'The New NHS': Financial Incentives for Quality?

Irene Papanicolas

London

May, 2011

A thesis submitted to the Department of Social Policy at the London School of Economics for the degree of Doctor of Philosophy.

The London School of Economics and Political Science

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

Irene Papanicolas

Abstract

In April 2002, five years after the Blair government's proposals to create a 'New NHS [National Health Service]', the government outlined the key priorities that would mark the NHS reform. The main reforms involved patient choice supported by a system of 'Payment by Results' (PbR) under which hospitals would be funded on the activity they undertook. PbR is a case based payment system, a type of system increasingly being adopted as the main form of provider payment across industrialised countries. The literature on this type of payment system and experiences from other countries identifies many different behavioural incentives that can have both positive and negative impacts on quality of care. This thesis investigates the quality implications observed so far in England, for seven conditions which represent a spectrum of important clinical areas that are admitted through both emergency and elective admissions.

In order to identify changes in quality, this thesis first considers how to construct an appropriate measure of quality. The first part of the thesis utilizes two different methodological techniques used for quality measurement; a latent variable approach and a technique put forward by McClellan and Staiger (1999) using Vector Autoregressions. The results from these techniques indicate that quality measurement approaches differ markedly with regards to how much measurement and systematic error they are able to filter out of raw outcome data. Finally, the new indicators created by these techniques are used to evaluate the quality impact the introduction of PbR as the main form of hospital payment has had in England. The analysis indicates that since the policy's implementation, there have been differential quality effects on the different conditions. However, for the most part this indicates an improvement in mortality outcomes, and a reduction in the variation of outcomes across hospitals. As found, the interpretation of readmissions has to be approached with caution as more severe patients being kept alive through quality improving measures on mortality create more mixed signals for the readmission indicators. In two conditions we find changes in activity that are indicative of efficiency gains, in the form of better coding and adoption of new technology, both as a result of differences in reimbursement categories.

Acknowledgements

I would like to express my gratitude to all those who offered their expertise, guidance and support throughout the many years it has taken me to complete this Thesis. A special thanks goes to my supervisors, Alistair McGuire and Elias Mossialos, for contributing their expertise and giving me focus, always with patience and support. I would like to thank Peter Smith and Alastair Gray who have both been supportive of my research efforts and provided me with direction over these years. Particular thanks also goes to Champa Heidbrink for her encouragement and support.

A special thank you goes Lucia Kossarova, Sotiris Vandoros, Nikolas Koutroumanos, Azusa Sato, Jon Cylus and Corinna Sorenson for giving me their input as well as lending me their ears and enduring my countless ramblings on quality measures, regressions and figures. My collegues and friends have made this experience much more enjoyable and at times endurable, in particular I would like to thank my flatmates Jen Yang and Emily LaBarge for all their invaluable support, and my friends Thania Drakopoulou, Alexia Vasilikou, Pavlina Papailiopoulou, Sarah-Jane Fenton, Zeba Hirji, Julia Stafford, Ezra Konvitz, Dominic Rose, Ali Sharaf, Tim Hicks, Dickon Ausden and Daniel Hadas. The wonderful formatting would not have been the same without Dimitrios Korres, who I would also like to thank for his great support, ethusiasm and patience.

Finally, I would like to thank my sister and my parents for all their encouragement, guidance and support.

Contents

List of Figures

List of Tables

List of Abbreviations

$D30_{ht}$ 30-Day In Hospital Mortality

$D365_{ht}$ 365-Day Mortality

$E11$ HRG Case Group for AMI with Complications

$E12$ HRG Case Group for AMI without Complications

$H80$ HRG Case Group for Cemented Hip Replacement

$H81$ HRG Case Group for Uncemented Hip Replacement

$R28_{ht}$ 28-Day Readmissions

$R365_{ht}$ 365-Day Readmissions

AMI Acute Myocardial Infarction

CABG Coronary Artery Bypass Graft

CCF Congestive Cardiac Failure

CCF Diagnostic Resource Group

GMM Generalized Method of Moments

GP General Practitioner

HCQI Healthcare Quality Indicators

HES Hospital Episode Statistics

HPS Hospital Provider Spell

HRG Healthcare Resource Group

HSMR Hospital Standardized Mortality Ratios

ICD-10 International Statistical Classification of Diseases, Tenth Revision Procedures

IHD Ischemic Heart Disease

ISTC Independent Specialist Treatment Centre

LOS Lenght of Stay

MFF Market Forces Factor

MI Myocardial Infarction

NHS National Health Service

NICE National Institute for Health and Clinical Excellence

OECD Organization of Economic Co-operation and Development

OPCS4.3 Office of Population Censuses and Surveys Classification, Version 4.3

PbR Payment by Results

PCT Primary Care Trust

PROMS Patient Reported Outcome Measures

SUR Seemingly Unrelated Regression

TIA Transient Ischemic Attack

UK United Kingdom

US United States of America

VAR Vector Autoregression

Part I

Introduction

1 | Introduction

1.1 Introduction

In April 2002, five years after the Blair government's proposals to create a 'New NHS [National Health Service]' (Secretary of State for Health, 1997), the White Paper, 'Delivering the NHS Plan' was published. This publication described key priorities that would mark the NHS reform, namely patient choice supported by a system of 'Payment by Results' (PbR) under which hospitals would be funded on the activity they undertook. These mechanisms were intended to ensure that money would follow the patient to providers, such that stronger incentives existed to improve performance and ensure that the NHS worked for patients (Ham, 2009). The ultimate vision of the [NHS] plan was to provide "prompt, convenient, high quality services, which treat patients as partners" (Secretary of State for Health, 2002).

The payment of health care providers represents one of the most essential tools of the health care system. Not only is provider payment historically one of the largest areas of total hearth care expenditure for most developed countries, but it also has the power to create powerful incentives which can ultimately affect the quantity, quality, equity, efficiency and costs of health interventions. Indeed it was large health care expenditures in the seventies and eighties that led to the development case payment systems, one of the most commonly adopted payment systems today. Pure case payment systems, or activity-based payment, are fixed payment systems used to pay hospitals according to patient characteristics, often measured by Diagnostic Related Groups (DRGs). PbR is an example of a case payment system for hospitals. Under PbR, Primary Care Trusts (PCTs) reimburse hospitals for each procedure they perform though a nationally set tariff that takes into account the diagnosis, mix and complexity of patients receiving care as measured by Healthcare Resource Group (HRG), the English equivalent to the DRG.

DRGs are essentially a classification tool, proposed by Robert Fetter et al. in 1980 as a way of comparing and controlling hospital costs. By dividing patients into diagnostic groups which are weighted according to factors influencing the cost of treatment, relative case groups can be constructed to reflect the difference in the resource utilization of hospitals. These case groups form the basis of 'case payment', as they provide a means with which hospitals can be reimbursed according to indices of their case weighted admissions

adjusting for other hospital factors such as location and status of the hospital. Under a case based system hospitals are paid for the activities they perform, thus encouraging them to respond to patient preferences and demands, and operate more efficiently. Moreover this type of payment system allows costs to become much more transparent. Yet, while this system is beneficial when designed correctly, it also carries with it many risks.

Under this type of system payment for each case is determined ex ante. The payment received by hospitals is thus unrelated to the actual costs incurred during the patients stay. Thus, hospitals may have a financial incentive to discharge patients early, or skimp on quality, as they will not be paid for any additional costs they incur. Under the same logic providers have the incentive to seek out treating low risk (cost) patients, and avoiding high risk (cost) patients. To some extent this can be avoided if the DRG classifications become more specific, and account for the heterogeneity in patient case mix (Dranove and White, 1987; Shleifer, 1985). The financial incentives that derive from such a payment mechanism also depend on how high the case payments are set (Ginsburg and Grossman, 2005; Newhouse, 2002). If payments are relatively generous (above marginal costs), hospitals will try to attract more patients, conversely if payments are low (below marginal costs) hospitals will discourage patients. This financing mechanism may also create motivation to 'game' the system in order to profit. Providers can do this by '' patients, that is classifying them into a higher (more expensive) DRG category in order to receive a bigger payment. For this reason it is necessary to regularly audit case-based systems, and carefully design payment rates and diagnostic codes so that they accurately reflect the costs of providing different types of medical services (Ginsburg and Grossman, 2005; Schreyögg et al., 2006; Street and Maynard, 2007).

The United States first introduced a case-based system, when adopting Diagnosis Related Groups (DRGs) in Medicare and Medicaid (Newhouse, 2002). As this type of system motivates more efficient and cost-effective behaviour (Audit Commission, 2004; Jegers et al., 2002) and it has become increasingly popular. More recently, this system of payment has been adopted in European countries such as England, Austria, Sweden, Germany and others. Evidence from the United States (US) and European systems indicate increased efficiency (Kahn et al., 1990), changes in coding and admissions behaviours (Duckett and Jackson, 2000; Keeler et al., 1990; Kahn et al., 1992; Rogers et al., 2005), as well as incentives that may create instances of behaviour that can adversely influence quality (Newhouse, 2002). Yet, the same underlying payment system can differ between different health systems either because of different design features, such as the size of payment, the breadth of payment or the timing of payment, or because of differences in existing political, organizational and institutional structures. The variation in these parameters across

national health systems make if very difficult to extrapolate findings from one country to the other. Many countries are still in the early years of adopting case-based systems and given the inherent institutional and organizational differences between health care systems there are differences in the effects these policies are having, and will have.

Indeed while case-payments have been, and are being, applied in many health care systems, the English case is unique. Apart from Germany, who is also in the process of applying a case-based system, in no other country are hospital incomes completely determined by activity related payments. Moreover, the national tariff in the English PbR system reflects average costs alone, which apart from France, is not the case for any other country. Most countries apply a more complex pricing system to provide all hospitals with an incentive to improve their performance (Street and Maynard, 2007). These two differences, make the recent PbR policy being applied to England a very interesting case study for both the academic literature in activity-based funding and policymakers alike. It is of interest to examine the English experience in order to determine whether the effectiveness of the payment system differs, given the organizational and institutional differences of the English health system and the PbR policy. The results of such an analysis can be useful not only for informing policy within England, but to be used comparatively to understand how the different structural and design features influence the effectiveness of this type of payment system.

Some work has already being undertaken to examine the effects of the English PbR policy on length of stay, readmissions and volumes of inpatient and emergency activity (Farrar et al., 2009). While case-based reforms are usually focused on improving the efficiency of health care delivery, they do raise concerns about their effects on quality of care, which may be adversely effected as evidence from the US experience has suggested (Cutler, 1995; Kahn et al., 1992; Shen, 2003). Moreover, in 2008, the Audit Commission noted that 53% of doctors were wary of the quality effects the PbR policy would have, and while they also report increases in readmission rates between 2003/04 – 2006/07, yet do not find evidence to attribute these to the PbR policy. Similarly a case study of South Yorkshire (Mannion and Street, 2006) indicated increased tensions between commissioners and providers surrounding issues of 'supplier induced demand and the cost implications of poor clinical coding and other information imperfections that could result under this type of system. Little evidence has been published to suggest that PbR has had any effect on quality of care, however the few studies that address this issue have used crude proxies for quality: in-hospital mortality, 30 day post surgical mortality and emergency re-admissions, metrics that the authors themselves note are insufficiently sensitive (Farrar et al., 2009).

In order to determine how effective PbR has been at creating "prompt, convenient, high quality services, which treat patients as partners", this thesis aims to investigate the effects PbR has had on the quality of health care providers. However, as demonstrated by previous efforts, the measurement of quality of care has proved to be a hurdle in this undertaking. Thus, in order to determine the effects of this reform it is first necessary to be able to first establish a suitable metric for quality of care. Part II is thus devoted to this endeavor, Part III then uses the information generated in Part II to evaluate the effects PbR has had on quality. Before jumping into Parts II and III of the thesis, this chapter will go on to provide some background to both areas. The following section will review the methodological techniques commonly used for quality measurement today. This gives us an opportunity to consider what tools are available to conduct an analysis of this sort, and justify our selection for the methods used. Before concluding we will also revisit the discussion on PbR, going into more detail on why and how we expect it to influence quality in the English setting. Finally, the last section will review in more detail what will be done in each of the thematic chapters, the data used in the analysis, and the choice of conditions selected for this study.

1.2 Background

Measuring Quality

The desire to measure the quality of hospital care dates back to the advent of medicine itself. Yet, the measurement of hospital quality is no easy feat. Health care is complex, multidimensional and the link between clinical practice and patient outcomes is often tenuous at best. Many hurdles face those who attempt to measure quality starting with the seemingly simple task of defining it. As far back as ancient Greece, the challenge in defining quality of care resulted in using list of attributes, categories or features to aid in its conceptualizations. The ancient civilizations of Egypt and Babylon recognized that poor quality care can lead to harm, and good quality care to the absence of harm, however still struggled with a better way to measure it than simply focusing on the final outcome of care (Reerink, 1990). Indeed, up until the pioneering work of Nightingale, Codman and Donabedian, the notion of quality of care while very real in terms of being recognized and appreciated, was a mystery in terms of how to palatably define or measure it.

The first proponents of routine clinical outcome measurement were Florence Nightingale (circa 1860) and Ernest Codman (circa 1900). Nightingale pioneered the systematic and rigorous collection of hospital outcomes data in order to understand and improve performance. While Codman advocated the "end results idea", essentially the common sense

notion of following every patient treated for long enough to determine whether their treatment was successful, and if not to understand and learn from the failures which occurred. Unfortunately, political and practical barriers prevented both these ideas from becoming fully adopted until the last twenty years. Currently, quality of hospital care is often conceptualized with regards to the performance in different domains, and in its measurement indicators range beyond clinical outcomes, such as clinical process measures and resource utilization measures. Avedis Donabedian, whose name is synonymous with quality measurement, advocated the measurement of structure process and outcome rather than the use of only outcomes to measure quality. He argued that "good structure increases the likelihood of good process, and good process increases the likelihood of good outcome" (Donabedian, 1988). Indeed many of the indicators used for quality measurement are often thought of in terms of this framework, and increasingly quality management policies use combinations of the three types of indicators.

Although clinical outcome measures are the gold standard for measuring effectiveness in health care, their use can be problematic, for example if the outcomes cannot realistically be assessed in a timely or feasible fashion, or when trying to understand the contribution of health services to health outcomes. Thus many health services performance initiatives use measures of health care process instead of, or in addition to, measures of outcome. Process measures have certain distinct advantages, for example, they are quicker to measure, and easier to attribute directly to health service efforts (Brook et al., 1996). In addition they are commonly considered a better measure of quality as they examine compliance with what is perceived as best practice. However, they may have less value for patients unless they are related to outcomes, and may be too specific focusing on particular interventions or conditions. Moreover, process measures may ultimately ignore the effectiveness or appropriateness of the intervention and pre-judge the nature of the response to a health problem, which may not be identical in all settings, such as for patients who have multiple morbidities (Klazinga, 2011). In recent years another important development in the assessment of health service performance has been the growing use of patient reported outcome measures. These type of measures typically ask patients to assess their current health status, or aspects of health problems (Fitzpatric, 2009). In England, the routine use of Patient Reported Outocome Measures (PROMS) is growing, with wide-scale adoption in the NHS from 2009 for certain elective procedures.

However, amongst these different measures and dimensions, clinical outcome measures arguably carry the most weight as they are often the most meaningful for stakeholders and more clearly represent the goals of the health system. Even Donabedian himself concluded that, "outcomes, by and large, remain the ultimate validation of the effectiveness

and quality of medical care" (Donabedian, 1966). In the past decades, many industrialized countries have invested large amounts in the development and routine collection of hospital outcome indicators. Indicators are being developed, tested and used in countries such as Austria, Finland Spain, Italy, France, Germany, Australia, the UK and the US, where administrative databases and medical records are able to provide large-scale sources of individual patient level data. These databases allow researchers to easily and relatively cheaply calculate hospital-specific mortality rates which often serve as outcome-based measures of quality. It is easy to see why this type of measure is desirable. A simple indicator that allows the identification of 'good' and 'bad' hospitals can serve as instruments to direct policy and or to inform patient decisions. Indeed for some conditions, routinely available data or this sort has been shown to be as good a predictor of death as some expensive clinical databases (Aylin et al., 2007).

As measures of health outcome are increasingly used to inform policy, statistical researchers have made efforts to address some of the methodological issues associated with them. For example, it is well known that a patient's outcome will be influenced by the severity of their condition, their socio-economic status as well as the resources allocated to their treatment. In such cases, it is critical to employ methods of risk adjustment when using and comparing indicators to help account for these variations in patient populations. Failure to risk adjust outcome measures before comparing patient performance may result in misinterpretation of data which can have serious implications for quality improvement and policy (Iezzoni, 2009). Typically, some sort of risk adjustment technique is employed to address these attribution problems, and control for the other influencing factors. However, many different risk-adjustment mechanisms exist, and are applied differently by different users (Iezzoni, 1994). Thus risk-adjusted measures may not always be comparable with one another (Iezzoni et al., 1996).

Hospital standardized mortality ratios (HSMR) are common risk-adjusted measures used to evaluate overall hospital mortalities. Initially developed by Jarman (Jarman et al., 1999), HSMRs compare the observed numbers of deaths in a given hospital with the expected number of deaths based on national data, after adjustment for factors that affect the risk for in-hospital death, such as age, diagnosis and route of admission (Shojania and Forster, 2008). However, despite their prolific use, many authors express concerns as to the degree of true quality information these indicators hold and implore users of this information to exercise caution in drawing conclusions from them (Birkmeyer et al., 2006; Dimick et al., 2004; Lingsma et al., 2010; Mohammed et al., 2009; Normand, Wolf, Ayanian, and McNeil, Normand et al.; Powell et al., 2003; Shahian et al., 2010). In part, these concerns represent skepticism about how good risk adjustment techniques are at

controlling for differences in for case mix or chance variation. But also, mortality may not always be a valid indicator of quality (Iezzoni, 2009; Shojania and Forster, 2008). For, even when outcome measures are risk adjusted they still run the risk of not accounting for factors that cannot be identified and measured accurately.

