Dranove, 2012).

One solution is to smooth capital costs over time, although how to do this exactly poses issues (CEPA, 2014). Another is to model operating costs and capital costs separately, OPEX and CAPEX modelling. This, however, requires data to be available for both costs. As with TOTEX modelling, this has been carried out by several regulators (Parker et al., 2006; NERA, 2008; CEPA, 2014). OPEX-type patient level costing is being introduced amongst NHS hospitals, and may serve as a solution to the issue of CAPEX allocation (Monitor, 2015). It is, of course, important to model both of these features to prevent gaming – focussing solely on OPEX modelling may encourage firms to 'dump' costs in CAPEX. It is for this reason that Ofwat used a TOTEX approach (CEPA, 2014). The final drawback with the separate approach is that modelling OPEX restricts conclusions on the scale properties of production and on Total Factor Productivity, which has been identified in the NHS context (Buckell et al., 2015).

4.5.4 Engagement with Industry

A key feature of the benchmarking process for regulators is engaging with the industry. For example, Ofwat's approach involved consultations with its own engineers, company board directors and the industry research body, UK water industry research (UKWIR) on models (CEPA, 2013). In doing so, regulators seek guidance on features of the model to ascertain whether the model is a good reflection of reality. Features may include, amongst other things, the signs and magnitudes of model coefficients, the implied economies of scale properties and the rankings of firms. Not only does this process help to guide the process of modelling, it is a mechanism to both engage stakeholders in the regulatory process and to increase transparency. Engaging end users has been recognised as key in encouraging the uptake of

efficiency studies in health markets (Hollingsworth, 2008; Smith, 2015). Promisingly, engagement with providers is now mandated in Monitor's NTPS process (Monitor, 2014). Indeed, the approach must be approved by at least half of the providers to be imposed; otherwise the proposal is referred to the CMA, as is currently the case.

4.5.5 Range of Methods

It is considered good practice for regulators to use a range of efficiency measures in making assessments of firms' efficiency. Ofcom used both SFA and DEA in predicting British Telecom's performance. In using a range of methods, the regulator presents themselves with a choice: either to choose a preferred model, as in Smith (2012), based on some predefined criteria (statistical tests, expected signs/magnitudes of coefficient values, the underlying assumptions, or other); or to use an approach based on averaging across the models, known as triangulation[44]. As suggested by Bauer et al. (1998), consistency between models' predictions of efficiency, rankings of firms and common outliers denotes reliability. Studies in health markets have shown mixed results; in some studies using mixed methods efficiency estimates coalesce, elsewhere they do not (Hollingsworth and Peacock, 2009). In the case that different methods yield inconsistencies, the regulator may be faced with a difficult choice between models, particularly when they are subject to challenge by the regulated firms and/or the public.

Haney and Pollitt (2013) take the view that the choice of technique should be made on the grounds of it being the most appropriate for the task at hand and the prevailing conditions (e.g. analytical problem, data availability, etc.). Other reasons are posited for choosing between techniques, such as the influence of peers' (i.e. other regulators) methodological choices, that techniques were in the process of being implemented or human resource constraints (Haney & Pollitt, 2009).

Non-frontier econometric methods have also been used in academic studies in the NHS context. These include multi-level modelling and seemingly unrelated regression (SUR) approaches (Jacobs et al., 2006) as well as multi-stage approaches (Laudicella et al., 2010) and difference-in-differences regression (Cooper et al., 2012). The SUR framework may be particularly useful in the context of NHS hospitals since there are often multiple departments within hospitals for which joint modelling is likely appropriate. The SUR framework has

[44] For example, Coelli and Perelman (1999) used the geometric mean of model predictions

been incorporated into frontier efficiency methods directly, although not in health (Lai and Huang, 2012). The multi-level method has been applied to decompose cost variation at various organisational levels to isolate the effect of the hospital on its costs (Dormont and Milcent, 2004; Gutacker et al., 2013a). Here, a hospital-specific effect is taken as the measure of hospital performance meaning there is parallel that can be drawn between this approach and panel data frontier approaches (e.g. Schmidt and Sickles, 1984). Of course, there is an issue around the composition of the hospital-specific effect – this effect comprises any unobserved, firm-specific, time-invariant factors, one of which could well be inefficiency, but could also be a number of other factors.

The application of policy evaluation tools, e.g. difference-in-differences regression, to judge efficiency rests on the measure under analyses. In Cooper et al. (2012), the authors use the pre surgery length of stay for a single procedure. This approach has been criticised as performance may vary across various departments within a hospital, casting doubt over the reliability of measures of this kind as representative of the entire hospital (Bevan and Skellern, 2011). However, these methods do allow specific policy questions to be answered, as is important to policy makers (Hollingsworth, 2008). That is not to say, of course, that other efficiency techniques are unable to answer policy questions – Ferrari (2006) examines the effect of price competition on efficiency in Scottish NHS hospitals, for example.

4.5.6 Panel Data and Temporal Efficiency

The availability of panel data will, of course, be largely determined by the nature of the sector. Those sectors that have a limited number of firms, but perhaps have observed firms over a number of years, often look to international comparison to assemble a panel of data for analysis (NERA, 2008; Smith et al. 2010). Panel data is particularly useful in a regulatory environment, where advanced panel data techniques allow the decomposition of inefficiency and unobservable firm-specific heterogeneity (Kumbhakar et al., 2014). In addition, advanced methods are available to analyse firms' efficiency over time (Coelli et al., 2005). These features of models can be statistically tested, making regulators' efficiency predictions (and corresponding pricing determinations) defensible. Using more sophisticated modelling approaches requires large data. In some cases, sample sizes dictate that only the most basic panel data techniques can be used (Ofgem, 2011). In the literature, more sophisticated models have been found to be difficult to estimate on small datasets (Farsi et al., 2007, pp. 68).

As noted above, another approach to making use of panel data techniques is to exploit the hierarchical structure in organisations (Smith and Wheat, 2012; Smith et al., 2015). In doing so, it is possible to decompose efficiency into company and sub-company components (or equivalents in the NHS hospital context). This approach has a number of advantages for regulators. First, it allows the regulator to precisely locate the source of inefficiency within organisations. Second, efficiency estimates may be biased if the structure is not taken into account. Third, it allows the regulator to expand the size of the data set. Fourth, data need only be collected once. This also obviates issues with data consistency over time. See Smith and Wheat (2012) for an application to European rail network managers.

4.5.7 Heterogeneity

The incorporation of environmental, quality and input price variables into cost function analyses follows from production theory (Coelli et al., 2005; section 3.3.3). The extent to which regulators have managed to incorporate these into their respective analyses varies. The more comprehensive inclusion of these variables is Ofwat's data, which has a number of variables for all three of these features. Indeed, Ofwat's analysis includes these variables jointly in its modelling, whereas other regulators' analyses have not, e.g. CAA for air traffic control, where each are considered in isolation as indicators. Of course, failure to include these variables brings into question the validity of analyses. Weyman-Jones (2012) is critical of Ofgem's RIIO[45] approach, in part, on the grounds of the omission of these variables.

Heterogeneity has been identified as a major issue in the analysis of costs in health (Dormont and Milcent, 2004). Heterogeneity arises in several forms, we examine these in the health context according to three categories: organisational, patient level and quality/outcomes. We consider each in turn below.

4.5.8 Heterogeneity: Organisational

A common approach in academic studies to hospital efficiency analysis is at an aggregated level, such as whole hospitals. This is also the approach employed currently by Monitor (Deloitte, 2014b). There are a number of issues with analysis here which restrict the usefulness of the results. The first issue is that aggregate measures are of limited use to both

[45] Acronym that denotes the approach to regulation: Revenue = Incentives + Innovation + Outputs

managers and policy makers, who are interested in making gains at the level of individual services within hospitals (Hollingsworth and Street, 2006; Hollingsworth, 2008; Mutter et al., 2011). Knowing whether a particular hospital is itself efficient is doubtless useful, but targets are more likely to be enforced if applied at disaggregate levels of activity, since managers can more ably respond at this level. Scott and Parkin (1995) used hospital-aggregate data to estimate cost functions. They did not draw conclusions on their results, partly due to aggregation issues. Indeed, aggregating outputs may cause downward bias for scale economies and likely overlook scope properties of production (Gaynor et al., 2015); a production index approach to estimating the hospital cost function is proposed as a solution. Alternatively, recent studies have been conducted at lower levels of aggregation, for example in specialised services (Diadone and Street, 2013), maternity services (Laudicella et al., 2010), mental health (Moran and Jacobs, 2015) and pathology (Buckell et al., 2013; Buckell et al., 2015).

To the extent that hospitals offer a range of services and specialisations, it is unlikely that two are the same. Indeed, hospitals are commonly in various stages of investment cycles, under differing ownership regimes, providing varying levels/types of teaching, and to varying extents are part of service networks, inter alia (Mutter et al., 2011). Unless these features are controlled for, assigning common cost or production functions is questionable. Thus capturing these differences is critical in making credible assessments of performance. Some features can be readily incorporated into efficiency analysis, ownership status for example (Tiemann et al., 2012) or teaching (Buckell et al., 2015).

For other sources of heterogeneity, data are continually refined and developed for various aspects of service heterogeneity in health. For instance, the ways in which healthcare diagnoses and procedures are coded - by ICD[46] or OPCS[47] coding – are subject to regular updates to reflect developments in practice (WHO, 2004; HSCIC, 2013a). However, even in the case that highly granular data are to hand, there are likely many differences that remain unobserved, the age of hospital buildings or their physical layout, for example. This implies controlling for unobservable heterogeneity is critical.

[46] International Classification of Diseases
[47] Office of Population Censuses and Surveys (the body from which the coding system is granted its name; its full name is "OPCS Classification of Interventions and Procedures)

4.5.9 Heterogeneity: Patient Level

Patient level heterogeneity has long been an issue in comparative analysis of health care metrics. Going back, the following is aimed at Florence Nightingale's comparison of hospital mortality rates in 1863,

"Any comparison which ignores the difference between the apple-cheeked farm-labourers who seek relief at Stoke Pogis, and the wizzened, red-herring-like mechanics of Soho or Southwark, who came into a London Hospital, is fallacious." (Anonymous, 1864)

Patient-level heterogeneity is a clear issue when making cost comparisons between units in health (Iezzoni, 2010). Monitor is equipped with large, granular data on patient characteristics in HES. Diadone and Street (2013) used patient-level data in analysing costs of specialised care in the NHS, in part to make judgements on performance. Of course, there are many differences between patients for which data are unavailable. This implies controlling for unobservable heterogeneity is critical.

4.5.10 Heterogeneity: Quality and Outcomes

Capturing service quality in health efficiency analyses remains a vexing problem (Mutter et al., 2011). The fundamental issue is that service quality itself is both (a) multi-dimensional and (b) unobservable directly. Portrait et al., (2015) note that 'it is surprising that a large proportion of literature on measuring healthcare quality neglects the multidimensional nature of quality'. Indeed, the notion of quality may represent a range of aspects including waiting for services (access), the quality of the medical services, the quality of the environment in which the care is provided, any complimentary treatment, follow up treatment, etc. This issue is of particular focus in the context of the NHS, where, under policy pressure, Monitor had appeared to overlook aspects of service quality when granting foundation status to the Mid Staffordshire Trust (Francis, 2013). Within an econometric framework, service quality can be accommodated, either by direct incorporation of data or by appropriate treatment of unobserved heterogeneity.

Proxy measures for hospital quality (e.g. mortality rates) have been incorporated into efficiency analyses, see Romano and Mutter (2004) and Carey and Stefos (2011).

Additionally, there are stated preference measures, such as the 'Friends and Family Test' (Appleby, 2013). Another attempt to measure quality is through patient outcomes[48]. Smith and Street (2013) argue that the NHS's Patient Reported Outcome Measure (PROM) is a promising tool for capturing quality, albeit imperfect (there is no counterfactual, for example). Promisingly, this measure has been incorporated into efficiency analysis to distil out the effect of quality on costs for making efficiency comparisons between providers (Gutacker et al., 2013a). However, compressing quality into a single metric may cause a loss of information. Indeed, Gutacker et al. (2013b) found variation across different dimensions of outcomes. Further, recent research has begun to base analysis of performance on outcomes (Moran and Jacobs, 2015).

To reiterate, this complex issue underlines the importance of making allowances for unobservable heterogeneity between providers. We turn to this issue below.

4.5.11 Unobserved Heterogeneity

The conceptual appeal of making an allowance for unobservable heterogeneity is to allay concerns about differences in production environments between providers, across a number of dimensions, which are not captured by a set of explanatory regressors. This point is of particular importance in health (Mutter et al., 2011). Significant developments in the recent literature have been made regarding methods to control for unobservable heterogeneity.

In the frontier literature, much attention has been given to this topic, and a number of methods have been developed to accommodate unobservable heterogeneity. Approaches based on restrictions to the cost or production function have been applied in health. Simply adding dummy variables to account for unobserved characteristics is perhaps the simplest approach - Buckell et al. (2015) use regional dummies as one (of a range) control for unobserved heterogeneity. Next is to follow the approach of Mundlak (1978) and decompose a firm-specific (fixed or random) effect using group mean variables. This method has been applied in health markets to nursing homes (Farsi et al., 2005a).

Specific models that can account for unobservable heterogeneity include the "true" models (Greene, 2005), four-component, or "generalised true" models (Columbi et al., 2014;

[48] We differentiate outcomes from output as follows. Output denotes the level of service provision, whereas outcomes refers to patients' response to their treatment. Outcomes are considered a proxy for service quality where higher outcomes reflect higher service quality.

Kumbhakar et al., 2014; Tsionas and Kumbhakar, 2014; Filippini and Greene, 2015). Other approaches include those that take advantage of parameter heterogeneity – latent class or random parameters models (Besstremyannaya, 2011; Greene, 2012) and those that allow for correlation between joint production processes – SUR models (Jacobs et al., 2006; Lai and Huang, 2012). Finally, methodological approaches have been developed that can account for different forms of unobserved heterogeneity (Mundlak, 1978; Kumbhakar et al., 2014) – we have developed a framework for examining these forms of unobserved heterogeneity in multi-level models in the empirical work in this thesis (Smith et al., 2015; chapter 6).

4.5.12 Uncertainty and Sensitivity

Dealing with uncertainty is an important facet of efficiency benchmarking. In regulated industries, the ORR were interested in capturing uncertainty around efficiency predictions in stochastic frontier models, leading academics to reconsider this issue (Wheat et al., 2014).

In stochastic frontier models, to the extent that there is uncertainty surrounding the decomposition of inefficiency and noise, interval estimation is appropriate for inefficiency prediction (Wheat et al., 2014). A known property of cross-sectional models is that the variance of the conditional distribution of inefficiency does not tend to zero as the sample size increases. Here, the central intervals remain wide[49][50] (Street, 2003). This is a key aspect of efficiency analysis for policy makers, being one reason for which efficiency studies have not been widely used (Hollingsworth and Street, 2006; Hollingsworth, 2008). This is not the case, however, in the panel data setting, where the variance does tend to zero (Murillo-Zamorano, 2004). However, this does not provide a perfect solution; there is always uncertainty in separating of inefficiency from noise. Moreover, the ultimate quantity of interest to the researcher is the interval itself, rather than the point estimate (Wheat et al., 2014).

In panel data Schmidt and Sickles (1984)-type approaches, firm-effects can be more precisely predicted as T increases. In the NHS setting, recently multi-level approaches have made use of patient level data, enabling precise estimates (Gutacker et al., 2013a).

[49] In addition, conventional intervals do not incorporate parameter uncertainty and are likely to be too narrow (Wheat et al., 2014).
[50] Conventional stochastic frontier prediction intervals (e.g. Horrace and Schmidt) are central, two sided intervals. Accordingly, they do not capture the asymmetry of the conditional distribution of the inefficiency, meaning they are not the minimum width intervals (Wheat et al., 2014)

Regulators have also been interested in sensitivity analysis. For example, Ofwat considered a large range of model specifications, functional forms and efficiency specifications (CEPA, 2014). Sensitivity analysis is also endorsed for efficiency measurement in health (Jacobs et al., 2006). Monitor made use of a number of specifications and variable definitions (Deloitte, 2014b). To accompany sensitivity diagnostics, a range of statistical testing procedures can guide model selection (Greene, 2012).

4.6 Discussion

4.6.1 Efficiency analysis for the regulation of NHS hospitals

As the existing economic regulator of foundation status hospitals – and therefore as financial arbiter –Monitor is well disposed to setting efficiency targets for NHS hospitals. Monitor has demonstrated its desire to take an approach to setting the efficiency factor which is in line with other regulators in Britain (based on econometric techniques), given that there is limited precedent in other health markets (Deloitte, 2014a).

Following this, Monitor has conducted analysis to set its efficiency factor for 2015/16 using two approaches: an econometric benchmarking exercise and a bottom-up modelling exercise (Deloitte, 2014b). The efficiency factor proposed on this analysis was 3.8%. As noted above, this was rejected by 75% of providers, and is thus being referred to the CMA. Given this, we discuss Monitor's modelling approach and how it accords with health-based efficiency measurement issues, given the issues raised in section 4.5. We derive an index of Monitor's approach to benchmarking which is in keeping with that of Haney and Pollitt (2009). In turn, we use this as a basis for setting some empirical goals for subsequent chapters. In doing so, we return to the question we set out in section 4.2,

(i) How to measure hospital efficiency whilst controlling for quality of care

We have augmented the regulatory best practice criteria of Haney and Pollitt (2009; 2011) to capture health-specific issues raised in academic studies, see table 4.3. We now reconcile Monitor's benchmarking against these issues to observe any potential areas for improving their methodology. In doing so, we draw from our own methodological discussion in chapter 3.

Monitor has made use of large data – their sample contains 832 observations. This is much larger than any other recent regulatory study, see Appendix B. Moreover, this was possible without making use of international comparison. The benefit of this is that assigning a common cost function across the sample is more defensible than would be the case in international comparison. Further, this could be expanded by making use of patient level data from HES. In this sense, Monitor can be seen as leading in regulatory terms.

Data quality was accounted for in Monitor's analysis by making use of a number of variables and through sensitivity analysis. Of course, there remain issues around data quality over time. For example, clinical codes are updated annually, casting doubt over the consistency of data over time (HSCIC, 2013b). However, there have been no major overhauls to the coding system in the period under analysis (e.g. from HRG-4 to HRG-5). As noted, hierarchical modelling is one possible solution to avoid data inconsistency over time. We examine this issue in detail in chapter 6 of this thesis. A further answer to this issue is to make allowances for unobserved heterogeneity to account for these inconsistencies. We examine this issue in both chapters 5 and 6 of this thesis.

Allocating capital costs is not an issue for Monitor's approach to modelling since the reference costs include capital costs. The disadvantage is that the allocation of capital costs is unknown from the data – Reference Costs – meaning that separate OPEX and TOTEX modelling would not be possible as it is for other regulators.

As noted, Monitor has, by mandate, to engage with the industry, as set out in the Health and Social Care Act (2012). It does this via a series of workshops, consultations and through the final response of providers for its efficiency factor determination. In this sense, Monitor is in keeping with this criterion of best practice.

Monitor has also adopted a range of methods to its analysis. It has considered the full spectrum of economic tools (Deloitte, 2014a), and from recommendations of the report, opted for a combination of econometric benchmarking and bottom-up modelling. Within its econometric benchmarking, it has further used a range of methods, namely a random effects model as per Kumbhakar and Lovell (2000), a Pitt and Lee (1981) time-invariant stochastic frontier model and a Battesse and Coelli (1992) time-varying stochastic frontier model[51]. In this sense, it can be said that Monitor has fulfilled this aspect of best practice. However, given its large sample size, it has the potential to use a more sophisticated range of models.

[51] For specification see chapter 3

We see this as a useful extension to Monitor's analysis. We pursue this issue in chapters 5 and 6 of this thesis by using a wider set of inefficiency models in our analysis.

As noted above, Monitor has used panel data and has estimated a model which allows temporal variation in efficiency. However, it is possible for Monitor to make greater use of both of these aspects of its data to enhance its analysis. Panel data techniques can be used to account for unobserved heterogeneity, to which we turn below. Next, the issue of time-varying efficiency is very important in this setting, and in regulatory settings more widely (see for example Smith (2012)). The assumption of time-invariant efficiency over a 5 year period – especially in light of policy pushes for efficiency and productivity gains over the period (see chapter 2; fig 4.1) – is unpalatable in the context of NHS hospitals. Monitor's approach makes use of one model, that of Battesse and Coelli (1992), which does allow for temporal change. However, the treatment of temporal change is fairly limited, for example it applies the same direction of change to all hospitals in the sample, which is unrealistic.

There are a number of models available that allow a more sophisticated approach to modelling temporal efficiency change. Given the large sample size, it would be possible to estimate more advanced models. Moreover, change in efficiency over time is an aspect of efficiency analysis that NHS staff have indicated as being useful (Hollingsworth and Peacock, 2008). For these reasons, we have paid close attention to this topic both in our methodological discussion (chapter 3) and our empirical work (chapter 5).

Organisational heterogeneity is again a key issue in health efficiency measurement and has featured in the analysis of Monitor. Monitor's approach is to maintain a whole hospital level approach and use case-mix variables to account for organisational heterogeneity. This introduces a number of issues.

First, that having to index outputs requires that data are available for all outputs in order to capture them in the index. It is not clear that this is the case (Deloitte, 2014b, pp.7). The construction of the index itself may have bearing on the index value, the corresponding parameter estimates and therefore the estimates of efficiency. It is not clear that Monitor have checked for sensitivity to their method of case-mix adjustment. As noted in the academic literature, aggregate measures at the hospital level are of limited use to local management seeking to identify service-specific inefficiency. Lastly, there may well be unobserved heterogeneity that may influence estimates of efficiency.

The other major drawback of this approach to efficiency analysis is that it does not provide insights into the way in which services drive costs, which is useful information to managers and to policy makers.

We propose two major rectifications. First, to model at a disaggregate level of service. We take this approach in both chapters 5 and 6. Further, a multi-level approach could be used to disentangle the effect of upper tier hospital management on service-level efficiency. Moreover, this has been identified as a key way in which to extend efficiency analysis in health (Hollingsworth and Peacock, 2008). We adopt this approach in chapter 6. Secondly, we reiterate the importance of making allowances for unobserved heterogeneity. This is a central theme in chapters 5 and 6.

Other aspects of organisational heterogeneity have been included into Monitor's analysis. A set of dummy variables are included into the cost function to control for the size of the hospital, whether it provides specialist services, whether teaching services are provided and whether the hospital is a multiservice[52] provider. These help to account for heterogeneity, but may still be somewhat lacking. For example, in Buckell et al. (2015), foundation status was important. There may be other sources of provider heterogeneity, competition for example, that are important for the analysis of costs. Again, this motivates making allowances for unobserved heterogeneity is key when data are limited. We address these issues in chapters 5 and 6 with appropriate modelling approaches.

In terms of patient level heterogeneity, the story is similar. Monitor have partially captured this heterogeneity in their analysis via the use of variables age, gender, ethnicity and deprivation. However, there are arguably patient characteristics omitted; multimorbidity, for example. There are approaches which can make use of patient data, which allow the full information available to be used for analysis (Olsen and Street, 2008; Gutacker et al., 2013a). Equally, using controls for unobserved heterogeneity can mitigate concerns over bias in efficiency prediction.

For quality and outcomes, a specific quality variable was constructed by Monitor based on the NHS staff survey, entailing 15 questions of staff perceptions of their own level of quality. Therefore it can be considered that this aspect of benchmarking has been addressed. Of course, there are several other dimensions of quality, namely outcomes, access, etc. that are overlooked here. This approach is similar to that we have taken in our empirical chapters,

[52] Provide a wider range of service than secondary care, e.g. community services

except that, in recognising we have incompletely captured this feature, we have gone on to model unobserved heterogeneity. We discuss this issue in chapter 5.

We have already considered unobserved heterogeneity as it has emerged when considering other aspects of regulatory best practice. We have noted that we have sought to control for this in our empirical chapters, particularly in our latter chapter, 6.

For uncertainty and sensitivity, Monitor use a range of sensitivity analyses including testing the coefficients to different specifications of the dependent variable (using different deflators for costs over time); using random effects and stochastic frontier models; using samples that have extreme observations removed; and models with and without insignificant variables. Coefficient values appeared to be robust to these specifications. Thus, Monitor has captured some of the features identified for sensitivity in their analysis. However, we propose several extensions.

First, to test whether the imposed Cobb-Douglas functional form is the most appropriate in statistical terms (the best fit of the data). Using more sophisticated functional forms helps to capture unobserved heterogeneity and may have appealing economic properties – for example the translog allows economies of scale to vary across the output range (see chapter 3; chapter 5; Buckell et al., 2015). Second, although some statistical testing of individual variables was conducted, this process can be improved by testing inefficiency models against each other. This helps to justify model selection. We adopt this approach in chapters 5 and 6.

A summary of Monitor's approach is given below in table 4.4 as a Haney and Pollitt (2009)-esque benchmarking index.

Issue	Score	Reasoning
Large Data	1	data set large
Data Quality	0.5	issues around consistency of collection
Allocating capital costs	1	capital costs allocated in reference costs
Engagement with industry	1	in a number of ways
Range of methods	1	econometric benchmarking and bottom-up modelling
Panel methods and temporal efficiency	0.5	panel data used; time dynamics could be explored further
Heterogeneity: Organisational	0.5	aggregate measure of output used; could disaggregate, could adopt multi-level approach
Heterogeneity: Patient level	0.5	some controls but not comprehensive coverage
Heterogeneity: Quality/Outcomes	0.5	some controls but not comprehensive coverage
Unobservable Heterogeneity	0	no controls for unobserved heterogeneity
Uncertainty and sensitivity	0.5	some sensitivity; little by way of uncertainty
Total	7 of 11	

Table 4.4: Benchmarking Index for Monitor's 2015/16 NTPS Analysis. 0 indicates that the issue is not addressed in the analysis; 0.5 means that some control for the issue has been made; 1 denotes the issue is captured in the analysis. The scores are the author's judgements, based on the preceding section.

An index value of 7 from a possible 11 (table 4.4) indicates that there are many satisfactory elements of the approach, but also that there are some areas in which improvements could be made.

In Haney and Pollitt (2009), international regulators' benchmarking index values ranged from 0 to 7 on a scale of 8; there was a mass of regulators with 0 score. Regulators in the UK obtained scores in the range of 3 to 6, with an average of 4.5. If Monitor's benchmarking was subject to this index, it would receive a full score of 8, indicating that it has fulfilled all of the criteria.

However, given that (a) its current analysis has been soundly rejected by providers; (b) there are health-specific issues that require specific attention (e.g. patient level heterogeneity), there is a clear need to tailor the index for health care markets[53]. In so doing, this index may be of use to health regulators around the world who are engaged in benchmarking in health. These form the basis of our empirical analysis, as indicated below.

(i) Analysis at a disaggregate level of service (Ch 5 & 6);
(ii) Dual-level efficiency analysis (Ch 6);
(iii) Accounting for unobserved heterogeneity (Ch 5 & 6);
(iv) Extending the analysis of temporal efficiency change (Ch 5);
(v) Functional form (Ch 5 & 6); and
(vi) Statistical testing (Ch 5 & 6).

4.6.2 Encouraging Efficiency in NHS Hospitals

Efficiency measurement is an important first step for Monitor. On this, to a greater or lesser extent (depending on regulatory judgement), Monitor will set the efficiency factor. The next stage is to ensure that corresponding savings are achieved. NHS hospitals appear to have, in general, responded well to various targets that they have been set. This is in contrast to the savings set out by the efficiency factor during the same period, which have not been met. We have reviewed evidence to draw out lessons for Monitor, to which we now turn.

We have identified from the literature features of hospital targets which appear to effective (box 4.1); and symptoms of regime failure (box 4.2).

As regards box 4.1, many features of the revised system are conducive to Monitor successfully implementing efficiency savings amongst NHS hospitals. Monitor, the existing economic regulator, will have good knowledge of hospitals' data and financial performance, given that this is already within its remit. Moreover, Monitor has a much narrower agenda than that the Department of Health; they do not have to provide and maintain health services as well as to regulate them. This means that they are able to assign greater priority and focus to the efficiency factor. In order for targets to be met effectively, it appears important features include ensuring targets are effectively communicated, prioritised and applied consistently;

[53] Equally, regulators in other sectors may also benefit from reviewing their sector-specific issues and developing their own indices accordingly

that the results are widely disseminated; and that comprehensibility is key (Mannion et al., 2005; Bevan, 2006). Further, Monitor has the authority to impose sanctions for failure to meet targets; having sanctions for failure is effective for improving performance (Smith and Sutton, 2013).

As regards box 4.2, using an econometric approach has key advantages with regard to some of the issues encountered in applying other targets such as ambulance response times. They are easily interpreted (bounded by zero and one); they allow ranking of providers; and can be used in a time series dimension. Modelling based on cost functions can help to obviate gaming, unintended consequences, partial measures of performance and diversion of resources from elsewhere, since the entire production process is modelled (Buckell et al., 2015). However, the extent to which this is possible in reality is unclear; cost (and other forms of) data can be gamed (e.g. Moran and Jacobs, 2015). They do not require aggregation of several indicators and multiple measures can be included in the analysis as variables. Some performance indicators, for example readmission rates, may themselves give misleading accounts of provider performance (Laudicella et al., 2013). Persistence can assuage concerns over myopia and tunnel vision (Mannion et al., 2005).

From our review of regulatory pricing mechanisms (section 4.4), an important step is to lengthen the current regulatory lag in line with regulatory theory. Regulatory theory suggests that when regulatory lag is short, the incentives to improve efficiency via reducing costs are reversed from being high powered to low powered. Whilst it is not clear what an optimal lag would be, the current lag of a single year appears too short based on current performance[54]. To follow the regulatory norm would be to set the lag at 5 years; alternatively, Monitor may wish to align the regulatory lag to specific policies such as the *five year forward view* (NHS England, 2014a). Moreover, the price-cap mechanism can be incorporated into a more complicated system which embeds service level requirements from the Department of Health. We gave an example of the ORR's tripartite approach. This may be particularly useful given the current stand-off between Monitor and NHS providers.

4.7 Conclusions

The responsibility for setting the efficiency target has recently changed from the Department of Health to the economic regulator, Monitor. Monitor has sought a more evidence-based

[54] Whilst it would be surprising if this was the only reason for missed targets, we submit that it is a contributory factor

approach to setting the efficiency target (Deloitte, 2014a). Monitor has used econometric benchmarking as part of a series of methods to estimate the efficiency factor for NHS hospitals. This determination has been rejected by hospitals and is currently being reviewed by the CMA.

We have reviewed recent methodological advances in efficiency measurement in health and beyond. We recommended the extension of econometric techniques to assess NHS hospitals' efficiency. We have identified particular features of the analysis that may be used to enhance Monitor's approach, based on practical and statistical issues.

In addition, we have reviewed the application of alternative policies for performance management, and observed features of these policies that were effective; and those that proved problematic. This should be useful information for Monitor in seeking to reduce leakage. We have further reviewed regulatory pricing and suggested alterations to the current pricing mechanism to foster efficiency.

Having paid attention to features of and issues with efficiency analysis in the NHS setting (and in health more generally), we move to our empirical analysis. Here, we seek to estimate efficiency within NHS hospitals. Specific focus is given to pathology laboratories within NHS hospitals. Therefore, we move to the next two chapters which present empirical work in NHS pathology.

5. Efficiency over time, economies of scale, multi-factor productivity and mergers in National Health Service Pathology

In our introductory chapters, we noted that on account of the scale of savings required, there is a need to examine new areas of hospital services to identify areas in which efficiency savings can be derived. This chapter, based on Buckell et al. (2015), represents such work.

In the previous chapter, we reviewed the regulation of efficiency amongst NHS hospitals. A central aspect of this process is the measurement of efficiency itself, as a guide to setting efficiency targets. We described how the measurement of inefficiency had been problematic in health markets, for a variety of reasons, both technical and practical. Therefore, our methodological approach has been designed to factor in these issues.

This chapter is one of two empirical studies of hospital efficiency in this thesis. There is evolution in our analysis, across the two studies, as follows. The focus of this study is on describing pathology services and on primal economic issues regarding the level of inefficiency in pathology services, the drivers of laboratory costs, scale and so on. In the next study, we move to more sophisticated methodological questions by considering multi-level organisational structures and the nature of unobserved heterogeneity.

The main features of this analysis are in keeping with issues raised for health-based efficiency analyses in prior chapters. First, the study picks up on several technical themes identified: using an econometric approach based on the cost function, the use of panel data, analysis at an appropriate level of disaggregation, sensitivity analysis and statistical testing, allowing for unobserved heterogeneity and allowing for service quality. Further, we answer several pathology-based policy questions pertaining to: the extent of potential efficiency savings, economies of scale, the cost implications of mergers, change in efficiency over time, reconciling technological progress and efficiency change.

This study demonstrates the challenging trade-off that regulators face as regards disaggregate analysis. On the one hand, as has been argued in the literature, concerns around heterogeneity are abated with disaggregate studies, and, data permitting, a more granular approach to characterising heterogeneity can be taken (see chapter 4). The downside, on the contrary, is that estimates of inefficiency are limited to the service themselves, in this case pathology

services (this again has been noted in the literature, see chapter 4). Whilst the goal of this study is not to reconcile this issue, it is important to make note of it. Of course, the regulator, having access to a wealth of data, is able to conduct similar studies across a range of hospital services.

We begin in section 5.1 by describing pathology services and their impact on wider NHS services. We then consider how pathology services have been measured and detail the advantages of taking an econometric approach (section 5.2). In the following section, 5.3, we discuss the methods and data used in detail. We then present and discuss various aspects of our results, in sections 5.4 and 5.5, respectively. Finally, in section 5.6, we conclude.

5.1 Introduction

Pathology services account for an estimated 3-5% of the overall NHS budget, costing an estimated £2.5bn in 2005 (Department of Health, 2006). Although relatively small as a proportion of total health care spend, potential efficiency gains in these services are not confined to pathology itself. Pathology activity supports many front-line services and so savings in pathology services promote further gains elsewhere in the healthcare system (Veronesi et al., 1997; Buckell et al., 2013). The Carter Review (Department of Health, 2006) estimates 70-80% of all clinical decisions are affected by pathology analyses; thus good pathology practice can lead to cost savings along a patient's treatment pathway (Department of Health, 2006). There is evidence of unnecessary repeat testing (Department of Health, 2006), suggesting that inefficient practice is present in these services. Lastly, there is variation in the uptake of lean practice initiatives[55] meaning that there is likely variation in the magnitudes of efficiency in these services. Therefore, there are likely significant gains to be made by encouraging best practice in pathology services to contribute to the policy objective of achieving efficiency savings. This study aims specifically to identify the level of inefficiency in pathology services in order to measure the extent of savings possible in this area.

The current approach to measuring inefficiency in pathology in the NHS is performance indicator analysis (such as cost per test carried out); (Healthcare Commission, 2007;

[55] NHS Institute for Innovation and Improvement: Pathology lean practice case studies, http://www.institute.nhs.uk/quality_and_value/lean_thinking/leean_case_studies.html

Department of Health, 2008; Liebmann, 2011; Holland et al., 2012). These are partial measures which do not fully reflect all the factors affecting the costs of provision under different circumstances (for example, scale properties or sources of operational heterogeneity between providers). This point has been established in the wider health context (Goddard and Jacobs, 2009; Street et al., 2011). We use the data collected and analysed by the Keele University Benchmarking Unit (Holland et al., 2012), but extend the analysis by utilising an econometric framework to give a single measure that captures the overall efficiency of pathology services. Our model takes account of a range of factors influencing costs, whilst controlling for unobservable heterogeneity.

We use stochastic frontiers which have been applied widely in health at the micro level (Street, 2003; Farsi et al., 2005a, 2008; Herr, 2008; Hollingsworth, 2008; Olsen and Street, 2008; Rosko & Mutter, 2008; Sorensen et al., 2009; Herr et al., 2011). We adopt a particular stochastic frontier method with attractive properties in respect of analysing efficiency change over time; this method has been applied by economic regulators outside health for that reason (Smith, 2012). To our knowledge, no stochastic frontier (or other efficiency measurement tool such as DEA) work has been conducted on pathology laboratories, meaning that our application is the first of its kind[56].

5.2 Performance Measurement in Pathology

Pathology services are increasingly recognised as key support for a range of services across the NHS. As demand for NHS services increases in general, demand for pathology services increases (as derived demand). Faced with increasing demand and falling income (Department of Health, 2006), the performance of pathology services is coming under ever-increasing scrutiny. Therefore, rigorously measuring the performance of laboratories is critical. Typically, pathology laboratories are situated within NHS trusts (see below).

[56] If pathology is classed as diagnostic medicine, then there exists some stochastic frontier work in this area (Dismuke & Sena, 1999). However, this study concerns patient-based, in-hospital activity such as computerised axial tomography (CAT) scans, whereas our study involves pathology laboratories – which are independent of their host hospitals and do not have direct patient contact – conducting blood and tissue tests. We therefore view pathology services as distinct from this kind of diagnostic medicine.

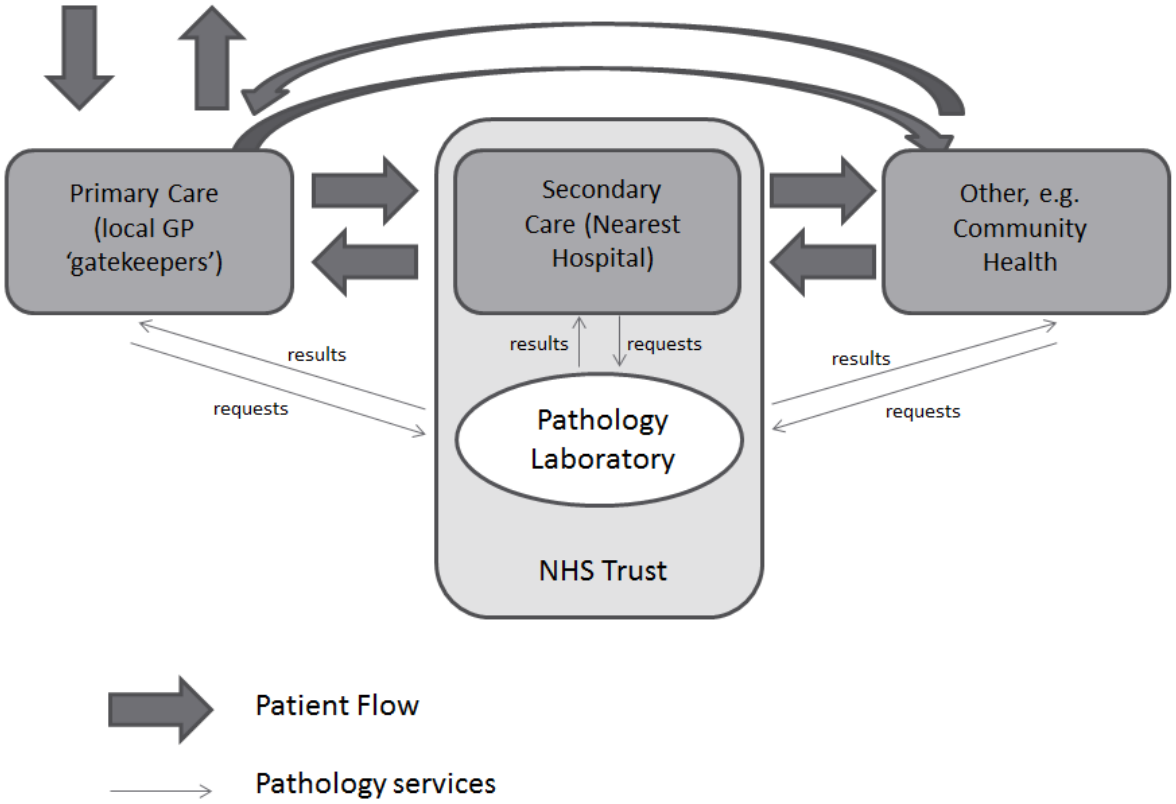

Figure 5.1: Schematic of Pathology Services

As can be seen from Fig. 5.1, as patients move around the healthcare system, diagnostic services are requested and performed. As activity occurs, information is recorded and used for analysis of these services.

Major reviews of NHS pathology services include the Carter Report (Department of Health, 2006), and the associated follow up report which included pilot studies of services (Department of Health, 2008); the Healthcare Commission's study (2007); the NHS confederation (2010); and the Keele University Benchmarking project (Holland et al., 2010; 2011; 2012)[57]. There is a growing body of evidence on these services, and good quality data available; a summary of these studies' analyses is provided in Table 5.1.

[57] Some key performance indicators are being introduced, but have not yet been employed (Liebmann, 2011).

Study	Year	Number of Sites	Type of study	Summary of Key Points
Department of Health	2006	163	Qualitative	Full qualitative analysis of pathology services. Identified key areas for performance improvement - workforce balance, economies of scale, information systems adoption, out of hours working, network activity. Recommended pilot studies conducted. Noted that geographical location may be a source of cost heterogeneity.
Healthcare Commission	2007	163	Quantitative	Breakdown by pathology discipline comparative cost per test analysis; requests:staff and tests:staff ratios used; descriptive statistics for out of hours operation, information systems adoption, use of automated services, network activity; recognised that tests for primary care may be cheaper than for secondary care; noted the issue of tests:requests as a potential source of performance variation. Foundation trusts may take a commercial approach to service provision.
Department of Health	2008	12	Quantitative	Breakdown by pathology discipline (e.g. biochemistry) comparative cost analysis; some economies of scale observation; little control for heterogeneity; savings estimate £250m (extrapolated results nationally from 12 pilot studies).
NHS Confederation	2010	163	Qualitative	Identifies variation in practice; difficulty in monitoring staff leads to variation in practice; workforce balance, IT systems adoption, leadership and network activity as key areas for performance improvement.
Keele Benchmarking	2012	84	Quantitative	Breakdown by pathology discipline (e.g. biochemistry, hystocytology); test volumes descriptive statistics; productivity indicators; 5 year trend analysis of outputs and productivity indicators; expenditure of laboratories; quality indicators (e.g. turnaround times)

Table 5.1: Pathology Studies

Table 5.1 describes the outcomes of each of the studies. The quantitative analyses above use performance indicators to judge the performance of NHS pathology laboratories (e.g. cost per test ratios, staff per test, turnaround times, test to request ratios). The use of these indicators is widespread in NHS pathology and across the world (Valenstein et al., 2001; Kiechle and Main, 2002; Price, 2005; France and Francis, 2005), but there are limits to their ability to reflect the entire operation of a laboratory. Moreover, in health markets, indicators can be targeted for gaming (Propper and Wilson, 2003; Propper et al., 2008; Mutter et al., 2008; Palangkaraya and Yong, 2013), or relying solely on indicators can lead to unintended consequences (Bird et al., 2004; Cots et al., 2011). Lastly, judging a single unit's performance across several indicators may be difficult if the values conflict.

An econometric framework is proposed to overcome these issues. Our measure of cost efficiency yields a single efficiency score capturing overall performance which is easily interpreted (bounded by zero and one). Gaming is no longer an issue since the entire production process is modelled[58] .

A further key advantage of the econometric approach is that it is underpinned by economic theory and stochastic frontier analysis is used widely across many sectors, including health (Kumbhakar and Lovell, 2000; Hollingsworth, 2008). In addition, we can analyse the temporal pattern of laboratory inefficiency, which NHS staff have indicated as a desirable feature of performance analysis (Hollingsworth and Peacock, 2008). Finally, econometric analysis allows us to value the impact of some of the issues noted in the qualitative studies (Table 5.1), such as the ratio of primary care tests on costs – as raised in the Healthcare Commission study (2007), which is useful information in the policy context.

5.3 Methods

Stochastic frontiers (Aigner et al., 1977; Meeusen and van Den Broeck, 1977) are econometric tools used to estimate the level of inefficiency of firms or decision making units (DMU) in a sample. Laboratory costs are our metric of interest. Our economic stochastic

[58] We use operating costs rather than total costs (including charges), meaning the production process is not strictly entirely modelled. Capital costs are budgeted centrally at trust (hospital) level, rather than laboratory level, meaning assigning specific capital charges to laboratories can only be estimated. We note that this has been found in pathology elsewhere, e.g. New Zealand (France and Francis, 2005). Moreover, this is not particular to pathology (Drummond et al., 2005, pp. 64).

frontier model for pathology, derived from a basic cost function (see chapter 3), takes the form,

$$c = f(y, w, t, z, q) + u + v \tag{5.1}$$

Where c are costs, y represents output, w represents input prices, z represents the observable heterogeneity, q represents quality and t represents time. As standard for stochastic frontiers, u represents the inefficiency and v represents random statistical noise.

As standard in the literature, output and input prices are considered exogenous, which is obvious for input prices and reasonable for output levels given that the laboratories do not choose their level of output. In the case of pathology, using the work of previous studies (see table 5.1), the operational characteristics of the pathology operating environment can be identified and variables are used to capture these where data are available (the *z* vector). Otherwise, methods for capturing unobservable heterogeneity are employed.

For service quality, although measures of quality in pathology services are not as complex as in the treatment of patients (Smith and Street (2013) note the multi-dimensional nature of patient treatment quality), this remains an issue for our study. Each of the laboratories in our sample has acquired quality accreditation[59]. Our understanding of accreditation is that it represents a baseline level of quality. Therefore, we recognise that there may well be laboratory-specific variation in quality over and above this baseline level. This is one reason for which we apply empirical controls for unobserved heterogeneity; that is, quality that is not captured in the accreditation is absorbed into the control for unobserved heterogeneity rather than absorbed by the inefficiency component of the model.

A set of five models stochastic frontier is used to model inefficiency. These include a generalised least squares random effects model[60], see Kumbhakar and Lovell (2000). We refer to this as REM. We use a Pitt and Lee (1981) stochastic frontier with time invariant inefficiency, which we refer to as P&L. Next, we use a Battesse and Coelli (1992) stochastic frontier with time varying inefficiency. We refer to this as BC92. Our penultimate model is that of Cuesta (2000), which is a stochastic frontier with firm-specific (or in our case, lab-specific) time-varying inefficiency. We refer to this as Cuesta. Finally, we use a true random

[59] Clinical Pathology Accreditation: http://www.cpa-uk.co.uk/
[60] Hausman tests (1978) consistently favoured RE over FE estimation; we are also interest in examining time-invariant variables which we are unable to do in a FE framework.

effects model (Greene, 2005). We refer to this as TRE. See table 5.2 for econometric specification; see section 3.5.4 for discussion of the models.

The REM is used to give 'baseline' values for both parameter estimates and for inefficiency (using the GLS procedure outlined in Kumbhakar and Lovell (2000)). Parameter estimates from these models do not rely on the distributional assumptions of the stochastic frontiers[61] and so parameter estimates are used to validate those derived from the frontiers.

The P&L model assumes time-invariant inefficiency. The BC92 fits a time trend to the inefficiency - the η parameter (table 5.2) - which subjects all firms' efficiency scores to a common direction of change over time. The Cuesta model is a generalisation of this, allowing estimation of independent firm efficiency time trends: individual ηs for each laboratory[62]. This means firms can 'catch up' relative to others over time and the efficiency rankings of the laboratories can change over time, which are realistic features. This point is particularly relevant in a policy context, and this model has been used by regulators in other sectors, e.g. rail (Smith, 2012). Alvarez et al. (2006) further note that a key advantage of this model is that it enables the unrealistic assumption of independence in inefficiency over time (a problem that plagues many comparator models) to be relaxed.

The TRE model claims to delineate efficiency from unobservable heterogeneity by including a time-invariant, firm-specific term in the model to capture unobserved factors, in addition to the inefficiency term (Greene, 2005). A potential drawback of this model is that efficiency scores are independent over time, meaning that time trends of firms cannot be tested statistically. Additionally, this model assumes that all the time-invariant variation in the cost function that is not explained by the regressors is unit-specific heterogeneity and not inefficiency; this is not necessarily the case as some time invariant persistent inefficiency may also be present.

To these models, we test three alternative specifications to examine heterogeneity. First, a basic cost function with output, input prices and time is estimated. By including a time trend in the cost function, we separate exogenous change in costs over time from cost inefficiency (Kumbhakar and Lovell, 2000).

[61] Due to an unbalanced panel, a Baltagi & Li (1990) adaptation of the Breusch-Pagan (1980) test has been used and confirms the use of panel methods.
[62] Within this framework, the temporal pattern of inefficiency can be tested statistically, which is a key advantage over alternative approaches such as Cornwell et al. (1990).

In the second, we add the vector, *z*, of observable heterogeneity variables. These include the number of primary care tests (which are thought to be less costly than other tests), and the test to request ratio which captures the variation in the number of tests per request, which varies between laboratories, and is therefore a source of heterogeneity. Another source is the geographical setting of the laboratory: metropolitan, urban or rural (following Department of Health, 2006, see table 5.1). This will be referred to as the TYPE of laboratory. It has been suggested that pathology demands of inner city laboratories are much different to those in rural areas. Further, the foundation status[63] of a trust is seen to motivate it to act more commercially (Healthcare Commission, 2007, see table 5.1; Marini et al., 2008), which is expected to be extended to their pathology services. Lastly, data are available on whether the laboratories provide teaching services.

The third specification finally adds dummy variables to capture unobservable heterogeneity (e.g. IT infrastructure/maturity, network activity) (Arocena et al., 2012). We use the strategic health authority dummy variables and then group them by region for parsimony.

We refer to the specifications as s(i), s(ii) and s(iii).

Finally, after having used this testing process to select a model, we exploit the fact that the stochastic frontier framework is based on a cost function to examine the cost elasticity properties across the output range and derive average and marginal costs in pathology production (AC and MC hereafter). We note that this is a key advantage of this method over DEA as an alternative. Focus is given to this aspect of production because this is a popular theme of interest throughout the literature (table 5.1), because there is little empirical evidence on this issue, and because of the growing membership of laboratories to local networks, which is encouraging the pooling of output); see Department of Health (2011).

5.3.1 Empirical Specification

First, for functional form, we test between a Cobb-Douglas and a translog specification to approximate our economic model in eqn. (5.1). A translog nests a Cobb-Douglas and we can readily test down. A translog has some appealing empirical and economic features: its flexible nature means it provides a second-order differential approximation to any unknown

[63] Foundation status of a NHS trust (a trust is a hospital or small group of hospitals) means that it operates under an independent, not-for-profit regime, allowing it financial autonomy which it does not have without having foundation status (Marini et al., 2008). Trusts apply for foundation status, which is granted by the regulator, monitor, if the trust has satisfied the regulator of its financial competence. Foundation status has not been awarded to all NHS trusts.

function $f(.)$ (as in Equation (5.1)) (Kumbhakar and Hjalmarsson, 1995); it does not impose restrictions on substitution possibilities; and allows economies of scale to vary with output levels (Christensen and Greene, 1976).

Logarithms are taken to give Farrell (1957)-type radial measures of inefficiency[64]. The translog representation is estimated for each model,

$$
\begin{aligned}
&ln\, c_{it} \\
&= \alpha_0 + \beta_1 \ln y_{it} + \frac{1}{2}\beta_{11} (\ln y_{it})^2 + \beta_2 \ln wl_{it} \;\; + \frac{1}{2}\beta_{22} (\ln wl_{it})^2 \\
&+ \sum_{a=1}^{2} \beta_3 lnz_{it} + \frac{1}{2}\sum_{a=1}^{2} \beta_{33} (\ln z_{it})^2 + \beta_{12} lny_{it}.\, lnwl_{it} + \sum_{n=1}^{1}\sum_{a=1}^{2} \beta_{13}\, lny_{it}.\, \ln z_{it} \\
&+ \sum_{b=1}^{1}\sum_{a=1}^{2} \beta_{23} ln\, wl_{it}.\, \ln z_{it} + \beta_{34} lnz_{1t}.\, lnz_{2t} + \sum_{c=1}^{4} \beta_5 z_i + \sum_{d=1}^{3} \beta_6 \omega_r + \beta_7 t \\
&+ \varepsilon_{it}
\end{aligned}
\tag{5.2}
$$

Where c_{it} are operating costs; y_{it} is output; wl_{it} are labour input prices; z_{it} are exogenous variables including tests for primary care and the test to request ratio; z_i are laboratory-specific, time-invariant dummy variables for the following: foundation status, teaching status and laboratory type[65]; ω_r are regional dummy variables to capture unobservable heterogeneity; and t is a time trend capturing real cost changes over time (in this sample). Then, ε_{it} is decomposed into u_{it} and v_{it} which are inefficiency and statistical noise, respectively (see table 5.2 below for detailed specifications of each model).

To decide on a preferred model, a number of statistical tests are applied[66]. We test functional form using a Wald test[67].

Next, we test between the three specifications from above, by which we mean either no heterogeneity variables s(i); observable heterogeneity variables only s(ii); and observable and unobservable heterogeneity variables[68] s(iii). We use LR tests for this. We refer to this as TEST 1.

[64] Variables are mean-scaled to allow direct interpretation of the first order terms; see Appendix A for derivation.
[65] Types of laboratory include rural, urban and metropolitan; rural is the reference case for modelling.
[66] Lai and Huang (2010), pp. 3, lament that "there are only limited systematic treatments of tests or model selection criteria in the existing stochastic frontier literatures."
[67] H_0: additional translog terms (squared and cross terms) are jointly equal to zero.
[68] H_0: observable or unobservable heterogeneity variables are jointly equal to zero.

We then test between each efficiency model, by which we mean one of the 5 different efficiency models (REM, P&L, BC92, Cuesta, TRE), using a LR test[69] for nested models (which we refer to as TEST 2) and a Vuong test (1989) for non-nested models[70] (which we refer to as TEST 3).

In total, there are 30 models to be estimated[71]. 15 models are reported for comparison which represents our full set of models once the test for functional form has been applied. LIMDEP software (Greene, 2012a) is used for estimation.

5.3.2 Inefficiency Models

Table 5.2 below shows the econometric specifications of our range of models estimated.

	REM	P&L	BC92	CUESTA	TRE
Firm-specific component, α_i	$iid(0,\sigma_\alpha^2)$	$iid(0,\sigma_\alpha^2)$	$iid(0,\sigma_\alpha^2)$	$iid(0,\sigma_\alpha^2)$	$N(0,\sigma_\alpha^2)$
Random Error, ε_i	$iid(0,\sigma_\varepsilon^2)$	$\varepsilon_{it}=u_{it}+v_{it}$ $u_{it}\sim\lvert N(0,\sigma_u^2)\rvert$ $v_{it}\sim N(0,\sigma_v^2)$	$\varepsilon_{it}=u_{it}+v_{it}$ $u_{it}\sim\lvert N(0,\sigma_u^2)\rvert$ $v_{it}\sim N(0,\sigma_v^2)$	$\varepsilon_{it}=u_{it}+v_{it}$ $u_{it}\sim\lvert N(0,\sigma_u^2)\rvert$ $v_{it}\sim N(0,\sigma_v^2)$	$\varepsilon_{it}=u_{it}+v_{it}$ $u_{it}\sim\lvert N(0,\sigma_u^2)\rvert$ $v_{it}\sim N(0,\sigma_v^2)$
Inefficiency	$\hat{\alpha}_i - min\{\hat{\alpha}_i\}$	$E[u_{it}\lvert u_{it}+v_{it}]$	$E[u_{it}\lvert u_{it}+v_{it}]$	$E[u_{it}\lvert u_{it}+v_{it}]$	$E[u_{it}\lvert \alpha_i+\varepsilon_{it}]$
Time Trend			$u_{it}=\exp[\eta(t-T)].u_i$	$u_{it}=\exp[\eta_i(t-T)].u_i$	

Table 5.2: Econometric Specifications of Models

5.3.3 Merging Laboratories

A feature of recent pathology services is that, following recommendations from the Carter Review, laboratories in close proximity are increasingly beginning to pool their production (Department of Health, 2006; 2009). A natural question arises as to what happens to the costs of production when laboratories merge. This is, of course, tied closely to the issue of

[69] H_0: log likelihood model (a) is equal to log likelihood model (b)
[70] H_0: model (a) is equal to model (b)
[71] 2 (functional forms) x 3 (heterogeneity variable specifications) x 5 (types of efficiency model)

economies of scale, which is of great interest to NHS policy makers and policy makers more widely.

In our data, there are no examples of laboratory mergers. However, it is possible to use the model to simulate the effects of laboratories merging to shed some light on this issue: we can simply compare the sum of the predicted merged laboratory costs and the sum of the predicted unmerged laboratory costs. We do this for laboratories in the final year of the dataset.

To operationalise the merged scenario, we merge the smaller laboratories with each other. We define a "small laboratory" as one whose output (number of requests) is lower than the sample median. We then merge the largest "small laboratory" with the smallest "small laboratory", the second largest with the second smallest, and so on. We assume the larger laboratory absorbs the smaller; we thus assume the characteristics (i.e. foundation status, teaching status, region, etc.) of the larger laboratory for computing merged cost estimates. We are interested in the proportional change in total costs that would occur if small laboratories were to merge, thus we compute the following ratio,

$$\frac{\sum_{i=1}^{I} E\left(c_{i,T}\middle|{x_{it}}'\beta\right) - \sum_{j=1}^{J} E\left(c_{j,T}\middle|x_{it}'\beta, y > \tilde{y}\right)}{\sum_{i=1}^{I} E\left(c_{i,T}\middle|{x_{it}}'\beta\right)} \tag{5.3}$$

where $E\left(c_{i,T}\middle|{x_{it}}'\beta\right)$ is the conditional expectation of costs for laboratory *i* in its final year, *T*. The ${x_{it}}'\beta$ is the estimated cost function, y is output and $y > \tilde{y}$ denotes all output is greater than the (original) sample median, that is, laboratories with output lower than the median have merged. $\sum_{i=1}^{I} E\left(c_{i,T}\middle|{x_{it}}'\beta\right)$ is the sum of the predicted costs across all unmerged laboratories and $\sum_{j=1}^{J} E\left(c_{j,T}\middle|x_{it}'\beta, y > \tilde{y}\right)$ is the sum of predicted costs across all merged laboratories. As a result of simulation, of the full sample of 57 laboratories, 28 "small" laboratories are merged into 14, thus reducing the number of laboratories from 57 to 43. Therefore, *I*, the number of unmerged laboratories, is 57 and *J*, the number of merged laboratories, is 43.

Given the specification of our model (see equation (5.2)), there is an issue around retransformation of logged (predicted) costs (Manning, 1998). When the disturbance of the error term is normal, $\hat{\varepsilon} \sim N(0, \sigma^2(x))$, then a straightforward correction can be made,

$$E\left(c_{i,T}\middle|x_{it}'\beta\right) = e^{x_{it}'\beta+0.5\sigma^2(x)} \tag{5.4}$$

where the uncorrected estimate is an underestimate since,

$$e^{x_{it}'\beta+0.5\sigma^2(x)} > e^{x_{it}'\beta} \tag{5.5}$$

However, normality is an invalid assumption in our case as the stochastic frontier model does not, by definition, assume a normally distributed disturbance. Thus, an approach is required that can account for non-normally distributed errors. Therefore, as suggested by Greene (2012c, pp. 123), we use the smearing estimator proposed by Duan (1983). Thus our predictions of laboratory costs are,

$$E\left(c_{i,T}\middle|x_{it}'\beta\right) = h^0 e^{x_{it}'\beta} \tag{5.6}$$

where,

$$h^0 = \frac{1}{n}\sum_{i=1}^{I} e^{\hat{\varepsilon}_i} \tag{5.7}$$

where *n* denotes the number of observations and $\hat{\varepsilon}_i$ are the fitted residuals.

5.3.4 Data

Annual pathology benchmarking data (Keele Benchmarking) is used to compile an unbalanced panel of 57 English NHS pathology laboratories during a 5 year period from 2006/7 to 2010/11[72] (187 observations); accordingly we use maximum likelihood estimation (Baltagi, 2008) (except the REM which uses GLS and the TRE which uses simulated maximum likelihood). The sample represents approximately one third of the 163 NHS pathology laboratories in England. From table 5.3, there is considerable variation in the range and standard deviation of the costs, tests and requests variables, giving us confidence that we have a broad sample of laboratories. There is an almost even spread of laboratories amongst strategic health authorities (and therefore across England).

Our data is for biochemistry services only. Biochemistry is one of five disciplines of pathology (the other four being haematology, hystocytology, immunology and microbiology).

[72] In our sample, 27 laboratories are observed twice, 7 are observed 3 times, 2 are observed 4 times and 21 are observed in every year – 5 times.

Biochemistry is chosen because it is highly mechanised thus diminishing the issue of heterogeneity for modelling. It is the largest area of pathology (around 70% total activity (Holland et al., 2011)) and all laboratories run biochemistry services. A three stage process of data validation between the laboratories and Keele Benchmarking Unit is applied to ensure the data is accurate.

Variables include total operating costs (net of capital charges), output (for which two measures are available: the number of tests and the number of requests), input prices of labour (from the UK labour force survey) and exogenous variables including the number of tests for general practice (primary care) and dummy variables for the foundation status of the host trust, for the pathology service providing teaching, for the laboratory type (metropolitan, urban, rural) and for the strategic health authority in which the pathology service is located. Service quality is assumed given that laboratories have been accredited as noted earlier.