Indeed, measures of risk may not be uniformly related to patient outcomes across all hospitals. Certain systematic factors which bias results when these differences are not taken into account. Mistaking such errors for differences in quality is known as "case-mix fallacy". Systematic errors of these sort will lead to erroneous conclusions concerning a variables true value. For example, patterns of use of emergency services may indicate higher degrees of illness in some areas, but poor availability of alternative services in others (Wright and Shojania, 2009). It would me misleading to adjust the data across hospitals according to only one of these assumptions. Mohammed et al. (2009) find systematic associations between hospital mortality rates and the factors used to adjust for case-mix in English Dr. Foster data. Thus, using these measures for case-mix adjustment may actually increase the bias that they are intended to reduce (Lilford et al., 2007; Powell et al., 2003). In these cases standardized mortality ratios, or other risk-adjustment methods may also be misleading. In order to avoid these types of errors it is critical that data collection methods are carefully designed and implemented (Terris and Aron, 2009). Most recently Shahian et al. (2010) present evidence suggesting that the methodology used to calculate hospital wide mortality rates is instrumental in determining the relative 'quality' assigned to a particular hospital. The authors note that rather than suggesting a particular preferred technique for the calculation of hospital mortality, they call into question the very concept of the measurement of hospital-wide mortality.

Moulton (1990) notes that using aggregate variables, such as average death rates, in combination with individual observations by trust or site to determine relationships through regressions or other statistical models runs the risk of producing downwards biased standard errors, and possibly exaggerating the significance of certain effects based on spurious associations. Moreover, while some deaths are preventable, or more dependent on treatment, it is not sensible to look for differences in preventable deaths by comparing all outcomes from one provider. Focusing on mortality rates associated with procedures where the quality of care is known to have a large impact on patient outcomes, such as those that are heavily dependent on technical skill, is in fact more informative (Lilford and Pronovost, 2010).

Indeed, focusing on certain conditions could be considered an extreme form of risk adjustment, where measures focus only on particular conditions, rather than creating organization wide outcome measures. Surgical mortality rates for specific conditions or

procedures have become more popular as they are able to identify key areas where health system quality is more likely to influence outcomes, and where medical progress has been instrumental in improving outcomes. Popular outcome indicators of this sort are 30-day mortality rates for acute myocardial infarction (AMI) and Stroke. Better treatment of AMI in the acute phase has led to reductions in mortality (Capewell et al., 1999; McGovern et al., 2001). The last few decades have seen a dramatic change in care for AMI patients (Klazinga, 2011) first with the introduction of coronary care units in the 1960s (Khush et al., 2005) and then with the advent of treatment aimed at restoring coronary blood flow in the 1980s (Gil et al., 1999). Aside from the contributions from medical technology, improved processes have also contributed to the improvement in outcomes. Research showed that the time from AMI occurrence to re-opening the artery is a key driver of prognosis, and since care processes were changed radically. It is now common for emergency medical personnel to administer drugs, such as aspirin, during patients transport to hospital and emergency departments have instituted procedures to ensure that patients receive definite treatment with thrombolysis or catheterisation within minutes of arrival (Klazinga, 2011). Moreover the proven link between identified care processes and patient outcomes, for conditions such as AMI, allow researchers to be more confident in making judgements about quality and the end result of care. Indeed, there has been considerable work that has used AMI as a proxy for quality both in England (Bloom et al., 2010; Propper et al., 2004, 2008) and internationally (Kessler and McClellan, 1996, 2011; McClellan and Staiger, 1999; Shen, 2003).

The Organization for Economic Co-operation and Development (OECD) Health Care Quality Indicators (HCQI) project, initiated in 2002, which aims to measure and compare the quality of health service provision in the different countries identifies key quality variables that can be used at the acute care level[1]. These indicators include case-fatality rates for AMI and Stroke (OECD, 2010). The Agency for Healthcare Research and Quality in the US, identified seven operations for which they recommended surgical mortality as a quality indicator: Coronary Artery Bypass Graft (CABG) surgery, Repair of Abdominal Aortic Aneurysm, Pancreatic Resection, Esophageal Resection, Pediatric Heart Surgery, Craniotomy and Hip Replacement (Dimick et al., 2004). However, even in cases where there is an established link between treatment and quality, it is not necessarily the case that surgeries are performed frequently enough, in all hospitals, to reliably identify hospitals with increased mortality rates. Indeed, Dimick et al. (2004), attempted to identify

[1]As the HCQI project is concerned with overall health system quality it also identifies suitable quality indicators in other health system domains, including patient safety, health promotion, protection and primary care, patient experiences, cancer care and mental health care. For more information see http://www.oecd.org/health/hcqi.

how many hospitals had an appropriate sample size to determine quality based on these seven conditions. They found that apart from CABG surgery, the remainder of operations for which surgical mortality was advocated as a suitable indicator were not performed frequently enough to make valid assessments of quality. Indeed further work on the relationship between hospital volumes and outcome indicate that mortality rates are poor measures of quality when small numbers of procedures are performed, unfortunately most procedures are not performed frequently enough to allow valid assessment of procedure-specific mortality at the individual hospital level (Birkmeyer et al., 2002). Indeed, most observed variation across hospitals and across time is actually, as a consequence, from random variation (good or bad luck) and does not reflect meaningful changes in quality (Dimick and Welch, 2008).

Another common outcome measure at the hospital level are readmission rates. The measure has become increasingly popular despite the fact that it cannot always be attributed to the quality of care delivered by the hospital. Indeed, McClellan and Staiger (1999) note that high readmissions may be easily misinterpreted as indicators of poor quality when in some cases they may indicate good quality treatment of severe patients. Moreover, readmissions may be the result of poor quality care of other parts of the health system (primary care), behavioural factors (poor adherence), or even the result of good quality care. Benbassat and Taragin (2000) conclude that readmission indicators are not good measures of quality of care for most conditions, as there is large variation in the percentage of the indicator that can be attributed poor quality care. Their own study using reports of different readmission indicators for various conditions indicated a range between 9%-50%. They note that readmissions for specific conditions, such as Child Birth, Coronary Artery Bypass Grafting and Acute Coronary Disease as well as approaches that ensure closer adherence to evidence based guidelines, may be more appropriate.

However, after initial use in the US, there are now a growing number of European countries that measure readmission rates more systematically as a health service outcome (Klazinga, 2011). A recent literature review conducted by Fischer et al, (2010), indicated that of the 360 studies reviewed which used readmission rates as an outcome indicator, only 23 focused on the validity of the indicator and only 14 looked at the specific source of data used to calculate the indicators. The authors concluded that routinely collected data on readmissions alone is most likely insufficient to draw conclusions about quality. Some of the major problems linked to this conclusion was evidence of inaccurate and incomplete coding of the indicator, and little evidence to indicate that readmissions are related with quality of care carried out.

While investigating mortality and readmission rates by different condition may allow a

clearer relationship between outcome and quality of care, other challenges such as random error data quality still persist. Powell et al. (2003) note that variations in outcome will be influenced by change variability which can manifest itself in type 1 or type 2 errors as well as data quality. Both these issues are important, and while the former can be accounted for to some degree using statistical tools the later can seriously undermine conclusions made using the data. The best way to reduce the likelihood of both these types of errors is to have more data, or more precision in the way they are collected. As routine data collection mechanisms are still being developed and improved there is no way to completely avoid this issue. Yet, as Spiegelhalter et al. (2002) note it would be advantageous to have better data on morbidity collected, as mortality data is in most circumstances sufficiently rare, and thus of limited value in monitoring. Regardless, known limitations in the data should always be made explicit when it is used.

Over the past two decades, much empirical research has been done to create improved adjustment mechanisms to make the best use of this information (Iezzoni, 2003). As more organizations begin to use performance systems to make judgements about health service quality and support decision making, more work has been concerned with methodological techniques that can be used to create suitable profiles of provider quality (Landrum et al., 2000). Different statistical techniques have been used to this end, investigating one dimension of care, including Bayesian hierarchical regression models (Normand et al., 1997; Christiansen and Morris, 1997) and maximum likelihood estimates (Silber et al., 1995). These models control for differences in cases per hospital, thus reducing the noise which may produce large differences between observed and expected mortality between hospitals with different sample sizes – due primarily to sampling variability.

However, as quality is multidimensional, this type of focus will limit the focus of comparison across providers, and result in misleading results. However, reporting on too many different types of indicators may create confusion or overwhelm users of performance information, when there are contradictory indicators or simply too much information. So called composite measures, or aggregated measures, may address some of these problems. However there is often much controversy surround them because of the methods required to construct them which often involve weighing different aspects of performance. Yet, different methodological studies have been undertaken to try to find suitable methods to address these issues (Landrum et al., 2000).

Latent variable models have been used to account for the correlation among performance measures and to measure the quality of providers. This type of methodology assumes an unobservable (latent) trait, such as quality, contributes to the attainment of an ultimate outcome. Correlation among different measures is induced by variability in the

latent trait of any one provider, which represents the summary of the unobserved quality they are able to deliver (Landrum et al., 2000). Originally these types of models were used in psychology research (Bentler, 1980; Cohen et al., 1990), but have been applied to many disciplines, including economics where they have been used to measure areas that are not directly observable, such as quality of life (Theunissen et al., 1998). One of the advantages to using this methodology is that it can deal with multidimensionality of data, as it is able to aggregate a large number of observable variables to represent an underlying concept. Chapter 2 uses this approach to measure the quality of different English NHS hospitals in providing services over the period 1996-2008 for seven different conditions. These latent quality measures are then studied cross-sectionally and longitudinally and compared to the raw data in order to try and understand trends in performance across providers and over time.

However, the variability present in latent measures, or in our case in latent quality across providers, will include both a systematic component and a random component. The former can be explained by provider specific covariates and the latter by chance. While the systematic components will also include measures of quality, they may also include other systematic differences that contribute to outcome, such as deprivation or severity, which may bias the measures (Mohammed et al., 2009). Such bias is referred to as systematic error, as discussed previously. In order to correct for these biases, as well as some of the noise still present in the estimates, and create better measures of quality McClellan and Staiger (1999) proposed using multivariate autoregression methods. These models are able to create smoothed out hospital rates of mortality and complications over time as well as to forecast future performance. Chapter 3 applies this technique to the latent estimates calculated in Chapter 2 and assesses the performance of the two measures as compared to one another.

An important question when using any metric of quality that is created specifically from the data of a single condition or procedure is how generalizable these findings are to the organization as a whole. In order to better understand the relationship of quality within a hospital, across conditions, Chapter 4 examines how different outcome measures constructed from these methodologies for different condition are related to each other. We find interesting relationships across the indicators for the different conditions, suggesting that interpretation of the metrics may be more complex and requires careful consideration. The findings of the first four chapters provide us with a thorough understanding of the difficulties in measuring quality, but also with metrics that are sensitive enough to evaluate quality. This allows us to proceed to the second part of the thesis which evaluates quality change since the implementation of the 'New NHS' reforms.

Evaluating Quality

With suitable quality metrics we are able to return to our initial question, and examine the effect PbR has had on hospital quality. International experience and theory suggest the policy can have a number of positive and negative effects (Table 1.1). In theory case-based payments are designed to make providers indirectly compete with one another by setting prices in such a way that they reflect efficient performance. This is usually the main driver behind the adoption of such a policy, as it will help health systems to contain costs by forcing providers to become more efficient. Moreover, this type of system is relatively easy to operate, and if there are not too many case-groups it also can be relatively cheap to administer. International experience suggests that the payment mechanism is successful containing costs, decreasing length of stay, and increasing technical efficiency in many countries who have adopted it (Table 1.2).

However, the cost reductions that make this system so desirable, may not always be linked to only positive behaviours. Indeed, the very pressure put on providers to match the price the are reimbursed for each case group can lead to undesired effects. These often manifest themselves in terms of undesirable changes in activity, adverse effects on quality and gaming.

Table 1.2 indicates the positive and negative effects case payment systems have on different areas of performance. By definition, a case-based payment system removes the economic incentive to over-provide services for any single case, as providers will only be reimbursed a pre-determined tariff for each case-group. However, it may also encourage providers to increase the number of unnecessary admissions, as they will be reimbursed for every extra case (Mannion and Street, 2006). Results form international experience indicates the latter effect, as activity has increased in many countries, but also the former effect in the United States. This may be related to wider structural factors, such as the organization of the health system, but also the payment mechanisms prior to the adoption of case-payment.

In a similar vein, theory predicts that case-based payments will have a negative effect on length of stay (LOS) and a positive effect on quality. Providers are encouraged to minimize costs and be most effective, which can result in lower length of stay and higher quality of care. However, depending on how the payment is set, providers may be encouraged to discharge patients earlier so as not to make losses, in which case there will be a reduction in LOS and also quality. This phenomenon has been referred to as discharging patients 'quicker but sicker' (Duckett and Jackson, 2000) The evidence from international literature suggests that LOS has indeed decreased in most settings. There is also evidence of increases in LOS, often in other areas of care (such as long term care), which has been attributed to

cost-shifting. Moreover, there have been increases in hospital readmissions following the adoption of this type of system, suggesting some element of quality skimping.

Table 1.1: Effects of case-based payments.

Domain	Theory	Evidence
Activity	+/-	• Activity Increase: Australia (Healy et al., 2006; Duckett, 1995); Sweden (Diderichsen, 1995); England (Audit Commission, 2005; Farrar et al., 2009; Sussex and Farrar, 2008; Rogers et al., 2005) • Activity decrease: US (Guterman et al., 1988; Davis and Rhodes, 1988; Kahn et al., 1992; Rosenberg, 2001)
LOS	-	• Reductions in LOS: Austria (Theurl and Winner, 2007); Germany (Schreyögg et al., 2005); USA average LOS (Feder et al., 1987; Newhouse and Byrne, 1988; Shen, 2003), England (Audit Commission, 2005; Sussex and Farrar, 2008) • Increase in LOS: Sweden hospitalization rates (Diderichsen, 1995); USA long-stay patients (Newhouse and Byrne, 1988);
Efficiency	+	• Reductions in costs: Australia (Duckett, 1995); USA (Cutler, 1995; Shen, 2003), England (Rogers et al., 2005; Farrar et al., 2009); • Technology improvement: Austria (Sommersguter-Reichmann, 2000)
Quality	+/-	• Improved Quality of care: USA (Kahn et al., 1990, 1992; Wells et al., 1993) • Higher Readmissions: Australia (Duckett and Jackson, 2000); Austria (Rauner et al., 2003); England (Kahn et al., 1992); USA Kahn et al. (1992) • Decline in Quality of care: Sweden (Forsberg et al., 2001); • No effect on Quality: England no evidence (Farrar et al., 2009)
Gaming	+	• Patient Selection: USA (Ellis and McGuire, 1996; Newhouse, 1989; Meltzer et al., 2002) • Upcoding: Australia (Ellis and Vidal-Fernandez, 2007); Sweden (Mikkola et al., 2002); USA (Chulis, 1991; Ginsburg and Carter, 1986; Sloan et al., 1988) • Cost Shifting: Austria (Sommersguter-Reichmann, 2000); USA (Cutler, 1995; Ellis and Vidal-Fernandez, 2007; Newhouse and Byrne, 1988)

Other behaviours that can adversely effect quality are those which are unwittingly

incentivised by the payment mechanism. These include patient selection, cost-shifting, '', as well as fraud. All of these behaviours are ways to game the payment system and profit. Patient selection occurs when providers try to minimize costs by selecting to treat the less severe patients over the more severe. Cost-shifting deals with the more expensive patients by shifting them to another part of the system, typically funded through a different payment mechanism, such as long-term care or social-care. Upcoding refers to the shifting of patients to similar case-categories which are reimbursed at a higher rate, such as the same case with complications. While fraud is simply lying about the activity recorded in order to increase revenues. The evidence for all of these is limited, but especially so for upcoding and fraud. Indeed, some cases which have been identified as possible instances of upcoding have later been attributed to other factors (Carter et al., 1990).