Costs and wage data are in real terms (2007 prices) using the consumer prices index. Labour force survey data is chosen over other sources (NHS staff census data, for example). This is firstly to ensure the exogeneity of the data: because the labour force survey data is collected and constructed independently from our study data, which would not be the case using the NHS-based data[73]. In addition, this data is a reflection of the true labour market conditions, which is not necessarily the case with the NHS data. Lastly, the NHS equivalent data is constructed using staff numbers which implies the measure may be correlated with output, which may lead to undesirable statistical issues such as collinearity. Secondly we aim to better reflect the regional variation in labour input prices than would be possible using alternative data. The ratio of tests to requests is calculated from the data[74]. Strategic health authorities are, following initial modelling, combined to form regional dummy variables for London, the South, the Midlands and the North using a Wald test procedure (Greene, 2012b).

One available measure of clinical quality was available for analysis: turnaround times of tests. We did not use this for three major reasons. First, as an indicator, this is an incomplete measure of clinical quality (i.e. there are other dimensions of quality which may vary). This may induce measurement error if used to capture quality in our cost function. Second, some laboratories, although recording turnaround times, do not make efforts to reach targets as they

[73] Mutter et al. (2013) demonstrate using healthcare data that endogeneity can bias efficiency scores.
[74] As this variable is constructed using a variable that is also in the models, we check the correlation of the two variables for collinearity concerns. The correlation between the two variables is -0.34. We therefore do not see this as an issue. In any case, we note that collinearity is less an issue in panel data models than in cross-sectional or time series alternatives (Baltagi, 2008).

are not enforced. This means that this measure is likely to give a skewed reflection of this (partial) measure of quality. Third, the data completeness and validity is much lower than for the remainder of collected data (partly as some labs do not pay a great deal of attention to turnaround times).

We note that we could have also used Reference Costs data for this analysis. For two major reasons we have not. First, we do not have the allocation of capital costs in Reference Costs data. We therefore do not know if inefficiency derives from inconsistencies in the allocation of capital costs or inefficiency itself. Secondly, the only output variable available is the number of tests; this, as explained, means that inefficiency estimates are vulnerable to gaming.

Variable	Mean	S.D.	Min	Max
Operating costs (adjusted)	3617320	2058358	963875	11741895
Number of tests	5037362	2990846	1380384	30199502
Number of requests	714125	465535	191078	4423531
Input prices (Labour) (adjusted)	24551	4160	15834	49955
Number of primary care tests	2059689	932794	380790	5480395
TYPE: Metropolitan	0.27			
TYPE: Urban	0.36			
TYPE: Rural	0.37			
Foundation Trusts	0.32			
Teaching Laboratories	0.46			
REGION: London	0.17			
REGION: South	0.25			
REGION: Midlands	0.29			
REGION: North	0.29			

Table 5.3: Descriptive Statistics

5.4 Results

5.4.1 Cost Function Parameters

Across the range of models estimated (table 5.4), a number of general observations can be made. Cost elasticity with respect to output implies economies of scale (which we refer to as size – see later section) in pathology production (the first order parameters are elasticities at the sample mean; we go on to explore how these vary with output later in this section). Real operating costs appear to be decreasing over time as indicated by the negative coefficient on the time trend variable. Operating costs in pathology laboratories are higher for those which have high test to request ratios, are located in metropolitan and urban locations (relative to rural laboratories), provide teaching services and are in the Midlands (relative to the Northern laboratories). Operating costs are lower for foundation trust laboratories and for those located in London or the South (relative to the North). There was no clear finding as to the effect of GP tests on laboratory operating costs, where the effect appears negative in two models, positive in another and not statistically significant in any other.

5.4.2 Statistical Testing and Inefficiency Model Selection

Wald tests strongly and consistently favoured the translog functional form (the null being the Cobb-Douglas). Test 1 finds the s(ii) and s(iii) heterogeneity variables jointly significant additions to the models in all cases (table 5.5). Test 2 strongly favours the Cuesta model over the BC92 and P&L. Test 3 favours the Cuesta model over the TRE model[75]. Therefore our preferred inefficiency model is Cuesta s(iii) based on statistical criteria. Indeed, this model is preferred a priori because of how it deals with efficiency change over time (see section III for details). A significant lambda value (table 5.4) confirms the presence of inefficiency[76].

[75] We are aware that the Vuong test has no degrees of freedom restriction, meaning that it imposes no penalty for additional parameters estimated and so is likely to, in this case, favour the Cuesta model which has more parameters than the TRE model. Therefore, as a robustness check, we have also tested the P&L (which has fewer parameters than the TRE) against the TRE, and the test favours the P&L. Because our LR test preferred the Cuesta to the P&L, and the Vuong preferred the P&L to the TRE, we prefer the Cuesta to the TRE.
[76] In addition, we have tested the presence of inefficiency using the LR test procedure outlined in Coelli et al. (2005) pp.258, which also confirms our result, but we do not report the test results here.

Dependent Variable: OPEX	Specification														
	s(i) c = y, w, t					s(ii) c = y, w, t, z – observable					s(iii) c = y, w, t, z - observable, z - unobservable				
	Model														
	REM	P+L	BC92	CUESTA	TRE	REM	P+L	BC92	CUESTA	TRE	REM	P+L	BC92	CUESTA	TRE
PARAMETER VALUES															
CONSTANT	6.55***	6.32***	6.31***	6.31***	6.60***	6.55***	6.40***	6.39***	6.39***	6.61	6.53***	6.42***	6.42***	6.35***	6.54***
	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.06)	(9.11)	(0.04)	(0.03)	(0.03)	(0.07)	(0.00)
OUTPUT	0.43***	0.29***	0.31***	0.30***	0.74***	0.67***	0.55***	0.62***	0.35***	0.99***	0.67***	0.58***	0.64***	0.44***	0.93***
	(0.05)	(0.06)	(0.06)	(0.04)	(0.05)	(0.07)	(0.10)	(0.09)	(0.07)	(0.04)	(0.07)	(0.09)	(0.08)	(0.07)	(0.00)
INPUT PRICES	0.61***	0.52**	0.54**	0.68***	0.64***	0.64***	0.53***	0.49**	0.59***	0.84***	0.83***	0.80***	0.89***	1.30***	1.04***
	(0.21)	(0.21)	(0.22)	(0.14)	(0.14)	(0.21)	(0.20)	(0.22)	(0.13)	(0.09)	(0.22)	(0.23)	(0.24)	(0.21)	(0.01)
YEAR	-0.01**	-0.01	-0.01	-0.01**	-0.02***	-0.01**	-0.01*	0.01	-0.01	-0.01	-0.01**	-0.01*	0.01	-0.01***	-0.01***
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GP_TESTS						0.01	0.03	0.02	0.07*	-0.14***	0.01	0.02	0.03	0.06	-0.08***
						(0.05)	(0.06)	(0.06)	(0.04)	(0.03)	(0.05)	(0.06)	(0.05)	(0.06)	(0.00)
TES:REQ						0.23***	0.19**	0.21***	0.02	0.47***	0.24***	0.21***	0.21***	0.12**	0.48***
						(0.07)	(0.09)	(0.08)	(0.05)	(0.06)	(0.07)	(0.08)	(0.07)	(0.05)	(0.00)
TYPE: METROPOLITAN						0.13***	0.14***	0.14***	0.15***	0.09***	0.14***	0.15***	0.15***	0.16***	0.10***
						(0.04)	(0.03)	(0.03)	(0.05)	(0.01)	(0.03)	(0.03)	(0.03)	(0.03)	(0.00)
TYPE: URBAN						0.03	0.04*	0.04	0.05*	0.01	0.02	0.02	0.01	0.02	0.01***
						(0.03)	(0.02)	(0.02)	(0.02)	(0.01)	(0.03)	(0.02)	(0.02)	(0.03)	(0.00)
FOUNDATION						-0.06**	-0.07***	-0.07***	-0.11***	-0.06***	-0.04	-0.06**	-0.06***	-0.07***	-0.04***
						(0.03)	(0.03)	(0.02)	(0.03)	(0.01)	(0.03)	(0.03)	(0.02)	(0.03)	(0.00)
TEACHING						0.03	0.04*	0.03	0.01	0.01	0.04	0.05**	0.03	0.02	0.03***
						(0.03)	(0.02)	(0.03)	(0.03)	(0.01)	(0.03)	(0.02)	(0.02)	(0.02)	(0.00)
REGION: LONDON											-0.02	-0.05	-0.07**	-0.16***	-0.02***
											(0.04)	(0.03)	(0.02)	(0.05)	(0.00)
REGION: SOUTH											-0.03	-0.04	-0.05*	-0.01	-0.01***
											(0.03)	(0.03)	(0.03)	(0.03)	(0.00)
REGION: MIDLANDS											0.08**	0.10***	0.09***	0.10***	0.08***
											(0.03)	(0.03)	(0.03)	(0.04)	(0.00)

Table 5.4: Estimation Outputs. Standard errors in parentheses. *Notes:* *,** and *** denote significance at the 10%, 5% and 1% level, respectively. OPEX - operating expenditure.

	REM	P+L	BC92	CUESTA	TRE	REM	P+L	BC92	CUESTA	TRE	REM	P+L	BC92	CUESTA	TRE
EFFICEINCY FIGURES															
mean	0.71	0.81	0.81	0.79	0.99	0.76	0.87	0.88	0.82	1	0.77	0.9	0.9	0.87	1
s.d.	0.1	0.11	0.11	0.12	0	0.08	0.08	0.08	0.11	0	0.07	0.07	0.07	0.1	0
lambda		5.11***	5.17***	13.01***	3974.52		3.15***	3.33***	11.97***	0		2.67***	3.04***	8.38***	552028
		(1.64)	(0.03)	(0.01)	(5.71*10^7)		(0.94)	(0.05)	(0.01)	(205.77)		(0.71)	(0.05)	(0.02)	(5.69*10^7)
eta			-0.01					-0.07*					-0.11**		
			(0.02)					(0.04)					(0.05)		

Table 5.4 (cont.): Estimation Outputs. Standard errors in parentheses. *Notes:* *,** and *** denote significance at the 10%, 5% and 1% level, respectively. OPEX - operating expenditure.

LR Statistic Tests for Heterogeneity Variables: TEST 1

Model		P&L	BC92	CUESTA	TRE
Restriction of S(ii) to S(i): Observable heterogeneity variables	(d.f.: 13,13,13,12)	44.6***	48.00***	44.82***	91.04***
Restriction of S(iii) to S(ii): Unobservable heterogeneity variables	(d.f.: 3,3,3,4)	14.86***	17.70***	8.38***	38.60***

LR Statistic Tests for Model Selection (nested models only): TEST 2

		CUESTA v. P&L	CUESTA v. BC92
Specification (i): Basic Cost function	(d.f.: 57, 56)	166.84***	166.70***
Specification (ii): Observable Heterogeneity	(d.f.: 57, 56)	167.00***	163.32***
Specification (iii): Regional Dummies for Unobserved Heterogeneity	(d.f.: 57, 56)	160.52***	154.00***

Vuong Test Statistic: TEST 3

TRE specification (iii) vs. Cuesta model specification (iii)	V = -9.066***

Model Log Likelihood Function Values and degrees of freedom (K)

Model	P&L	BC92	CUESTA	TRE
Specification (i): Basic Cost function	198.80	198.97	282.22	135.81
K	9	10	66	10
Specification (ii): Observable Heterogeneity	221.13	222.97	304.63	181.33
K	22	23	79	23
Specification (iii): Regional Dummies for Unobserved Heterogeneity	228.56	231.82	308.82	200.63
K	25	26	82	26

Table 5.5: LR Specification and Model Selection

Notes: *,** and *** denote significance at the 10%, 5% and 1% level, respectively.

5.4.3 Inefficiency Predictions

From table 5.4, the mean efficiency estimate from our preferred model is 0.87. On average, efficiency is computed as decreasing slightly amongst pathology laboratories over time (which is in agreement with the BC92 s(iii) model[77] in table 5.4, given their eta coefficients) from 0.87 in 2007 to 0.86 in 2011. Fig 5.2 shows the cost efficiency estimates of laboratories over time. The bar in Fig. 5.2 is at efficiency = 1, i.e. full efficiency. Groups of points correspond to each individual laboratory, e.g. observations 1-5 are the efficiency estimates for laboratory 1 in years 1 to 5, observations 6 to 10 are laboratory 2 in years 1-5, and so on. We do not find the problem of efficiency scores dropping off the frontier in the final year of the sample, which has been a concern for other applications of this model (Wheat and Smith, 2012). In addition, we find that many of the laboratory-specific etas are statistically significant. Those that were not tended to be the firms that are on the frontier (and thus have little or no inefficiency change over time), which can be seen in figure 5.2.

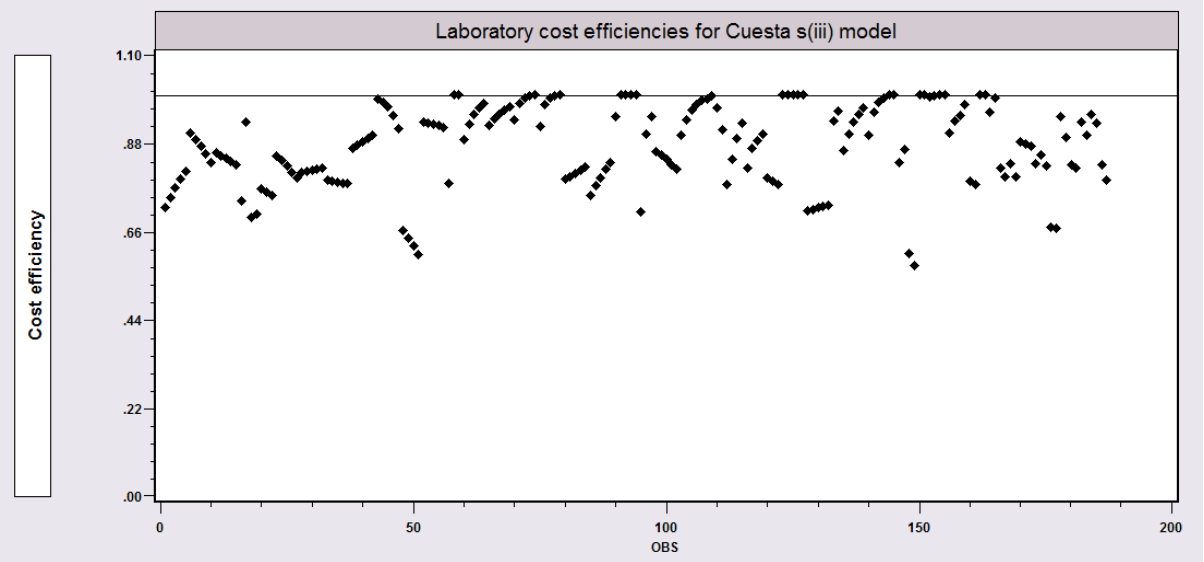

Figure 5.2: Laboratory Cost Efficiency Estimates Over Time

[77] Which is preferred of the three candidate BC92 models, see table 5.5

5.4.4 Elasticity of cost, Average and Marginal Costs

Our set of models give estimates of the elasticity of cost with respect to output at the sample mean in the range of 0.29-1.04 (table 5.4) and is 0.44 in the preferred model. However, a more informative approach is to examine how this elasticity changes with the scale of the operation, proxied by output (Fig. 5.3), using our preferred model. Using this elasticity, we are able to further estimate AC and MC per request using fitted values from the model (see Wheat and Smith, 2008, for details) (Figs 5.4 and 5.5).

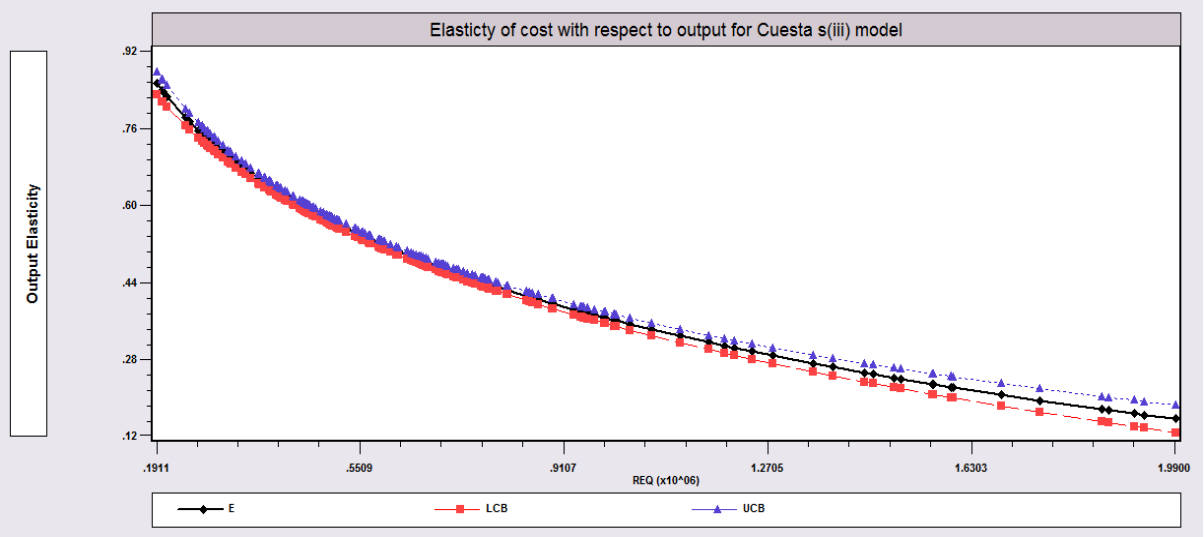

Figure 5.3: Elasticity of Cost with respect to Output for Cuesta s(iii) Model

Note to Figure 5.3: LCB – lower confidence bound, UCB – upper confidence bound. Requests are varied, all other variables are held at the sample mean.

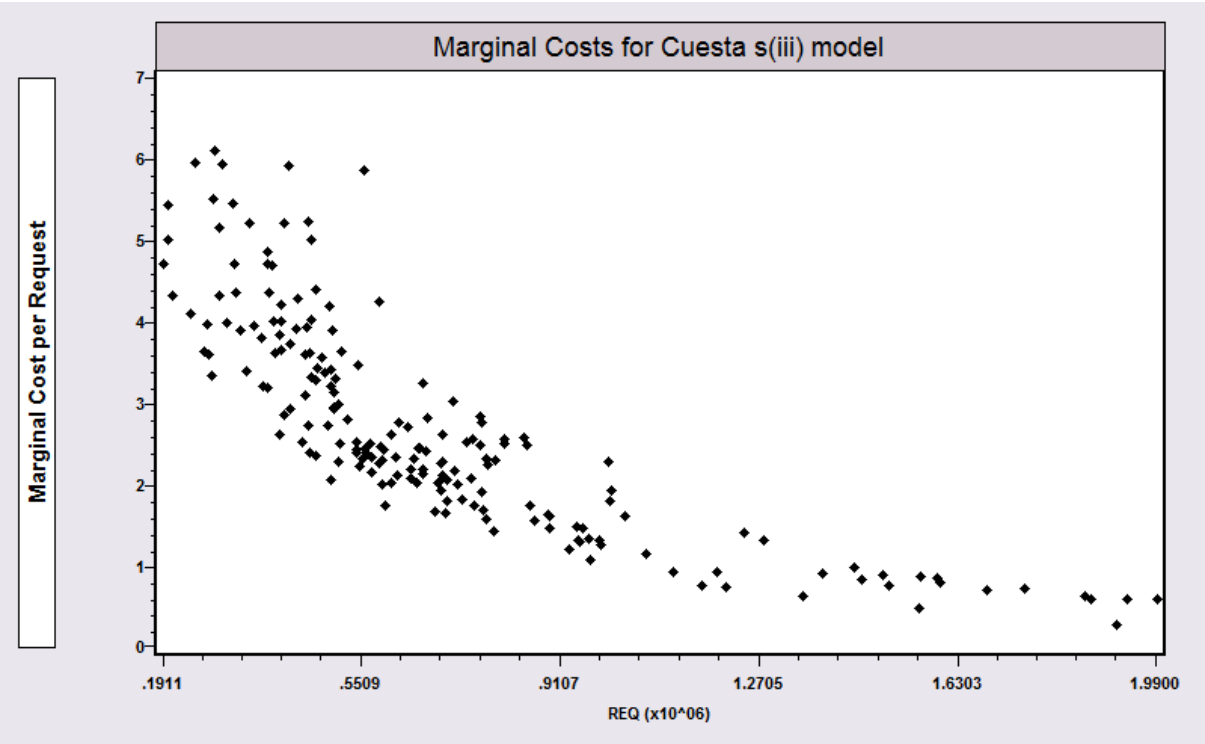

Figure 5.4: Marginal cost (MC) for Cuesta s(iii) Model

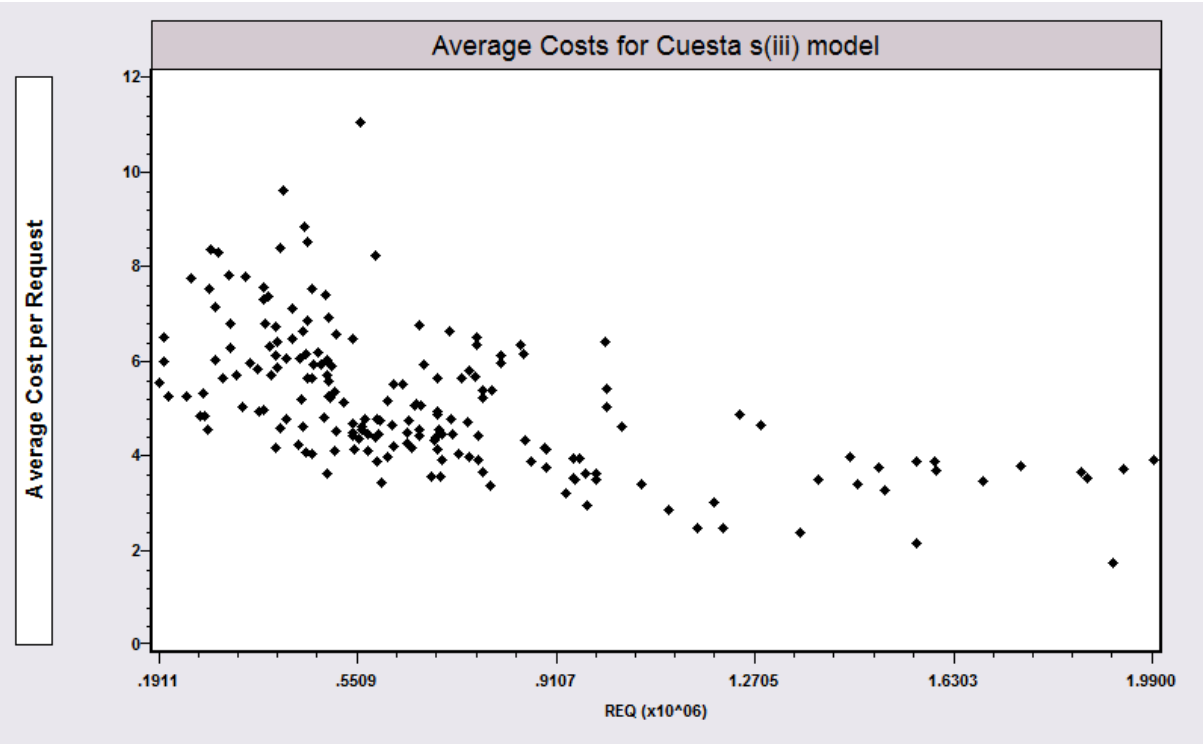

Figure 5.5: Average cost (AC) for Cuesta s(iii) Model

5.5 Discussion

5.5.1 Cost Function Parameters

This section draws on all models to examine the parameters of the cost function. The parameter estimates in the frontier models show reasonable concordance with each other and with the REM model, giving us confidence in our models.

The coefficient on input prices appears to be highly significant and in the range of 0.52-0.89, aside from two models, the Cuesta s(iii) and the TRE s(iii), which have values of 1.30 and 1.04, respectively. These estimates appear to be out of line with the remaining estimates. If the value of this coefficient was truly greater than 1, it would imply that operating costs were rising more quickly than input prices. However, we note that the 95% confidence intervals for both of these estimates include 1, meaning that we are unable to confirm that estimate of the coefficient, based on either of these models, exceeds 1. Of course, we only have data for labour input prices, and are thus unable to impose linear homogeneity of degree one on input prices, which gives rise to the possibility of beta estimates in excess of 1. We emphasise that the remaining models, including our benchmark REMs (which do not impose the distributional assumptions of the stochastic frontier models), all appear to have estimates of the coefficient on labour input prices within a plausible range. Lastly, we note that other studies have shown large labour cost shares for biochemistry operating costs - approximately 80-90% (Department of Health, 2008 pp.44). This may explain the reported coefficients.

The time trend coefficients suggest a reduction in real laboratory operating costs of 0-2% per year. The 0-2% figure can then be seen as the shifting of the frontier over time. The frontier may exhibit downward shift if, for example, productivity in pathology production is increasing, which would support the findings of Holland et al., (2012).

Moving to the observable heterogeneity parameter coefficients (s(ii) variables), there was no clear finding of the impact of GP tests (the parameter was not statistically significant). From the healthcare commission (2007), a negative coefficient value was expected because primary care tests are thought to be cheaper than other tests.

The tests to requests ratio coefficients are in line with a priori expectations (positive and less than 1) from the literature (table 5.1). The estimated elasticity from this sample is in the range

0.12-0.48. The implications depend on the interpretation of this practice – it may be considered gaming by laboratories to inflate their performance figures; on the other hand it may be a reflection of a better quality of service since more patient information is being supplied per request.

The type of laboratory is found to be a source of cost heterogeneity, which matches previous literature (table 5.1). In our analysis, we were able to investigate this issue further. Laboratories situated in metropolitan areas are on average 9-17%[78] more costly than laboratories in rural areas. The findings for urban-based laboratories are that on average they are 0-5% more costly than rural laboratories. We caveat this finding by noting that the coefficient was significant in only three of fifteen models.

The foundation status of the host trust appears to be associated with a 4-10% reduction in operating costs for pathology laboratories. From the literature, profit incentives motivate hospitals to reduce costs to a greater extent than non-profit hospitals (Sloan, 2000), which is the aim of granting foundation status to a trust and should mean pathology services act commercially (Healthcare Commission, 2007).

Lastly, laboratories which provide teaching activities are found to have higher operating costs, in the range 0-5%, to those which do not; coefficients in only three of ten models were statistically significant. This is in line with expectations, firstly because of the activity itself, but also because pathology services which are more specialised (and generally more expensive) tend to be associated with teaching activities, which may also be driving costs up (Department of Health, 2006). Moreover, this finding is in line with other health care studies (Gutacker et al., 2013).

The unobservable heterogeneity variable parameters (s(iii)) suggest that laboratories in London and the South are in the range 0-15% (statistically significant in 3 of 5 models) and 0-5% (statistically significant in 2 of 5 models), respectively, less expensive than laboratories from the North (the omitted dummy); and that operating costs of laboratories in the Midlands are on average 8-11% higher than those of laboratories in the North. From the literature, unobservable heterogeneity amongst these laboratories likely derives from information

[78] Because our model is estimated in logarithms, we have applied an exponential retransformation to recover our estimate of the effect on costs. To illustrate, for the Cuesta s(iii) model, exp(0.16) = 1.17, meaning that the beta on TYPE: Metropolitan from this model implies that costs are 17% higher than for TYPE: Rural laboratories.

systems adoption, network activity and peer contact (Department of Health, 2006; Healthcare Commission, 2007; Eijkenaar, 2013)[79].

5.5.2 Inefficiency Predictions

Our efficiency estimates are based on results from our preferred efficiency model: Cuesta s(iii).

To calculate our estimates of the potential savings we use laboratories' efficiency estimates in their final observed year. We calculate the potential cost of production if each laboratory adopted best practice (of that observed in the sample, denoted by each laboratory's efficiency estimate). Then, we subtract this estimate from the observed costs of laboratories to yield the potential available savings. We find potential savings of £32.8m in our sample (average cost efficiency in final year = 0.86).

We extrapolate to NHS pathology services (that is, all laboratories outside this sample and all other remaining pathology disciplines), giving an estimate of £390m per year of potential savings available to contribute to the Nicholson Challenge. This is around double the savings estimate that was proposed in the grey literature based on a much smaller sample – extrapolated comparably - of around £250m (Department of Health, 2008). Recalling that this data is for biochemistry services - the most mechanised of the five major pathology disciplines - we envisage that our estimates may well underestimate the true level of inefficiency, since mechanised pathology services are more homogenous than other disciplines (Kiechle and Main, 2002). We thus conclude that this is more likely a minimum efficiency saving than a maximum, which underlines the importance of pathology services for policy makers if expenditure reduction is high on their agenda.

However, driving out inefficiency may be more of a challenge amongst the more heterogeneous disciplines, such as hystocytology. First, not all laboratories conduct these services, meaning that there are fewer opportunities to compare practice and share knowledge. Second, that there is a paucity of available data in these disciplines means that measuring inefficiency may be more challenging (Buckell et al., 2013).

[79] According to anecdotal evidence from pathologists, these features are more prevalent in London and the South and thus are likely driving this variation in costs.

The average efficiency score over time is decreasing slightly. However, we find that individual etas imply that some laboratories are becoming more efficient over time, some are constant over time, and some are becoming less efficient over time (Fig. 5.2); many of the laboratory-specific etas were found to be statistically significant. Information on the efficiency profiles of the individual laboratories is a powerful output of this type of top-down benchmarking as it indicates where further attention needs to be focused to drive out efficiency improvements. As noted earlier, the approach used to model efficiency change over time has been applied in economic regulation in other sectors. We do not identify individual laboratories for confidentiality reasons.

5.5.3 Multi-Factor Productivity

Given that we have reduced efficiency over time and technical change (falling costs) as per the time trend coefficient in our preferred model (i.e. frontier shift), it is informative to compute the Total Factor Productivity (TFP) Index (Coelli et al., 2005) to give an overall account of pathology performance (chapter 3.6). However, we do not observe costs which include capital, nor an output mix effect, meaning that it would be inappropriate to describe our measure as a TFP index. We therefore define a Multi-factor Productivity (MFP) Index as our measure of overall pathology performance.

Year	Average cost efficiency	Cost efficiency index	Frontier Shift	Overall MFP Index	change MFP
2007	0.868	1	1	1	0
2008	0.839	0.967	1.014	0.981	-1.9%
2009	0.857	0.987	1.029	1.016	3.5%
2010	0.847	0.976	1.044	1.020	0.3%
2011	0.858	0.989	1.059	1.048	2.8%

Table 5.6: Multi-Factor Productivity Pathology Laboratories

As can be seen in table 5.6, the overall MFP for pathology is increasing over time, from 1.000 in 2007 to 1.048 in 2011. The annual change is positive for three of the years and negative for one year. Overall, MFP increases by 4.8% over the period of study. Thus, the small reduction in the efficiencies of laboratories away from the frontier is more than offset

by the gains in costs by the efficient firms (the frontier shift), yielding the overall MFP increase.

5.5.4 Economies of Size in Pathology

Due to our measure of costs not incorporating capital charges, we are, strictly speaking, unable to interpret changes in the relationship between output and costs as economies of scale. Accordingly, we refer to 'economies of size', and interpret this as the way in which operating costs change across the output range.

The cost elasticity estimates with respect to output indicate economies of size properties in pathology production (Fig. 5.3). Further, MC is falling faster than AC (Figs 5.4 and 5.5), meaning that the elasticity is falling (Fig 5.3), so the extent of economies of size is increasing as the scale of production increases; this will continue as long as MC falls faster than AC. This suggests that the growing formation of local pathology networks may help to lower costs for laboratories where production is pooled, which corresponds to pathology analysis elsewhere (Kiechle and Main, 2002). Encouragingly, this is being recognised by policy makers at the top level (Department of Health, 2011). It is of course possible that the economies of size be exhausted at some point, though we cannot conclude that based on our sample[80].

With regard to comparisons with previous studies, a direct comparison with the economies of scale finding in the Department of Health study (Department of Health, 2008) is difficult given that our measure does not incorporate capital costs. However, it is not clear that their measure did either, given that no empirical results on this issue are presented. Although capital cost information is collected (Department of Health, 2008, pp. 37) their only analysis (of unit costs) presented does not include these costs (Department of Health, 2008, pp. 44, 46, 48, 49). Therefore, on this issue, our study appears to be the first to present empirical evidence.

[80] We note that the AC curve appears to be flattening towards the extreme of the sample (Figure 5.5). However, given that MC remains lower than AC at this point, this must be being driven by factors other than size which are associated with higher costs when size increases. However, further research with different data would be needed to draw any conclusion on the point at which size economies are exhausted.

5.5.5Merged Laboratories

Using our preferred model, Cuesta specification (iii), and equation (5.3), we were able to simulate the effects of mergers between small laboratories. We find that, if the smaller laboratories in the sample merged, the sum of the implied predicted costs would be approximately 17% lower than those previously incurred by these laboratories separately. This suggests that there are potential considerable cost savings available via laboratories pooling production. Indeed, this estimate suggests that these potential savings are greater than those available through efficiency improvements.

While we consider this to be a useful indicative valuation of the effect of potential pooling, we attach a number of caveats to this estimate; this exercise is ultimately a stylised scenario. Firstly, we note that there is no consideration to the additional costs incurred through merging (e.g. the costs to transport samples, the costs of service delays, etc.). Second, we do not take into account any effect on the quality of the service, the effect of specialisation or the interaction with other hospital services. Thirdly, we do not consider whether these laboratories are contiguous, which could potentially be limiting to mergers. On the other hand, this estimate is based on a small number of laboratories merging, in practice there is no limit to the number that can merge. In addition, we have assumed pairwise mergers; it is, of course, possible that multiple laboratories will merge. Thus, based on the last two caveats, the potential savings could be even larger than estimated here (as long as the subadditivity of costs continues).

5.6 Conclusions

We have applied econometric efficiency estimation techniques to an under-researched area in health care literature: pathology. In doing so, we have developed performance measurement in this field beyond existing indicators benchmarking techniques. We have found, having controlled for cross-unit heterogeneity, 13% inefficiency in pathology services in the NHS in England. If this is indicative of NHS pathology as a whole, there could be £390m per year of available savings from pathology to contribute to the Nicholson Challenge of NHS efficiency savings. In addition, we found that the pooling of production looks to induce substantial gains

in pathology cost savings. If smaller laboratories merged their production, they could save around 17% in their operating costs.

We have found that overall efficiency in pathology has decreased over time. The particular method that we have adopted also allows the time paths of efficiency for individual laboratories to be studied. We have also found frontier shift which decreases costs over time. Overall, MFP for the laboratories in our sample has increased by around 5% between 2007 and 2011.

We have estimated the magnitudes of various drivers of laboratory costs which were identified from previous pathology studies. Some of these drivers have not previously been quantified (e.g. the costs of teaching or the effect of the host trust having foundation status). We have paid particular attention to the elasticity of cost with respect to output. We have found economies of size, which is encouraging from a policy perspective because local networks are being formed in pathology services which increase the scale of production. We note, however, that our measure of costs does not include a component of capital, and thus are findings are limited to this extent. We also note that, although discussed in previous studies, no empirical evidence has been presented in previous literature on this issue (Department of Health, 2006; 2008). Therefore, on this issue, our study appears to be the first to present empirical evidence.

We believe these findings are important to policy makers because it provides them with the evidence needed to make informed decisions on the allocation of resources and on the management of pathology services. The method that we have adopted highlights performance variation both between decision making units (in our case, pathology laboratories) and over time. It has been applied by economic regulators outside health as a means of driving out efficiency improvements and we consider that it also has the potential to be applied much more widely in the health sector.

We now turn to our second empirical study in which we proceed to examine some further issues raised in our review of efficiency analysis in health.

6. Dual-level inefficiency and unobserved heterogeneity in NHS pathology

This chapter is based on Smith et al. (2015). In keeping with chapter 4, we focus on pathology for a number of practical reasons: first, focussing on a speciality within health services (as opposed to broader entities such as whole hospitals or regions as the unit(s) of analysis) is more likely to be useful to policy makers and thus likely to encourage the use of efficiency predictions; and second, efficiency studies in health should target specific policy objectives: this study feeds into the policy of promoting pathology efficiency (Department of Health, 2006: and more generally, as discussed at length in opening chapters of this thesis, the NHS's 'a call to action' for efficiency improvements (NHS England, 2013).

This is the second empirical study of this thesis, and the second that is focussed on pathology services. In the first application (chapter 5), we examined several features of pathology production, including efficiency and how it changed over time; the drivers of pathology costs, including a particular focus on economies of scale in pathology; and we took an overall account of pathology performance by estimating MFP change in pathology.

This study builds on the prior chapter by examining two further aspects of pathology production, and the estimation of efficiency. These are the location of inefficiency in hierarchical organisational structures; and the incorporation of unobservable heterogeneity of varying forms into models. These issues are not confined to pathology, findings and methodological advances here are relevant more widely in health and to other sectors beyond that; indeed, the model on which our study is based was developed in transport markets (Smith and Wheat, 2012). Further, the identification of efficiency within organisational hierarchies has been identified as a key area for advancing health-based efficiency analysis (Hollingsworth and Peacock, 2008).

The remainder of this chapter is as follows. 6.1 establishes the policy context, the structure of pathology services and our economic rationale. In 6.2, our models, model features, statistical testing and estimation strategy are discussed. 6.3 details the data set used to estimate the models. 6.4 presents our results in terms of the estimated model parameters, model selection, predictions of inefficiency, implications for health policy and the implications for modelling with multi-level data sets. Section 6.5 concludes.

6.1 Introduction

An important aspect of measuring performance is being able to locate the source of inefficiency. This allows initiatives to adopt best practice to be targeted effectively. In health markets, organisations – particularly the NHS - are typified by hierarchical managerial structures where inefficiency may arise at different points within the (vertical) hierarchy as well as horizontally between organisational units at the same organisational level.

Recent health literature has begun to recognise that organisational structure should be incorporated into performance analysis (Adams et al., 2003; Olsen and Street, 2008; Sorensen et al., 2009; Castelli et al., 2013; Zhang et al., 2013). However, the previous health efficiency literature focuses its attention on horizontal comparisons, albeit at different levels of aggregation depending on the study (see Murillo-Zamorano et al., 2011; D'Amico et al., 2012; Felder et al., 2013).

In this chapter, we carry out a multi-level efficiency analysis that seeks to identify where inefficiency resides within a vertical organisational hierarchy in NHS pathology services. Pathology services are conducted in laboratories providing diagnostic medicine to primary care (local GPs) and secondary care (hospitals) within the NHS.

Pathology services are organised hierarchically, where groups of laboratories are under the direction of Strategic Health Authorities[81] (SHAs hereafter). SHAs dictate central policy, corporate culture and have some degree of control over pathology services (e.g. the configuration of services) (Department of Health, 2006); leaving some managerial autonomy at laboratory level. Thus, there is a component of overall inefficiency attributable to each SHA (which is persistent across laboratories within the SHA). Lower down in the organisation, inefficiency is likely to vary according to the relative ability of laboratory-level management. From a policy and management perspective, it is important to understand both sources of inefficiency so that appropriate incentives can be offered to drive improvements in efficiency. By combining the inefficiency estimates from the two hierarchical levels (persistent and lab-varying), an overall measure of inefficiency for the higher level (SHA) can be computed.

[81] The NHS has recently undergone a substantial reorganisation under which the SHAs have been abolished. However, they were in place during the period under study.

Measuring multi-level performance may be of greater practical use than single level measures, which should encourage their uptake amongst policy makers, which has hitherto been limited in health markets (Hollingsworth, 2012). Moreover, multi-level performance has been identified as a key future direction for health care-based efficiency analysis (Hollingsworth and Peacock, 2008).

To obtain inefficiency measures at these different organisational levels, and an overall inefficiency measure, we adopt the dual-level stochastic frontier model (DLSF; see Smith and Wheat (2012)), which has been applied in other sectors to measure multi-level firm inefficiency. The advantage of this model is firstly that it enables inefficiency at different organisational levels to be identified. Smith and Wheat (2012) use the terminology sub-company, or internal inefficiency, which in our case corresponds to inefficiency at the laboratory level; and persistent or external inefficiency, which in our case refers to persistent inefficiency at the SHA level. A further key finding of their paper is that, when the organisational structure is not accounted for, inefficiency predictions can exhibit a downward bias. Thus there is motivation in adopting a DLSF model both to yield insight into the level of inefficiency variation at different levels and to eliminate bias in the overall prediction.

Another form of bias, which is particularly problematic in health, results from the failure to appropriately model unobservable heterogeneity (Greene, 2004; Smith et al., 2012). In the case of pathology services, there are significant differences in laboratories' production processes. These may include factors typically studied by economists such as outputs and input prices; but also specificities such as patient mix or service quality, inter alia (Department of Health, 2006; Buckell et al., 2013; NHS England, 2014b; Buckell et al., 2015). Some of these features are difficult or even impossible to measure directly, and so accounting for unobservable heterogeneity is of paramount importance.

Smith and Wheat (2012) recognise that persistent inefficiency at the higher level of aggregation (which corresponds, in our case, to SHA level inefficiency) could also reflect time and laboratory invariant unobserved heterogeneity. However, they leave that issue for future research. At the same there has been considerable interest in the wider panel data stochastic frontier literature (Farsi et al., 2005a; Kumbhakar et al., 2014) on how to separate inefficiency from unobserved, time invariant heterogeneity. We therefore augment the Smith and Wheat (2012) DLSF approach to reflect developments in the wider panel data efficiency literature with regard to the vexing problem of disentangling inefficiency from unobserved

heterogeneity. We compare our findings to a model without any attempt to separate inefficiency and unobservable heterogeneity, to demonstrate the importance of accounting for the latter case of multi-level data structures.

This chapter therefore contributes to the literature in two ways. It is the first application of the Smith and Wheat (2012) DLSF inefficiency model in a health context. It is also the first time the approaches set out in Farsi et al. (2005a) and Kumbhakar et al. (2014) have been applied to a multi-level data structure. We thus apply and develop state-of-the-art models to draw policy conclusions on pathology services within the NHS in England, and also offer insights on the relative merits of different approaches to separating inefficiency and unobserved heterogeneity when applied to multi-level data structures.

6.2 Methods

We begin our methodological discussion with the general form of the dual-level stochastic frontier proposed by Smith and Wheat (2012). We next discuss the issue of unobservable heterogeneity and its relevance to efficiency estimation in our application. We then outline two further models, each of which adds a component to the model to distil out the unobservable heterogeneity from the efficiency prediction, though with differing assumptions. We finally consider a fully generalised model comprising the features of the preceding models. In total, four models are considered. Estimation, econometric specification and statistical testing are described.

Our starting point is the dual-level stochastic frontier (DLSF) model proposed by Smith and Wheat (2012). This model is derived from panel data stochastic frontier models, with the exception that the structure of the panel is amended from firm and time to firm and sub-company, where the sub-company units are repeat observations of their respective firms. In this way, the structure of the organisation is embodied in the model. This allows the decomposition of inefficiency at the two organisational levels in the hierarchy. In this application, SHAs are equivalent to firms, and laboratories are equivalent to the sub-company units.

Smith and Wheat (2012) outline the advantages of this model and its application to multi-level data structures. First, multi-level data structures increase the number of observations for analysis, which can be a major benefit for economic regulators who often have to work with small cross-sections and limited time periods. Second, it permits a clearer understanding of where inefficiency resides in the vertical hierarchy, allowing regulators to target the elimination of persistent differences between SHAs (external inefficiency) and differences in performance of laboratories within the same SHA (internal inefficiency). Finally, it is beneficial to conduct performance analysis at the level of disaggregation that relates to how SHAs/laboratories actually organise themselves, in particular allowing the true scale properties of the cost function to be established.

The imposed form of inefficiency is well suited to the multi-level model. Smith and Wheat (2012) note that, in traditional panels, having an overall inefficiency comprising a component of SHA inefficiency that is time-invariant and a component that varies randomly over time may not accurately capture the natural temporal evolution of inefficiency. In contrast, imposing a SHA-invariant component and a laboratory-varying component to the structure of inefficiency befits the aim of vertically decomposing inefficiency.

The DLSF model takes the general form,

$$C_{i,s} = X_{i,s}'\beta + \delta_i + \varepsilon_{i,s} \quad (6.1)$$

$$\varepsilon_{i,s} = \tau_{i,s} + v_{i,s} \quad (6.2)$$

$$\tau_{i,s} \sim N^{+}(0, \sigma_{\tau}^{2}) \quad (6.3)$$

$$v_{i,s} \sim N(0, \sigma_{v}^{2}) \quad (6.4)$$

Where $C_{i,s}$ is the cost of laboratory s in SHA i. $X_{i,s}$is a vector of outputs, input prices and environmental variables; β is a vector of parameters to be estimated. δ_i is the SHA-specific effect. τ_{is} is laboratory-specific inefficiency and v_{is} is random statistical noise. The notation in (6.1) highlights the tiered structure of the data only; in the empirical work presented below, there is also a time dimension to the data.

Estimation proceeds via estimation of a SHA-stratified random effects model (REM) by Generalised Least Squares (GLS) (as in equation (6.1)), yielding estimates of β ($\hat{\beta}$), predicted

values of SHA effects (to which we turn our attention in the following sections), and residuals, $\hat{\varepsilon}_{i,s}$.

The prediction of laboratory-specific inefficiency is conducted in a second stage. It is common for all four models. We take the model residuals from the first stage (which have had the SHA effect removed), stratify by laboratory and apply the Jondrow et al. (1982) procedure to retrieve laboratory-specific predictions of inefficiency,

$$\hat{\tau}_{i,s} = E\left[\tau_{i,s} \middle| \tau_{i,s} + v_{i,s}\right] \tag{6.5}$$

We assume time-invariance for the predicted efficiency at laboratory level, given that our panel is both short in its time dimension and unbalanced. The competing models are then distinguished according to the treatment of the SHA-specific effect, δ_i.

6.2.1 The Dual-Level Stochastic Frontier (Model 1)

The DLSF treats the SHA-specific effect as inefficiency, which in the case of GLS estimation yields,

$$\delta_i = \alpha_0 + \mu_i \tag{6.6}$$

$$\mu_i \sim N\left(0, \sigma_\mu^2\right) \tag{6.7}$$

Where the prediction of SHA inefficiency is a Schmidt and Sickles (1984)-type correction, $\hat{\mu}_i = \hat{\mu}_i - \min(\hat{\mu}_i)$.

6.2.2 Accounting for Unobservable Heterogeneity

A simplifying assumption of the DLSF proposed by Smith and Wheat (2012) was that the SHA effect is interpreted as the SHA inefficiency. This is consistent with the received literature such as Kumbhakar and Heshmati (1996). However, Smith and Wheat (2012) acknowledged that this interpretation may not be appropriate in all cases. In particular, any heterogeneity that is not captured by the regressors is incorporated into this effect, which

biases inefficiency estimates (Kumbhakar and Lovell, 2000). Ultimately, the SHA effect is a mixture of unobservable effects, one of which being SHA invariant inefficiency.

In the case of pathology production, there are features of laboratories' production environments for which no data are available, e.g. the service quality (which is known to vary between laboratories and SHAs), implying the DLSF may be an inappropriate specification. As such, the DLSF model is extended to examine two approaches to incorporate the influence of unobservable heterogeneity, namely the use of the Mundlak (1978) transformation and the residual decomposition approach of Kumbhakar et al. (2014). In addition, we estimate a model which incorporates both of these approaches. We utilise statistical testing to determine an appropriate approach. Results are compared between models to demonstrate differences.

6.2.3 The Mundlak-Transformed DLSF (Model 2)

One way to introduce a control for unobservable heterogeneity into the DLSF model follows Farsi et al. (2005a), which was first extended to the DLSF by Wheat (2014) and subsequently implemented on a railways dataset in Smith and Wheat (2014). The approach makes use of Mundlak's (1978) recognition of the link between random and fixed effects in panel data models. This approach is operationalised via a direct insertion of group means of the regressors into the random effects model[82]. In this way, this model nests model 1.

This model assumes that inefficiency is uncorrelated with the regressors whilst unobserved heterogeneity is assumed to be correlated with the regressors. Correlation between the SHA effects and the regressors is modelled explicitly by using the variable group means. Under the assumption that this correlation represents unobservable heterogeneity, it is removed from the SHA effects. Then the SHA effects that remain are treated as before and efficiency predictions are derived.

$$\delta_i = \alpha_0 + \bar{X}_i'\rho + \mu_i \tag{6.8}$$

$$\mu_i \sim N\left(0, \sigma_\mu^2\right) \tag{6.9}$$

[82] There is an alternative approach using a fixed effects model and an auxiliary regression on the SHA effects (Farsi et al., 2005a). In linear models, this method returns identical parameter estimates, but underestimates standard errors in the auxiliary stage, so the random effects approach is preferred (see Baltagi, 2006, pp.1192, for the variance of the group means in the REM).

Here, $\bar{X}_i'\rho$ captures unobservable heterogeneity that is correlated with the regressors. SHA inefficiency predictions, $\hat{\mu}_i$, are: $\hat{\mu}_i = \hat{\mu}_i - \min(\hat{\mu}_i)$.

Model 2 has a number of appealing features. First, the separation of inefficiency from unobservable heterogeneity (that is correlated with the regressors) is achieved. Second, consistent, unbiased within estimators for the frontier parameters are recovered through application of GLS to this model[83]. Third, it is possible to examine the relationship between the unobservable heterogeneity and the variables via the group mean coefficients ($\hat{\rho}_{GLS}$) (Farsi et al., 2005a; Farsi et al., 2005b). Fourth, this model does not require any additional stages; the model is estimated exactly as the DLSF with the addition of the group mean variables. Fifth, the restriction (no correlation between the regressors and unobserved heterogeneity) can be readily tested using a Wald test on the joint hypothesis: $\rho = 0 \; \forall \; \bar{X}_i$ (which is referred to as the Wu test (Greene, 2008)).

There are some drawbacks to using this method. First, the model relies on the assumption that the unobservable heterogeneity is correlated with the regressors while inefficiency is assumed to be completely uncorrelated with regressors. Thus any unobservable heterogeneity that is uncorrelated with regressors is interpreted as inefficiency and, conversely, any inefficiency correlated with regressors (but firm invariant) is interpreted as unobserved heterogeneity. Finally, relative to the simpler DLSF, the Mundlak transformation proliferates parameters, which will reduce the precision of parameter estimates.

6.2.4 The Four-Component DLSF (Model 3)

A second approach to amend the DLSF to account for unobservable heterogeneity is to follow the approach of Kumbhakar et al. (2014) based on their four-component model. The application to our hierarchical data is similar to model 1, except for an additional stage to separate the firm inefficiency from the unobservable heterogeneity (the latter now assumed uncorrelated with the regressors). In this way, model 3 nests model 1.

In this additional stage, the SHA effects are decomposed by imposing distributional assumptions and applying a stochastic frontier to them. Thus, unobservable heterogeneity is assumed to embody the features of statistical noise in traditional stochastic frontiers (SFs) (equations (10)-(13) below). Unobservable heterogeneity is assumed to be uncorrelated with

[83] We note that within estimators are in some cases imprecise, which to some extent diminishes their appeal

the regressors. This is in direct contrast to the Mundlak approach (Wheat, 2014). Here, SHA inefficiency is computed using the Jondrow et al. (1982) method, rather than the Schmidt and Sickles (1984) approach used in models 1 and 2.

$$\delta_i = \alpha_0 + \mu_i \tag{6.10}$$

$$\mu_i = \alpha_i + w_i \tag{6.11}$$

$$\alpha_i \sim N^+(0, \sigma_\alpha^2) \tag{6.12}$$

$$w_i \sim N(0, \sigma_w^2) \tag{6.13}$$

Where w_i represents unobserved heterogeneity that is uncorrelated with the regressors and inefficiency is calculated as: $\hat{\alpha}_i = E[\alpha_i | \alpha_i + w_i]$.

The benefits of this model are that, firstly, it is possible to control for unobservable heterogeneity. Second, it is possible to test the decomposition of the inefficiency and the unobserved heterogeneity by applying routine tests in the SF literature. Third, although full distributional assumptions are made to predict inefficiency, the parameter estimates of the frontier are estimated using much weaker (and thus robust) assumptions in the first stage, which is a noteworthy advantage over a single stage alternative (Smith and Wheat (2012)).

There are disadvantages to implementing this model. First, relative to the simple DLSF, there are additional assumptions on the error components necessary to enable separation of inefficiency from unobserved heterogeneity, and these are arbitrary. Second, the SF procedure to obtain SHA persistent inefficiency predictions is conducted on the number of SHAs, which may be small in empirical applications (in our case 10); and, in turn, may yield imprecise parameter estimates, particularly with respect to the variances of the SHA invariant error components. In traditional panels it is also the case that this part of the procedure faces limitations if the cross-section is small. In addition, the multi-stage approach yields standard errors of second stage parameter estimates smaller than their true magnitude owing to the use of first stage residuals in the second stage (Kumbhakar et al. (2014) note that this issue is routinely disregarded). However the fundamental limitation of this approach is the assumption that unobserved heterogeneity is uncorrelated with the regressors, which in turn requires reliance on distributional assumptions to separate inefficiency from the unobserved heterogeneity.

6.2.5 The Mundlak-Transformed Four-Component DLSF (Model 4)

Our final model is a DLSF that is augmented for unobservable heterogeneity by combining the three approaches above. In this model, inefficiency is purged of both types of unobserved heterogeneity, that is, unobserved heterogeneity that is correlated with the regressors, and that which is not. Thus the appeal of this specification is that the somewhat restrictive assumptions about the correlation between unobservable heterogeneity and the regressors in the two prior approaches can be (a) relaxed and (b) tested. We therefore specify the following,

$$\delta_i = \alpha_0 + \bar{X}_i'\rho + \mu_i \tag{6.14}$$

$$\mu_i = \alpha_i + w_i \tag{6.15}$$

$$\alpha_i \sim N^+(0, \sigma_\alpha^2) \tag{6.16}$$

$$w_i \sim N(0, \sigma_w^2) \tag{6.17}$$

Where $\bar{X}_i'\rho$ captures unobservable heterogeneity that is correlated with the regressors and $w_i$represents unobserved heterogeneity that is uncorrelated with the regressors. Inefficiency is calculated as: $\hat{\alpha}_i = E[\alpha_i|\alpha_i + w_i]$.

Model 4 nests its component models - it is possible to test down to arrive at a preferred model. In particular, it is possible to test each of the components individually, and examine the presence and/or form of unobservable heterogeneity, and to remove it from the estimates of inefficiency.

Overall, four models are estimated and tested: the dual-level stochastic frontier (DLSF) of Smith and Wheat (2012) (Model 1); the DLSF with the Mundlak adjustment applied (Model 2); the four-component DLSF model based on Kumbhakar et al. (2014) (Model 3); and the Kumbhakar-DLSF model with the Mundlak adjustment applied (Model 4). Table 6.1 below shows the econometric specifications of these models.

Model	Stage 1: RE GLS	SHA inefficiency	Laboratory (sub-company) Inefficiency
(1)	$C_{i,s} = \alpha_0 + X_{i,s}'\beta + \varepsilon_{i,s}$	$\hat{\mu}_i = \hat{\mu}_i - \min(\hat{\mu}_i)$	$\hat{\tau}_{i,s} = E[\tau_{i,s}\|\tau_{i,s} + v_{i,s}]$
(2)	$C_{i,s} = \alpha_0 + \bar{X}_i'\rho + X_{i,s}'\beta + \varepsilon_{i,s}$	$\hat{\mu}_i = \hat{\mu}_i - \min(\hat{\mu}_i)$	$\hat{\tau}_{i,s} = E[\tau_{i,s}\|\tau_{i,s} + v_{i,s}]$
(3)	$C_{i,s} = \alpha_0 + X_{i,s}'\beta + \varepsilon_{i,s}$	$\hat{\alpha}_i = E[\alpha_i\|\alpha_i + \omega_i]$	$\hat{\tau}_{i,s} = E[\tau_{i,s}\|\tau_{i,s} + v_{i,s}]$
(4)	$C_{i,s} = \alpha_0 + \bar{X}_i'\rho + X_{i,s}'\beta + \varepsilon_{i,s}$	$\hat{\alpha}_i = E[\alpha_i\|\alpha_i + \omega_i]$	$\hat{\tau}_{i,s} = E[\tau_{i,s}\|\tau_{i,s} + v_{i,s}]$

Table 6.1: Econometric Specifications of Models 1-4

From table 6.1, we note that stage 1 is identical for models 1 and 3; and for models 2 and 4. In model 1 and model 2, the predicted SHA inefficiencies are derived from $\hat{\mu}_i$. In models 3 and 4, the SHA effects are decomposed to yield inefficiency predictions according to the distributional assumptions specified.

Laboratory inefficiency predictions for models 1 and 3 are identical as a corollary of the common first stage. Similarly, models 2 and 4 have identical predicted laboratory inefficiencies.

We now turn the choice between models 1-4. We are able to use statistical tests to guide model selection. Table 6.2 summarises our model testing. We first test the SHA effects using a Moulton-Randolph test (a Standardised Lagrange Multiplier test (SLM)), which is better suited to unbalanced panels (as in our panel) than the standard LM test (Moulton and Randolph, 1989). We then move to testing the decomposition of inefficiency and unobservable heterogeneity. The unobservable heterogeneity that is correlated with the regressors is testing using a Wald test on the group mean variables jointly. This test is applied to models 2 and 4. To test unobservable heterogeneity that is uncorrelated with the regressors, we use a LR test on the SHA SF. These tests apply to models 3 and 4. Finally, the test of the

presence of inefficiency at the laboratory level is tested using a LR test on the laboratory level SF.

	Model 1	Model 2	Model 3	Model 4
Test of firm effects				
Firm effects (vs. pooled model)	Moulton-Randolph H_0: no firm effects	Moulton-Randolph H_0: no firm effects	Moulton-Randolph H_0: no firm effects	Moulton-Randolph H_0: no firm effects
Decomposition				
Inefficiency and UOH correlated with regressors		Wald test on $\bar{X}_i$ H_0: $\rho = 0 \; \forall \; \bar{X}_i$		Wald test on $\bar{X}_i$ H_0: $\rho = 0 \; \forall \; \bar{X}_i$
Inefficiency and UOH uncorrelated with regressors			LR of firm effect SF H_0:no inefficiency	LR of firm effect SF H_0:no inefficiency
Test of laboratory inefficiency				
Test of sub-company inefficiency	LR on sub-company SF H_0:no inefficiency	LR on sub-company SF H_0:no inefficiency	LR on sub-company SF H_0:no inefficiency	LR on sub-company SF H_0:no inefficiency

Table 6.2: Statistical Tests on Models 1-4. UOH – Unobserved heterogeneity, LR - likelihood ratio, SF – stochastic frontier

6.2.6 Overall Inefficiency

Finally, having retrieved the two efficiency predictions at the separate hierarchical levels, it is necessary to compute an overall efficiency for the SHA – our persistent, top-level inefficiency measure - which is the sum of its SHA-specific inefficiency and the (cost) weighted average of its constituent laboratories' inefficiencies. We use this measure to compute our overall savings estimates. Taking model 1 as an example,

$$\bar{u}_i = \hat{\hat{\mu}}_i + \frac{\sum_{\forall s} C_{i,s} \cdot \hat{\tau}_{i,s}}{\sum_{\forall s} C_{i,s}} \tag{6.18}$$

6.3 Data

Annual pathology benchmarking data is used to compile an unbalanced panel of 57 English NHS pathology laboratories amongst 10 Strategic Authorities during the 5 year period from 2006/7 to 2010/11. The sample represents approximately one third of the 163 NHS pathology laboratories in England.

Our dependent variable is the laboratory's total operating costs (net of capital charges).

Output is measured by the number of requests for tests. We could, of course, use the number of tests actually carried out as our output measure. However, laboratories are known to conduct varying numbers of tests per request, which may distort the measure of output if it is based on tests. We further capture this variation by including a variable defined as the ratio of tests to requests (variable name Tests:Requests), in addition to our output measure. Input prices for labour are based on data from the UK labour force survey. Labour force survey data is chosen over other sources (NHS staff census data, for example) to ensure the exogeneity of the data[84]. In the absence of other input prices data, this variable is considered a proxy for labour and materials.

Variables capturing exogenous characteristics include: a binary variable for the foundation status of the host trust[85], meaning that it has financial autonomy (variable name Foundation).

[84] Mutter et al. (2013) demonstrate using healthcare data that endogeneity can bias efficiency scores.
[85] The term 'trust' in the NHS refers to a single hospital or a small group of hospitals in close proximity (e.g. in an urban area) which operate as a single entity.

It is expected that foundation status trusts will have lower operating costs than their non-foundation counterparts owing to a more commercial outlook towards service provision (Healthcare commission, 2007). We also include a binary variable (variable name Metropolitan) denoting within an urban area or city; the null case is rural. This is to capture the differences in service provision between rural and urban patient populations and their differing pathology demands, e.g. a broader range of diseases in larger cities (Department of Health, 2006).

As in chapter 5, we neglect the use of Reference Costs data for reasons of capital cost allocation and lack of key variables.

Descriptive statistics are presented in table 6.3. Costs and wage data are in real terms (2007 prices), adjusted using the consumer prices index (CPI). The ratio of tests to requests is calculated from the data, as are variable group means for the Mundlak transformation. For estimation, natural logarithms of variables are taken. We use a Cobb-Douglas functional form[86]. LIMDEP software is used for estimation (Greene, 2012).