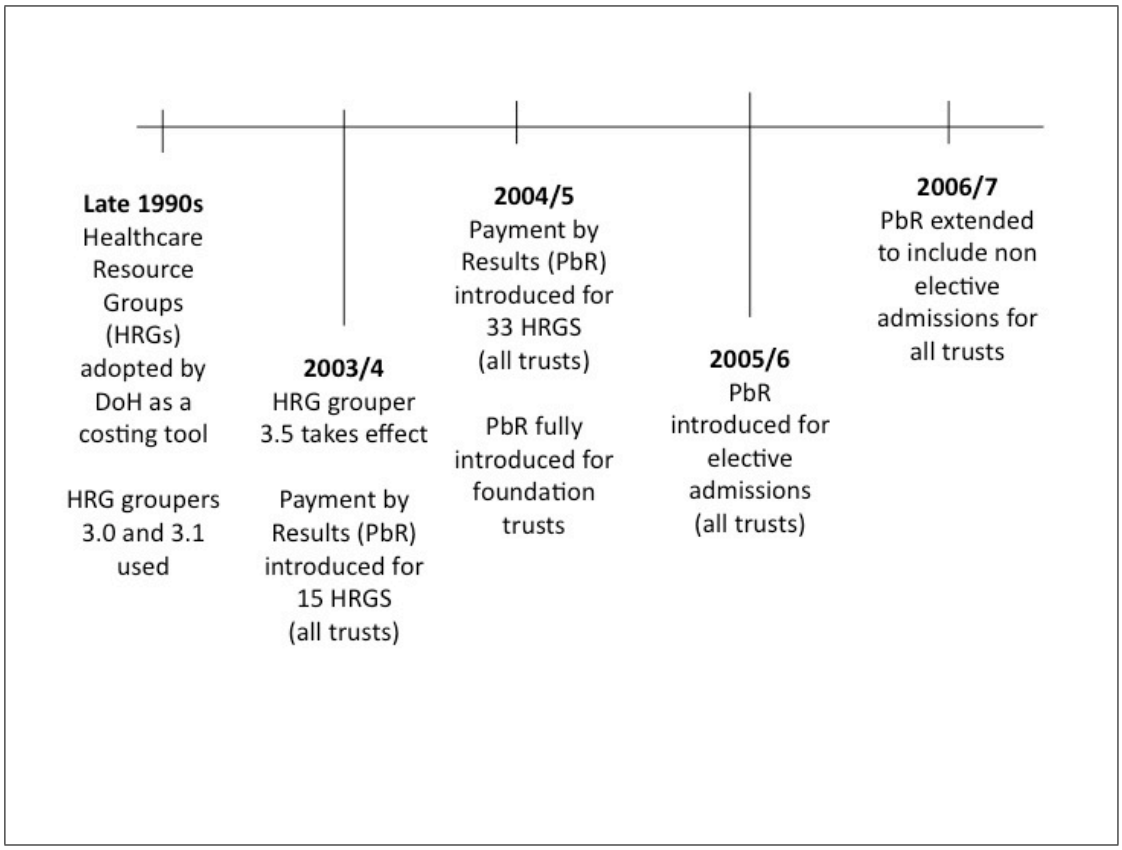
We have reviewed much of the experiences of other countries, in Table 1.1 and also in the introduction, yet, we can only infer so much of that to the English setting. The same underlying payment system can differ between different health systems either because of different design features, such as the size of payment, the breadth of payment or the timing of payment, or because of differences in existing political, organizational and institutional structures. The variation in these parameters across national health systems make if very difficult to extrapolate findings from one country to the other. There are some key features, regarding the organization of the English NHS as well as the implementation of the case payment system, which make it unique, and should be taken into account before attempting to analyse the effectiveness of PbR.

The basis of the English national tariff is an average of all hospital costs for the HRG case-group. HRGs are an English measure of case-mix which allow a clinically meaningful grouping where resource use can be expected to be roughly the same, and thus can have a particular cost ascribed to it. The HRG case-mix is constructed using ICD-10 codes for diagnosis and the OPCS4 classification for procedures, while HRG costs are derived from national reference cost exercises. Street and Dawson (2002) note that due to the organizational structure of the NHS and its lack of a substantial private insurance sector that would require detailed billing data, there is no history of routine patient level cost data collection. Moreover efforts in the mid-eighties to encourage such activities towards this end failed to spark an interest and were abandoned. As a result most hospitals cost activity on the basis of top-down allocations. This is an important difference from many other systems which had detailed billing data prior to the introduction of case payment systems, as it requires providers to make considerable efforts to in coding and collecting cost information.

At the time of their development in the early 1990s, HRGs were not used to reim-

burse providers, but primarily for benchmarking exercises and to set targets to encourage unit cost reductions (Street and Dawson, 2002). Currently the PbR tariff is payable for admitted patient care (elective, non-elective and emergency), outpatient attendances and accident and emergency admissions, while specialist work is excluded. Hospitals also receive a separate payment, the Market Forces Factor, which is based on the geographical price indices for land, labour and building costs. PbR started being implemented in April 2004 to NHS foundation trusts, and was extended to elective activity for all other NHS trusts in April 2005, and to non-elective and outpatient care from April 2007 (Audit Commission, 2004, 2005) (Figure 1.1). The two key differences between the English PbR tariff and that adopted in other settings has already been mentioned previously, namely that hospital incomes are completely determined by this type of payment and that the national tariff reflects the average costs of all hospitals.

Figure 1.1: Timeline of PbR and HRGs.



The academic literature on case payment systems discusses the different incentives associated with alternative tariff setting mechanisms (Ellis and McGuire, 1996; Schreyögg et al., 2006; Street and Maynard, 2007). As discussed previously the English tariff reflects average costs alone, and is calculated from cost data gathered from all hospitals. This

will encourage hospitals to become 'average' rather than to strive to become considerably more efficient (Street and Maynard, 2007). The English case reports the tariff in monetary units, unlike other countries which separate price and underlying cost by using a points system where policy makers decide how much to pay per point (Schreyögg et al., 2006). Ellis and McGuire (1996) warn of the possibility that this type of system will encourage providers to practice upcoding, however they note that this will only result in temporary gains as the higher price level will be factored into future revisions of the tariff. Similarly Street and Maynard (2007) predict there will be a short-term bout of upcoding, but that it will not last as future tariff revisions incorporate this behaviour into the price.

In addition to the reforms of PbR and competition in the time period being investigated there are other changes that may contribute to change in quality of care. Between 1997 and 2008 health expenditure on the English NHS more than doubled in cash terms, going from £55.1 billion to £125.4 billion. Expenditure on health care per capita increased from £231 in 1980 to £1,168 in 2000, and by 2008 it was £1,852 (Boyle, 2011). Moreover, other reforms in this period have been key priorities English health policy since 2000, and will have had an impact on quality. Amongst these are the expanded use of private-sector provision; the introduction of more autonomous management of NHS hospitals through Foundation Trusts; the new General Practitioner GP, Hospital Consultant and Dental Services contracts; the establishment of the National Institute for Health and Clinical Excellence (NICE) and the expansion of its remit to include the development of comprehensive guidelines for all services; and the establishment of the Care Quality Commission to regulate providers and monitor quality of services as well as the expansion of the NHS workforce to include over 50,000 more doctors, including 10,000 more GPs, and almost 100,000 more nurses and midwives (Boyle, 2011). As much as possible these factors need to be controlled for in any analysis of change during this period, and where this is not possible they need to be acknowledged.

Based on the theory of case payments, the design of the English PbR policy, and the overall organizational system and reform of the English NHS, we expect to see quality improvements over the period analysed. In part this will be expected due to the sheer rise in money spent on hospitals, and the greater workforce. However, given the way the tariff is set we do expect to see less variation amongst hospitals, as they are incentivised to deliver 'average' performance. Moreover, we do expect to see some short-term 'upcoding', especially as there is little experience of providers to code for payment historically.

In order to investigate the effects of PbR on quality we conduct two analyses. The first, in Chapter 5, uses a simple fixed effects model to understand how the policy has changed quality, as measured in the previous chapters, in each of the seven conditions.

We are interested in exploring what has happened to average quality over the period as well as relative quality between different hospitals. We find that the policy has had variable effects by condition, but where quality has changed it is mostly improved. However, we do find that there is less variation across providers. The second analysis, in Chapter 6, conducts an in-depth investigation is performed for two pairs of HRG groups where activity appears to have risen for the most expensive case groups at the expense of the cheaper alternative. However, similarly to other studies in the literature we find that what appears to be 'upcoding' behaviour is indeed the result of other phenomena. In our case it is a combination of improved coding and movement to a previously less adopted technology.

1.3 Data

The data used to conduct this analysis is Hospital Episode Statistics (HES) which documents hospital activity in England. The HES database has been in existence since 1987 and is/has been used by the Department of Health to provide performance information at the hospital level (Spiegelhalter et al., 2002). Hospital episode statistics (HES) contain records for all NHS patients admitted to English hospitals in each financial year (April 1 to March 31), with information on all medical and surgical specialties, including private patients treated in NHS hospital trusts. The HES data holds over 15 million patient records each year, stored according the financial year in which the period of care was completed. Each NHS hospital is required to submit data items for each episode in every patient's stay in that hospital. Data is entered from patient's notes onto hospital administration systems by trained clinical coders (Aylin et al., 2007).

Diagnosis of patients are coded using ICD-10 (international statistical classification of diseases, tenth revision) codes while procedures use the UK Office of Population Censuses and Surveys classification (OPCS4). Since the introduction of the internal market, HES data has also been used for contracting between purchasers and providers. The data available in the HES database contains patient characteristic data (e.g. gender, age), clinical information (e.g. diagnoses, procedures undergone), mode of admission (emergency, elective), outcome data (mortality, readmission, discharge location) as well as details on the amount of time spent in contact with the health system (waiting times, date of admission, date of discharge) and details of which hospital the patient was treated in. The HES data we used was accessed through Dr. Foster Intelligence, an independent association dedicated to providing high quality health information.

While HES data is a rich source of information, it requires some manipulation in order to ensure that the total care received by a patient is measured under the same episode. HES data measures the care received under one consultant during the course of

the patient's treatment, in the case that the patient is treated by more than one consultant it is important to identify such patients and link their records of care to provide a complete picture of their care experience. Dr. Foster has done the matching within the HES data and is able to provide information on the complete patient experience. In addition they have linked to other data sources such as the death registries, to provide additional information such as death rates at different intervals (30-days and yearly), readmission rates and further details on the patient, such as further information on their co-morbidities and on some socioeconomic characteristics.

Data on gender and age are used as explanatory variables in the analysis, as is a variable indicating whether the treatment undergone was an elective procedure. The Charlson co-morbidity index which predicts the 1 year mortality for a patient who may have a range of co-morbid conditions was used to control for severity of patients. This index is constructed by assigning a score to each condition depending on the risk of dying associated with it, and summing these scores up (Charlson et al., 1987). Finally, socio-economic status was measured using the Carstairs index of deprivation. This index is based on four census indicators: low social class, lack of car ownership, overcrowding and male unemployment, which are combined to create a composite score. The deprivation score is divided into seven separate categories which range from very low to very high deprivation.

There has been much discussion in the literature regarding the quality of HES data (Aylin et al., 1999; McKee and James, 1997; McKee et al., 1999; Stark et al., 2000; Westaby et al., 2007; Williams and Mann, 2002) as well as that provided by Dr. Foster (Hawkes, 2010a,b; Mohammed et al., 2004, 2009). Aylin et al. (1999) concluded that the HES were reasonably reliable at the broad level of procedure groups but judged that data before 1991 were unreliable. However, they that reported that a number of admissions had missing outcomes, which may be due to failure to link episodes within an admission or simply that no outcome was recorded. Stark et al. (2000) argues that HES data are unreliable and in particular undercount activity compared with departmental records, while McKee and James (1997); McKee et al. (1999) notes the lack of secondary diagnosis coding throughout the database, which may have serious implications for case-mix adjustment. Some authors caution that even mortality may not always be recorded accurately (Lilford et al., 2004; Westaby et al., 2007), although not to such a degree that it would influence standardized hospital mortality rates (Mohammed et al., 2004). Another issue in the data is the increase use of the code for palliative care. In the past five years there has been a big increase in this code, ranging from 7% in some hospitals to 50% in others (Hawkes, 2010a). Patients coded this way are assumed to have come to the hospital to die and are coded differently to prevent putting the blame of their death on the hospital. However, this change in coding

will influence quality as is measured by risk-adjusted mortality.

With regards to the Dr. Foster data in particular, Mohammed et al. (2009) notes that there are systematic differences in the associations between hospital mortality and the factors we use to adjust for patient case-mix, such as age, emergency admissions and co-morbidity. While these differences will play a role in influencing standardized mortality ratios (Wright and Shojania, 2009) they will be accounted for by the McClellan and Staiger (1999) methodology adopted in Chapter 3. Finally, some of the clinical audit literature from the US (Hsia et al., 1988) and England (Cox and Koutroumanos, 2010) suggests that there will also be error in the coding of patients, which will vary from hospital to hospital and by condition. Again, the McClellan and Staiger (1999) technique corrects for measurement error and so is arguably best suited for data facing this sort of variation and inaccuracy.

We requested data for seven conditions for the financial years 1996-2008, namely: Acute Myocardial Infarction (AMI), Myocardial Infarction (MI), other Ischemic Heart Disease (IHD), Congestive Cardiac Failure (CCF), Stroke, Transient Ischemic Aattack (TIA) and Hip Replacement. The data for these conditions was extracted based on the ICD-10 and OPCS 4.3 classification codes indicated in Table 1.2. In most cases there were problems with the sample sizes of some of the years before 2000, and so these years were not included in the analysis. Any hospital trust that had less than 10 admissions throughout the entire period of analysis was dropped from the analysis. Moreover, any primary care trusts, private trusts acting as NHS providers and social care trusts were also excluded. For the sample of patients admitted with AMI, only emergency admissions were examined, and only for patients with a length of stay greater than two days. For the patients admitted with Stroke and Congestive Cardiac Failure, all patients admitted as day cases were excluded.

Seven conditions were chosen in order to be able to evaluate how well the quality measures performed for different clinical areas, different sample sizes and different type of admissions. AMI and MI, otherwise referred to as a heart attack, both refer to an acute blockage of an artery that provides blood to the heart muscle, it is a major health event that almost always results in hospitalization. MI is classified separately in our analysis as it refers to a recurrent heart attack, thus this population should have modifiable risk factors addressed/treated, and are therefore a separate epidemiological subgroup. AMI is commonly used for quality assessment purposes because of the large sample size available, the established link between treatment and survival, and because it is usually an emergency admission which makes patient selection by providers difficult.

Table 1.2: Summary Statistics of the Sample.

Condition	ICD-10/OPCS4.3 codes	Years Analyzed	Mean cases per year	Number of hospitals
AMI	ICD-10: 121	2000-2008	399,560	139
MI	ICD-10: 122,123	2000-2008	7,641	150
IHD	ICD-10: 120,125	2000-2008	142,638	119
CCF	ICD-10: I11.0, I13.0, I25.5, I50.0, I50.1, I50.9, J81X	2000-2008	3,717	122
Stroke	ICD-10: I60 - I67	2000-2008	66,866	167
TIA	ICD-10: G45.0-G45.4, G45.8-G45.9, G46.0-G46.8	2000-2008	12,433	139
Hip	OPCS4.3: W37-W39 W46-W48 W58	1996-2008	40,564	125

IHD refers to other forms of Ischemic Heart Disease, excluding AMI and MI. Hospitalization for these conditions involve similar symptoms to the MI cases but somewhat less severe illness, characterized by inadequate blood flow to the heart that does not actually case death of heart muscle. Hospital treatment in these instances is provided to assure that a heart attack has not occurred, and to attempt to improve blood flow and reduce heart workload to prevent future heart attacks, chest pain or breathing problems. As a performance indicator, IHD is selected for similar reasons to AMI and MI. It is an important clinical domain, accounting for a large burden of illness in England. For this reason it is a well recognized condition and easily coded. Moreover, as many patients are admitted with IHD and thus it offers a large sample size which we have already noted is important in any methodological technique. Finally by including IHD in the analysis we will also be able to investigate any cross correlation of quality at the specialty level.

CCF is characterized by the inability of the heart to supply sufficient blood flow to meet the body's needs. It is a serious condition which needs and responds to conservative (or eventually palliative) management. Treatment for CCF commonly involves lifestyle measures (such as smoking cessation, light exercise including breathing protocols, decreased salt intake and other dietary changes) and medications, and sometimes devices or even surgery. CCF is a chronic condition but is very readily hospitalized when it gets bad

enough or when its in its acute form, mostly Pulmonary Oedema. Patients with Heart Failure are frequently readmitted to hospital and high quality care for these patients have been identified as a key priorities for decreasing morbidity and costs (Chin et al., 1997). As PbR aims to reduce costs but also improve quality it is interesting to include a condition with patients where clinical decline is inevitable. Moreover, especially because this condition differs from the other cardiovascular conditions included in this way, it is of interest to investigate the cross correlation of quality at the specialty level.

A Stroke is characterized by rapidly developing loss of brain function(s) due to disturbance in the blood supply to the brain arising from acute disruption of one of its feeding black vessel (artery). The larger the distribution of the pathological blood vessel, the greater the symptoms. The quality and severity of symptoms however also depend on the assigned function of that brain's area. A Stroke is a medical emergency and can cause permanent neurological damage, complications, and lead to death. They are defined in two main categories, the commoner Ischemic Strokes, where a blood vessel is acutely blocked/obstructed for a variety of reasons (commonest a blood clot or cholesterol-containing vessel wall debris) and Haemorhagic Stroke (much less common) where the diseased blood vessel sustains a rupture/break and the lack of blood supply to the distal tissue is further complicated by the pressure effect of leaking blood within the skull. A Transient Ischemic Attack (TIA), also referred to as a 'mini Stroke', is a temporary disruption/disturbance in the blood supply to a particular area of the brain, resulting in brief neurological dysfunction. If these symptoms persist for more than 24 hours, it is categorized as a Stroke. TIAs are not typically of the haemorhagic type. Stroke case fatality rates are increasingly used as performance indicators, and thus we though it important to include them in the study. They are also well recognized conditions, with a large sample size. TIA was included in the analysis to investigate any cross correlation of quality between the two conditions, as treatment upon admission will follow the same protocol.