Variable	Mean	S.D.	Min	Max
Operating costs (adjusted)	3617320	2058358	963875	11741895
Number of tests	5037362	2990846	1380384	30199502
Number of requests	714125	465535	191078	4423531
Input prices (Labour) (adjusted)	24551	4160	15834	49955

Table 6.3: Descriptive Statistics

[86] We tested a Translog specification, however, the coefficients on some key variables were not significant. Therefore, we prefer a Cobb-Douglas specification which gives a credible set of parameter estimates, and a more credible model from which our efficiency predictions are derived.

6.4 Results

In this section our results are summarised and discussed. We begin with our parameter estimates from the first stage of models 1-4. Next, we discuss model selection and select our preferred model. We then move to the efficiency predictions and our savings estimates. Finally, we comment on the health policy and wider modelling implications of our empirical results.

6.4.1 Cost Function Parameters

Models 1-4 use a random effects model as the first stage in estimation. Models 2 and 4 extend the model with the Mundlak group mean variables. Therefore, two model outputs are reported: one with a Mundlak adjustment (models 2 and 4), and one without a Mundlak adjustment (models 1 and 3). Table 6.4 reports the model outputs.

	REM with Mundlak Model 2 || Model 4			REM without Mundlak Model 1 || Model 3		
	Beta	s.e.	Sig	Beta	s.e.	Sig
Constant	1.285	5.497		-5.833	1.712	***
Output (requests)	0.897	0.043	***	0.897	0.043	***
Input prices	0.892	0.161	***	0.774	0.153	***
Tests:Requests	0.549	0.066	***	0.547	0.069	***
Metropolitan	0.196	0.046	***	0.198	0.047	***
Foundation	-0.065	0.041		-0.081	0.041	**
Time	-0.021	0.012	*	-0.019	0.013	
REQBAR	-0.334	0.194	*			
INPBAR	-0.287	0.451				
TESBAR	-1.070	0.552	*			
METBAR	0.097	0.203				
FOUBAR	0.222	0.169				
TIMBAR	0.234	0.130	*			

Table 6.4: Model Outputs for Mundlak Adjusted and non-Mundlak Adjusted Random Effects Models. *, **, *** denote statistical significance at the 10%, 5% and 1% level, respectively. s.e. – standard errors. The Mundlak group mean variables are denoted "XXXBAR" and correspond to their respective variables above.

Table 6.4 shows the parameter estimates from both of the first stage models. The $\hat{\beta}$ are similar between models and similar to findings in other studies of pathology services (Buckell et al., 2013; Buckell et al., 2015)[87]. The within estimators do not appear to exhibit imprecision (which was a concern of adopting this approach, see section 6.2).

In both models the output coefficients are positive and significant, suggesting, at the sample mean, increasing returns to scale (RTS) properties in pathology production (since RTS = $1/\hat{\beta}_{output}$ = 1/0.897 = 1.115). This corresponds to results and/or predictions from other pathology studies (Department of Health, 2006; 2008; Healthcare Commission, 2007; Holland et al., 2011; Buckell et al., 2013).

We find that laboratories facing higher input prices have higher costs; that laboratories with higher tests-to-requests ratios have higher operating costs; and that laboratories in urban settings have higher operating costs (coefficient on the Metropolitan variable), which is in agreement with other pathology studies (Department of Health, 2006). The within estimator suggests that the foundation variable is not significant, whilst this variable is found to be statistically significant at the 5% level in the REM without Mundlak. The study of the Healthcare Commission (2007) suggested that the foundation of the host trust may lead to lower operating costs, although no empirical results were presented.

The coefficient on the time variable, representing technical change (frontier shift), is significant only in the REM with Mundlak (the within estimator). The coefficient suggests that pathology costs are, on average across the market, decreasing annually by around 2% owing to technical change. This finding is in keeping with the empirical findings of Holland et al. (2011). Moreover, this result is intuitively sound, as a heavily mechanised industry such as pathology is likely to be characterised by technological change over time, leading to cost reductions, even in the short run (as in this data).

We now discuss the Mundlak group mean coefficients. There appears to be divided opinion in the literature as to their interpretation individually, although most authors do not comment on them in isolation. Of those that do, Farsi et al. (2005a) and Farsi et al. (2005b) take the view that the group means indicate correlation between the variable and unobservable heterogeneity. Conversely, Filippini and Hunt (2012) state that the interpretation of these variables is not straightforward, and do not assign any interpretation to these coefficients. In

[87] We note that similar data is used for these studies so this result is not surprising

our application, we are interested in the decomposition of efficiency and unobservable heterogeneity, thus the interpretation of these variables is of no specific interest to us.

Overall, the Mundlak-transformed model is considered a better reflection of the economic reality than its non-transformed counterpart on a priori grounds as it permits inefficiency estimates to be purged of unobserved heterogeneity that is correlated with the regressors. We discuss model selection based on appropriate statistical testing below.

6.4.2 Model Selection

We now move to our discussion on model selection. To begin, we consider the testing procedure outlined in table 6.2. We discuss the results from these tests, which are reported in table 6.7. We also draw on the model efficiency predictions, which are presented in table 6.5.

The first issue is whether the multi-level structure is appropriate. From the significant Moulton-Randloph statistic, the panel specification of the first stage formulation is preferred to the pooled model, supporting the presence of SHA effects. Of course, as noted above, we then need to consider the interpretation and decomposition of these SHA effects.

For all models, the LR statistic on the laboratory level SF is significant, supporting the presence of inefficiency at laboratory level.

We now turn to the unobservable heterogeneity test statistics. The Wald test of 6 linear restrictions – the Wu test - indicates that the variable group means are jointly statistically significant additions to the model[88]. There is thus evidence to support the correlation between the SHA effects and the regressors, which we interpret as unobservable heterogeneity. On this basis, we prefer model 2 to model 1 and model 4 to model 3.

As expected, model 1 appears to confound unobservable heterogeneity with inefficiency. This issue is well known in the health context (Greene, 2004; Farsi et al., 2005a).When the Mundlak adjustment is applied, the average predicted efficiency increases significantly from 0.625 to 0.715 (table 6.5). This finding, combined with the results of the Wu test, suggests that there is a substantial amount of unobservable heterogeneity that is correlated with the regressors.

[88] We have also used the more familiar Hausman test. In this case, however, the test statistic could not be computed because the variance-covariance matrix is not positive definite. We thus revert to the Wu test (Greene, 2012b) and note that, in any case, reliance on the Hausman statistic alone is discouraged (Baltagi, 2008).

Model 3 is unable to detect any inefficiency at the SHA level (table 6.5) – the SHA effects exhibited wrong skew. As noted, the Wu test result suggests that there is a high amount of unobservable heterogeneity that is correlated with the regressors, which model 3 does not allow for. Therefore, the finding of zero inefficiency is likely more a matter of model misspecification than of economic reality. This finding suggests that controlling for unobservable heterogeneity that is correlated with the regressors is vital: had we estimated only models 1 and 3, we might have concluded that there is no inefficiency at the SHA level and that SHA effects were driven by heterogeneity. We therefore prefer model 2 to models 1 and 3.

The final model selection decision is then a choice between model 2 and model 4. This choice hinges on the result of the attempt to decompose the SHA effect into inefficiency and unobserved heterogeneity that is correlated with the regressors (stage 2 in model 4). Although inefficiency was detected at the SHA level in model 4 (which was not the case in model 3), the result was not statistically significant (table 6.7). The conclusion, at face value then, is that once purged of unobserved heterogeneity (correlated and uncorrelated with the regressors) there is no statistically significant SHA-level inefficiency.

However, we note that stage 2 of the multi-stage approach is based on only 10 observations (as we have only 10 SHAs). As a result, the failure to find inefficiency in this model is unsurprising (this is likely to be an issue for this model on any dataset, like ours, where the number of firm observations is low).

We further note a striking concordance between the predicted SHA efficiencies of models 2 and 4 with respect to rank (Kendall's tau = 0.600**, see table 6.5 for ranks), absolute correlation (=0.92[89]) and mean predicted efficiency (model 2 = 0.921; model 4 = 0.944, see table 6.5). So, although the inefficiency effects are not statistically significant when making the final decomposition of inefficiency and unobserved heterogeneity (uncorrelated with the regressors), the inefficiency predictions and ranks are scarcely affected. It appears that much of the unobservable heterogeneity is correlated with the regressors and it is then difficult to disentangle the remaining effect.

Overall, we conclude that there is some remaining inefficiency at the SHA level and in the discussion that follows, we use model 4 as our preferred model. This is on the grounds that it

[89] Farsi et al (2005a) suggest, as a rule of thumb, any score greater than 0.9 can be considered as similar; our result is well in excess of this.

takes account of unobserved heterogeneity that is uncorrelated with the regressors, noting that results are very similar if we were to revert to model 2.

6.4.3 SHA, Laboratory level and Overall Efficiency Predictions

Table 6.5 shows the efficiency predictions from the four models. For each model there are four columns corresponding to the SHA-specific efficiency, the laboratory-specific efficiency, the overall efficiency and the rank of the SHA in terms of its overall efficiency. For the first three columns, the means of the predicted efficiencies and corresponding standards deviations are provided.

Table 6.6 shows the rank correlations between the predicted overall efficiencies for models 1-4[90]. As can be seen, there is very little concordance between almost all of the models' predicted ranks. This is not entirely surprising given that model 1 makes significantly different assumptions to the remaining models and that model 3 failed to recognise any inefficiency at the SHA level. The exception to the trend is that the predicted ranks of model 2 and model 4 are statistically significantly correlated.

Model 1 exhibits the lowest predicted efficiency with a mean overall efficiency of 0.625. This is as expected given that, by construction, this model makes no allowance for the effect of unobservable heterogeneity on efficiency prediction. Thus, the unobservable heterogeneity is encompassed in the inefficiency component of the model. This issue is well known in the health context (Greene, 2004; Farsi et al., 2005a).

In model 2, it is assumed that unobservable heterogeneity is correlated with the regressors. As such, we are able to use the procedure outlined in section 2 to remove it from the SHA effects. Here, the mean overall efficiency increases significantly to 0.715.

In model 3, the unobservable heterogeneity is assumed to be uncorrelated with the regressors and assumed to embody a set of assumptions (section 6.2). In this application, the SHA effects that had a SF applied to them (stage 2 of the multi-stage approach) exhibited wrong skew. Thus, no inefficiency was detected at the SHA level; that is, the firm effect is entirely composed of unobservable heterogeneity. In this sense, model 3 predicts the highest SHA efficiency. As noted, we believe this model to be misspecified.

[90] We use Kendall's tau to measure rank correlation, which is well suited to small samples (Kendall & Gibbons, 1990).

Model 4 combines both the assumptions and procedures of the preceding three models: unobservable heterogeneity is assumed to be, in part, correlated with regressors and, in part, uncorrelated with the regressors.

As can be seen, as expected, the predicted mean overall efficiency in model 4, 0.732, is higher than that of model 2, 0.715. The difference is slight in contrast to the predictions of model 1 versus the predictions of model 2, suggesting that there is less unobservable heterogeneity that is uncorrelated with the regressors than that which is correlated with the regressors. This indicates that the Mundlak adjustment appears to capture almost the full extent of the unobservable heterogeneity. However, there was a small difference between the predicted efficiency ranks and SHA efficiencies, suggesting that the additional control is worth retaining.

There is a more fundamental point when comparing model 3 with models 2 and 4, which is that there is potential for model misspecification, which may have serious implications for findings. In our case, this could lead to what we believe to be an incorrect conclusion about the performance of the SHAs: zero inefficiency. This underlines the importance of accounting for both forms of unobservable heterogeneity discussed here.

As discussed in section 2, the predicted laboratory efficiencies are identical in pairs: the laboratory efficiency predictions of models 1 and 3 are one pair; and of models 2 and 4 are the other pair. That is, there are two ‘sets’ of laboratory efficiency predictions. These two sets of efficiency predictions are very similar with regard to their averages, 0.771 and 0.776 (table 6.5), their absolute correlation (=0.98) and their rank correlation (Kendall’s tau = 0.956***). This suggests that efficiency predictions at the laboratory level are robust to the specification of unobservable heterogeneity (or indeed whether it is assumed away, as in model 1). This result likely arises from the similarity between the estimated model parameters in the first stage(s).

We note in passing that there may be a residual amount of unobservable heterogeneity between laboratories within SHAs; we did not investigate this issue and are not aware of models that would permit this. We therefore leave this for future research.

6.4.4 Implications for Health Policy

To begin, the overall inefficiency predicted by our model for pathology services in the NHS is around 27% (see table 6.5). Therefore, through appropriate target setting, it should be possible to make substantial efficiency gains in services as a whole (that is, even the best performing SHAs can improve). By overall region, the most efficient SHA is B[91] with an overall inefficiency of around 20% (see equation (18) for derivation); and the least efficient region in SHA H with an overall inefficiency of 30%. It should be noted that even the most efficient SHAs have room to improve because of variations in the laboratory performance within them (discussed below). The efficiency gap between the best and worst performing regions is around 10%. SHAs I and J are also close to the SHA H level of inefficiency. Thus, pathology policy makers should look to these SHAs for maximum gains.

To calculate potential monetary savings, we take the efficiency prediction of each laboratory in its final year, apply its cost weight and compute the potential saving per laboratory. When this is aggregated across all of the laboratories, we find £54m of potential annual savings in the sample. If this is applied to all NHS pathology services, this would suggest potential savings of around £675m per annum[92]. This is significantly more than found in other empirical studies (£250-500m in Department of Health, 2008; £390m in Buckell et al., 2015).

Next, our model enables policy makers to look within SHAs to locate the source of overall inefficiency. As envisaged at the outset, we find inefficiency at both levels, but laboratory inefficiency dominates. The mean inefficiency at the SHA level is relatively low at 6%, where the least well performing SHA has 12% inefficiency. In contrast, the mean inefficiency at the laboratory level is much greater at 22%, and the least well performing group of laboratories appears to be 27% inefficient. Thus targets and policy mechanisms would appear to be better aimed at reducing or exploring differences in performance between laboratories within SHAs, rather than looking at persistent efficiency differences between different SHAs.

A further advantage of this model is that it allows policy makers to observe inefficiency differences between individual laboratories; variation that is concealed when considering average laboratory inefficiency for each of the SHAs (which can be seen from table 6.5 do

[91] Due to data confidentiality we are unable to reveal the identity of SHAs.
[92] In our sample, we have only one third of English laboratories, none of the laboratories in Wales, Scotland or Northern Ireland and one of five pathology disciplines. We thus follow other pathology studies and apply our overall savings to total pathology expenditure to arrive at our estimate.

not vary enormously). In figure 6.1, we see that two laboratories have inefficiency that >40%: laboratories 12 and 38. Laboratory 38 in particular should be singled out by policy makers to improve its performance given an inefficiency of 56%. We note that these predictions do not encompass the effects of the SHAs, which have been removed. Of course, as noted earlier, further examination of those laboratories would be needed as it may be that part of the efficiency gap is explained by other factors not taken account of in our model.

	Model 1				Model 2				Model 3				Model 4			
SHA	SHA	Lab	Overall	Rank	SHA	Lab	Overall	Rank	SHA	Lab	Overall	Rank	SHA	Lab	Overall	Rank
A	0.902	0.726	0.655	2	0.907	0.734	0.666	9	1.000	0.726	0.726	9	0.983	0.734	0.721	5
B	1.000	0.827	0.827	1	1.000	0.814	0.814	1	1.000	0.827	0.827	1	0.987	0.814	0.803	1
C	0.800	0.785	0.628	3	0.888	0.791	0.703	5	1.000	0.785	0.785	5	0.905	0.791	0.716	6
D	0.797	0.772	0.615	6	0.974	0.782	0.762	2	1.000	0.772	0.772	6	0.978	0.782	0.765	2
E	0.669	0.805	0.538	10	0.813	0.808	0.657	10	1.000	0.805	0.805	2	0.882	0.808	0.712	7
F	0.755	0.767	0.579	8	0.952	0.772	0.735	4	1.000	0.767	0.767	7	0.964	0.772	0.744	4
G	0.727	0.786	0.571	9	0.920	0.799	0.735	3	1.000	0.786	0.786	4	0.935	0.799	0.748	3
H	0.841	0.736	0.619	5	0.936	0.739	0.691	7	1.000	0.736	0.736	8	0.951	0.739	0.702	10
I	0.792	0.786	0.622	4	0.871	0.793	0.691	8	1.000	0.786	0.786	3	0.889	0.793	0.705	8
J	0.825	0.719	0.593	7	0.952	0.730	0.695	6	1.000	0.719	0.719	10	0.964	0.730	0.704	9
Mean	0.811	0.771	0.625		0.921	0.776	0.715		1.000	0.771	0.771		0.944	0.776	0.732	
s.d.	0.092	0.035	0.078		0.054	0.031	0.047		0.000	0.035	0.035		0.039	0.031	0.033	

Table 6.5: Efficiency Predictions at SHA Level, Laboratory Level and Overall Efficiency with Overall Ranks. Models 1-4.

	Model 1	Model 2	Model 3	Model 4
Model 1				
Model 2	0.022			
Model 3	-0.022	0.156		
Model 4	0.156	0.600**	0.289	

Table 6.6: Rank Correlation (Kendall's tau) between Overall Inefficiency Predictions, models 1-4. *, **, *** denote statistical significance at the 10%, 5% and 1% level, respectively.

	Model 1	Model 2	Model 3	Model 4
Firm effects (vs. pooled model)				
Moulton-Randolph	2.696***	4.168***	2.969***	4.168***
Decomposition of inefficiency and unobserved heterogeneity				
Wald test of 6 linear restrictions, $\rho = 0 \; \forall \bar{X}_i$		12.89**		12.89**
LR of firm effect SF (vs OLS)			0	1.147
Test of sub-company inefficiency				
LR of laboratory SF (vs OLS)	64.689***	58.170***	64.689***	58.170***

Table 6.7: Test Statistics, models 1-4. *, **, *** denote statistical significance at the 10%, 5% and 1% level, respectively.

For several reasons, the use of efficiency studies by health policy makers, despite their prevalence, has been limited (cf. Hollingsworth 2008; Hollingsworth, 2012). We have addressed three of these issues in this study. First, as is clear from our analysis, this modelling framework gives a complete top-to-bottom view of pathology services. In doing so, we are able to indicate the precise location of the inefficiency in these services, which is not possible with single level approaches, making our results of greater use in a practical sense. Second, we have purposefully focussed on a speciality of health services (as opposed to more aggregated entities such as whole hospitals or health regions) – pathology - again to make our findings of use to policy makers. Third, we have targeted a specific policy: the NHS's "A Call To Action" to make efficiency gains (NHS England, 2013).

6.4.5 Implications for Modelling multi-Level Data Structures

We now turn to the wider modelling implications of our work. We have already noted the advantages of adopting the multi-level model as it is possible for policy makers to observe inefficiency at different levels (Smith and Wheat, 2012). The alternatives, namely pooling laboratory level data, or modelling at the SHA level of aggregation do not (by construction) allow this decomposition (Smith and Wheat, 2012). In preliminary analysis we estimated

both of these alternatives and found that overall inefficiency was underestimated, which is in keeping with Smith and Wheat (2012). This is likely contributing to the differences in efficiency savings between our findings and other pathology studies (section 6.4.4).

However, the contribution of this analysis – apart from being the first health application of the Smith and Wheat (2012) DLSF – is to augment that model to control for unobserved heterogeneity in a multi-level context. We have shown the importance of accounting for unobserved heterogeneity in our study. This has clear implications for policy. Of course, we found that inefficiency is overestimated when unobservable heterogeneity is disregarded (which is well known in the health literature). We also found that unobservable heterogeneity arises in various forms; specifically, we find that it is important to take account of unobserved heterogeneity that is correlated with the regressors as well as that which is not. Indeed, we find that models that do not take account of the former, such as the recently developed approach by Kumbhakar et al. (2014) (model 3 in this chapter), may lead to unrealistic predictions and erroneous conclusions.

Further, in the context of multi-level structures, we noted that it may be hard to distinguish inefficiency from unobserved heterogeneity that is uncorrelated with regressors. This is because this part of the decomposition is based on the number of observations at the SHA level, which in our case is only 10. Thus there may be limits to the degree to which unobserved heterogeneity can be separated from inefficiency in data structures of this nature. As a caveat to this statement, a finding of no inefficiency when applying the Kumbhakar et al. (2014) model could be a reflection of underlying economic reality, and not necessarily because of misspecification or lack of data points (though we believe the latter to be the case in our example). Of course, the same problem, namely lack of observations to decompose inefficiency and unobserved heterogeneity that is uncorrelated with the regressors, also arises in traditional panels with a small cross-section.

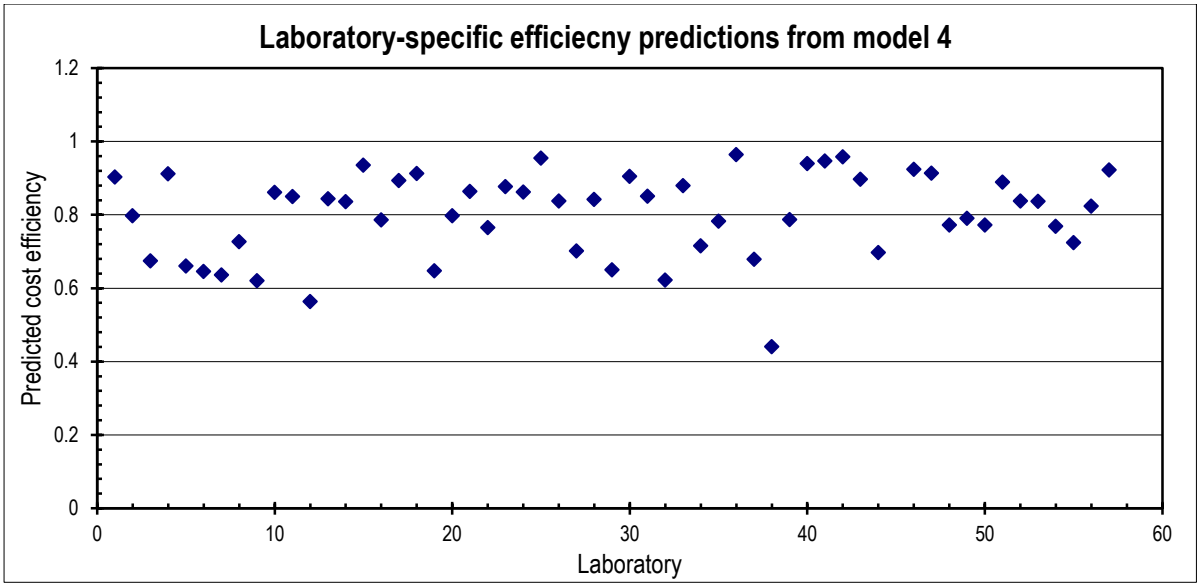

Figure 6.1: Laboratory Efficiency Predictions from Model 4. NB – to preserve the anonymity of the SHAs and laboratories, we do not assign the laboratories to their SHAs in this graph.

6.5 Conclusions

The study in this chapter is the first application of the Smith and Wheat DLSF (2012) in a health context and the first time vertically distinct measures of inefficiency have been simultaneously estimated in health markets. It is also the first time the approaches set out in Farsi et al. (2005a) and Kumbhakar et al. (2014), to control for unobserved heterogeneity, have been applied to a multi-level data structure.

Our results suggest overall inefficiency in pathology services in England of around 27%. This would correspond to annual savings of approximately £675m if applied to all NHS pathology services. This estimate exceeds previous studies' savings estimates, thus suggesting the scope for further improvements than have previously been envisaged (which is a conclusion in keeping with that of other application of this model; see Smith and Wheat (2012)).

The source of the inefficiency is visible in our study, which was not the case in previous studies. The results show that the dominant source of inefficiency is variation at the laboratory level inefficiency within SHAs, though the SHA-level persistent inefficiency effects are also important. This illumination of the location of inefficiency should provide a

useful guide for policy makers. Our results further show that some individual laboratories have particularly high inefficiency, which is worthy of further investigation.

With respect to the method, we find that it is important to consider both sources of unobservable heterogeneity (correlated and uncorrelated with the regressors). In our case, unobserved heterogeneity that is correlated with the regressors dominates. We note that the Kumbhakar et al. (2014) model (model 3 in this study) did not detect SHA-level inefficiency (wrong skew), which we attribute to model misspecification given that it neglects an important source of unobserved heterogeneity. Model 4, which takes account of both sources of unobserved heterogeneity, struggled to disentangle inefficiency from unobserved heterogeneity that is uncorrelated with the regressors. We attribute this problem to the fact that this stage of the decomposition relied on only 10 observations (as we have 10 SHAs). This could be a limitation to the degree to which unobserved heterogeneity can be separated from inefficiency in data structures of this nature, where there may be a small number of observations for the top-level of the hierarchy. Of course, the same problem would occur in traditional panel models with a small cross-section. We do note, however, that failure to separate inefficiency from unobserved heterogeneity that is uncorrelated with the regressors could simply reflect economic reality rather than caused by model misspecification and/or lack of data points (though we believe the latter to be true in our example).

Whilst the different approaches produced different results for SHA-level inefficiency, the inefficiency predictions at the laboratory level were largely the same across all models. It appears, then, that the inefficiency estimates at this lower level are robust to the treatment of unobserved heterogeneity. This is likely due to the estimated parameters being similar between models. However, we consider that further research might incorporate how unobserved heterogeneity at the lower level might be incorporated into the modelling framework.

With this, we conclude our empirical work in this thesis. We therefore turn, in the following and final chapter, to the discussion and overall conclusion of this thesis.

7. Conclusions

In this chapter we draw together the findings from our introductory and empirical chapters. In 7.1 we summarise the NHS policy context in which this thesis resides. In 7.2 we review the objectives that were set out in the introduction, and consider the extent to which our chapters have answered these issues. In 7.3 we take a synoptic view of our contribution. In section 7.4 we evaluate our empirical work. Finally, in section 7.5, we set a research agenda for future work.

7.1 Policy Context

We have argued that there is long- and short-run pressure on the health care budget in England. We identified efficiency as one possible way to relieve budgetary pressure. Noting that the majority of expenditure is on secondary care in England, we have paid particular attention to this area. Under a revised governance structure, NHS hospitals are now subject to efficiency targets set by the sector regulator, Monitor, as opposed to the Department of Health before it; Monitor are thus tasked with promoting efficiency amongst NHS hospitals.

We have reviewed the process of setting efficiency targets from an economic perspective: we have examined how to measure efficiency; and how to incentivise hospitals to reach those targets set. This is likely to be key information going forward, not least due to the current contention surrounding the proposed efficiency target. In this sense, we feel we have made a valuable contribution to current policy issues.

Of course, we are also interested in the academic contribution of our endeavours. We therefore turn to look in detail at the findings from our empirical chapters, and reconcile these against our research objectives to demonstrate the contribution to the literature of this thesis.

7.2 Revisiting the Research Objectives

In chapter 1 (section 1.3), the objectives of this thesis were set out. In this section, we revisit these objectives and comment on how this thesis has answered each.

(i) To inform the process of setting efficiency targets for NHS hospitals, by reviewing germane literature and conducting efficiency analysis; and to set out some empirical issues to which we are able to provide solutions.

In chapter 4, this objective was achieved in several ways.

First, we reviewed a number of policies that have been applied to NHS hospitals across its recent history. These policies were applied with the intention of improving hospitals' performance. They had varying degrees of success. We used the available evidence to identify features of performance management regimes that were effective, and those that were not. In doing so, we provided some lessons on what appears to be effective when looking to induce improved performance. This is likely useful for Monitor who is responsible for enforcing hospital efficiency targets.

Second, we reviewed the pricing mechanism for hospital reimbursement. We established the link between the National Tariff Payment System (NTPS) and Price-Cap regulation. We focussed our discussion on the incentives around price-control periods. We observed that price-cap regulation had been effective in a number of industries, but required a longer regulatory lag than is in place under NTPS. We therefore concluded that by lengthening regulator lag for the NTPS, Monitor could encourage efficiency amongst hospitals. Further, we considered the regulatory mechanism that is employed in the rail industry as an example of alternative pricing models when there are multiple parties involved in financing and delivering services. We drew the parallel to health services.

Finally, we reviewed efficiency studies in the health sector and in other regulated industries. We identified health-based issues with efficiency analysis. There were both methodological issues and practical issues that have hindered the use of these studies at the policy level. We therefore set out some ways in which these might be resolved to encourage uptake. This review was used to frame the empirical work.

In chapters 5 and 6, we conducted econometric efficiency analysis in NHS hospitals. In particular, we studied pathology laboratories. In doing so, we identified the level of inefficiency within pathology services. Although useful, efficiency analysis at this level provides information in a limited sense in that it would be difficult to justify setting targets based on efficiency analysis at the individual service level. However, it is perfectly possible – and data are available – to conduct several such studies across a number of services, to enable a more detailed picture of how efficiency varies across areas of services within hospitals. This approach, although more involved, has two major advantages. First, a more detailed level of analysis, compared to a whole hospitals approach, means that targeting managerial effort to for capturing the gains identified is a straightforward process. Secondly, analysis at a more detailed level of service enables the researcher to better control for service and patient level heterogeneity. This is in keeping with contemporary thought in the academic literature.

(ii) To provide new economic evidence for an area of NHS hospital activity for which empirical evidence is scant: pathology laboratories. This is, in turn, to feed into the top-level policy goal of making efficiency savings across the NHS.

In chapters 5 and 6 this objective was achieved. In a first analysis, we analysed fundamental economic issues pertaining to costs and efficiency in pathology laboratories. We estimated the level of inefficiency in pathology services, and estimated that around £390m savings annually would be possible if all laboratories improved performance in line with best practice. We used a flexible model which allowed us to estimate how individual laboratories' efficiency changed over time. We found that this was important, as some laboratories were making efficiency gains over time, some losses and others were unchanged over the observed period. On average, laboratories became less efficient over time.

Next, we conducted a simulation exercise to estimate the cost implications of merged laboratories. We found potential savings of a similar magnitude to potential gains from inefficiency reduction. We also examined a number of factors

that drive costs which have been raised in policy discourse. We paid particular attention to economies of scale, which is of significant interest to policy makers.

Lastly, we took an overall account of pathology production by combing the various aspects of the cost function into an estimate of multi-factor productivity and its change in over time. We found an improvement of around 5% over the period studied, driven largely by technical change.

In chapter 6, we developed our analysis to consider the way in which pathology services are organised hierarchically. Indeed, we found that were components of inefficiency at vertically separate organisational levels in pathology services. We found the majority of inefficiency resided at the lower tier of the organisation and a smaller amount at the upper tier. Further, we found that accounting for this structure was important for overall estimates of inefficiency. We therefore revised our potential savings estimate for pathology services to £675m, which is considerably more than was found in previous studies, including those of this thesis.

(iii) To advance the measurement of efficiency in health markets and beyond

This objective has been achieved across chapters 4, 5 and 6.