Hip Replacement occurs in instances where the hip joint is surgically replaced by a prosthetic implant. Hip Replacement surgery can be performed as total replacement or a half replacement. Total Hip Replacement is defined as the surgical removal of the entire hip joint (ball and socket) and its replacement by a prosthetic (metal or synthetic plastic) implant. A half replacement or hemiarthroplasty, being a slightly less invasive procedure, refers to the replacement of only the ball part of the joint (femoral head). Hip Replacements can be undertaken as elective conditions, with the vast majority of such cases being undertaken to treat Arthritis. The procedure may be the result of an emergency to replace a broken hip. This condition was selected as it is not dominated

by emergency admissions, indeed elective hip arthroplasty are extremely common and extremely successful. Also Hip Replacement is an easily identifiable condition, which will be easily coded and provide us with a large sample of patients to work from.

1.4 Organization of Thesis

A challenge inherent to any study assessing quality change is the measurement of quality. In order to be able to properly evaluate whether quality has changed since the introduction of PbR we believe it is necessary to look beyond readily available outcome measures, or even risk adjusted measures, and identify tools that can be used to produce robust metrics. While ultimately the challenge set forward by this body of work is to evaluate the PbR policy and its impact on quality, we find this complex undertaken lies in an area where statistical methodology and substantive issues are tightly interwoven. For this reason we spend Part II of the thesis replicating a method we feel is sensitive enough to measure quality during this time period and separate it from the noise and systematic bias inherent in the data. Once these metrics have been produced and carefully analysed, we are able to use them to evaluate quality, and examine the effectiveness of the PbR policy in achieving high quality services for NHS patients. The detailed breakdown of the chapters that make this body of work is presented in Table 1.3.

Table 1.3: Breakdown of thematic chapters.

The New NHS: Incentives for improved quality in English hospitals?
Part 1: Introduction
Chapter 1: Introduction
Part 2: Measuring quality
Chapter 2: Using a latent variable approach to measure the quality of English NHS hospitals
Chapter 3: Using a Vector Autoregression Framework to measure the quality of English NHS hospitals
Chapter 4: Examining the persistence of hospital quality across conditions
Part 3: Evaluating Quality
Chapter 5: The effect of Payment by Results on quality of care
Chapter 6: The effect of payment by results on the activity of related procedures
Part 4: Conclusions and Policy Recommendations
Chapter 7: Conclusions and Policy Recommendations

Part II

Measuring Quality

2 Using a latent variable approach to measure the quality of English NHS hospitals

2.1 Introduction

Often timely and relevant data on quality of care does not exist, because it is costly and difficult to collect. Yet, even when such data is available it is not straightforward to use. We are still far from having complete data sets informing us of all the factors that influence outcomes, or measures for the appropriate clinical processes of care required to obtain good outcomes. Indeed, while work on the evidence base of medicine is growing, and many conditions have been identified as amenable to health care (Holland, 1988; Nolte et al., 2004), there still remains considerable uncertainty about the clinical effectiveness of over half of current medical practice (Tovey, 2007; Maynard, 2008). Even with good data, multidimensionality is a problem - as quality of medical care has many dimensions: outcomes, processes and others - ideally all of which would be integrated into a quality evaluation (McClellan and Staiger, 1999).

Increasingly measures of health outcome are used to inform policy, whether it is through investigations into surgical performance (Spiegelhalter et al., 2002), to produce publicly available indicators of performance for hospitals (Healthcare Commission, 2004; New York State Department of Health, 2004), or to produce research to evaluate health care reforms (Farrar et al., 2009; Jarman et al., 1999). Use of outcomes to compare quality of care assumes that variation attributable to other factors can be properly accounted for, such that any residual variation in outcomes, such as observed mortality and morbidity, is indicative of variation in quality of care (Lilford et al., 2004).

While outcome measures are influenced by quality of care they are also a result of data collection techniques, data quality, patient case-mix, and chance. Definitions of outcomes can vary considerably across institutions influencing the comparability of data. Even with a simple outcome such as death, systematic differences can arise with definitions and the

way they are applied across institutions, for example by classifying patients who come to the hospital to die under the palliative code Z51.5 (Hawkes, 2010b). Moreover, trying to identify groups, or cases, of patients whose outcome is being compared can also be difficult. Some cases such as Child Birth or Fractured Neck of Femur are very easy to identify, but other areas such as Stroke or Infertility are harder to classify, and data coding and collection may vary across institutions (Lilford et al., 2004). Both these issues pose challenges in using data for to compare providers.

The type of information will also differ markedly when obtained from different sources such as case records or administrative data. Routine administrative data often contains little information on co-morbidity and severity of disease which will have a large impact on outcome. Clinical databases, run by various bodies, may record more detailed clinical information, but are also likely to vary considerably with regards to data quality (Aylin et al., 2007). In England, HES is often regarded as unreliable by clinicians because of problems in its early years, notably it's inability to secondary diagnoses for most patients (McKee and James, 1997; McKee et al., 1999). However, since then data quality has improved considerably (Audit Commission, 2004). Indeed Aylin et al. (2007) note that, "if suitable predictive models could be developed using this routinely collected information source, they would be a valuable tool for generating measures of performance adjusted for case mix".

The most commonly discussed challenge of using outcome measures as a quality metric is accounting for differences in patient case-mix. Fair comparison of quality between providers needs to consider the differences in patient factors that will influence outcomes, such as patient severity or co-morbidity. Typically, some sort of risk adjustment is employed to address the attribution problems associated with outcome measures, and control for the other influencing factors. Yet, even when outcome measures are risk adjusted, as is the case for hospital standardised mortality ratios (HSMRs) for example, they still run the risk of not accounting for factors that cannot be identified and measured accurately. Mistaking such errors for differences in quality is known as "case-mix fallacy" (Lilford et al., 2004).

Finally, we should not forget the presence of random error, which is inherent to any measurement. Random error can take the form of type 1 (false negatives) or type 2 (false positives) error. The only way to limit the amount of both types of error is to ensure the sample size is large enough. Unfortunately this is not always possible when dealing with health care, as it depends on the prevalence of particular treatments and conditions. Many studies have noted the challenges associated with sample size and outcomes, which limit the number of procedures that can be assessed (Birkmeyer et al., 2002; Dimick et al.,

2004).

In using outcome measures as quality metrics, the challenge lies in controlling for all these factors in order to extract the true quality signal from the measure. With the exception of chance, every other factor has components that can be measured and others that cannot (Lilford et al., 2004). Case-mix adjustment is often used to control for variance in outcome, indeed there is a wide literature on different statistical methods used to produce case-mix adjusted outcomes (Jarman et al., 2005; Iezzoni, 1994; Iezzoni et al., 1996; Iezzoni, 1997; Shahian et al., 2001, 2010). However, worryingly enough different case-mix adjustment methods can produce different results, identifying different providers as good or bad performers depending on the adjustment technique used (Iezzoni et al., 1996; Shahian et al., 2010).

Even if a case-adjustment technique was uniformly applied to providers outcomes could still vary systematically across providers due to differences in the other areas. For example, outcomes could differ because of systematic factors such as systematic differences in data recording, or systematic differences to patient behaviours before and after treatment. For example, hospitals that treat less educated patients may have worse outcomes because they have lower adherence to medications after treatment. An alternative method to control for noise was put forward by McClellan and Staiger (1999) who used information on individual patient characteristics at the individual hospital level to adjust hospital quality measures at the hospital level and thus take account of any systematic biases which are embedded within these measures. The approach, in recognition that the measurement of hospital quality is difficult, begins with latent variable approach to address measurement issues, before going on to refine this measure with a vector autoregression framework.

A latent variable is a variable that can not be observed directly, such as hospital quality. Latent variable assessment takes the observable data and combines it to make assumptions about the unobservable, latent, phenomenon. The latent variable measured by this method will consist of the 'true' variance with both random and systematic errors. This type of technique has been used extensively in the areas of psychological and education testing (Hambleton and Cook, 1977), political science (Treier and Jackman, 2008), and increasingly in epidemiology (Muthén, 1989). In educational testing, information is collected about the subject's ability through their answers on various questions which are indicative of their underlying ability. Conditional on the latent trait, the multiple responses are assumed to be independent observations, and thus correlation amongst the responses is induced by variability in the latent ability amongst the subjects. Thus by modelling their responses the latent ability can be estimated (Landrum et al., 2000).

This model can be extended to health, where the latent trait being modelled is provider

quality. By using hospital-specific intercepts derived from a patient level equation which maps quality of outcome (e.g. 30-day mortality) against patient characteristics we are able to create latent measures of quality, as they pick up the unmeasured systematic aspects which are retained after controlling for observable variation in patient outcomes. While these latent outcome measures will still be noisy, they will filter out much of the estimation error that is otherwise present due to systematic differences in patient mix across hospitals rather than differences in care. Additionally, as these estimates are normally calculated using large samples of patients it is possible to eliminate much of the noise inherent in raw outcome measures, that make quality measures difficult to interpret. While a latent variable approach is essentially another case-mix adjustment technique it is an improvement to other methods as the latent measure provides a composite measure of quality for each provider than result in different outcomes.

These latent outcome estimates of hospital quality thus have a number of attractive features. They can incorporate information on quality measures in a systematic manner, are relatively easy to compute from available data and overcome the risk of over-estimation which is common when aggregate data are combined with individual observations. These hospital intercepts, which estimate the mean value of quality measure holding patient characteristics constant, are therefore less noisy and less likely to be inconsistent estimates than crudely observed aggregate measures of hospital quality. In their analysis, McClellan and Staiger (1999) take an additional step to further control for noise that is present across time periods, and also to address issues of the multidimensional nature of quality. This involves using all the information provided, across time and across dimensions, to create better single point quality estimates. These steps will be discussed in more detail in Chapter 3, which will use the results of this chapter in an attempt replicate their entire method using the latent measures in a Vector Autoregressive framework to analyse the quality of English hospitals for a longer time period and across a wider range of conditions.

In this chapter, we will explore the use of latent variable modelling to measure quality of health care providers. We find that this methodology can be useful for quality measurement in situations where random or systematic measurement error is a problem, where the phenomenon under study is not directly observable, and where many indicators are needed to describe the different aspects of the phenomenon. The next section will outline the empirical methods used in the chapter, before we go one to describe the data used for the model. Finally, we will present the results of the analysis and conclude.

2.2 Empirical Model

This chapter introduces the basic approach which analyses the determinants of hospital quality through a two-step process, with hospitals being the unit of analysis. In the first step, individual patient level data is used to create latent outcome measures at the hospital level using multiple individual outcome measures (mortality and readmission rates), adjusted for individual level patient characteristics. This process allows the amount of noise surrounding these outcome measures to be minimized, thus creating more robust quality measures at the hospital level. The latent outcome measures are then used in the second step to examine how different hospital and social characteristics influence quality of care.

Creating Latent Outcome Measures

The first step of the analysis uses the quality measures provided at the individual patient level over a given amount of years to estimate the relative difference in the mean value of outcomes of each hospital holding patient characteristics constant. These hospital intercepts are estimated using the following equation:

$$Y^k_{iht} = \beta q^k_{1h} + \sum \phi X_{jht} + u_{iht}\,, \tag{2.1}$$

where Y^k represents the quality outcome measure (mortality and readmission rates at different time intervals), with i denoting the individual patient, h the hospital and t the year. $\sum \phi$ represents a group of individual control variables for patient characteristics (age, gender, socioeconomic deprivation, co-morbidities and type of admission). While β represents the fixed effects of all NHS hospitals in which treatment was provided to the sample of patients. The model is run with no constant term, and thus a variable for every hospital can be included. As there is no reference case, the β^k coefficients will represent the intercept value of the hospital's mortality/readmission regression.

This patient level regression is run separately for each year t and quality outcome measure k. By saving the β^k coefficients for each regression, we obtain a vector of hospital intercepts for each quality outcome measure k, and year t. These estimates of quality are appealing as they provide relative rankings of hospital performance contributing to the quality outcomes, controlling for other observable influences. These estimates thus allow some of the noise to be removed from the outcome measures, and thus enable a more appropriate comparison of hospital trusts to one another.

Panel Data Estimation with Lagged Variables

Just as the performance of hospitals in obtaining a desired outcome for their patients will be influenced by the characteristics of patients, it will also depend on other factors such as the characteristics of the hospital itself. The second step of this analysis uses the latent quality measures calculated in the manner explained above and examines how they are associated with key hospital characteristics. This will inform us about what factors our quality metrics are associated with, but can also help draw conclusions about how well the latent indicator performs.

Hospital performance is likely to be influenced by hospital characteristics such as the type of hospital. Different types of hospitals (teaching, acute, specialist, foundation) may be associated with different quality care because of different underlying incentives or management models. For example, in addition to delivering medical care to patients, a teaching hospital will also provide clinical education and training to future health professionals, and also invests in research and technology. These functions may result in different objectives and managerial style which can also contribute to quality differences. Specialist trusts and foundation trusts were introduced to the NHS in 2004. Specialist trusts are dedicated to providing elective care, while a foundation trust is a high performing hospital that has received more managerial and financial freedoms. These differences are likely to have an effect on hospital performance.

Other factors such as the number of patients treated (caseload) may also contribute to overall performance. The relationship between cases and outcomes is not clear. Increased caseload may result in lower quality due to overcrowding, or it can result in higher caseload as doctors become more experienced. Moreover, higher quality may lead to more cases as demand increases, or it can be the result of selecting fewer cases. Similarly average deprivation and co-morbidity is likely to be correlated with the latent outcome measure if there is a confounding relationship between them, that is if they are both correlated with another variable that influences quality. For example if more deprived patients are less likely to adhere to treatment after being discharged and thus are more likely to have bad outcomes.

However, it is unlikely that hospital performance is instantaneously influenced by a change in any of these variables, because of institutional, technological or even psychological reasons (Gujarati, 2003). In an institution structural forces dominate, at least in the short term. For example contractual obligations may prevent hospitals from switching sources of labour, on ancillary services immediately. Thus in certain regards institutions are 'locked in' to current conditions, at least until the medium term. Technological reasons may relate to the adoption of medical innovations. There will be a time gap between

the introduction of a new technology and its adoption in routine health service provision. Finally, the psychological factors refer to the inertia of the status quo, managers and employers take time to adjust to change, thus even if some other factor such as changes in prices occur it may take a transition period for this change to translate into behaviour. Moreover, the performance of a hospital is likely to depend on its past performance. Thus it is important to consider how lags of hospitals own performance, as well as of the other explanatory variables influence the latent variable.

In order to examine the relationship between the exogenous factors discussed above and hospital performance, as measured by the latent variable, the following equations are estimated:

$$D30_{ht} = \alpha + \beta_1 D30_{h(t-n)} + \beta_2 \sum X_{ht} + \beta_3 \sum X_{h(t-1)} + \beta_4 H_{ht} + \epsilon_{ht} \quad (2.2)$$

$$D365_{ht} = \alpha + \beta_1 D365_{h(t-n)} + \beta_2 \sum X_{ht} + \beta_3 \sum X_{h(t-1)} + \beta_4 H_t + \epsilon_{ht} \quad (2.3)$$

$$R28_{ht} = \alpha + \beta_1 D30_{ht} + \beta_2 R28_{h(t-n)} + \beta_3 \sum X_{ht} + \beta_4 \sum X_{h(t-1)} + \beta_5 H_t + \epsilon_{ht} \quad (2.4)$$

$$R365_{ht} = \alpha + \beta_1 D365_{ht} + \beta_2 R365_{h(t-n)} + \beta_3 \sum X_{ht} + \beta_4 \sum X_{h(t-1)} + \beta_5 H_t + \epsilon_{ht} \quad (2.5)$$

where $D30_{ht}$ and $D365_{ht}$ denote the 30-day and 365-day hospital mortality intercepts gained from the first stage analysis, representing the latent mortality for each hospital h at year t, and where $R28_{ht}$ and $R365_{ht}$ denote the 30-day and 365-day term hospital readmission intercepts, representing latent readmissions for each hospital h at year t. The lag variables $D30_{h(t-n)}$ and $D365_{h(t-n)}$ take into account the latent mortality measures of n years prior to year t, while variables X_1 and X_2 control for hospital type and treatment characteristics of the hospital such as average length of stay of patients, number of cases admitted and average waiting time). H_t represents yearly dummies which are intended to capture any contemporaneous shocks that may influence quality.

In any model that includes a lagged dependent variable there is an inherent problem of autocorrelation, which is magnified when the time-series dimension of the data is small (Nickell, 1981). The problem arises because of the correlation of the lagged dependent variable and the error term, and results in making the estimators inconsistent. Including additional regressors does not remove this bias, and if they are correlated with the lagged dependent variable their coefficients may also be seriously biased. This problem can by addressed by estimating a dynamic panel data model, which uses the first differenced Generalized Method of Moments (GMM) estimator (Arellano and Bond, 1991). However, the Arellano-Bond estimator may not perform well if the autoregressive parameters are too large and the time series observations are moderately small. This problem is addressed by the later work of Arellano and Bover (1995) and Blundell and Bond (1998) which impose

additional restrictions on the initial conditions process. The Blundell-Bond estimator is used in this analysis, given the small time-series component available in the newly constructed panel dataset. In addition the Blundell-Bond estimator is able to incorporate lagged levels as well as lagged differences, which increases the efficiency of the model by allowing us to add additional instruments such as hospital characteristics.

The `xtabond2` command in statistical package STATA is used to perform the analysis. A two-step estimator was used, as it is asymptotically more efficient. Robust standard errors were requested, which ensures that the `xtabond2` command includes a finite-sample correction to the two-step covariance matrix derived by Windmeijer (2005), which can make two-step robust estimations more efficient than one-step robust, especially for system GMM, with lower bias and lower standard errors. Following the recommendations made by Roodman (2006) the model is specified so that every repressor is included in instrument matrix, with endogenous variables, such as the lagged dependent variable, specified from two lags, and exogenous predetermined variables specified from one or two lags. Finally, the `xtabond2` command reports the results of the Arellano-Bond AR(1) and AR(2) tests for autocorrelations, as well as the Sargan and Hansen test statistics which indicate how well specified the model is.

2.3 Data

HES data accessed via Dr. Foster Intelligence is used to conduct this analysis. This data is reviewed in detail in the data section of Chapter 1. To undertake the first analysis mortality and readmission rates at different intervals are used as dependent variables. These variables are directly reported in the data, an represent noisy estimates of hospital outcome. The model was estimated for 30-day within hospital mortality rates, designed to measure if a patient dies within up to 30 days in hospital after their initial admission for treatment (with a value of 1 in the case of death and 0 otherwise), 365-day overall mortality rates, 28-day readmission rates which measure whether a patient is readmitted for the same condition in a 28 day period (with a value of 1 for readmission and 0 otherwise), and 365 day readmission rates.

A trust code is used to distinguish each hospital in the data, allowing us to identify which hospitals are performing better or worse. However, we do not report on individual hospitals in this chapter. Data on gender and age are used as explanatory variables in the analysis, as is a variable indicating whether the treatment undergone was an elective procedure. For AMI specifically we only ran the model for patients who were admitted as emergency admissions, as very few patients were admitted, as very few patients were admitted as elective. The Charlson co-morbidity index was used to control for patient co-

morbidity, and the Carstairs index of deprivation was used to control for socio-economic status.

The second part of the analysis considers how explanatory variables at the hospital level influence the latent quality measures. In order to do this we take the latent measures constructed for every outcome measure, for every year and create a new panel data set at the hospital level. This data set contains the newly created latent outcome measures for each year, along with the hospital characteristic information that was available in the individual data. In addition, it is possible to estimate a set of aggregate variables for each year in the data corresponding to the average socioeconomic and demographic characteristics of the patients treated by each hospital, such as mean co-morbidity of patients treated, mean deprivation of patients treated, number of admissions, mean length of stay and mean waiting times.

2.4 Results

This chapter examines the measurement of quality at the hospital level using the two models described previously. Model 1 refers to the model used to construct the latent measures (equation (2.1)) and Model 2 to the analysis of the latent measures (equations (2.2)–(2.5)). The results are presented for the two models, separately by condition. The results for three conditions, AMI, Stroke and Hip Replacement, are presented in the results section and the results for the remaining four conditions, MI, Ischemic heart disease, CCF and TIA, are presented in Appendix A.

For each condition, the results begin with a graphical representation of the the average value of the four outcome measures being studied (30-day and 365-day mortality; and 28-day and 365-day readmissions) to indicate the hospital average, over time, before any analysis is undertaken. In each of these figures, the dashed lines represent the 95% confidence intervals of the estimates representing variation among hospitals. These graphs are useful in order to visually represent the difference between the latent measures and the raw measures.

Model 1

In Model 1, the unit of analysis is the individual patient, and the main focus of interest is the relationship between individual patient's death rates/ readmission rates and the quality of the hospital at which they were treated, controlling for patient characteristics. In the first model a linear regression is used (estimated separately for each year and each outcome indicator) to measure the hospital specific effects that contribute to mortality and

readmission rates, controlling for age, gender, deprivation, co-morbidities and whether the procedure undertaken was elective or emergency. The model is estimated as a regression-through-the-origin, thus R^2 value is not meaningful and not reported. The results of these regressions are presented for each condition and indicate how each of the control variables influence the outcomes measured as the dependent variables. The sign and magnitude for each variable in each regression is as expected, such that age, gender, co-morbidity, deprivation and type of admission significantly influence outcomes for most conditions.

From the Model 1 regressions, the hospital intercepts are extracted and saved, and used as a 'latent outcome measures' of the unobservable hospital specific effect on mortality and readmissions. The latent outcome measure averaged across hospitals is graphically illustrated over time, separately for each outcome measure and condition studied. In these figures, the solid line represents the mean latent outcome measure of all hospitals in the sample, over the time period being evaluated, while the dashed lines indicate the 95% confidence intervals for these estimates representing variation among hospitals. If we imagine hospital quality to be the true underlying signal, surrounded by random, systematic and measurement error, than each hospital's intercept indicates the slope of that curve when graphed over time. As the latent variables are the hospital coefficients in each of the outcome equations, a negative value indicates or a fall in the raw outcome attributable to unobserved hospital effects, while a positive value indicates a rise in the raw outcome. While the 95% confidence intervals represent the variation across hospitals. These diagrams allow us to visualise the average latent measures over time and draw some conclusions about trends in quality. The latent measures show a clearer change in quality over time than apparent from the noisy outcome data. This is consistent with the trends as shown by other risk adjusted measures, such as HSMRs, which indicate that quality is improving (Hawkes, 2010a).

In order to observe the trend in latent outcome measures at the hospital level, the time trend of latent values are illustrated for a selection of four hospitals in each condition, represented in four panels. Of the four hospitals included, there is always a small hospital (upper left), a large hospital (lower right), and two midsize hospitals, where size of hospital is determined relative to the average caseload per year per condition. These hospitals are not a random sample, but chosen to illustrate the data in different settings. The solid line in each of these panels indicates the latent outcome measure estimated from the linear model for the selected hospital, while the two dashed lines indicate the 95% confidence intervals of these estimates. The latent outcome measures have been normalized such that the mean aggregate outcome value of all hospitals for each year is equal to zero, and any deviation from the mean indicates above or below average performance for that hospital.

Negative values indicate lower mortality than average, and positive values indicate higher mortality than average, controlling for patient characteristics. The solid lines can be interpreted as absolute outcome differences, for example a value of 0.02 indicates that the hospital's mortality was 2% above the average hospital in that year. These figures allow an interpretation of how the individual hospitals are performing relative to their peers in all areas evaluated. They indicate that the latent measures are easy to use as performance indicators at the individual hospital level.

AMI

Figure 2.1 indicates the trends over time in raw average mortality and readmission rates for AMI. The trend in average AMI 30-day mortality across hospitals is downward, with short term mortality falling gradually over time, at a gradual pace. Average 365-day mortality is also falling, yet while this trend is gradual for most years there is a large sudden drop from 2005 to 2006, which is not present in the 30-day mortality trend. The trend in average 28-day readmissions indicates an almost negligible increase over time. While, average 365-day readmissions stay relatively constant, rising and falling marginally over the time period studied, such that readmissions at the end of the time period are around the same level as they were in the beginning.

The regression results from Model 1, presented in Table 2.1, indicate that patient characteristics such as age, gender, deprivation and co-morbidities are almost always significant at high levels for all four outcome indicators. Gender significantly impacts mortality and readmissions, such that women have a marginally higher mortality, and higher readmission rates, than men. Existing co-morbidities, as measured by the Charlson Co-morbidity index, significantly increase mortality and readmissions, as do increased deprivation as measured by the Carstairs score. Only emergency admissions are considered for AMI, and so type of admission is not controlled for. Trust dummies included for each hospital and year are highly significant for all four outcome measures.

Figure 2.1: Trends across years in average AMI outcome measures across hospitals.

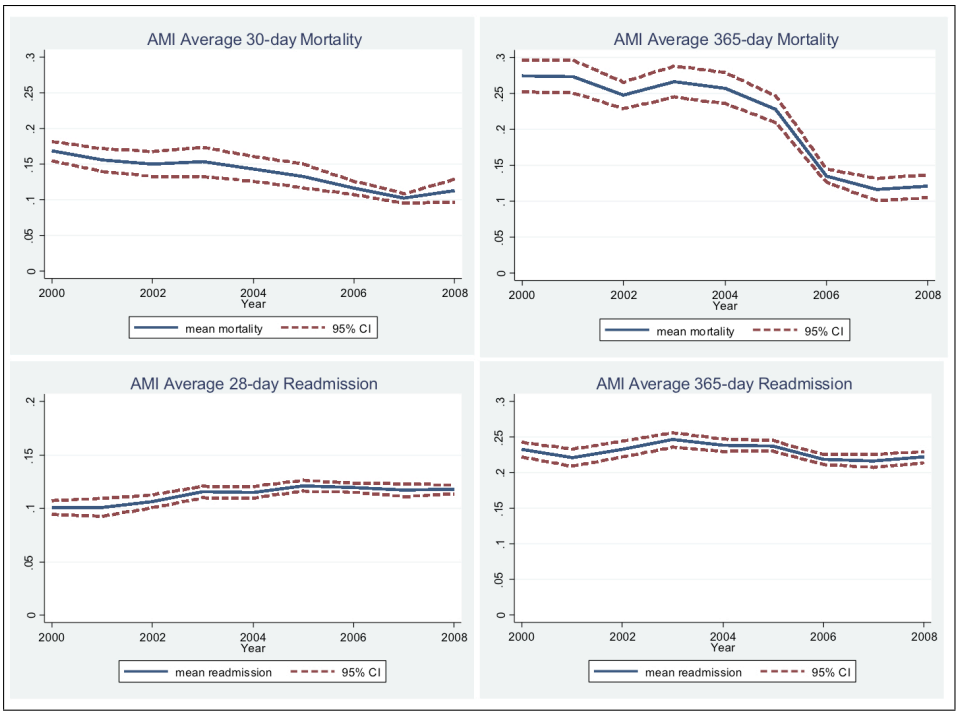


Table 2.1: Regression results for AMI Model 1.

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Trust dummies
30-Day Mortality						
		(0.000)	(0.003)	(0.000)	0.001	
2001	43986	0.004***	0.011***	3.82E-04	0.037***	yes
		(0.000)	0.003	(0.000)	-0.001	
2002	44619	0.004***	0.007***	0.001***	0.039***	yes
		(0.000)	0.003	(0.000)	0.001	
2003	44160	0.004***	0.009***	8.16e-04*	0.040***	yes
		(0.000)	0.003	(0.000)	0.001	
2004	43426	0.004***	0.009***	-2.18E-05	0.035***	yes
		(0.000)	0.003	(0.000)	0.001	
2005	40186	0.004***	0.019***	5.35E-04	0.035***	yes
		(0.000)	0.003	(0.000)	0.001	
2006	37743	0.003***	0.002	0.001***	0.031***	yes
		(0.000)	0.003	(0.000)	0.001	
2007	36240	0.003***	0.001***	1.61E-04	0.032***	yes

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Trust dummies
2000	32950	0.004***	0.006*	0.001**	0.043***	yes
		(0.000)	0.003	(0.000)	0.001	
2008	33607	0.003***	0.001	4.29E-04	0.031***	yes
		(0.000)	0.003	(0.000)	0.001	
365-Day Mortality						
2000	37346	0.011***	0.010**	0.004***	0.090***	yes
		(0.000)	-0.004	-0.001	-0.002	
2001	49780	0.011***	0.016***	0.003***	0.086***	yes
		(0.000)	-0.004	-0.001	-0.002	
2002	50711	0.011***	0.017***	0.004***	0.084***	yes
		(0.000)	-0.004	-0.001	-0.002	
2003	50202	0.011***	0.027***	0.004***	0.086***	yes
		(0.000)	-0.004	-0.001	-0.002	
2004	49762	0.011***	0.014***	0.003***	0.078***	yes
		(0.000)	-0.004	-0.001	-0.002	
2005	46914	0.010***	0.020***	0.004***	0.075***	yes
		(0.000)	-0.004	-0.001	-0.002	
2006	45133	0.006***	0.012***	0.003***	0.042***	yes
		(0.000)	-0.003	-0.001	-0.002	
2007	44026	0.005***	0.017***	0.002***	0.040***	yes
		(0.000)	-0.003	-0.001	-0.001	
2008	41474	0.005***	0.009***	0.001**	0.039***	yes
		(0.000)	-0.003	-0.001	-0.001	
28-Day Readmission						
2000	32950	4.80e-04***	0.007*	0.002***	0.010***	yes
		(0.000)	-0.004	-0.001	-0.002	
2001	43986	5.71e-04***	0.013***	0.002***	0.009***	yes
		(0.000)	-0.003	-0.001	-0.002	
2002	44619	5.40e-04***	0.011***	0.001**	0.011***	yes
		(0.000)	-0.003	-0.001	-0.002	
2003	44160	6.64e-04***	0.022***	0.001**	0.008***	yes
		(0.000)	-0.003	-0.001	-0.002	
2004	43426	9.75e-04***	0.008***	0.002***	0.011***	yes
		(0.000)	-0.003	-0.001	-0.002	

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Trust dummies
2000	32950	0.004***	0.006*	0.001**	0.043***	yes
2005	40186	0.001***	0.017***	0.002***	0.013***	yes
		(0.000)	-0.004	-0.001	-0.002	
2006	37743	0.001***	0.015***	0.001	0.014***	yes
		(0.000)	-0.004	-0.001	-0.002	
2007	36240	0.001***	0.016***	0.003***	0.015***	yes
		(0.000)	-0.004	-0.001	-0.002	
2008	33607	0.001***	0.017***	0.002***	0.013***	yes
		(0.000)	-0.004	-0.001	-0.002	
365-Day Readmission						
2000	37346	0.001***	0.008*	0.005***	0.018***	yes
		(0.000)	-0.005	-0.001	-0.003	
2001	49780	0.001***	0.022***	0.007***	0.020***	yes
		(0.000)	-0.004	-0.001	-0.002	
2002	50711	0.001***	0.017***	0.005***	0.021***	yes
		(0.000)	-0.004	-0.001	-0.002	
2003	50202	0.002***	0.026***	0.005***	0.018***	yes
		(0.000)	-0.004	-0.001	-0.002	
2004	49762	0.002***	0.023***	0.005***	0.020***	yes
		(0.000)	-0.004	-0.001	-0.002	
2005	46914	0.001***	0.026***	0.005***	0.022***	yes
		(0.000)	-0.004	-0.001	-0.002	
2006	45133	0.001***	0.030***	0.004***	0.022***	yes
		(0.000)	-0.004	-0.001	-0.002	
2007	44026	0.002***	0.027***	0.006***	0.022***	yes
		(0.000)	-0.004	-0.001	-0.002	
2008	41474	0.002***	0.029***	0.004***	0.020***	yes
		(0.000)	-0.004	-0.001	-0.002	

* Significant at $p \leq 0.1$

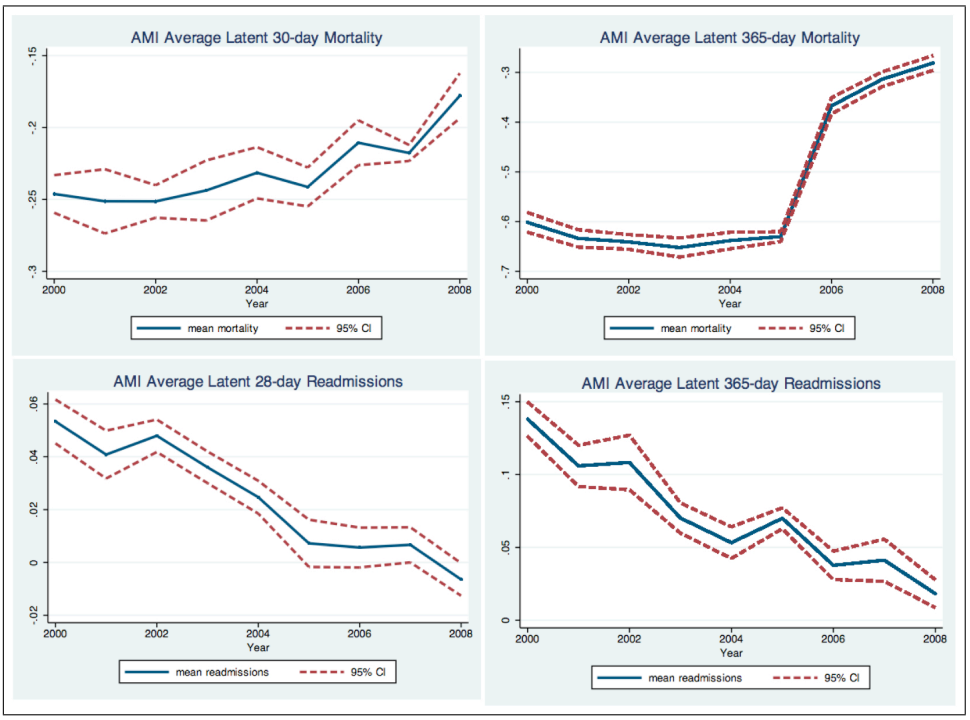
** Significant at $p \leq 0.05$

*** Significant at $p \leq 0.01$

Figure 2.2 shows the trends over time in average latent mortality and readmission rates for AMI, such that the curve in each panel represents the rate of change in the

raw outcomes over time controlling for patient characteristics. The mean hospital latent mortality outcomes are negative for both 30-day and year-long mortality, indicating that on average rate of change in mortality attributable to each hospital is decreasing. However, for both outcomes the values are becoming more positive over time, meaning that the mean is decreasing at an increasing rate. The confidence intervals for 30-day mortality show a variation of just under 2% in the beginning of the sample, which narrow after 2005 to about 1%. This indicates less variation in the quality of hospitals towards the end of the time period. The panels for readmissions are mostly positive, indicating increasing readmissions attributable to hospital performance are increasing, on average. However, in both readmission panels the points are approaching zero indicating that they are increasing at a decreasing rate. Indeed for 28-day readmissions the value falls below zero for 2007-2008, indicating decreasing readmissions in this period.

Figure 2.2: Trends across years in average latent AMI outcome measures across hospitals.