In terms of efficiency analysis in health markets, we have made several steps. Firstly, the literature review of methods in chapter 4 provides a review of the literature based on current issues in the estimation of efficiency in health markets. In addition, focus is given to policy-based issues that have been raised. We also contrasted the methodological issues in health with those of other regulated industries, providing richer insights into efficiency measurement for regulatory practice in health care. We extended the regulatory best practice criteria for benchmarking to health markets. In these regards, we conclude that we have made a useful contribution to the literature.

We then, based on our review, conducted empirical work on NHS hospitals. In doing so, we introduced a number of new empirical efficiency models in the health context. In chapter 5, we, for the first time, adopted the model of Cuesta (2000) to examine efficiency change over time of individual laboratories. This is of particular use as NHS staff have indicated that knowing how performance changes over time is of use to them (Hollingsworth and Peacock, 2008).

In chapter 6, we, for the first time, employed the dual-level stochastic frontier (DLSF) of Smith and Wheat (2012) to make a vertical decomposition of inefficiency. We note that modelling hierarchical efficiency has been identified as a key area for empirical development in health care (Hollingsworth and Peacock, 2008).

These models are of use in the wider health context, and allow both researchers and policy makers/regulators important insights into inefficiency variation.

In chapter 6, we were able to go beyond the health context in our methodology. We have extended the DLSF model of Smith and Wheat (2012) to allow for the presence of unobserved heterogeneity. In particular, we have presented a modelling approach that can account for unobserved heterogeneity that can manifest in several forms. In doing so, we have derived a model that is of general use. This is approach is particularly of use in the health context, where unobserved heterogeneity – as argued - is a significant concern.

7.3 Synopsis of Empirical Research

We have commented on how each of the individual chapters contributes in both policy and academic senses, we now discuss potential holistic benefits of the thesis combined.

Our central theme is NHS hospital efficiency. We have conducted analysis in an intense fashion, applying due academic rigour to our analysis. Whilst this is, naturally, of high importance in the academic setting, there is a question as to the application of our findings in

the practical setting: we ran complex econometric methods, does it matter? Our view is, broadly, yes.

There are some distinct advantages to the approach that we adopted, relative to Monitor's, which we have already noted – disaggregate analysis, efficiency of individual units over time, multi-level efficiency analysis and unobserved heterogeneity. The multi-level analysis was of particular significance: when the structure was taken into consideration, there appeared to be substantially more available savings than in the single-level setting (see section 6.4).

Further, our multi-factor productivity index allowed us valuable insights into how other aspects of productivity changed. In our prediction of merging laboratories, we demonstrated significant potential in scale benefits – comparable in magnitude to potential efficiency gains. These analyses enrich the information available on hospital performance. As such, they are worth pursuing in the regulatory setting. Incorporating these features into future efficiency analyses may help resistance from providers, at least in methodological terms. For these reasons, our view is that a more involved methodological approach by Monitor is likely to yield significant benefits.

These benefits, however, come at the cost of resource: more complex models are difficult to construct and run, require expertise, require testing/sensitivity diagnostics, etc. In a single year when resources are constrained, these more complicated approaches become increasingly difficult to adopt. To take an approach which incorporates these features would be more realistic in a longer period. Lengthening the control period would enable deeper engagement with providers, in particular with regard to methods. Moreover, these changes could be cost effective if the estimated additional cost savings in our empirical work are in excess of the likely cost of employing extra resources to enable a more sophisticated approach. (Whether or not this is true in reality is unknown without data to support it.)

In terms of the pricing mechanism, we suggested lengthening the regulatory lag as theory and empirical evidence suggest that doing so enhances efficiency incentives. Suppose this was brought in line with other regulators in England to five years. It might reasonably be expected that, due to the strengthened incentives mentioned, efficiency gains are realised. This would fit neatly with adopting a more sophisticated approach to econometric benchmarking. Of

course this may be complicated by policy change, which has been common throughout the lifetime of the NHS.

In addition, if the lag is lengthened, it would be easier to introduce adjustments to the mechanism to allow arbitration by a central body. This would help with the matter of disputes, as are ongoing at the time of writing. We suggested adopting the tripartite system as in the rail industry. A useful piece of academic research in this regard would be an economic model to predict the potential effects of making changes. We discuss this in the following section.

With regard to other incentives, lengthening regulatory lag would allow Monitor to take advantage of the beneficial features of target setting as set out in box 4.1. For example, it would allow targets to be clearly set and communicated; to be applied over longer periods avoiding short-term quick-fixes; targets would be prioritised; and sanctions for failure could readily be imposed. Importantly, it would allow the regulator to engage with providers on a long-term basis and have greater interaction with the process of efficiency factor setting; the current rejection of the 2015/16 NTPS determination suggests discontent amongst providers with the current approach.

Whilst helpful for promoting positive aspects of performance management schemes, lengthening lag would, equally, be conducive to avoiding some of the potential pitfalls in setting targets, as presented in box 4.2. In chapter 5, we discussed that an econometric approach based on a cost function avoids a number of potential issues such as gaming. A five year lag relative to a single year's lag has two clear advantages in relation to box 4.2. First, the issue of myopia is mitigated by lengthening the lag.

Second, given that multiple policies are still imposed (such as waiting times targets), it is difficult to delineate the specific effect of each. Lengthening the lag would give a clearer idea as to what drives performance change - the pricing system or the other targets in place. That is, if there was a significant change to the pricing system and changes to performance, it is likely that these were driven through the change. On the other hand, if no discernible difference in performance was observed, it could be argued that performance was scarcely affected by the pricing mechanism. This information would be of use whatever the outcome.

Whilst we have proposed a number of advances for Monitor, we have also identified some areas for useful future research. We therefore move to evaluate our approach before discussing directions for research in the future.

7.4 Reconciliation of Empirical Results

7.4.1 Reconciliation Against NHS Hospital Efficiency Studies

In chapter 3, we reviewed efficiency studies in health and focussed on applications of econometric efficiency analysis techniques to NHS hospitals. We reported results in table 3.3. We noted that not all studies explicitly measured levels of inefficiency or indeed reported results. However, for those that did, it is a useful exercise to compare and contrast the results of prior studies to the results of the empirical work in this thesis.

Of the studies which explicitly report efficiency estimates for NHS hospitals, of which there are 4, there is a range of efficiency reported. SFA studies report inefficiency in the range of 8% to 39%; in our empirical work, we report average inefficiency of 13% and 27%. This appears, then, to be in keeping with the findings of prior studies. In terms of technical change, one study (Ferrari, 2006) reports that hospitals experienced technical progress of around 3% a year. In our analysis, we found comparable progress of around 2%.

In terms of efficiency change over time, the results are mixed. In one study (Ferrari, 2006), whilst some analysis of efficiency over time was conducted, there was no temporal change found, which is in direct contrast to our findings. However, we note that the model employed is rather restrictive – that of Battesse and Coelli (1992) - which may be an explanation as to why. Elsewhere, results indicate that, across a number of specifications, efficiency appears to be decreasing over time (Jacobs et al., 2006). This is consistent with our results, however the authors note some issues with sensitivity (pp. 86-9). We did not find such issues.

Overall, our results appear to be in line with the received literature in terms of efficiency predictions, technological change and temporal change.

7.4.2 Reconciliation of Efficiency Predictions in Chapters 5 and 6

In chapter 5, the average prediction of inefficiency was 13% which implied cost savings of £390m p.a. in pathology services. Using the same data, the average of predicted inefficiency

was 27%, which, in turn, implies potential cost savings of £675m in NHS pathology. There is thus fairly sizeable discrepancy between the two chapters' estimates of efficiency (and therefore the corresponding policy implications) which bears comment.

First is to highlight the differences in approach - and therefore the models - applied to the data. In chapter 5, our preferred model allowed inefficiency to vary between laboratories over time; in chapter 6 time-invariant efficiency was assumed. Second is the stratification. In chapter 5, the data was stratified by laboratory and time. In chapter 6, the first stratification was by SHA and laboratory; and in the second stage by laboratory. Further, the two studies' specifications varied in terms of both variables and functional form. In chapter 5 there were additional variables in the cost function and a translog functional form was adopted. In chapter 6, some of these variables were not significant (and so dropped) and the stratification did not support a translog form, and we thus reverted to a Cobb-Douglas form. For these reasons, efficiency predictions are likely to vary which may explain the observed discrepancy.

However, the most likely driver of the difference stems from the organisational structure and the differing economic interpretations between the two analyses. Considering this issue highlights the link between the two approaches. In chapter 5, we used a set of regional dummies to control for unobserved differences between laboratories. These are SHA dummies that have been collapsed for parsimony. Our economic interpretation of these variables was that they represented regional differences that are unobserved between pathology laboratories. In chapter 6, we used this structure in the data, but adopted both a different approach and structure. First, we stratified by SHA and laboratory to identify the effect of SHA. Here, our interpretation of these effects was that they were partly, as per chapter 5, comprising unobserved heterogeneity but also that there was SHA-specific inefficiency present in these effects. We duly applied a number of specifications and tests to determine if this was the case, concluding that it was. Therefore, this effect is likely contributing to the observed difference.

In passing, we note that and alternative, indeed simpler, strategy for the analysis in chapter 6 would be to estimate a model that is stratified by laboratory and time and use SHA dummies to estimate the SHA effects. In this sense, the framework is akin to a true fixed effects formulation (Greene, 2005). We did not adopt this approach for two reasons. First, TFE are known to suffer from the incidental parameters problem, particularly when the panel length, t,

is small, as is the case in our application. The second follows from the findings of chapter 5. Whilst we could estimate this model, we are unable to identify the SHA effects (i.e. not all of them are statistically significantly different from zero). This is not the case in our approach. For this reason, we retained the methodology detailed in chapter 6.

7.4.3 Reconciliation Against NHS Policy

In the opening chapters of this thesis, we identified that, under current policy, an efficiency savings target of £30bn p.a. by 2020 has been established, following from the NHS's "A Call to Action" (NHS England, 2013). Of this, central government has made a commitment of £8bn. The remaining £22bn is to be found across NHS services. The "Five Year Forward View", or is as known elsewhere as the "Stevens Plan" after the NHS chief executive Simon Stevens, has identified that 2-3% annual productivity improvements across the NHS will ensure that the gap is met.

As regards the findings of this thesis, we have estimated that productivity in pathology services has increased by, on average, 1.18% per year over the period studied. This is, in crude terms, rather behind the required savings. However, we note that our MFP measure did not encompass scale change, which was estimated to be a powerful element of cost reduction and so, by extension, productivity change. In light of this, it may well be possible that the required productivity gains are delivered. We further note that the disaggregate approach we have adopted allows us to identify where these gains can be, or at least have been, made – in this case technological and scale change – and so where additional gains can be made. That is, given our estimates of efficiency, the productivity gains can be boosted if managerial focus is given to capturing the efficiency gains identified.

Of course, looking at crude figures, whilst useful, is only part of the issue. There is, of course, the issue of what is possible in reality. Whilst we have identified the potential for efficiency gains, a question arises as to how much of these can be realised. This may limit the potential gains available. There may be an acceptable level of inefficiency that is achievable, beyond which seeking further gains may be extremely difficult. It is, of course virtually impossible to make such a judgement on empirical evidence; this is an issue for the regulator to resolve itself. This may be guided by prevailing policy – the global NHS and/or specific hospital efficiency targets are clear candidates for 'satisfactory' levels of performance.

The other issue is the interactive element of pathology service provision. This is important in terms of both the formation of pathology networks and pathology's interaction with other NHS services. There may be a joint effect here – does the formation of laboratory networks – whilst helping to make gains in scale – jeopardise the service provided to primary or secondary care? This is a clear issue for future research.

Overall, given that technological progress has improved productivity substantially, that efficiency gains are possible to drive further gains and that the formation of local networks appears to be inducing scale gains, our view is that the 2-3% productivity challenge for NHS pathology laboratories is perfectly possible. Whether it is achieved, which it has not been in recent years, depends on managerial performance.

7.5 Evaluation of Empirical Research

We have argued that this thesis has made a number of advances for measuring efficiency in health care and applied techniques to hospitals in the NHS. Of course, there are outstanding issues that remain. By way of evaluation, we highlight areas in which our analysis was not able to provide answers. In addition, we offer some general remarks around how we conducted the research. We go on to discuss future directions for research in the following section.

One issue is sample size. The data set used here, as noted, contains 187 observations. In health terms, where HES data contains 18.2 million patient records annually, this data would not be considered as large. In regulatory terms, this number, although not as large as some, will be much larger than others. Appendix B shows a table of regulators' efficiency analyses and sample sizes used, which range from 5 to 560 observations. Here, 187 would be the second largest of the data sets, and at the higher end of the scale. So whether this is considered large appears to be a largely contextual matter.

One important aspect of sample size is the ability or otherwise to estimate models. In our empirical work, with the exception of the true random effects model in section 5.4.2, which although did estimate, gave questionable efficiency predictions, 187 observations was sufficiently large to allow estimation. Moreover, there were difficulties in using the Kumbhakar et al. (2014) approach to making the decomposition of SHA-level inefficiency

and unobserved heterogeneity (section 6.4). We ascribed this to sample size issues where the small cross-section of SHAs appeared to cause difficulties in estimating the model. A larger sample size (in terms of the cross section of firms, or as in our case, SHAs) may have helped to estimate the models.

In chapter 4, we identified patient-level cost heterogeneity as a key aspect of health care efficiency analysis. By analysing pathology services which do not deal directly with patients, there is less concern around patient-level heterogeneity. We were therefore unable to deal directly with this issue.

In addition, in chapter 4, we identified quality and outcomes as a key aspect of health care efficiency analysis. We addressed the issue of pathology quality in chapter 5.3. Laboratories in our sample have obtained a base level of quality by virtue of having been accredited; excess variation in quality is absorbed in our controls for unobserved heterogeneity. However, we were not satisfied that the quality of available data to include this feature directly into our analysis. Therefore, again, we were unable to deal directly with this issue.

On account of both of these features, and the organisation of pathology laboratories, there is a question around whether pathology can be considered as a department of a hospital, or whether the practice of pathology in laboratories is a separate area of health care services entirely. This has implications for the generalizability of the results from our empirical analysis. Certainty, that there is in no direct patient contact brings in to question that pathology laboratories are indeed a hospital department. On the other hand, that they are situated in hospital trusts, that they provide direct secondary care services and are integrated with other hospital services makes the case for their inclusion. It would be useful to survey trust managers to seek their opinion on this matter.

Lastly, we feel that this work was conducted in a manner that engaged the pathology industry. First, the data collection is based on long running relationships with laboratories. As noted in section 5.2, our work is an extension of existing indicator benchmarking; that benchmarking occurs at all suggests that laboratories are interested in their efficiency. Next, we have presented the work at a number of pathology conferences, for example the National Pathology Benchmarking Conference in Birmingham, 30th November 2012. Lastly, we have

published preliminary analyses of this work in pathology journals, with pathologists as co-authors (see Buckell et al., 2013).

We now turn to discuss some potential areas for future research.

7.6 Directions for future research

Following from the discussion in the empirical chapters, the following directions for future research are set forth.

7.6.1 Single Stage Estimation Dual-Level Stochastic Frontier for Unobserved Heterogeneity

In chapter 3, we described the recently developed four component models. These models have the capacity to separate out time-invariant unobserved heterogeneity, time-invariant efficiency, time-varying inefficiency and random statistical noise. We noted several approaches to estimating these models, including single- and multi-stage approaches.

In chapter 6, we adapted the four component model for use in a dual-level efficiency capacity. In this setting, we employed only the multi-stage approach to estimation. There may be advantages to estimating a single stage equivalent.

Firstly, using the approach of Filippini and Greene (2015), there is the benefit of computational ease (although, we note that, in small samples, this model may be difficult to estimate). It would further serve as a robustness check for estimates of inefficiency and/or bias in the estimates of parameters, given differing estimation procedures.

Additionally, it has the advantage that it is easier to test against other models. That is, although we demonstrated that each of the components can be individually tested in the multi-stage approach, we would not be able to test the multi-stage model against non-nested alternatives, such as a latent class stochastic frontier model. This would be possible with a single stage approach. Indeed, in chapter 5 we made use of such a test, the Vuong test, for non-nested models. There are of course, other approaches to testing non-nested models, for example Pollak and Wales's (1991) likelihood dominance criterion.

Another advantage is around prediction intervals of lower tier inefficiency. In a multi-stage approach, one could naively estimate prediction intervals in the lower tier SF stage (e.g. standard Horrace and Schmidt (1996) central intervals that are built into most modern software packages). However, doing so would be to neglect that the dependent variable – the adjusted residual from the first stage – is itself uncertain. Then, these intervals underestimate the true uncertainty around the inefficiency predictions. Therefore, a correction would need to be made to introduce the additional uncertainty. In doing so, it would be likely that the intervals would be rather wider than the naïve intervals. This is not the case in the single stage approach, where intervals around single stage lower tier estimates capture this uncertainty.

For these reasons, in our view the application of the single-stage generalised true random effects model to the dual-level stochastic frontier is of significant empirical relevance.

7.6.2 Lower Tier Unobserved Heterogeneity in the Dual-Level Stochastic Frontier

In chapter 6, we extended the dual-level stochastic frontier model in a number of ways to account for, and test, different forms of unobserved heterogeneity. The controls we used were implemented in either the first stage (i.e. Mundlak), or the second stage at upper tier level (i.e. Kumbhakar et al.). We noted in passing that we were not able to capture unobserved heterogeneity at the lower tier level in our approach.

Before we address this, we note that this is perhaps not entirely true. Of course, we applied controls exclusively at the upper tier level in estimation but, in the case of the Mundlak adjustment, these controls were applied to lower tier data, albeit stratified according to the upper tier. Therefore, the argument could be made for having controlled for unobserved heterogeneity at the lower tier that is correlated with the regressors. Of course, even if this is accepted, there remains unobserved heterogeneity at the lower tier that is uncorrelated with the regressors.

If the data are permitting, it would be possible to address this issue. That is, in our data, we have a time series dimension. By making use of this, it would be possible to examine any remaining unobserved heterogeneity at the lower tier level. In some preliminary work, we have attempted this via the Wang and Ho (2010) approach to the true fixed effects model.

The initial results appear promising, and suggest there is some residual unobserved heterogeneity at this level, although we will conduct further testing before reporting any empirical results. There may be other approaches to this issue, the latent class model, for example.

We see this as a very useful area of future research, so that lower tier unobserved heterogeneity is not misinterpreted as inefficiency.

7.6.3 Larger NHS Data Set for the Application of the Mundlak-Transformed Four-Component DLSF

Whilst we have defined and tested a model for varying forms of unobserved heterogeneity, we found difficulty in estimation. We attributed this difficulty to the small cross section of SHAs in our data. Therefore, a clear extension would be to deploy this approach using a data set with a larger cross section which we believe would rectify this issue. Further, if a dataset could be found for some disaggregate NHS hospital activity (e.g. proton beam therapy), then the findings could be expanded not only in methodological terms, but in terms of the policy goals identified in this thesis. This model is of particular relevance for application in this setting, since hierarchical managerial structures exist in the provision of hospital services. This would further allow us to investigate some of the issues that we were unable to do with the data here, e.g. patient level cost variation. This objective can be coupled with other proposed directions for future research, in particular single stage estimation given that comparably complex models in chapter 5 were found to be difficult to estimate on the data at hand.

References

Abbott, M. & Cohen, B. 2009. Productivity and efficiency in the water industry. *Utilities Policy,* 17**,** 233-244.

Adab, P., Rouse, A. M., Mohammed, M. A. & Marshall, T. 2002. Performance league tables: the NHS deserves better. *BMJ : British Medical Journal,* 324**,** 95-98.

Adams, G., Gulliford, M., Ukoumunne, O., Chinn, S. & Campbell, M. 2003. Geographical and organisational variation in the structure of primary care services: implications for study design. *Journal of Health Services Research & Policy,* 8, 87-93.

Aigner, D., Lovell, C. A. K. & Schmidt, P. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics,* 6, 21-37.

Alchian, A. 1965. some economics of property rights. *Il Politico,* 30, 816-29.

Alchian, A. & Kessel, R. 1962. Competition, Monopoly, and the Pursuit of Pecuniary Gain. *In:* Lewis, G. (ed.) *Aspects of Labor Economics.* Princeton: Princeton University Press.

Alvarez, A., Amsler, C., Orea, L. & Schmidt, P. 2006. Interpreting and Testing the Scaling Property in Models where Inefficiency Depends on Firm Characteristics. *Journal of Productivity Analysis,* 25, 201-212.

Alvarez-Rosete, A., Bevan, G., Mays, N. & Dixon, J. 2005. Effect of diverging policy across the NHS. *BMJ* 331.

Amado, C. & Santos, S.2009. Challenges for performance assessment and improvement in primary health care: The case of the Portuguese health centres. *Health Policy,* 91, 43-56.

Anonymous. 1864. Untitled. Response to letter by William Farr. *Medical Times and Gazette*: 13th February 1864, pp. 187-188

Appleby, J. 2012. A productivity challenge too far? *BMJ,* 344.

Appleby, J. 2013. Are "friends and family tests" useful: agree, disagree, neither, don't know? *BMJ 346*

Appleby, J., Humphries, R., Thompson, J. & Galea, A. .2013. *How is the Health and Social Care System Performing? Quarterly Monitoring Report,* London: The King's Fund. Available online at http://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/quarterly-monitoring-report-kingsfund-jun13.pdf. Accessed 12th January 2013

Appleby, J., Thompson, J. & Jabbal, J. 2014. *How is the Health and Social Care System Performing? Quarterly Monitoring Report July 2014.* London, The King's Fund. Available online: http://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/quarterly-monitoring-report-kingsfund-jun13.pdf. Accessed 25/10/2014

Armstrong, M., Simon, C. & Vickers, J. 1994. *Regulatory Reform,* Cambridge, MIT Press.
Audit Commission. 2004. Introducing payment by results. London.

Arocena, P., Saal, D. & Coelli, T. 2012. Vertical and Horizontal Scope Economies in the Regulated U.S. Electric Power Industry. *The Journal of Industrial Economics,* 60, 434-467.

Audit Commission. 2004. Introducing payment by results. London: Audit Commission.

Australian Competition and Consumer Commission. 2012. Regulatory Practices in Other Countries Benchmarking opex and capex in energy networks. Canberra: Australian Competition and Consumer Commission.

Averch, H. & Leland, L. J. 1962. Behavior of the Firm Under Regulatory Constraint. *The American Economic Review,* 52**,** 1052-1069.

Bailey, E. E. & Friedlaender, A. F. 1982. Market Structure and Multiproduct Industries. *Journal of Economic Literature,* 20**,** 1024-1048.

Baker, L. C. & Wheeler, S. K. 1998. Managed care and technology diffusion: the case of MRI. *Health Affairs,* 17, 195-207.

Baldwin, R. & Cave, M. 1999. *Understanding Regulation: Theory, Strategy and Practice,* Oxford, Oxford University Press.

Baltagi, B. H. 2006. An Alternative Derivation of Mundlak's Fixed Effects Results Using System Estimation. *Econometric Theory*, 22, 1191-1194.

Baltagi, B. 2008. *Econometric Analysis of Panel Data*. John Wiley & Sons, Chichester.

Baltagi, B. and Li, Q. 1990. A Comparison of Variance Components Estimators Using Balanced Versus Unbalanced Data. *Econometric Theory,* 6, 283-285.

Baltagi, B. H. & Moscone, F. 2010. Health care expenditure and income in the OECD reconsidered: Evidence from panel data. *Economic Modelling,* 27, 804-811.

Bates, L. J. & Santerre, R. E. 2013. Does the U.S. health care sector suffer from Baumol's cost disease? Evidence from the 50 states. *Journal of Health Economics,* 32, 386-391.

Bator, F. M. 1958. The Anatomy of Market Failure. *The Quarterly Journal of Economics,* 72**,** 351-379.

Battese, G. E. & Coelli, T. J. 1988. Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. *Journal of Econometrics,* 38, 387-399.

Battese, G. and Coelli, T. 1992. Frontier Production Functions, Technical Efficiency and Panel Data: With Application to Paddy Farmers in India. *Journal of Productivity Analysis,* 3, 153-169.

Battese, G. E. & Coelli, T. J. 1995. A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics,* 20**,** 325-332.

Bauer, P. W., Berger, A. N., Ferrier, G. D. & Humphrey, D. B. 1998. Consistency Conditions for Regulatory Analysis of Financial Institutions: A Comparison of Frontier Efficiency Methods. *Journal of Economics and Business,* 50**,** 85-114.

Baumol, W. J. 1967. Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis. *The American Economic Review,* 57, 415-426.

Baumol, W. J. 2012. *The Cost Disease: Why Computers Get cheaper and Health Care Doesn't,* New Haven, Yale University Press.

Baxter, K., Weiss, M. & Le Grand, J. 2007. Collaborative commissioning of secondary care services by primary care trusts. *Public Money and Management,* 27**,** 207-214.

Becker, G. 1983. A Theory of Competition Among Pressure Groups of Political Influence. *The Quarterly Journal of Economics,* 98, 371-400.

Berger, A. N. & Humphrey, D. B. 1991. The dominance of inefficiencies over scale and product mix economies in banking. *Journal of Monetary Economics,* 28**,** 117-148.

Besley, T, Bevan, G and Burchardi, K. 2009. 'Naming & Shaming: The impacts of different regimes on hospital waiting times in England and Wales'. London, Centre for Economic Policy
Research.http://www.cepr.org/active/publications/discussion_papers/dp.php?dpno=7306 accessed 12/8/2013.

Besstremyannaya, G. 2011. Managerial performance and cost efficiency of Japanese local public hospitals: A latent class stochastic frontier model. *Health Economics,* 20**,** 19-34.

Bevan, G. 2006. Setting Targets for Health Care Performance: Lessons from a Case Study of the English NHS. *National Institute Economic Review,* 197**,** 67-79.

Bevan, G. 2010. Performance Measurement of "Knights" and "Knaves": Differences in Approaches and Impacts in British Countries after Devolution. *Journal of Comparative Policy Analysis: Research and Practice,* 12**,** 33-56.

Bevan, G. 2015. Incentives and models of governance. *Health Economics, Policy and Law,* 10**,** 345-350.

Bevan, G. & Hamblin, R. 2009. Hitting and Missing Targets by Ambulance Services for Emergency Calls: Effects of Different Systems of Performance Measurement within the UK. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 172, 161-190.

Bevan, G. & Hood, C. 2006. What's Measured Is What Matters: Targets And Gaming In The English Public Health Care System. *Public Administration,* 84**,** 517-538.

Bevan, G. & Skellern, M. 2011. Does competition between hospitals improve clinical quality? A review of evidence from two eras of competition in the English NHS. *BMJ,* 343.

Bevan, G. & Wilson, D. 2013. Does 'naming and shaming' work for schools and hospitals? Lessons from natural experiments following devolution in England and Wales. *Public Money & Management,* 33, 245-252.

Bird, S. M., Cox, D., Farewell, V. T., Goldstein, H., Holt, T. & Peter C, S. 2005. Performance indicators: good, bad, and ugly. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168, 1-27.

Black, N. 2013. Myth of falling productivity in the UK NHS confirmed. *The Lancet* 381, 996.

Bloom, N., Propper, C., Seiler, S. & Reenen, J. V. 2010. The Impact of Competition on Management Quality: Evidence from Public Hospitals. *National Bureau of Economic Research Working Paper Series,* No. 16032.

Böhm, K., Schmid, A., Götze, R., Landwehr, C. & Rothgang, H. 2013. Five types of OECD health care systems: Empirical results of a deductive classification. *Health Policy,* 113, 258-269.

Bojke, C., Castelli, A., Street, A., Ward, P. and Laudicella, M. 2013. Regional variation in the productivity of the English National Health Service. *Health Economics,* 22(2): 194-211

Bojke C, Castelli A, Grasic K, Street A. Productivity of the English NHS: 2012/13/update. Centre for Health Economics, University of York; *CHE Research Paper 110* Available online at https://www.york.ac.uk/media/che/documents/papers/researchpapers/CHERP110_NHS_productivity_update_2012-13.pdf accessed 14/05/2015.

Bos, D. & Von Weizsacker, R. K. 1989. Economic consequences of an aging population. *European Economic Review,* 33, 345-354.

Brekke, K. R., Siciliani, L. & Straume, O. R. 2008. Competition and waiting times in hospital markets. *Journal of Public Economics,* 92**,** 1607-1628.

Brekke, K. R., Siciliani, L. & Straume, O. R. 2011. Hospital Competition and Quality with Regulated Prices. *Scandinavian Journal of Economics,* 113**,** 444-469.

Breusch, T. and Pagan, A. 1980. The LM Test and its Application to Model Specification in Econometrics. *Review of Economic Studies,* 47, 239-254.

Buckell, J., Jones, R., Holland, D. & Batstone, G. 2013. Efficient thinking: Introducing econometric techniques in pathology. *The Bulletin of The Royal College of Pathologists,* 164, 241-243.

Buckell, J., Smith, A., Longo, R. & Holland, D. 2015. Efficiency, heterogeneity and cost function analysis: empirical evidence from pathology services in the National Health Service in England. *Applied Economics,* 47, 3311-3331.

Burbidge, J. B., Magee, L. & Robb, A. L. 1988. Alternative Transformations to Handle Extreme Values of the Dependent Variable. *Journal of the American Statistical Association,* 83**,** 123-127.

Busse, R., Geissler, A., Quentin, W. & Wiley, M. 2011. *Diagnosis-Related Groups in Europe,* Maidenhead, Open University Press.

CAA .2006. Airports price control review – Initial proposals for Heathrow, Gatwick and Stansted. London: UK Civil Aviation Authority.

Carey, K. & Stefos, T. 2011. Controlling for quality in the hospital cost function. *Health Care Management Science,* 14**,** 125-134.

Castelli, A., Jacobs, R., Goddard, M. & Smith, P. C. 2013. Health, policy and geography: Insights from a multi-level modelling approach. *Social Science & Medicine,* 92, 61-73.

Castelli, A., Street, A., Verzulli, R. & Ward, P. 2015. Examining variations in hospital productivity in the English NHS. *The European Journal of Health Economics,* 16, 243-254.

CEPA. 2014. Ofwat Cost Assessment – Advanced Econometric Models. London: Cambridge Economic Policy Associates.

Chambers, R. 1988. *Applied Production Analysis: A Dual Approach,* New York, Cambridge University Press.

Chen, Y.-Y., Schmidt, P. & Wang, H.-J. 2014. Consistent estimation of the fixed effects stochastic frontier model. *Journal of Econometrics,* 181**,** 65-76.

Chernew, M. E. & Newhouse, J. 2012. Health Care Spending Growth. *In:* Pauly, M., McGuire, T. & Barros, P. (eds.) *Handbooks in Economics: Health Economics.* Oxford: North-Holland.

Christensen, L. R. & Greene, W. H. 1976. Economies of Scale in U.S. Electric Power Generation. *Journal of Political Economy,* 84, 655-676.

Clegg, A. & Young, J. 2011. The Frailty Syndrome. *Clinical Medicine,* 11**,** 72-75.

Clegg, A., Young, J., Iliffe, S., Rikkert, M. O. & Rockwood, K. 2013. Frailty in elderly people. *The Lancet,* 381**,** 752-762.

Coelli, T. & Perelman, S. 1999. A comparison of parametric and non-parametric distance functions: With application to European railways. *European Journal of Operational Research,* 117**,** 326-339.

Coelli, T., Rao, D., O'Donnel, C. & Battese, G. 2005. *An Introduction to Efficiency and Productivity Analysis,* Springer, New York.

Colombi, R., Kumbhakar, S., Martini, G. & Vittadini, G. 2014. Closed-skew normality in stochastic frontiers with individual effects and long/short-run efficiency. *Journal of Productivity Analysis,* 42**,** 123-136.

Cooper, W. and C. Lovell. 2011. History lessons. *Journal of Productivity Analysis,* 36, 193-200.

Cooper, Z., Gibbons, S., Jones, S. & Mcguire, A. 2010. Does Hospital Competition Improve Efficiency? An Analysis of the Recent Market-Based Reforms to the English NHS. *Centre for Economic Performance, LSE, CEP Discussion Papers.*

Cooper, Z., Gibbons, S., Jones, S. & Mcguire, A. 2011. Does Hospital Competition Save Lives? Evidence From The English NHS Patient Choice Reforms. *The Economic Journal,* 121**,** F228-F260.

Cooper, Z., Gibbons, S., Jones, S. & McGuire, A. 2012. Does Competition Improve Public Hospitals' Efficiency? Evidence from a Quasi-Experiment in the English National Health Service. *Centre for Economic Performance Discussion Paper 1125*. LSE.

Cornwell, C., Schmidt, P., and Sickles, R. 1990. Production Frontiers with Cross Sectional and Time Series Variation in Efficiency Levels. *Journal of Econometrics,* 46, 185-200.

Cots, W., Chiarello, P., Salvador, X., Castells, X. & Quentin, W. 2011. DRG-based hospital payment: Intended and unintended consequences. In: *Diagnosis Related Groups in Europe.* Open University Press, Maidenhead.

Cotton, M. L. 1976. A Theory of Government Enterprise. *Journal of Political Economy,* 84, 1061-1077.

Crew, M. & Parker, D. 2006. *International Handbook on Economic regulation*, Cheltenham, Edward Elgar.

Cuesta, R. 2000. A Production Model With Firm-Specific Temporal Variation in Technical Inefficiency: With Application to Spanish Dairy Farms. *Journal of Productivity Analysis,* 13, 139-158.

D'amico, F. & Fernandez, J.-L. 2012. Measuring Inefficiency in Long-Term Care Commissioning: Evidence from English Local Authorities. *Applied Economic Perspectives and Policy*, 34, 275-299.

Daidone, S. & Street, A. 2013. How much should be paid for specialised treatment? *Social Science & Medicine,* 84, 110-118.

Dawson, D., Gravelle, H., Jacobs, R., Martin, S. & Smith, P. C. 2007. The effects of expanding patient choice of provider on waiting times: evidence from a policy experiment. *Health Economics,* 16**,** 113-128.

Davis, K., Stremikis, K., Squires, D. & Schoen, C. 2014. Mirror, Mirror On The Wall: How the Performance of the U.S. Health Care System Compares Internationally. New York: The Commonwealth Fund. Available online at http://www.commonwealthfund.org/~/media/files/publications/fund-report/2014/jun/1755_davis_mirror_mirror_2014.pdf Accessed 10/12/2014.

De Alessi, L. 1974. An economic analysis of government ownership and regulation. *Public Choice,* 19**,** 1-42.

Debreu, G. (1951). The Coefficient of Resource Utilization. *Econometrica* ,19, 273-292.

Deloitte. 2014a. Methodology for efficiency factor estimation. London: Deloitte. Available online at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/317572/Supporting_document_A_-_Deloitte_Efficiency_Factor_for_publication352b.pdf Accessed 20/06/2014

Deloitte. 2014b. Evidence for the 2015/16 national tariff efficiency factor. final report. London: Deloitte. Available online at: https://www.gov.uk/government/consultations/nhs-national-tariff-payment-system-201516-engagement-documents accessed 12/12/2014.

Department of Health 2006. Report of the Review of NHS Pathology Services in England Chaired by Lord Carter of Coles. London: Department of Health. Available at http://www.connectingforhealth.nhs.uk/systemsandservices/pathology/projects/nlmc/carterreview2006.pdf Accessed 12th January 2013.

Department of Health. 2007. Payment by Results Guidance 2008/09 London: Department of Health. Available online at http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_081334.pdf accessed 13/5/2014.

Department of Health. 2008. Report of the Second Phase of the Review of NHS Pathology Services in England Chaired by Lord Carter of Coles. London: Department of Health. Available at http://microtrainees.bham.ac.uk/lib/exe/fetch.php?media=review_report_final_proof08.pdf Accessed 12th January 2013

Department of Health. 2009. Consolidation of pathology services. London: Department of Health. Available at http://www.evidence.nhs.uk/qipp Accessed 15th January 2013

Department of Health. 2011. The Operating framework for the NHS in England 2012/13. London: Department of Health. Available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/152683/dh_131428.pdf.pdf Accessed 12th January 2013

Department of Health.2012. Long Term Conditions Compendium of Information: Third Edition. London: Department of Health. Available online at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/216528/dh_134486.pdf accessed 15/10/2014.

Department of Health. 2014. Reference Costs Guidance 2013-14. London: Department of Health. Available online at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/289224/reference_costs_collection_2013-14_2.pdf Accessed 01/12/2014

Department of Health. 2015. Review of Operational Productivity in NHS providers: Interim Report June 2015. London: Department of Health. Available online at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/434202/carter-interim-report.pdf accessed 11/06/2015.

Diabetes UK. 2012. State of the Nation 2012 England. Available online at http://www.diabetes.org.uk/Documents/Reports/State-of-the-Nation-2012.pdf accessed 15/10/2014.

Dimakou, S., Parkin, D., Devlin, N. & Appleby, J. 2009. Identifying the impact of government targets on waiting times in the NHS. *Health Care Management Science,* 12**,** 1-10.

Dismuke, C. & Sena, V. 1999. Has DRG payment influenced the technical efficiency and productivity of diagnostic technologies in Portuguese public hospitals? An empirical analysis using parametric and non-parametric methods. *Health Care Management Science,* 2, 107-116.

Dormont, B., Grignon, M. & Huber, H. 2006. Health expenditure growth: reassessing the threat of ageing. *Health Economics,* 15**,** 947-963.

Dormont, B. & Milcent, C. 2004. The sources of hospital cost variability. *Health Economics,* 13**,** 927-939.

Dowling, B. 1997. Effect of fundholding on waiting times: database study. *BMJ* 315.

Dranove, D. 2012. Health Care Markets, Regulators and Certifiers. *In:* Pauly, M., Mcguire, T. & Barros, P. (eds.) *Handbook of Health Economics.* Oxford: North-Holland.

Dranove, D. & Satterthwaite, M. 2000. The industrial organization of health care markets. *In:* Culyer, A. J. & Newhouse, J. P. (eds.) *The Handbook of Health Economics.* Amsterdam: North-Holland.

Drummond, M., Sculpher, M., Torrance, G., O'Brien, B. & Stoddart, G. 2005. *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press, New York.

Dusheiko, M., Gravelle, H. & Jacobs, R. 2004. The effect of practice budgets on patient waiting times: allowing for selection bias. *Health Economics,* 13**,** 941-958.

Duan, N. 1983. Smearing Estimate: A Nonparametric Retransformation Method. *Journal of the American Statistical Association,* 78, 605-610.

Eakin, K. 2008. Do Physicians minimize Cost? *In:* Fried, H., Lovell, C. & Schmidt, S. (eds.) *The Measurement of Productive Efficiency.* New York: Oxford University Press.

Eijkenaar, F. 2013. Key issues in the design of pay for performance programs. *The European Journal of Health Economics,* 14, 117-131.

Evans, R. G. 1971. "Behavioural" Cost Functions for Hospitals. *The Canadian Journal of Economics / Revue canadienne d'Economique,* 4**,** 198-215.

Farrar, S., Yi, D., Sutton, M., Chalkley, M., Sussex, J. & Scott, A. 2009. Has payment by results affected the way that English hospitals provide care? Difference-in-differences analysis. *BMJ,* 339.

Farrell, M. J. 1957. The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society. Series A,* 120, 253-290.

Farsi, M., Filippini, M. & Kuenzle, M. 2005a. Unobserved heterogeneity in stochastic cost frontier models: an application to Swiss nursing homes, *Applied Economics,* 37, 2127-2141.

Farsi, M., Filippini, M. & Greene, W. 2005b. Efficiency Measurement in Network Industries: Application to the Swiss Railway Companies. *Journal of Regulatory Economics,* 28, 69-90.

Farsi, M., Filippini, M. & Kuenzle, M. 2007. Cost efficiency in the Swiss gas distribution sector. *Energy Economics,* 29**,** 64-78.

Farsi, M. & Filippini, M. 2008. Effects of ownership, subsidization and teaching activities on hospital costs in Switzerland, *Health Economics,* 17, 335-350.

Felder, S. and Tauchmann, H. 2013. Federal State Differentials in the Efficiency of Health Production in Germany: An Artefact of Spatial Dependence? *European Journal of Health Economics,* 14(1): 21-39.

Feldstein, M. 1967. Economic Analysis for Health Service Efficiency: Econometric Studies of the British National Health Service. Amsterdam: North-Holland Publishing Company.

Feng, G. & Serletis, A. 2009. Efficiency and productivity of the US banking industry, 1998–2005: evidence from the Fourier cost function satisfying global regularity conditions. *Journal of Applied Econometrics,* 24**,** 105-138.

Ferrari, A. 2006. The Internal Market and Hospital Efficiency: A Stochastic Distance Function Approach. *Applied Economics,* 38**,** 2121-2130.

Filippini, M. & Greene, W. 2015. Persistent and transient productive inefficiency: a maximum simulated likelihood approach. *Journal of Productivity Analysis***,** 1-10.

France, N. & Francis, G. 2005. Cross-laboratory benchmarking in pathology. *Benchmarking,* 12, 523-538.

Francis, R. 2013. Report of the Mid Staffordshire NHS Foundation Trust Public Inquiry, Volume 1: Analysis of evidence and lessons learned (part 1). London: The Stationery Office. Available online at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/279124/0947.pdf Accessed 12/12/2014.

Fried, H., Knox Lovell, C. A. & Schmidt, S. 2008. *The Measurement of Productive Efficiency,* Oxford, Oxford University Press.

Fried, L. P., Ferrucci, L., Darer, J., Williamson, J. D. & Anderson, G. 2004. Untangling the Concepts of Disability, Frailty, and Comorbidity: Implications for Improved Targeting and Care. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences,* 59**,** M255-M263.

Gallant, A. R. 1982. Unbiased determination of production technologies. *Journal of Econometrics,* 20**,** 285-323.

Gaynor, M., Moreno-Serra, R. & Propper, C. 2013. Death by Market Power: Reform, Competition, and Patient Outcomes in the National Health Service. *American Economic Journal: Economic Policy,* 5**,** 134-66.

Gaynor, M. S., Kleiner, S. A. & Vogt, W. B. 2015. Analysis of Hospital Production: An Output Index Approach. *Journal of Applied Econometrics,* 30**,** 398-421.

Gerdtham, U.-G. & Löthgren, M. 2000. On stationarity and cointegration of international health expenditure and GDP. *Journal of Health Economics,* 19, 461-475.

Gerdtham, U.-G., Søgaard, J., Andersson, F. & Jönsson, B. 1992. An econometric analysis of health care expenditure: A cross-section study of the OECD countries. *Journal of Health Economics,* 11, 63-84.

Geue, C., Briggs, A., Lewsey, J. & Lorgelly, P. 2014. Population ageing and health care expenditure projections: new evidence from a time to death approach. *The European Journal of Health Economics,* 15, 885-896.

Giuffrida, A. & Gravelle, H. 2001. Measuring performance in primary care: econometric analysis and DEA, *Applied Economics,* 33, 163-175.

Glenwright, J., Buckell, J. & Keenan, C. 2014. *New Congenital Heart Disease Review : Activity Analysis Update,* London, NHS England. Available online at http://www.england.nhs.uk/wp-content/uploads/2014/08/chd-7-amended.pdf accessed 25/10/2014.

Goddard, M. & Jacobs, R. 2009. Using Composite Indicators to Measure Performance in Health. *In:* Smith, P., Mossialos, E., Papanicolas, I. & Leatherman, S. (eds.) *Performance Measurement for Health System Improvement.* Cambridge University Press, Cambridge.

Gong, B.-H. & Sickles, R. 1989. Finite sample evidence on the performance of stochastic frontier models using panel data. *Journal of Productivity Analysis,* 1**,** 229-261.

Greene, W. 2004. Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Economics,* 13, 959-980.

Greene, W. 2005. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics,* 126, 269-303.

Greene, W. 2008. The Econometric Approach to Efficiency Analysis. in Fried, H., Lovell, K., and Schmidt, S. (eds.) *The Measurement of Productive Efficiency and Productivity Growth,* Oxford University Press, Oxford.

Greene, W. 2010. A stochastic frontier model with correction for sample selection. *Journal of Productivity Analysis,* 34, 15-24.

Greene, W. 2012a. *LIMDEP version 10*. Econometric Software Inc., New York.

Greene, W. 2012b. *LIMDEP Econometric Modelling Guide*. Econometric Software Inc., New York.

Greene, W. 2012c *Econometric Analysis.* Harlow: Pearson Education Limited.

Gujarati, D. 2003. *Basic Econometrics,* New York: McGraw-Hill.

Gutacker, N., Bojke, C., Daidone, S., Devlin, N., Parkin, D. & Street, A. 2013a. Truly Inefficient or Providing Better Quality of Care? Analysing the Relationship between Risk-Adjusted Hospital Costs and Patients' Health Outcomes. *Health Economics,* 22, 931-947.

Gutacker, N., Bojke, C., Daidone, S., Devlin, N. & Street, A. 2013b. Hospital Variation in Patient-Reported Outcomes at the Level of EQ-5D Dimensions: Evidence from England. *Medical Decision Making,* 33**,** 804-818.

Hajargasht, G. 2014. Stochastic frontiers with a Rayleigh distribution. *Journal of Productivity Analysis***,** online first.

Ham, C. 2013. Balancing budgets or protecting patient safety. *BMJ,* 347.

Hamilton, B. & Bramley-Harker, R. E. 1999. The Impact of the NHS Reforms on Queues and Surgical Outcomes in England: Evidence from Hip Fracture Patients. *The Economic Journal,* 109**,** 437-462.

Haney, A. B. & Pollitt, M. G. 2009. Efficiency analysis of energy networks: An international survey of regulators. *Energy Policy,* 37**,** 5814-5830.

Haney, A. B. & Pollitt, M. G. 2011. Exploring the determinants of "best practice" benchmarking in electricity network regulation. *Energy Policy,* 39**,** 7739-7746.

Haney, A. B. & Pollitt, M. G. 2013. International benchmarking of electricity transmission by regulators: A contrast between theory and practice? *Energy Policy,* 62**,** 267-281.

Hansmann, H. 1988. Ownership of the Firm. *Journal of Law, Economics, and Organization,* 4**,** 267-304.

Harker, R.2012. NHS Funding and Expenditure. House of Commons Social and General Statistics.

Harper, J., Hauck, K. & Street, A. 2001. Analysis of costs and efficiency in general surgery specialties in the United Kingdom. *HEPAC Health Economics in Prevention and Care,* 2**,** 150-157.

Harris, J. E. 1977. The Internal Organization of Hospitals: Some Economic Implications. *The Bell Journal of Economics,* 8**,** 467-482.

Harrison, A. & Appleby, J. 2009. Reducing waiting times for hospital treatment: lessons from the English NHS. *Journal of Health Services Research & Policy,* 14**,** 168-173.

Hartwig, J. 2008. What drives health care expenditure?—Baumol's model of 'unbalanced growth' revisited. *Journal of Health Economics,* 27, 603-623.

Hartwig, J. 2011. Can Baumol's model of unbalanced growth contribute to explaining the secularrise in health care expenditure? An alternative test. *Applied Economics,* 43, 173-184.

Hauck, K. & Street, A. 2007. Do Targets Matter? A Comparison of English and Welsh National Health Priorities. *Health Economics,* 16**,** 275-290.

Hauge, J. & Sappington, D. E. M. 2010. Pricing in Network Industries. *In:* BALDWIN, R., CAVE, M. & LODGE, M. (eds.) *The Oxford Handbook of Regulation.* Oxford: Oxford University Press.

Hausman, J. 1978. Specification Tests in Econometrics *Econometrica,* 46, 1251-1271.

Hawe, E. 2009. Office of Health Economics Compendium of Health Statistics 2009. Abingdon: Radcliffe Publishing Ltd.

Health and Social Care Act 2012, *c.7.* Available online at http://www.legislation.gov.uk/ukpga/2012/7/contents/enacted Accessed 15/10/2014.

Healthcare Commission. 2007. Getting results: Pathology services in acute and specialist trusts. London: Commission for healthcare audit and inspection. Available at http://www.bipsolutions.com/docstore/pdf/16479.pdf Accessed 12th January 2013

Health Select Committee Public Expenditure. 2010. Thirteenth Report of the Session 2010-2012. Available at www.publications.parliament.uk/pa/cm201012/cmselect/cmhealth/1499/149902.htm Accessed 12th January 2013

Hibbard, J. H., Stockard, J. & Tusler, M. 2003. Does publicizing hospital performance stimulate quality improvement efforts? *Health Aff (Millwood),* 22**,** 84-94.

Hicks, J. R. 1935. Annual Survey of Economic Theory: The Theory of Monopoly. *Econometrica,* 3**,** 1-20.

HM Treasury. 2014a. Statistical Bulletin: Public Spending Statistics February 2014. London: HM treasury. Available online at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/418351/PSS_February_2014.pdf accessed 12/12/2013

HM Treasury. 2015. Budget 2015. London: HM Treasury. Available online: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/416330/47881_Budget_2015_Web_Accessible.pdf accessed 23/02/2015.

Holland, D. & Trigg, G. 2010. Clinical Biochemistry/Chemical Pathology Core Report. National Pathology Benchmarking Review 2009/2010. Keele University.

Holland, D. & Trigg, G. 2011. Clinical Biochemistry/Chemical Pathology Core Report. National Pathology Benchmarking Review 2011/2012. Keele University.

Holland, D. & Trigg, G & Brookes, A. 2012. Clinical Biochemistry/Chemical Pathology Core Report. National Pathology Benchmarking Review 2010/2011. Keele University.

Hollingsworth, B. 2003. Non-Parametric and Parametric Applications Measuring Efficiency in Health Care, *Health Care Management Science,* 6, 203-218.

Hollingsworth, B. 2008. The measurement of efficiency and productivity of health care delivery. *Health Economics,* 17, 1107-1128.

Hollingsworth, B. 2012. Revolution, evolution, or status quo? Guidelines for efficiency measurement in health care. *Journal of Productivity Analysis,* 37, 1-5.

Hollingsworth, B., Dawson, P. & Maniadakis, N. 1999. Efficiency Measurement of Health Care: A Review of Non-Parametric Methods and Applications. *Health Care Management Science,* 2, 161-172.

Hollingsworth, B. & Peacock, S. 2008. *Efficiency Measurement in Health and Health Care*, Routledge, Abingdon.

Hollingsworth, B. & Street, A. 2006. The market for efficiency analysis of health care organisations. *Health Economics,* 15**,** 1055-1059.

Horrace, W. & Schmidt, P. 1996. Confidence statements for efficiency estimates from stochastic frontier models. *Journal of Productivity Analysis,* 7**,** 257-282.

HSCIC. 2013a. Summary of Changes OPCS-4.6-OPCS-4.7. Leeds: Health and Social Care information Centre. Available online at http://webarchive.nationalarchives.gov.uk/+/http://www.isb.nhs.uk/documents/isb-0084/amd-10-2013/0084102013summ.pdf Accessed 05/05/2014.

HSCIC. 2013b. HRG4 Summary of Changes. Leeds: Health and Social Care information Centre. Available online at: http://www.hscic.gov.uk/media/11483/HRG4-201314-Local-Payment-Grouper-Summary-of-Changes/pdf/HRG4_LP13-14_Summary_of_Changes_v1.0.pdf accessed 19/03/2014.

Hussey, P. S., De Vries, H., Romley, J., Wang, M. C., Chen, S. S., Shekelle, P. G. & Mcglynn, E. A. 2009. A Systematic Review of Health Care Efficiency Measures. *Health Services Research,* 44**,** 784-805.

Iacobucci, G. 2012. Performance data on all surgeons in England will be published within two years. *BMJ* 245

Iezzoni, L. 2009. Risk adjustment for performance measurement. *In:* Smith, P., Mossialos, E., Papanicolas, S. & Leatherman, S. (eds.) *Performance measurement for health system improvement: experiences, challenges and prospects.* Cambridge: Cambridge University Press.

Jacobs, R., Street, A. & Smith, P. 2006. *Measuring Efficiency in Health Care*. Cambridge University Press, Cambridge.

Jamasb, T. & Pollitt, M. 2007. Incentive regulation of electricity distribution networks: Lessons of experience from Britain. *Energy Policy,* 35**,** 6163-6187.

Jondrow, J., Knox Lovell, C. A., Materov, I. S. & Schmidt, P. 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics,* 19, 233-238.

Joy, R., Velagala, S. & Akhtar, S. 2008. Coding: an audit of its accuracy and implications. *Bulletin of The Royal College of Surgeons of England,* 90**,** 284-285.

Kendall, M. G., and J. D. Gibbons. 1990. Rank Correlation Methods. 5th ed. New York, Oxford University Press.

Kiechle, F. & Main, R. 2002. *The Hitchhiker's Guide to Improving Efficiency in the Clinical Laboratory*. American Association for Clinical Chemistry, Washington.

Kim, Y. & Schmidt, P. 2000. A Review and Empirical Comparison of Bayesian and Classical Approaches to Inference on Efficiency Levels in Stochastic Frontier Models with Panel Data. *Journal of Productivity Analysis,* 14**,** 91-118.

Koopmans, T.C. 1951. An analysis of Production as an Efficient Combination of Activities, In T.C. Koopmans eds., Activity Analysis of Production and Allocation, New York, Wiley.

Kumbhakar, S. C. & Heshmati, A. 1995. Efficiency Measurement in Swedish Dairy Farms: An Application of Rotating Panel Data, 1976–88. *American Journal of Agricultural Economics,* 77**,** 660-674.

Kumbhakar, S. 1990. Production frontiers panel data and time-varying technical inefficiency. *Journal of Econometrics,* 46**,** 201-211.

Kumbhakar, S. C. & Hjalmarsson, L. 1995. Labour-Use Efficiency in Swedish Social Insurance Offices. *Journal of Applied Econometrics,* 10, 33-47.

Kumbhakar, S., Lien, G. & Hardaker, J. B. 2014. Technical efficiency in competing panel data models: a study of Norwegian grain farming. *Journal of Productivity Analysis,* 41, 321-337.

Kumbhakar, S. & Lovell, C. A. K. 2000. *Stochastic Frontier Analysis*. Cambridge University Press, Cambridge.

Kumbhakar, S., Wang, H.-J. & Horncastle, A. 2015. *A Practitioner's guide to Stochastic Frontier analysis,* New York, Cambridge University press.

Kuosmanen, T. & Kortelainen, M. 2012. Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis,* 38**,** 11-28.

Lai, H.-P. & Huang, C. 2010. Likelihood ratio tests for model selection of stochastic frontier models. *Journal of Productivity Analysis,* 34, 3-13.

Laffont, J.-J. & Tirole, J. 2000. *Competition in telecommunications,* Cambridge, MIT Press.

Lago-Peñas, S., Cantarero-Prieto, D. & Blázquez-Fernández, C. 2013. On the relationship between GDP and health care expenditure: A new look. *Economic Modelling,* 32, 124-129.

Lai, H.-P. & Huang, C. 2012. Maximum likelihood estimation of seemingly unrelated stochastic frontier regressions. *Journal of Productivity Analysis,* 37**,** 1-14.

Laudicella, M., Olsen, K. & Street, A. 2010. Examining cost variation across hospital departments–a two-stage multi-level approach using patient-level data. *Social Science & Medicine,* 71, 1872-1881.

Laudicella, M., Li Donni, P. & Smith, P. C. 2013. Hospital readmission rates: Signal of failure or success? *Journal of Health Economics,* 32**,** 909-921.

Leibenstein, H. 1966. Allocative Efficiency vs. "X-Efficiency". *The American Economic Review,* 56**,** 392-415.

Leibenstein, H. & Maital, S. 1992. Empirical Estimation and Partitioning of X-Inefficiency: A Data-Envelopment Approach. *The American Economic Review,* 82**,** 428-433.

Lester, H. & Roland, M. 2009. Performance Measurement in Primary Care. *In:* Smith, P., Mossialos, E., Papanicolas, I. & Leatherman, S. (eds.) *Performance Measurement for Health System Improvement.* Cambridge University Press, Cambridge.

Liebmann, R. 2011. Key performance indicators in pathology - why now? *The Bulletin of The Royal College of Pathologists,* 155, 174-175.

Liston, C. 1993. Price-cap versus rate-of-return regulation. *Journal of Regulatory Economics,* 5**,** 25-48.

Longo, R., Hulme, C. & Smith, A. 2012. Measures of Performance in Primary Care - a Review of Efficiency Studies. *AUHE working paper series*: WP 12_03. University of Leeds.

Lovell, C. 2006. Frontier analysis in healthcare. *International Journal of Healthcare Technology and Management,* 7**,** 5-14.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B. & Leibowitz, A. 1987. Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment. *The American Economic Review,* 77**,** 251-277.

Manning, W. G. 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics,* 17, 283-295.

Mannion, R., Davies, H. & Marshall, M. 2005. Impact of star performance ratings in English acute hospital trusts. *Journal of Health Services Research Policy,* 10**,** 18-24.

Marini, G., Miraldo, M., Jacobs, R. & Goddard, M. 2008. Giving greater financial independence to hospitals—does it make a difference? The case of English NHS Trusts. *Health Economics,* 17, 751-775.

Marques, E., Noble, S., Blom, A. W. & Hollingworth, W. 2014. Disclosing Total Waiting Times For Joint Replacement: Evidence From The English NHS Using Linked HES Data. *Health Economics,* 23**,** 806-820.

Mason, A., Ward, P. & Street, A. 2011. England: The Healthcare Resource Group System. *In:* Busse, R., Geissler, A., Quentin, W. & Wiley, M. (eds.) *Diagnosis-Related Groups in Europe.* Maidenhead: Open University Press.

Maynard, A. 1991. Developing the Health Care Market. *The Economic Journal,* 101**,** 1277-1286.

Maynard, A. 2012. The powers and pitfalls of payment for performance. *Health Economics,* 21**,** 3-12.

Maynard, A. & Ludbrook, A. 1980. Budget Allocation in the National Health Service. *Journal of Social Policy,* 9**,** 289-312.

Meeusen, W. & Broeck, J. V. D. 1977. Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. *International Economic Review,* 18, 435-444.

Mitchell, K. & Onvural, N. M. 1996. Economies of Scale and Scope at Large Commercial Banks: Evidence from the Fourier Flexible Functional Form. *Journal of Money, Credit and Banking,* 28**,** 178-199.

Monitor. 2013a. Improvement opportunities in the NHS: Quantification and Evidence. London: Monitor. Available online: http://www.monitor.gov.uk/sites/all/modules/fckeditor/plugins/ktbrowser/_openTKFile.php?id=37844 accessed 23/02/2014.

Monitor. 2013b. Guidance for the Annual Planning Review 2014/15. London: Monitor. Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/283273/GuidanceAnnualPlanningReview2014-15Revised.pdf accessed 14/02/2015.

Monitor. 2013c. 2014/15 National Tariff Payment System. London: Monitor. Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/300547/2014-15_National_Tariff_Payment_System_-Revised_26_Feb_14.pdf accessed 14/02/15.

Monitor. 2014. 2015/16 National Tariff Payment system: national prices methodology discussion paper. London: Monitor. Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/310128/NT15-16MethodologyDiscussionPaper.pdf accessed 14/02/2015.

Monitor 2015. Approved Costing Guidance. London: Monitor. Available online at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/404708/Approved_costing_guidance_-_17_Feb_2015.pdf accessed 09/05/2015

Mooney, G. 1992. *Economics, Medicine and Health Care,* Hemel Hempstead, Harvester Wheatsheaf.

Moran, V. & Jacobs, R. 2015. Comparing the performance of English mental health providers in achieving patient outcomes. *Social Science & Medicine,* 140**,** 127-135.

Morris, S., Devlin, N. J. & Parkin, D. 2007. *Economic Analysis in Health Care,* Chichester, John Wiley & Sons.

Moulton, B. R. & Randolph, W. C. 1989. Alternative Tests of the Error Components Model. *Econometrica,* 57, 685-693.

Mundlak, Y. 1978. On the Pooling of Time Series and Cross Section Data. *Econometrica,* 46, 69-85.

Murillo-Zamorano, L. & Petraglia, C. 2011. Technical efficiency in primary health care: does quality matter? *The European Journal of Health Economics,* 12, 115-125.

Murray R., Imison, C. & Jabbal, J. 2014. Financial Failure in the NHS: What causes it and how best to manage it. London: King's Fund. Available online at http://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/financial-failure-in-the-nhs-kingsfund-oct14.pdf Accessed 10/12/2014

Murthy, N. R. V. & Ukpolo, V. 1994. Aggregate health care expenditure in the United States: evidence from cointegration tests. *Applied Economics,* 26, 797-802.

Mutter, R., Greene, W., Spector, W., Rosko, M. & Mukamel, D. 2013. Investigating the impact of endogeneity on inefficiency estimates in the application of stochastic frontier analysis to nursing homes, *Journal of Productivity Analysis,* 39, 101-110.

Mutter, R. L., Rosko, M. D., Greene, W. H. & Wilson, P. W. 2011. Translating Frontiers Into Practice: Taking the Next Steps Toward Improving Hospital Efficiency. *Medical Care Research and Review,* 68**,** 3S-19S.

Mutter, R., Wong, H. S. & Goldfarb, M. 2008. The Effects of Hospital Competition on Inpatient Quality of Care. *Inquiry,* 45, 263-279.

Nagendran M, Budhdeo S, Maruthappu M & Sugand K.2010. Should the NHS be Privatized? Annual Varsity Medical Debate – London, 22 January 2010. Philosophy, Ethics and Humanities in Medicine 2010, 5(7): Appendix 1.