Figures 2.3 and 2.4, show the trend in AMI latent 30-day and 365-day mortality for four selected hospitals. The confidence intervals for both figures, show more variation in latent mortality within hospitals than indicated by either of the averages plotted in Figure 2.2. For both short term and long term mortality, estimates within hospitals range from over 5% above average to more than 7% below average. There hospital specific panel indicate year-to-year variations of performance, commonly around 3-4% in either direction. Figures 2.5 and 2.6 show AMI latent 28-day and 365-day readmissions for the

same four hospitals. The confidence intervals for both readmission figures show about the same degree of within hospital variation among latent readmissions than was observed for latent mortality, of about 5%. The magnitude of the year-to-year variation, is also similar, but varies according to the hospital.

The latent outcome measures graphed by individual hospital allow quality comparisons to be made between different providers, and examination of their own quality trajectory through time. For example, in Figure 2.3, the midsize hospital in the bottom left hand panel is clearly providing above average quality for the entire time period, as the estimate and the entire 95% confidence interval lie below 0, or the average mortality rate of all hospitals. However, the trend of the estimate over suggests that while mortality rates are below average, they are steadily increasing relative to its peers. Looking at the latent long term mortality rates for the same hospital (Figure 2.4) indicates that the provider performs relatively worse on this outcome measure in the later time periods. While 365-day latent mortality rates started off below average in 2000, they have steadily increased throughout the sample until they are unequivocally above average in 2008.

Figure 2.3: Trends across years in latent AMI 30-day mortality for selected hospitals.

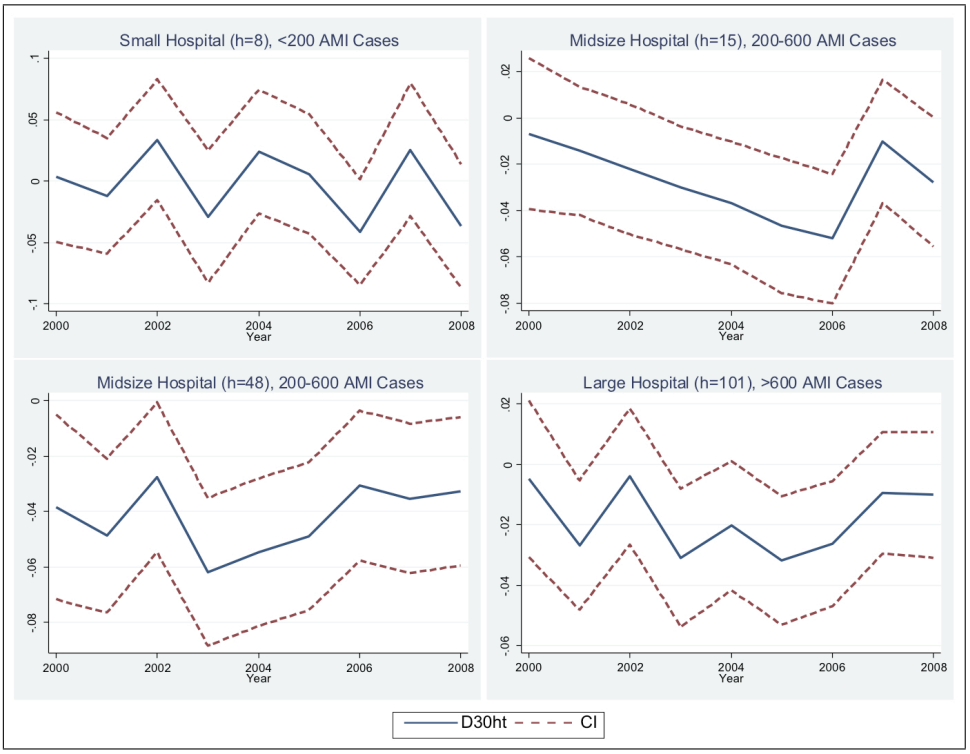


Figure 2.4: Trends across years in latent AMI 365-day mortality for selected hospitals.

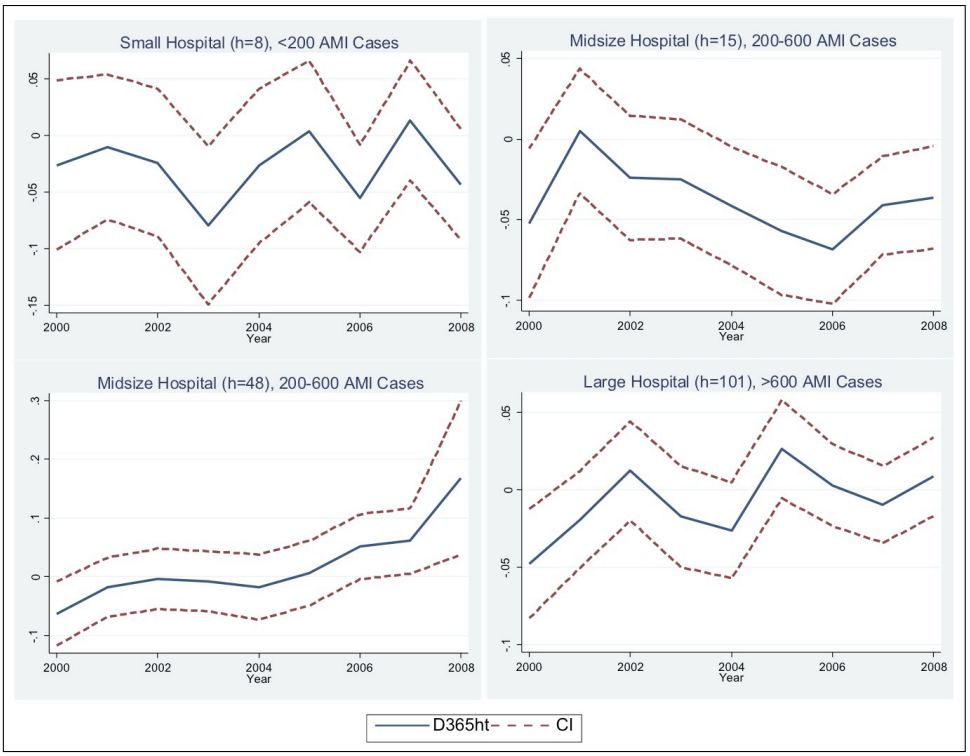


Figure 2.5: Trends across years in Latent AMI 28-day readmissions for selected hospitals.

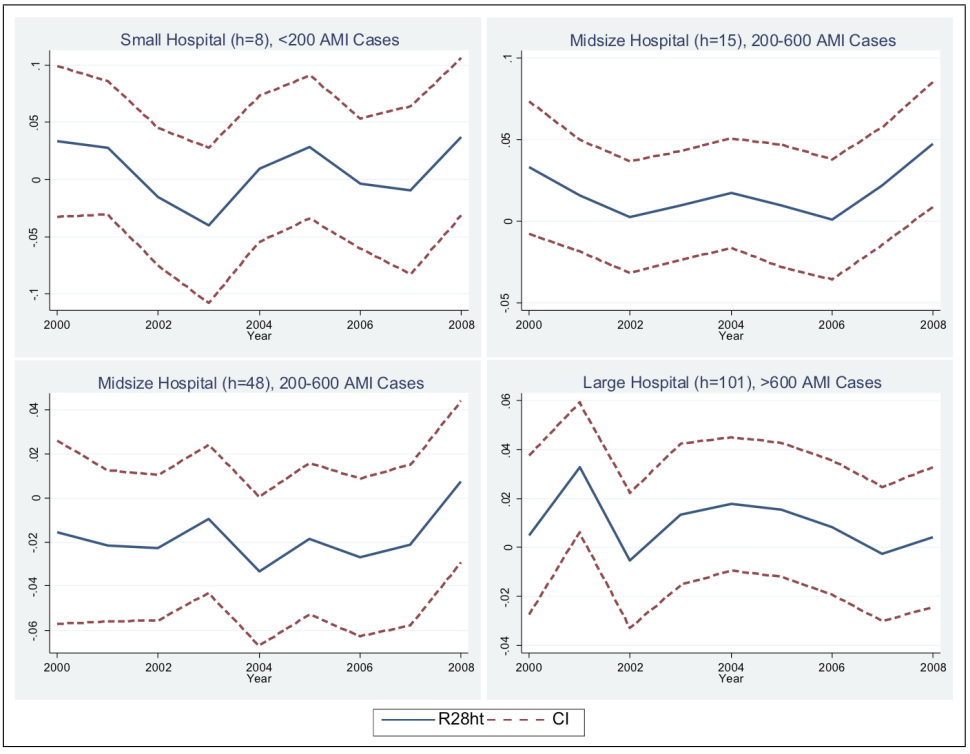
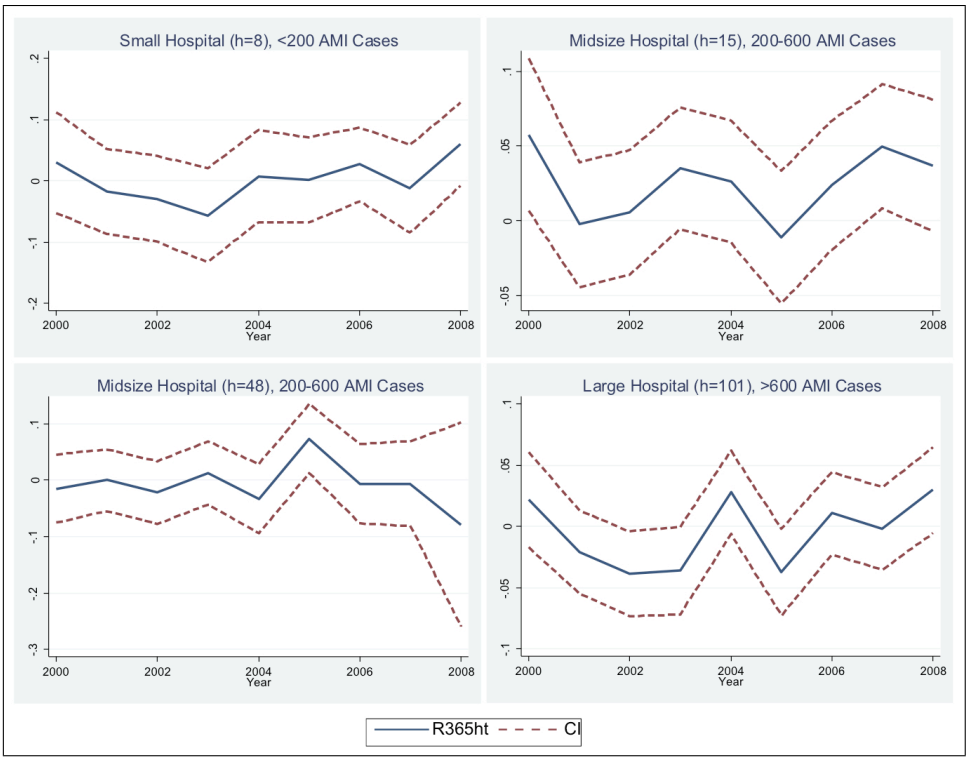


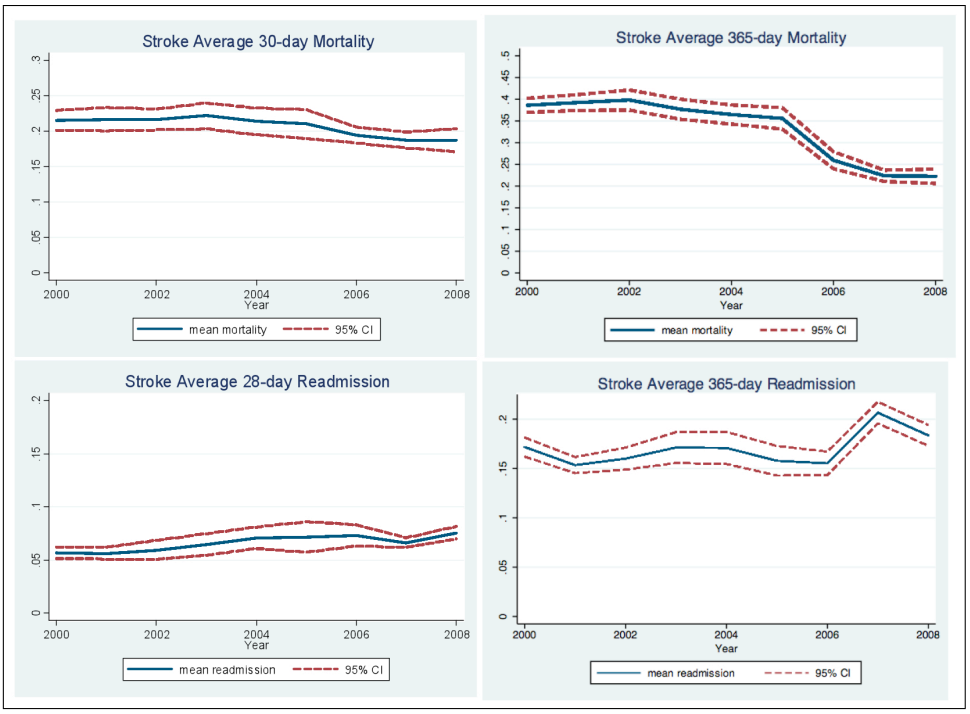
Figure 2.6: Trends across years in latent AMI 365-day readmissions for selected hospitals.



Stroke

The average 30-day mortality rates across hospitals, displayed in Figure 2.7, suggest that there has been a very modest decrease in Stroke mortality over the 2000-2008 period. The average 365-day mortality rates show a more distinct decline over the same period, that is especially pronounced over the years 2005-2006. The trends in average readmission rates over the same period indicate an increase in the average over time. Average 28-day readmissions undergo only a minor increase during the period, while average 365-day readmissions increase, especially over the 2006-2007 period. The confidence intervals for all four estimates is narrow in all figures.

Figure 2.7: Trends across years in average Stroke outcome measures across hospitals.



The results from the Model 1 mortality regressions, presented in Table 2.2 suggest that both short and long term Stroke mortality is significantly influenced by age, gender, co-morbidities and type of admission. Greater age, and increased co-morbidities are both associated with higher mortality, while women have slightly higher mortality than men, and elective admissions are significantly associated with lower mortality as compared to non-elective admissions. The results from the Model 1 readmission regressions, indicate that both short term and long term readmissions are significantly influenced by the same variables as mortality, as well as deprivation. Where higher deprivation is associated with higher readmissions, and the effect of all other significant variables is in the same direction as mortality. For both mortality and readmission, the trust dummies were highly significant.

Table 2.2: Regression results for Stroke Model 1.

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
	30-Day Mortality						
2000	52295	0.005***	0.028***	-6.12e-04	0.007***	0.098***	yes
		(0.000)	(0.004)	(0.000)	(0.002)	(0.011)	

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
2001	68864	0.005***	-0.031***	-1.46e-05	0.013***	0.086***	yes
		(0.000)	(0.003)	(0.000)	(0.002)	(0.010)	
2002	68243	0.005***	-0.030***	-2.06e-04	0.022***	0.095***	yes
		(0.000)	(0.003)	(0.000)	(0.002)	(0.012)	
2003	68385	0.005***	-0.033***	-4.38e-04	0.025***	0.080***	yes
		(0.000)	(0.003)	(0.000)	(0.002)	(0.011)	
2004	66396	0.005***	-0.034***	-2.19e-04	0.023***	0.084***	yes
		(0.000)	(0.003)	(0.000)	(0.002)	(0.012)	
2005	66825	0.005***	-0.039***	-9.15e-04	0.029***	0.094***	yes
		(0.000)	(0.003)	(0.000)	(0.002)	(0.011)	
2006	67705	0.005***	-0.039***	9.85e-04*	0.024***	0.094***	yes
		(0.000)	(0.002)	(0.000)	(0.002)	(0.011)	
2007	68403	0.005***	-0.035***	6.0e-04	0.023***	0.085***	yes
		(0.000)	(0.001)	(0.000)	(0.002)	(0.010)	
2008	69526	0.004***	-0.037***	2.05e-04	0.024***	0.078***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.010)	
	365-Day Mortality						
2000	52295	0.010***	-0.023***	-3.48e-05	0.033***	0.089***	yes
		(0.000)	(0.004)	(0.001)	(0.002)	(0.012)	
2001	68864	0.010***	-0.023***	0.001	0.033***	0.075***	yes
		(0.000)	(0.004)	(0.001)	(0.002)	(0.011)	
2002	68243	0.010***	-0.032***	0.001	0.047***	0.066***	yes
		(0.000)	(0.004)	(0.001)	(0.002)	(0.012)	
2003	68385	0.010***	-0.040***	0.001	0.052***	0.063***	yes
		(0.000)	(0.004)	(0.001)	(0.002)	(0.012)	
2004	66396	0.001***	-0.042***	0.001	0.053***	0.075***	yes
		(0.000)	(0.004)	(0.001)	(0.002)	(0.013)	
2005	66825	0.009***	-0.043***	0.001	0.055***	0.083***	yes
		(0.000)	(0.004)	(0.001))	(0.002)	(0.013)	
2006	67705	0.007***	-0.042***	0.001*	0.032***	0.102***	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.011)	
2007	68403	0.0059***	-0.039***	0.001*	0.028***	0.088***	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.011)	

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
2008	69526	0.006***	-0.039***	0.001	0.030***	0.084***	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.010)	
28-Day Readmission							
2000	52295	-3.58e-04***	0.006***	8.0e-04***	9.11e-04	0.010*	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.006)	
2001	68864	-2.15e-04***	0.007***	9.90e-04***	0.006***	0.013**	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.010)	
2002	68243	-2.53e-04***	0.007***	9.47e-04***	0.003**	0.006	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.012)	
2003	68385	-2.46e-04***	0.004**	9.64e-04***	0.004***	0.008	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.011)	
2004	66396	-0.001***	0.004*	0.001***	0.004***	0.019***	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.012)	
2005	66825	-3.17e-04***	0.008***	0.002***	0.003***	0.027***	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.011)	
2006	67705	-2.73e-04***	0.008***	0.001***	0.005***	0.016***	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.011)	
2007	68403	-1.37e-04***	0.008***	0.002***	0.005***	0.009	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.010)	
2008	69526	-2.81e-04***	0.008***	0.002***	0.006***	0.007	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.010)	
365-Day Readmission							
2000	52295	1.52e-04	0.014***	0.005***	0.007***	0.021**	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.010)	
2001	68864	1.07e-04	0.019***	0.004***	0.010***	0.027***	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.009)	
2002	68243	2.53e-04**	0.015***	0.004***	0.007***	0.022**	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.010)	
2003	68385	2.29e-04**	0.010***	0.003***	0.009***	0.027***	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.010)	
2004	66396	-3.67e-04***	0.012***	0.004***	0.009***	0.030***	yes
		(0.000)	(0.003)	(0.001)	(0.001)	(0.011)	
2005	66825	-1.71e-05	0.013***	0.004***	0.020***	0.036***	yes

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
		(0.000)	(0.003)	(0.001)	(0.001)	(0.010)	
2006	67705	1.03e-04	0.015***	0.004***	0.011***	0.016*	yes
		(0.000)	(0.003)	(0.001)	(0.001)	(0.010)	
2007	68403	5.62e-04***	0.013***	0.005***	0.009***	0.006	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.011)	
2008	69526	2.28e-04**	0.011***	0.004***	0.013***	0.023**	yes
		(0.000)	(0.003)	(0.001)	(0.001)	(0.010)	

* Significant at p $\leq$ 0.1

** Significant at p $\leq$ 0.05

*** Significant at p $\leq$ 0.01

Figure 2.8 shows the trend in latent outcome measures over the 2000-2008 time period. Both mortality panels show negative values on the hospital intercepts, indicating declining mortality. Over time the values become less negative, indicating that they are decreasing at an increasing rate. The confidence intervals for both latent short and long term mortality suggest a variation of about 2%. The hospital intercepts for both short and long term readmissions are positive, indicating that on average, controlling for patient factors, hospitals have positive readmissions. Throughout the period the slope is fluctuating for both readmission measures. The trajectory is such that readmissions initially increase, till about 2004, and then decreases back to its initial level. The confidence intervals suggest variation of about 0.5% for short term readmissions, and about 1% for long term readmissions.