National Audit Office. 2012 Progress in making NHS efficiency savings. Available at http://www.nao.org.uk/publications/1213/nhs_efficiency_savings.aspx Accessed 12th January 2013

National Audit Office. 2013. Managing the transition to the reformed health system. London: National Audit Office. Available online at http://www.nao.org.uk/wp-content/uploads/2013/07/10175-001-Managing-the-transition-to-the-reformed-health-system.pdf accessed 25/4/2014.

NERA. 2006. Cost Benchmarking of Air Navigation Service Providers: A Stochastic Frontier Analysis. London: NERA. Available online at http://www.eurocontrol.int/download/publication/node-field_download-4872-0 accessed 09/05/2015

NERA. 2008. The Comparative Efficiency of BT openreach: A report to Ofcom. London. Available online at http://stakeholders.ofcom.org.uk/binaries/consultations/llcc/annexes/efficiency.pdf Accessed 14/09/2013

Nerlove, M. 1963. Returns to Scale in Electricity Supply. *In:* Christ, C. (ed.) *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld.* Stanford: Stanford University Press.

Newhouse, J. P. 1992. Medical Care Costs: How Much Welfare Loss? *The Journal of Economic Perspectives,* 6, 3-21.

Newhouse, J. P. 1994. Frontier estimation: how useful a tool for health economics? *Journal of Health Economics,* 13.

NHS England. 2013. The NHS belongs to the people: A Call to Action - The Technical Annex. London, NHS England. Available at http://www.england.nhs.uk/wp-content/uploads/2013/12/cta-tech-Annex.pdf Accessed 15/12/2014

NHS England. 2014a. Five Year forward View. London: NHS England. Available online at http://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf Accessed 10/02/2015.

NHS England. 2014b. Pathology Quality Assurance Review. London, NHS England. Available at http://www.england.nhs.uk/wp-content/uploads/2014/01/path-qa-review.pdf Accessed 15/12/2014

NHS Confederation. 2010 Clinical responses to the downturn – pathology. Available at http://www.nhsconfed.org/Publications/Documents/Pathology.pdf Accessed 12th January 2013

Niskanen, W. 1971. *Bureaucracy and Representative Government,* Piscataway, American Political Science Association.

OECD. 2014. OECD statistics: Health. Paris OECD. Available online http://stats.oecd.org/ accessed 25/10/2014.

Office for Budget Responsibility. 2014. *Fiscal sustainability report,* London, Office for Budget Responsibility. Available online at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/328924/41298_OBR_Text.pdf accessed 25/10/2014.

Office for National Statistics. 2012a. Expenditure on health care in the UK: 2012. London: ONS. Available online at http://www.ons.gov.uk/ons/rel/psa/expenditure-on-healthcare-in-the-uk/2012/index.html Accessed 12/05/2014

Office for National Statistics. 2012b. Chapter 1: Background and Methodology, 2010-based NPP Reference Volume. London: ONS. Available online http://www.ons.gov.uk/ons/rel/npp/national-population-projections/2010-based-reference-volume--series-pp2/background-and-methodology.html accessed 21/02/2015.

Office for National Statistics.2012c. Population Ageing in the United Kingdom, its Constituent Countries and the European Union. London: ONS. Available online at http://www.ons.gov.uk/ons/rel/mortality-ageing/focus-on-older-people/population-ageing-in-the-united-kingdom-and-europe/rpt-age-uk-eu.html accessed 21/02/2015

Office for National Statistics. 2015. Public Service Productivity Estimates: Health care, 2012. Newport, Wales: Office for National Statistics. Available online at http://www.ons.gov.uk/ons/dcp171766_393405.pdf Accessed 05/01/2014

Ofgem 2009. Electricity Distribution Price Control Review Final Proposals - Allowed revenue - Cost assessment appendix. London: Ofgem. Available online at https://www.ofgem.gov.uk/ofgem-publications/46749/fp3cost-assesment-network-investmentappendix.pdf Accessed 01/05/2013.

Ofgem 2011. Decision on strategy for the next transmission price control - RIIO-T1. London: Ofgem. Available online at https://www.ofgem.gov.uk/ofgem-publications/53833/t1decision.pdf Accessed 03/03/2013

Okunade, A. A. & Murthy, V. N. R. 2002. Technology as a 'major driver' of health care costs: a cointegration analysis of the Newhouse conjecture. *Journal of Health Economics,* 21, 147-159.

Oliver, A. 2009. England. *In:* Rapoport, J., Jacobs, P. & Jonsson, E. (eds.) *Cost Containment and Efficiency in National health Systems.* Morlenbach: Wiley-Blackwell.

Oliver, A. 2012. Markets and Targets in the English National Health Service: Is There a Role for Behavioral Economics? *Journal of Health Politics, Policy and Law*.

Olsen, J. A. 2009. *Principles in Health Economics and Policy,* Oxford, Oxford University Press.

Olsen, K. R. & Street, A. 2008. The analysis of efficiency among a small number of organisations: How inferences can be improved by exploiting patient-level data. *Health Economics,* 17, 671-681.

Orea, L. & Kumbhakar, S. C. 2004. Efficiency measurement using a latent class stochastic frontier model. *Empirical Economics,* 29**,** 169-183.

ORR. 2008. Periodic review 2008: Determination of Network Rail's outputs and funding for 2009-14, London, Office of Rail Regulation. Available online at:

http://www.networkrail.co.uk/browse%20documents/regulatory%20documents/access%20charges%20reviews/pr2008/final%20conclusions/final%20conclusions%20for%20cp4,%20october%202008.pdf accessed 25/10/2014.

ORR. 2013. A guide to the rail programme for Network Rail 2014-19. London: Office of Rail Regulation. Available online at: http://orr.gov.uk/__data/assets/pdf_file/0019/487/guide-periodic-review-2013.pdf accessed 25/10/2014.

Oxford English Dictionary. 2015. *Thesis, n.* . Oxford: Oxford University Press. Available online at http://www.oed.com/view/Entry/200655?redirectedFrom=thesis#eid Accessed 04/03/2015.

Palangkaraya, A. & Yong, J. 2013. Effects of competition on hospital quality: an examination using hospital administrative data. *The European Journal of Health Economics,* 14, 415-429.

Parker, D., Dassler, T. & Saal, D. S. 2006. Performance benchmarking in utility regulation: principles and the UK's experience. *In:* Crew, M. & Parker, D. (eds.) *International Handbook on Economic Regulation.* Cheltenham: Edward Elgar.

Parmeter, C. F. & Kumbhakar, S. C. 2014. Efficiency Analysis: A Primer on Recent Advances. *Foundations and Trends in Econometrics,* 7**,** 191-385.

Peltzman, S. 1976. Towards a More general Theory of regulation. *Journal of Law and Economics,* 19**,** 211-48.

Peter, L. & Hull, R. 1969. *The Peter Principle,* New York, William Morrow.

Pett, P. I. & Clarke, N. M. P. 2012. Clinical Coding in Paediatric Orthopaedics and Its Cost Implications. *Bulletin of The Royal College of Surgeons of England,* 94**,** 1-4.

Pitt, M. and L. Lee, 1981. The Measurement and Sources of Technical Inefficiency in Indonesian Weaving Industry. *Journal of Development Economics,* 9, 43-64.

Polachek, S., W. & Yoon, B. J. 1987. A Two-Tiered Earnings Frontier Estimation of Employer and Employee Information in the Labor Market. *The Review of Economics and Statistics,* 69**,** 296-302.

Portrait, F. R. M., Van Der Galiën, O. & Van Den Berg, B. 2015. Measuring Healthcare Providers' Performances Within Managed Competition using Multidimensional Quality and Cost Indicators. *Health Economics***,** online first.

Pollak, R. A. & Wales, T. J. 1991. The likelihood dominance criterion: A new approach to model selection. *Journal of Econometrics,* 47**,** 227-242.

Price, C. 2005. Benchmarking in pathology medicine: are we measuring the right outcomes?. *Benchmarking,* 12, 449-466.

Propper, C. 1995. Agency and incentives in the NHS internal market. *Social Science & Medicine,* 40**,** 1683-1690.

Propper, C. 1996. Market structure and prices: The responses of hospitals in the UK National Health Service to competition. *Journal of Public Economics,* 61**,** 307-335.

Propper, C. 2012. Competition, incentives and the English NHS. *Health Economics,* 21**,** 33-40.

Propper, C., Burgess, S. & Green, K. 2004. Does competition between hospitals improve the quality of care?: Hospital death rates and the NHS internal market. *Journal of Public Economics,* 88**,** 1247-1272.

Propper, C., Burgess, S. & Gossage, D. 2008. Competition and Quality: Evidence from the NHS Internal Market 1991–99. *The Economic Journal,* 118**,** 138-170.

Propper, C., Croxson, B. & Shearer, A. 2002. Waiting times for hospital admissions: the impact of GP fundholding. *Journal of Health Economics,* 21**,** 227-252.

Propper, C. & Söderlund, N. 1998. Competition in the NHS internal market: an overview of its effects on hospital prices and costs. *Health Economics,* 7**,** 187-197.

Propper, C., Sutton, M., Whitnall, C. & Windmeijer, F. 2010. Incentives and targets in hospital care: Evidence from a natural experiment. *Journal of Public Economics,* 94**,** 318-335.

Propper, C. & Wilson, D. 2003. The Use and Usefulness of Performance Measures in the Public Sector. *Oxford Review of Economic Policy,* 19, 250-267.

Propper, C., Wilson, D. & Burgess, S. 2006. Extending Choice in English Health Care: The Implications of the Economic Evidence. *Journal of Social Policy,* 35**,** 537-557.

Rho, S. & Schmidt, P. 2015. Are all firms inefficient? *Journal of Productivity Analysis,* 43**,** 327-349.

Rice, T. 2003. *The Economics of Health Reconsidered,* Chicago, Health Administration Press.

Roberts, A., Marshall, L. & Charlesworth, A. 2012. *A Decade of Austerity: The Funding Pressures Facing the NHS from 2011/2012 to 2021/22,* London, The Nuffield Trust. Available at http://www.nuffieldtrust.org.uk/sites/files/nuffield/121203_a_decade_of_austerity_full_report_1.pdf Accessed 12th January 2013

Romano, P. S. & Mutter, R. 2004. The evolving science of quality measurement for hospitals: implications for studies of competition and consolidation. *Int J Health Care Finance Econ,* 4**,** 131-57.

Rosenman, R. & Friesner, D. 2004. Scope and scale inefficiencies in physician practices. *Health Economics,* 13, 1091-1116.

Rosko, M. D. & Mutter, R. L. 2008. Stochastic Frontier Analysis of Hospital Inefficiency. *Medical Care Research and Review,* 65, 131-166.

Rosko, M. D. & Mutter, R. L. 2011. What Have We Learned From the Application of Stochastic Frontier Analysis to U.S. Hospitals? *Medical Care Research and Review,* 68**,** 75S-100S.

Rothenburg, J.1951. *Welfare Implications of Alternative Methods of Financing Health Care*. American Economic Review. 41(2), pp 676-687.

Rossi, M. N. A. & Ruzzier, C. A. 2000. On the regulatory application of efficiency measures. *Utilities Policy,* 9**,** 81-92.

Sappington, D. M. & Weisman, D. 2010. Price cap regulation: what have we learned from 25 years of experience in the telecommunications industry? *Journal of Regulatory Economics,* 38**,** 227-257.

Schmidt, P. & Sickles, R. C. 1984. Production Frontiers and Panel Data. *Journal of Business & Economic Statistics,* 2, 367-374.

Scott, A. & Parkin, D. 1995. Investigating hospital efficiency in the new NHS: The role of the translog cost function. *Health Economics,* 4**,** 467-478.

Seshamani, M. & Gray, A. 2004. Ageing and health-care expenditure: the red herring argument revisited. *Health Economics,* 13, 303-314.

Siciliani, L. 2005. Does more choice reduce waiting times? *Health Economics,* 14**,** 17-23.

Siciliani, L. & Martin, S. 2007. An empirical analysis of the impact of choice on waiting times. *Health Economics,* 16**,** 763-779.

Siciliani, L., Sivey, P. & Street, A. 2013. Differences In Length Of Stay For Hip Replacement Between Public Hospitals, Specialised Treatment Centres And Private Providers: Selection or Efficiency? *Health Economics,* 22, 234-242.

Simar, L. & Wilson, P. 2000. Statistical Inference in Nonparametric Frontier Models: The State of the Art. *Journal of Productivity Analysis,* 13**,** 49-78.

Simon, H. A. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics,* 69**,** 99-118.

Shephard, R.(1953). Cost and Production Functions. Princeton: Princeton University Press.

Shleifer, A. 1985. A Theory of Yardstick Competition. *The RAND Journal of Economics,* 16**,** 319-327.

Skinner, B. 1938. *The Behaviour of Organisms,* New York, Appleton Century Crofts.

Sloan, F. 2000. Not-For-Profit Ownership and Hospital Behaviour. In *Handbook of Health Economics*. Elsevier, Amsterdam.

Smee, C. 2005. *Speaking Truth to Power,* Abingdon, Radcliffe Publishing.

Smith, A. 1976 (1776). *An inquiry into the nature and causes of the wealth of nations (ed. R.H. Campbell, A.S. Skinner, and W. B. Todd)*, Oxford University Press.

Smith, A.S.J. 2012. The application of stochastic frontier panel models in economic regulation: Experience from the European rail sector. *Transportation Research Part E: Logistics and Transportation Review,* 48, 503-515.

Smith, A. S. J., Buckell, J., Wheat, P. & Longo, R. 2015. Hierarchical performance and unobservable heterogeneity in health: A dual-level efficiency approach applied to NHS pathology in England. *In:* Greene, W., Sickles, R. C., Khalaf, L., Veall, M. & Voia, M. (eds.) *Productivity and Efficiency Analysis.* New York: Springer. Accepted for publication 16/04/2015.

Smith, A.S.J., Wheat, P. & Smith, G. 2010. The Role of International Benchmarking in Developing Rail Infrastructure Efficiency Estimates. *Utilities Policy,* 18**,** 86-93.

Smith, A.S.J. & Wheat, P. 2012. Estimation of cost inefficiency in panel data models with firm specific and sub-company specific effects. *Journal of Productivity Analysis,* 37, 27-40.

Smith, P. C. 2015. Performance management: the clinician's tale. *Health Economics, Policy and Law,* FirstView**,** 1-4.

Smith, P., Mossialos, E. and Papanicolas, I. 2012. Performance measurement for health system improvement: experiences, challenges and prospects. In: Figueras, J. and McKee, M., eds., *Health systems, health, wealth and societal well-being: assessing the case for investing in health systems.*, New York, McGraw-Hill, 247-280.

Smith, P. C. & Street, A. 2005. Measuring the efficiency of public services: the limits of analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168, 401-417.

Smith, P. & Street, A. 2006. Concepts and Challenges in Measuring the Performance of Health Care Organisatons. *In:* Jones, A. (ed.) *The Elgar Companion to Health Economics.* Cheltenham: Edward Elgar.

Smith, P. and Street, A. 2013. On The Uses of Routine Patient-Reported Health Outcome Data. *Health Economics,* 22, 119-131.

Smith, P. & Sutton, M. 2013. United Kingdom. *In:* Siciliani, L., Borowitz, M. & Moran, V. (eds.) *Waiting Time Policies in the Health Sector: What Works?* OECD Health Policy Studies: OECD Publishing.

Smith, S., Newhouse, J. P. & Freeland, M. S. 2009. Income, Insurance, And Technology: Why Does Health Spending Outpace Economic Growth? *Health Affairs,* 28, 1276-1284.

Söderlund, N. 1997. Impact of the NHS reforms on English hospital productivity: an analysis of the first three years. *British Medical Journal,* 315**,** 1126-1129.

Sørensen, T. H., Olsen, K. R. & Gyrd-Hansen, D. 2009. Differences in general practice initiated expenditures across Danish local health authorities—A multilevel analysis. *Health Policy,* 92, 35-42.

Spence, A. M. 1975. Monopoly, Quality, and Regulation. *The Bell Journal of Economics,* 6**,** 417-429.

Spinks, J. and Hollingsworth, B. 2009. Cross country comparisons of technical efficiency of health production: a demonstration of pitfalls. *Applied Economics,* 41, 417-427.

Stigler, G. 1971. The theory of economic regulation. *The Bell Journal of Economics,* 2, 3-21.

Stigler, G. J. 1976. The Xistence of X-Efficiency. *The American Economic Review,* 66**,** 213-216.

Street, A. 2003. How much confidence should we place in efficiency estimates? *Health Economics,* 12, 895-907.

Street, A. & Jacobs, R. 2002. Relative performance evaluation of the English acute hospital sector. *Applied Economics,* 34**,** 1109-1119.

Street, A., O'Reilly, J., Ward, P. & Mason, A. 2011. DRG-based hospital payment and efficiency: theory, evidence, and challenges. In: *Diagnosis Related Groups in Europe.* Open University Press, Maidenhead.

Szczepura, A., Davies, A., Fletcher, C. & Boussofiane, A. 1993. Efficiency and Effectiveness in General Practice. *Journal of Management in Medicine,* 7, 36-47.

Tandon, A., Murray, C. J. L., Lauer, J. A. & Evans, D. B. 2000. Measuring Overall Health System Performance For 191 Countries. *World Health Organization GPE Discussion Paper Series: No. 30 EIP/GPE/EQC.*

Tiemann, O., Schreyögg, J. & Busse, R. 2012. Hospital ownership and efficiency: A review of studies with particular focus on Germany. *Health Policy,* 104**,** 163-171.

Thaler, R. & Sunstein, C. 2008. *Nudge,* London, Yale University Press.

Train, K. 1991. *Optimal Regulation: The Economic Theory of Natural Monopoly*, Cambridge, MIT Press.

Tsionas, E. G. & Kumbhakar, S. C. 2014. Firm Heterogeneity, Persistent And Transient Technical Inefficiency: A Generalized True Random-Effects Model. *Journal of Applied Econometrics,* 29**,** 110-132.

Valenstein, P., Praestgaard, A. & Lepoff, R. 2001. Six-year trends in productivity and utilization of 73 clinical laboratories: a College of American Pathologists Laboratory Management Index Program study. *Arch Pathol Lab Med,* 125, 1153-61.

Veljanovski, C. 2010. Economic Approaches to Regulation. *In:* Baldwin, R., Cave, M. & Lodge, M. (eds.) *The Oxford Handbook of Regulation* Oxford: Oxford University Press.

Veronesi, E., Mambretti, C. & Gazzaniga, P. 1997. Health care expenditure, laboratory services and IVD market. *Int J Biol Markers,* 12, 87-95.

Viscusi, K., Harrington, J. & Vernon, J. 2005. *Economics of regulation and Antitrust,* Cambridge, MIT Press.

Vuong, Q. H. 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica,* 57, 307-333.

Wagstaff, A. 1987. Measuring technical efficiency in the National Health Service: a stochastic frontier analysis. Centre for Health Economics, University of York.

Walshe, K. 2014. Counting the cost of England's NHS reorganisation. *BMJ,* 349:g6430.

Wang, H.-J., Chang, C.-C. & Chen, P.-C. 2008. The Cost Effects of Government-Subsidised Credit: Evidence from Farmers' Credit Unions in Taiwan. *Journal of Agricultural Economics,* 59**,** 132-149.

Wang, H.-J. & Ho, C.-W. 2010. Estimating fixed-effect panel stochastic frontier models by model transformation. *Journal of Econometrics,* 157**,** 286-296.

Wanless, D. 2002. Securing our Future Health: Taking a Long-Term View. London: HM Treasury. Available online at: http://webarchive.nationalarchives.gov.uk/20130129110402/http://www.hm-treasury.gov.uk/consult_wanless_final.htm accessed 25/12/2013.

Weyman Jones, TG, Boucinha, J, Feteira Inacio, C, Latore, J .2006. Efficiency Analysis for Incentive Regulation. In Coelli, T and Lawrence, DE (eds.) *Performance Measurement and Regulation of Network Utilities*, Edward Elgar, pp.1-50.

Weyman-Jones, T. 2012. Benchmarking Procedures in RIIO for National Grid. London: Ofgem.

Westaby, S. 2014. Publishing individual surgeons' death rates prompts risk averse behaviour. *BMJ* 349.

Wheat, P. 2014. Econometric Cost Analysis in Vertically Separated Railways. Doctoral Thesis, University of Leeds.

Wheat, P. & Smith, A.S.J. 2008. Assessing the Marginal Infrastructure Maintenance Wear and Tear Costs for Britain's Railway Network. *Journal of Transport Economics and Policy,* 42, 189-224.

Wheat, P. & Smith, A.S.J. 2012. Is the choice of (t-T) in Battese and Coelli (1992) type stochastic frontier models innocuous? Observations and generalisations. *Economics Letters,* 116, 291-2.

Wheat, P., Greene, W. & Smith, A.S.J. 2014. Understanding prediction intervals for firm specific inefficiency scores from parametric stochastic frontier models. *Journal of Productivity Analysis***,** 42, 55-65.

WHO. 2004. International Statistical Classification of Diseases and Related Health Problems. Geneva: World Health Organisation.

Willemé, P. & Dumont, M. 2015. Machines that Go 'Ping': Medical Technology and Health Expenditures in OECD Countries. *Health Economics,* 24**,** 1027-1041.

Williamson, O. 1964. *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm,* Chicago, The University of Chicago Press.

Worthington, A. C. 2004. Frontier Efficiency Measurement in Health Care: A Review of Empirical Techniques and Selected Applications. *Medical Care Research and Review,* 61**,** 135-170.

Xu, K. & Holly, A. 2011. The determinants of health expenditure: A country-level panel data analysis. Paris: *OECD.*

Zhang, X., Hauck, K. & Zhao, X. 2013. Patient Safety In Hospitals – A Bayesian Analysis Of Unobservable Hospital And Specialty Level Risk Factors. *Health Economics,* 22**,** 1158-1174.

Zweifel, P., Felder, S. & Meiers, M. 1999. Ageing of population and health care expenditure: a red herring? *Health Economics,* 8, 485-496.

Appendix A: Mean-Scaled Translog

The cost elasticities derived from the translog functional form are, given the squared and cross terms, functions of the variables. This is the feature which allows the elasticities to vary across the range of values in the sample. The drawback of this is that the elasticities are not immediately obvious from the model's coefficients. One solution to this issue, which is commonly employed by researchers, is to mean-scale the variables. This allows the coefficients on the first order terms to be interpreted directly as elasticities at the sample mean. We derive this result below.

Suppose a cross-sectional cost function with two outputs, y_1 and y_2,

$$c_i = \alpha_0 + \beta_1 y_{1,i} + \beta_2 y_{2,i} + \varepsilon_i \qquad (A1)$$

The translog specification of which is,

$$\ln(c_i) = \alpha_0 + \beta_1 \ln(y_{1,i}) + \frac{1}{2}\beta_{11}[\ln(y_{1,i})]^2 + \beta_2 \ln(y_{2,i}) + \frac{1}{2}\beta_{22}[\ln(y_{2,i})]^2 + \beta_{12} \ln(y_{1,i}) . \ln(y_{2,i}) + \varepsilon_i \qquad (A2)$$

Then, the cost elasticity with respect to output y_1 is,

$$\frac{\partial \ln(c)}{\partial \ln(y_1)} = \beta_1 + \beta_{11} . \ln(y_{1,i}) + \beta_{12} . \ln(y_{2,i}) \qquad (A3)$$

Which reflects that the cost elasticity changes over values of y_1 and y_2. If the variables are mean-scaled, as ($\overline{y_{1,\iota}}$ are sample means of variables),

$$\ln(c_i) = \alpha_0 + \beta_1 \ln\left(\frac{y_{1,i}}{\overline{y_{1,\iota}}}\right) + \frac{1}{2}\beta_{11}\left[ln\left(\frac{y_{1,i}}{\overline{y_{1,\iota}}}\right)\right]^2 + \beta_2 \ln\left(\frac{y_{2,i}}{\overline{y_{2,\iota}}}\right) + \frac{1}{2}\beta_{22}\left[ln\left(\frac{y_{2,i}}{\overline{y_{2,\iota}}}\right)\right]^2 + \beta_{12} \ln\left(\frac{y_{1,i}}{\overline{y_{1,\iota}}}\right) . \ln\left(\frac{y_{2,i}}{\overline{y_{2,\iota}}}\right) + \varepsilon_i \qquad (A4)$$

Then the cost elasticity with respect to y_1 becomes,

$$\frac{\partial \ln(c)}{\partial \ln(y_1)} = \beta_1 + \beta_{11} . \ln\left(\frac{y_{1,i}}{\overline{y_{1,\iota}}}\right) + \beta_{12} . \ln\left(\frac{y_{2,i}}{\overline{y_{2,\iota}}}\right) \qquad (A5)$$

At the sample mean, the numerator and denominator are equal and the expressions in brackets in equation (A5) reduce to 1. Then, because $\ln(1) = 0$,

$$\frac{\partial \ln(c)}{\partial \ln(y_1)} = \beta_1 \qquad (A6)$$

This result holds for any number of additional regressors.

Sector	Market	Regulator	Regulatory Regime	Regulatory Period	Regulatory approach	Efficiency/performance modelling	Firms	Years	Observations	Data Years	Variables	Quality	Performance	Efficiency target set	Comments
Transport	Airports	CAA	RPI-X	2014-2019; Q6	Price-capping	TFP indices; RUOE; LEMS; output indices	3	11	33	1997-2006	weighted output; labour, capital and materials data	considered in other measures	TFP change 0.8-0.9 over period after being adjusted and assuming constant capital		sensitivity conducted for different weights attached to elements of output
Transport	Air Traffic Control	CAA	Performance Targets	2015-2019; RP2	Four areas of regulation: safety, environment, capacity and cost efficiency	internal benchmarking	2	5	10	2009-2013	total costs (inflation adjusted); service units (output)	considered in other measures	Ireland - unit cost change between -7.6% and 7%, negative for 3/5 years; UK - unit cost change between -7.1% and 5.6%, positive for 3/5 years	For Ireland, unit cost reductions of between -2.4% and 0.4%, negative for 4/5 years; UK - unit cost change between -10.3% and -2.9%, negative for 5/5 years	
Transport	Rail infrastructure	ORR	REEM	2009/10-2013/14	Overall goal comprising targets set by cost category and asset type	internal benchmarking	1	5	5	2009/10-2013/14	OPEX, CAPEX, revenue, financial information,	reflected in asset enhancement	close to efficiency goal for first 3 years, failed to make required savings for last 2 - 8% below target in 2013/14	23.5% REEM by 2013/14	
Energy	Electricity Distribution	Ofgem	RPI-X	2009/10-2014/15	Price capping based on three components: OPEX, CAPEX and real price effects (inflation and industry technical change)	Fixed effects panel (LSDV with year effects)	14	4	56	2004/05-2008/09	costs: OPEX, OPEX by group; independent variables (limited to two per model): Load, MEAV, CAPEX, no. faults, cables replaced, asset man-hours, spans cut/affected	no	OPEX efficiency 70-128%; Indirect cost efficiency 83-119%	indirect and non-operational CPAEX - upper quartile; network OPEX - upper third of efficiency scores	cost models at disaggregated service levels as well as total OPEX; firm-effects adjusted for through normalising

															variables; gave weights to models to reflect merits of each
Energy	Gas distributio n	Ofgem	RIIO	2013-2021; RIIO-GD1	Revenue allowance based on Incentives , Innovatio n and Outputs	OLS with year dummies; efficiency relative to bottom quartile, C-D specification	8	4	36	2008/09-2011/12	costs: TOTEX = controllable OPEX + shrinkage + smoothed CAPEX + REPEX; outputs: composite scale variables combining network scale (MEAV) and workload drivers (work management, emergency, repairs, maintenance, mains reinforcement, connections, repex)	not in models	efficiency estimates range from 0.89-1.06	cost allowances range from 4-11%	top-down, bottom-up using historical and forecast data. Range of modelling techniques and tests applied; used informatio n quality incentives (IQI) which examines firms' submitted costs vs assessed costs by OFGEM
Energy	Electricity and Gas transmissi on	Ofgem	RIIO	2013-2021; RIIO-T1	Revenue allowance based on Incentives , Innovatio n and Outputs	unit cost comparison	not provided	not provided	not provided	not provided	not provided	not provided	not provided	firms are able to earn up to £170m in additional revenues if perform well; will lose up to £220m in revenues if perform less well	insufficien t informatio n provided in reports; claim to have used an internally held database, but do not provide access; some figures in consultant reports are redacted

Utilities	Water	Ofwat	Menu regulation, including RPI-X and RCV (regulatory capital value) apsects; TOTEX cost performance incentives	2010-2015	Control of allowed revenue	COLS, SFA; C-D and Translog specifications	14 (WoCs); 10 (WaSCs)	5 Water; 7 Sewerage	90 Water; 70 Sewerage	2005-2011	costs: OPEX, CAPEX, TOTEX; outputs: length mains, usage; input prices: LFS, BCIS data; Network: density, metered properties (households and non-household); environmental: sources, pumping head, river sources, reservoir sources, new meters, new mains; quality variables	properties below reference pressure level; leakage; unplanned interruptions; planned interruptions	depends on model; triangulation on approach taken; typical mean around 80% for firms	(+-)2% return of regulatory equity (RoRE) for cap and collar; unique target per firm	
Communications	Postal services	Ofcom	Menu regulation	2011-2014	Regulatory goal is to preserve universal postal service. Four major areas to monitor: financial performance, efficiency, quality for consumers, competition	PVEO (price, volume, efficiencies and other (one-offs); Unit cost analysis; cost per staff; productivity	1	5	5	2009/10-2013/14	unit costs; cost per worker; revenue per worker; for productivity - input - hours, output - workload (volume & mix)	not included in efficiency but is part of wider review - on prices and service quality	PVEO suggests 0.2% improvement; cost per workload (adjusted) declined by 0.2%; productivity improved by 1.7%	none as yet; Royal Mail's own productivity target is 2-3%	working on developing more sophisticated measures; consultation has occurred
Communications	Wholesale broadband access (WBA)	Ofcom	RPI-X	2000-2006		SFA (C-D); DEA	70	8	560	1999-2006	TOTEX; outputs - switched lines, leased lines, total sheath; env - popn density, business residential ratio, fibre proportion, geographical dummies; time	used in later models, included orders completed in specified time (%), faults per 1000 switched lines, hours required to fix faults	one pseudo-firm 7.2% more efficient than upper decile (rank 2nd); other pseudo-firm 6.8% less efficient than upper decile (rank 19th)		no input prices variables; BT openreach added to data set of US firms for analysis; BT split into two 'pseudo-firms'; specification tests for model specification; difficult to make comparison because of how costs are recorded

Appendix B: Summary of Regulators' Efficiency Analyses