Figure 2.8: Trends across years in average latent Stroke outcome measures across hospitals.

Figures 2.9 – 2.12 present the latent mortality and readmission estimates for four selected hospitals treating Stroke patients. The variation within hospitals and year-to-year for all four outcome estimates is small for the midsize and large hospitals, ranging around or under 5% below or above average. However the confidence intervals of the small hospital show vary large within-hospital variation of over 50% in either direction for short term mortality rates, and up to 20% in either direction for the other three latent outcome estimates. This could be attributed to the sample size of the hospital. Similar to the previous conditions the figures also indicate large year-to-year fluctuations of hospital specific latent mortality and readmission measures, ranging around 2-3% for the short term outcome estimates and around 5% for the long term outcome estimates.

Figure 2.9: Trends across years in latent Stroke 30-day mortality for selected hospitals.

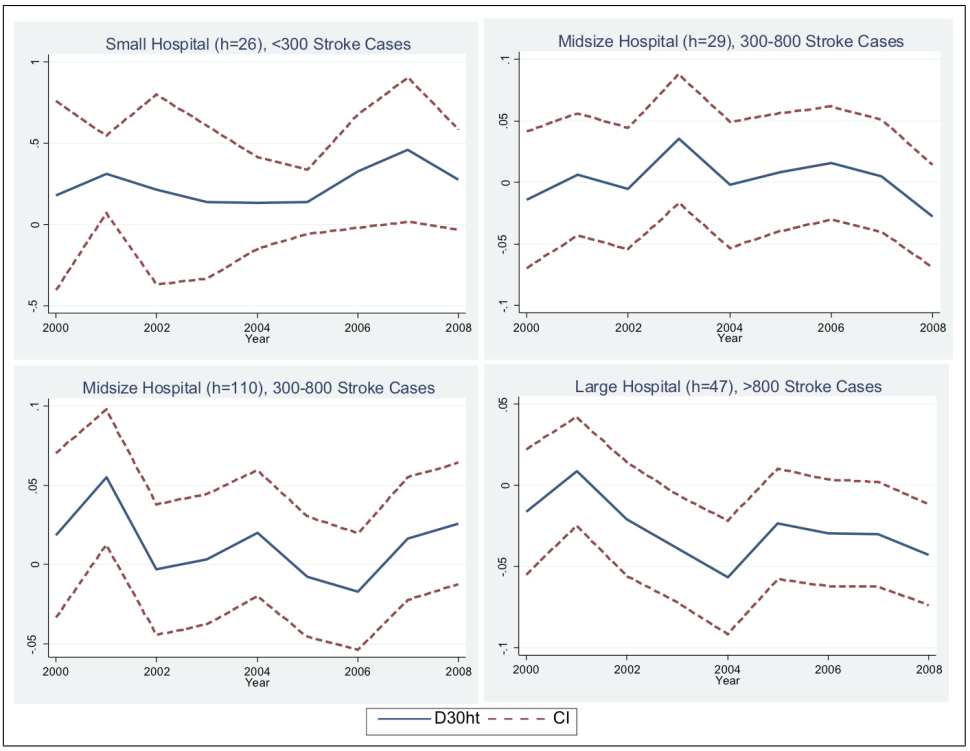


Figure 2.10: Trends across years in latent Stroke 365-day mortality for selected hospitals.

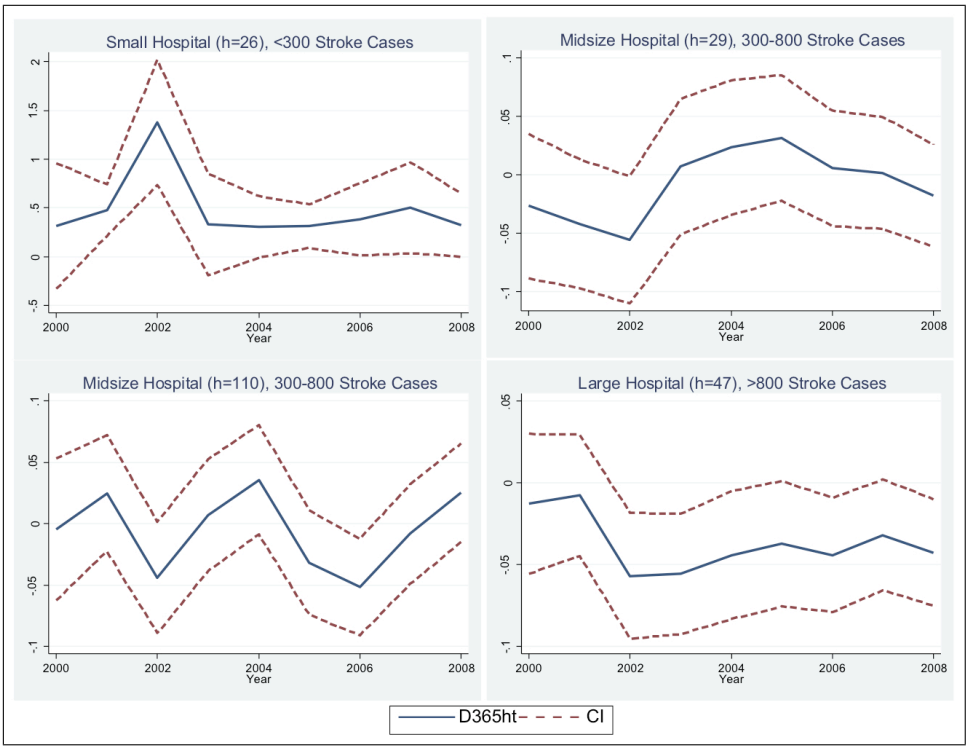


Figure 2.11: Trends across years in latent Stroke 28-day readmissions for selected hospitals.

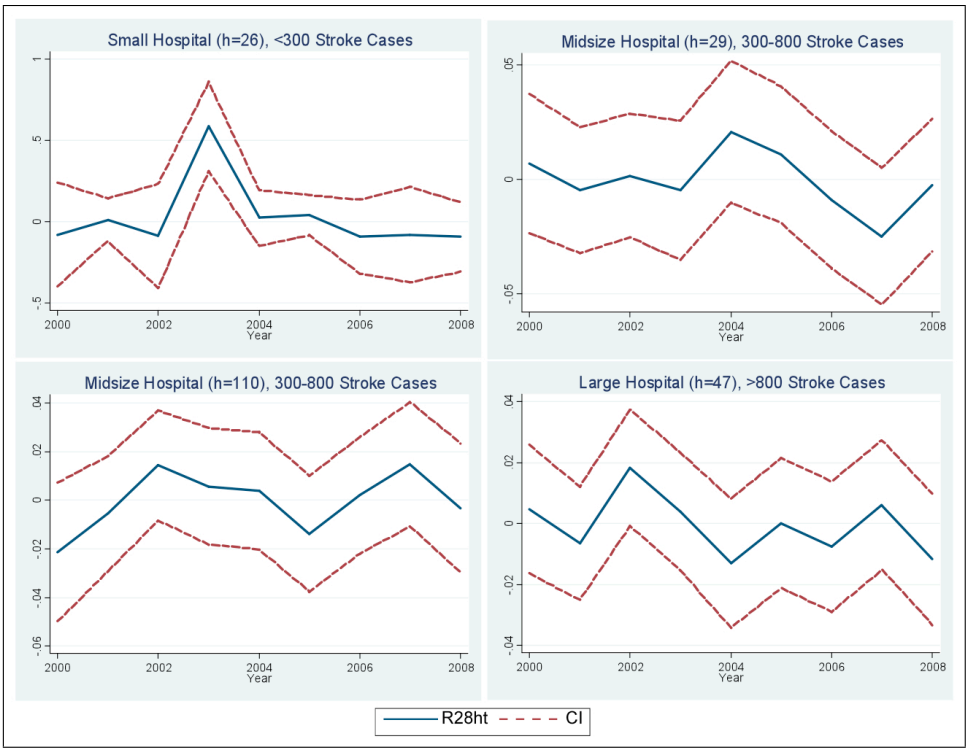
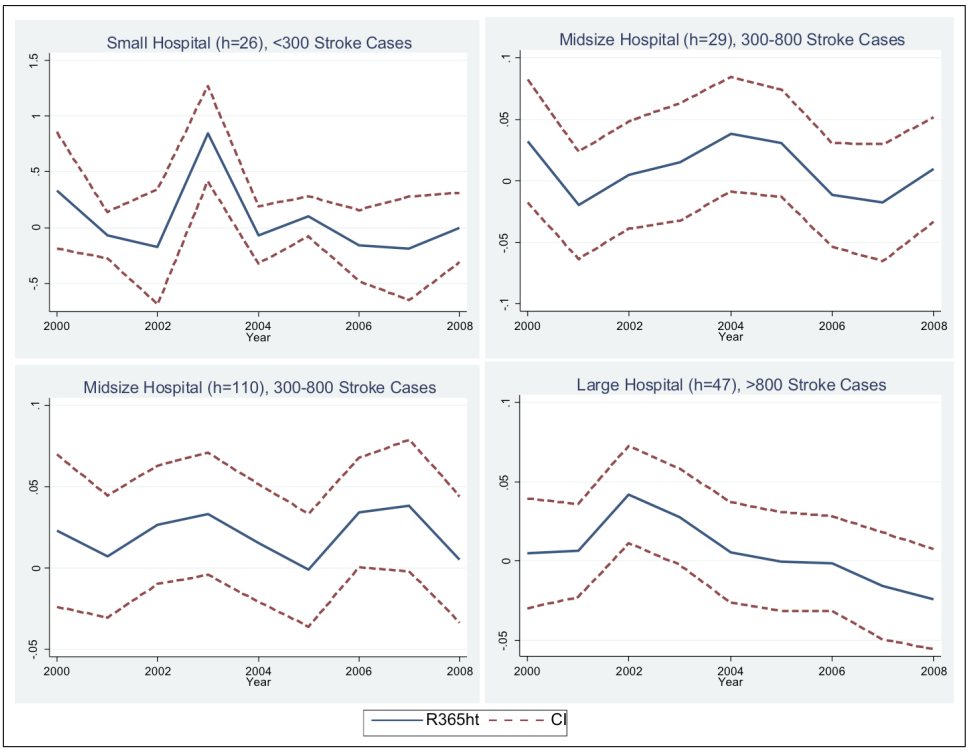


Figure 2.12: Trends across years in latent Stroke 365-day readmissions for selected hospitals.



Hip Replacement

Average 30-day mortality for Hip Replacement in the years 1996-2008, presented in Figure 2.13 exhibits a relatively constant trend during the entire time period. The confidence intervals surrounding the estimate narrow slightly from 2004 onwards, suggesting less variation in outcomes among hospitals from that point onwards. The average 365-day mortality rates for Hip Replacement shown in Figure 2.13 do not display a constant trend. Instead there is a noticeable increase from the year 1999 which is sustained until 2006, where mortality returns to is 2008 level. Similar to Figure 2.1, the confidence intervals surrounding the estimate narrow from 2005 onwards. Average 28-day and 365-day readmissions for Hip Replacement both show a slight upwards trend throughout the 2000-2008 period. In both figures there is a widening of confidence intervals in the year 2001, suggesting a larger variation in readmissions amongst hospitals for that year. For average 365-day readmissions only, there is sharp increase from 2006-2007.

Figure 2.13: Trends across years in average Hip Replacement outcome measures across hospitals.

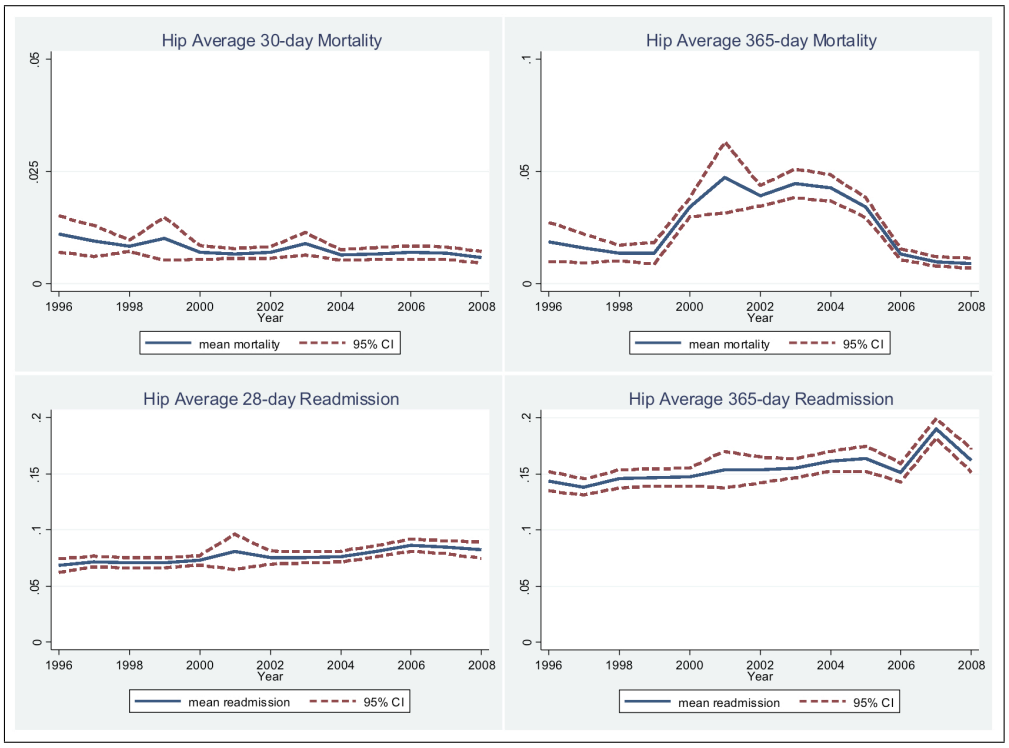


Table 2.3 presents the regression results from Model 1 for all four outcome indicators. In both mortality regressions age, gender, co-morbidity and type of admission is significant such that higher age and co-morbidity leads to increased mortality. In addition, women have a marginally higher mortality than men, and non-elective admissions have higher mortality than elective admissions. Deprivation is significant for only some of the years in

both regressions. Where significant, higher deprivation is associated with higher mortality. The trust dummies were always significant. All the explanatory variables included in the 28-day and 365-day readmissions models were significant, such that older patients, more deprived patients and patients with co-morbidities had higher rates of readmissions in addition to women and patients that where admitted for elective procedures.

Table 2.3: Regression results for hip Model 1.

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
30-Day Mortality							
1996	25835	4.29e-04***	0.004***	-2.12e-04	0.021***	0.015***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	
1997	29952	4.12e-04***	0.002**	-2.48e-04	0.019***	0.017***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	
1998	34559	5.34e-04***	0.004***	2.44e-04	0.020***	0.015***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	
1999	36527	4.41e-04***	0.002**	1.81e-04	0.020***	0.019***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	
2000	36864	4.55e-04***	0.004***	5.31e-05	0.017***	0.014***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	
2001	38745	4.47e-04***	0.002**	4.20e-04**	0.015***	0.016***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	
2002	41502	4.39e-04***	0.002**	4.73e-04***	0.017***	0.015***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	
2003	44759	4.68e-04***	0.001	9.38e-05	0.017***	-0.011***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	
2004	47124	3.87e-04***	0.003***	2.24e-04	0.012***	-0.012***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	
2005	46507	4.06e-04***	0.002***	1.52e-04	0.013***	-0.013***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	
2006	45438	3.76e-04***	0.003***	1.39e-04	0.013***	-0.013***	yes
		(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	
2007	47232	3.64e-04***	0.002***	1.26e-04	0.011***	-0.011***	yes
		(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	
2008	48243	3.35e-04***	0.002***	1.29e-04	0.008**	-0.008***	yes

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
		(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	
	365-Day Mortality						
1996	25835	0.001***	0.005***	-1.64e-04	0.033***	0.030***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	
1997	29952	0.001***	0.003***	-1.59e-04	0.032***	0.028***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	
1998	34559	0.001***	0.005***	3.10e-04	0.028***	0.026***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	
1999	36527	0.001***	0.002**	3.41e-04	0.026***	0.029***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	
2000	36864	0.002***	0.010***	3.00e-04	0.057***	0.064***	yes
		(0.000)	(0.002)	(0.000)	(0.002)	(0.003)	
2001	38745	0.002***	0.006***	0.002***	0.059***	0.089***	yes
		(0.000)	(0.002)	(0.000)	(0.002)	(0.003)	
2002	41502	0.002***	0.004**	0.002***	0.054***	0.079***	yes
		(0.000)	(0.002)	(0.000)	(0.002)	(0.003)	
2003	44759	0.002***	0.003*	0.001***	0.058***	0.082***	yes
		(0.000)	(0.002)	(0.000)	(0.002)	(0.003)	
2004	47124	0.002***	0.007***	0.002***	0.058***	0.077***	yes
		(0.000)	(0.002)	(0.000)	(0.002)	(0.003)	
2005	46507	0.002***	0.005***	0.001***	0.044***	0.070***	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.002)	
2006	45438	0.001***	0.004***	2.88e-04	0.019***	0.028***	yes
		(0.000)	(0.010)	(0.000)	(0.001)	(0.001)	
2007	47232	0.001***	0.003***	2.46e-04	0.014***	0.021***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	
2008	48243	0.001***	0.002***	3.76e-04**	0.010***	0.023***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	
	28-Day Readmission						
1996	25835	4.63e-04***	0.014***	0.001	0.008***	0.053***	yes
		(0.000)	(0.003)	(0.001)	(0.003)	(0.005)	
1997	29952	5.13e-04***	0.009***	0.002**	0.009***	0.046***	yes
		(0.000)	(0.003)	(0.001)	(0.003)	(0.004)	

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
1998	34559	4.36e-04***	0.018***	0.002**	0.008***	0.053***	yes
		(0.000)	(0.003)	(0.001)	(0.003)	(0.004)	
1999	36527	6.01e-04***	0.013***	1.23e-04	0.014***	0.055***	yes
		(0.000)	(0.003)	(0.001)	(0.003)	(0.004)	
2000	36864	5.99e-04***	0.012***	0.2e-06***	0.008***	0.052***	yes
		(0.000)	(0.003)	(0.001)	(0.003)	(0.004)	
2001	38745	9.41e-04***	0.0017**	0.002***	0.006***	0.059***	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.004)	
2002	41502	7.54e-04***	0.013***	0.002***	0.006***	0.070***	yes
		(0.000)	(0.002)	(0.001)	(0.002)	(0.001)	
2003	44759	8.63e-04***	0.014***	0.002***	0.011***	0.067***	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.001)	
2004	47124	9.01e-04***	0.016***	0.002***	0.006***	0.068***	yes
		(0.000)	(0.002)	(0.001)	(0.002)	(0.001)	
2005	46507	9.55e-04***	0.018***	0.002***	0.012***	0.066***	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.001)	
2006	45438	0.001***	0.019***	0.002***	0.011***	0.068***	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.001)	
2007	47232	0.001***	0.014***	0.003***	0.013***	0.075***	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.001)	
2008	48243	0.001***	0.014***	0.003***	0.011***	0.076***	yes
		(0.000)	(0.002)	(0.001)	(0.002)	(0.001)	
365-Day Readmission							
1996	25835	0.002***	0.013***	0.002**	0.026***	0.164***	yes
		(0.000)	(0.005)	(0.001)	(0.005)	(0.007)	
1997	29952	0.002***	0.017***	0.003***	0.021***	0.147***	yes
		(0.000)	(0.004)	(0.001)	(0.004)	(0.006)	
1998	34559	0.002***	0.025***	0.004***	0.023***	0.152***	yes
		(0.000)	(0.004)	(0.001)	(0.004)	(0.006)	
1999	36527	0.002***	0.016***	0.002***	0.023***	0.159***	yes
		(0.000)	(0.004)	(0.001)	(0.004)	(0.006)	
2000	36864	0.002***	0.011***	0.003***	0.022***	0.167***	yes
		(0.000)	(0.004)	(0.001)	(0.004)	(0.005)	

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
2001	38745	0.002***	0.0124***	0.004***	0.018***	0.165***	yes
		(0.000)	(0.004)	(0.001)	(0.003)	(0.005)	
2002	41502	0.002***	0.011***	0.003***	0.018***	0.185***	yes
		(0.000)	(0.004)	(0.001)	(0.003)	(0.005)	
2003	44759	0.002***	0.012***	0.004***	0.027***	0.184***	yes
		(0.000)	(0.003)	(0.001)	(0.003)	(0.005)	
2004	47124	0.002***	0.015***	0.005***	0.017***	0.178***	yes
		(0.000)	(0.003)	(0.001)	(0.003)	(0.005)	
2005	46507	0.002***	0.017***	0.004***	0.022***	0.164***	yes
		(0.000)	(0.003)	(0.001)	(0.003)	(0.005)	
2006	45438	0.002***	0.013***	0.003***	0.018***	0.165***	yes
		(0.000)	(0.003)	(0.001)	(0.003)	(0.005)	
2007	47232	0.003***	0.006*	0.006***	0.026***	0.186***	yes
		(0.000)	(0.003)	(0.001)	(0.003)	(0.005)	
2008	48243	0.002***	0.015***	0.005***	0.024***	0.174***	yes
		(0.000)	(0.003)	(0.001)	(0.002)	(0.005)	

* Significant at $p \leq 0.1$

** Significant at $p \leq 0.05$

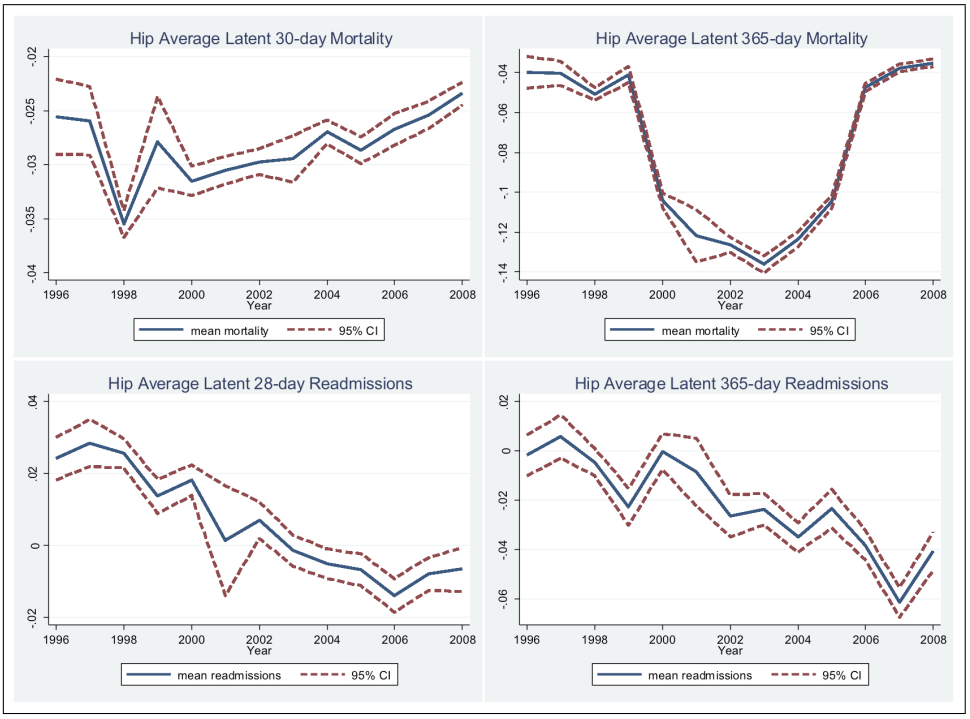
*** Significant at $p \leq 0.01$

Figure 2.14 shows the average Hip Replacement 30-day latent outcome estimates across hospitals in the time period 1996-2008. Both mortality estimates are negative throughout the period being investigated. This indicates that on average, the rate of change in mortality, controlling for patient characteristics is falling. For both short and long term mortality, the trajectory of the average hospital intercept indicates that initially the rate of change is decreasing at an increasing rate, but after 2000 it begins to decrease at an decreasing rate. This pattern is much more pronounced for year-long mortality than 30-day mortality. It is also apparent in both panels that the confidence intervals at the end of the period are narrower than in the first years of the sample.

The hospital intercepts for both readmission estimates vary around zero. The average latent outcome for 28-day readmissions indicates increasing average readmissions attributable to hospital performance, but at a declining rate. Around 2001, the average becomes negative, indicating that hospitals are contributing towards declining readmissions. The average latent outcomes estimated from 365-day readmissions are negative for

all years but 1997, indicating that throughout the period readmissions are falling. As the values are becoming increasingly more negative we can say that they are falling at a decreasing rate.

Figure 2.14: Trends across years in average latent Hip Replacement outcome measures across hospitals.



Figures 2.15 – 2.18 show the latent mortality and readmission estimates for four selected hospitals treating Hip Replacement patients. The within hospital and year-to-year variation in latent mortality is smaller than all other conditions aside from Hip Replacement, ranging around $1-5\%$ below and above both 30-day aggregate mortality and 28-day readmission measures, and around $5-10\%$ below and above the 365-day aggregate mortality and aggregate readmission measures. While there is year-to-year variation this is usually around $2-4\%$ in either direction. There is wider variation in the confidence interval and the year-to-year variation for the small hospital for all conditions.

Figure 2.15: Trends across hospitals in latent Hip 30-day mortality for selected hospitals.

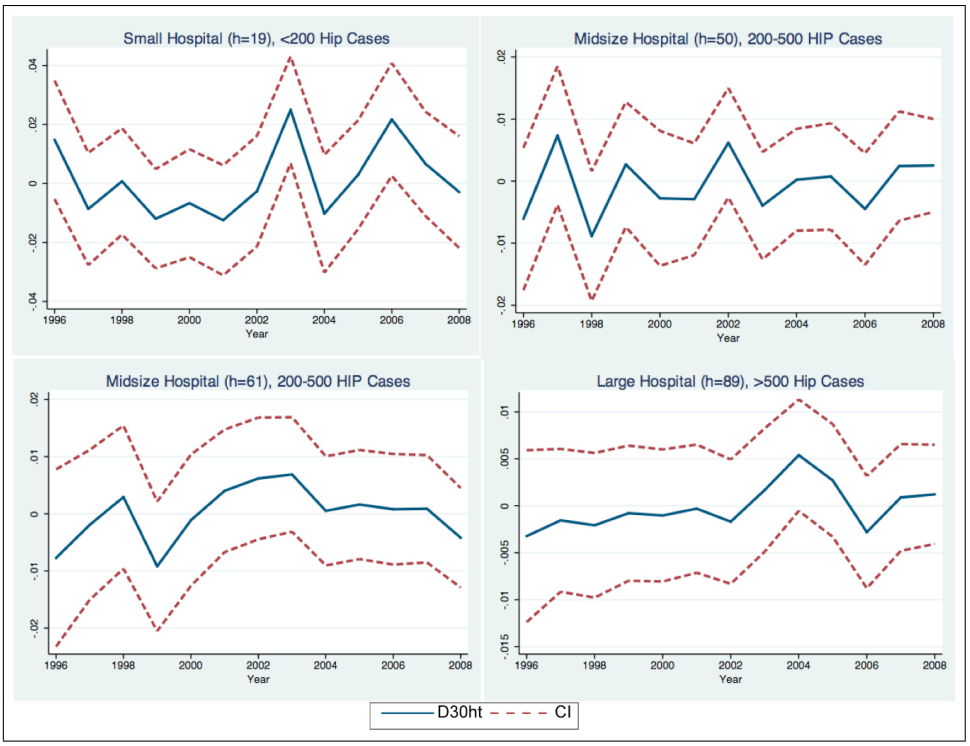


Figure 2.16: Trends across years in latent Hip 365-day mortality for selected hospitals.

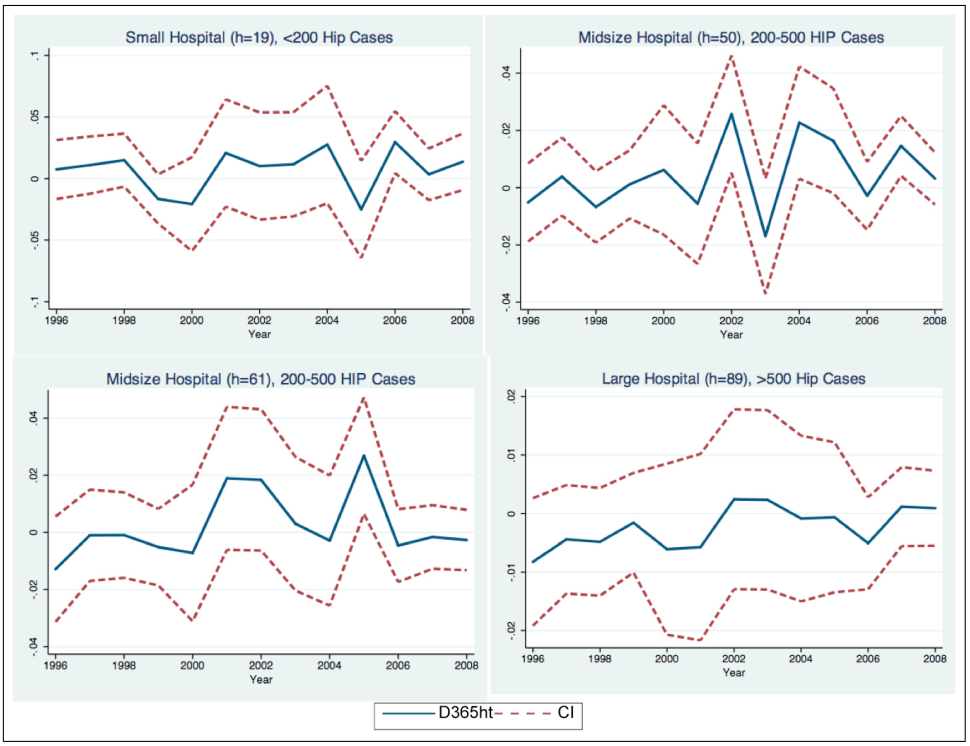


Figure 2.17: Trends across years in latent Hip 28-day readmissions for selected hospitals.

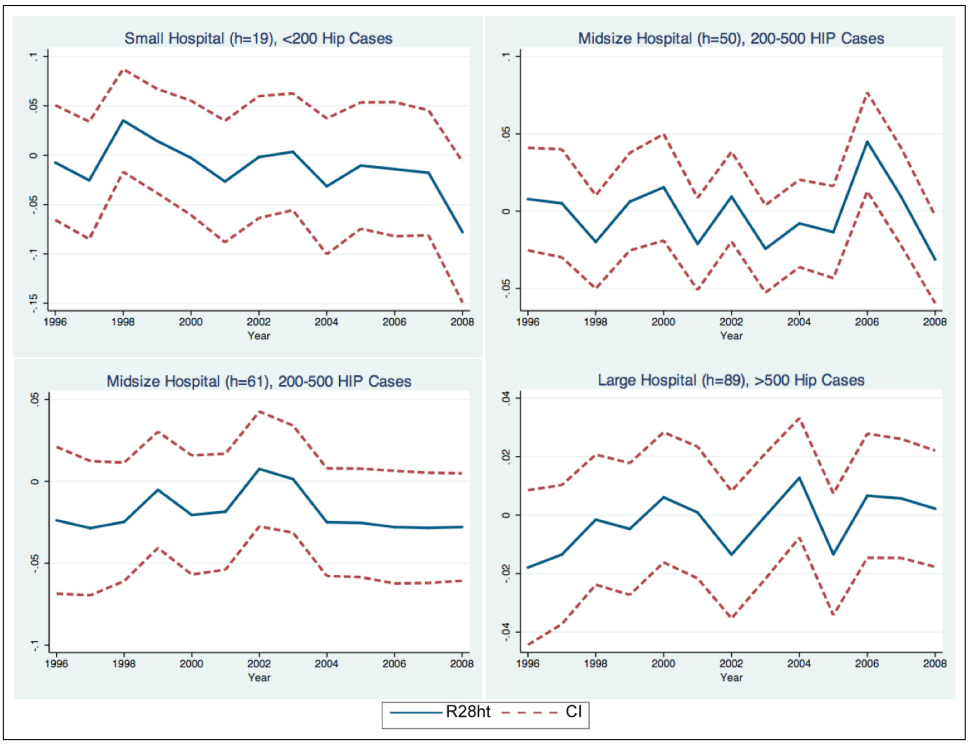
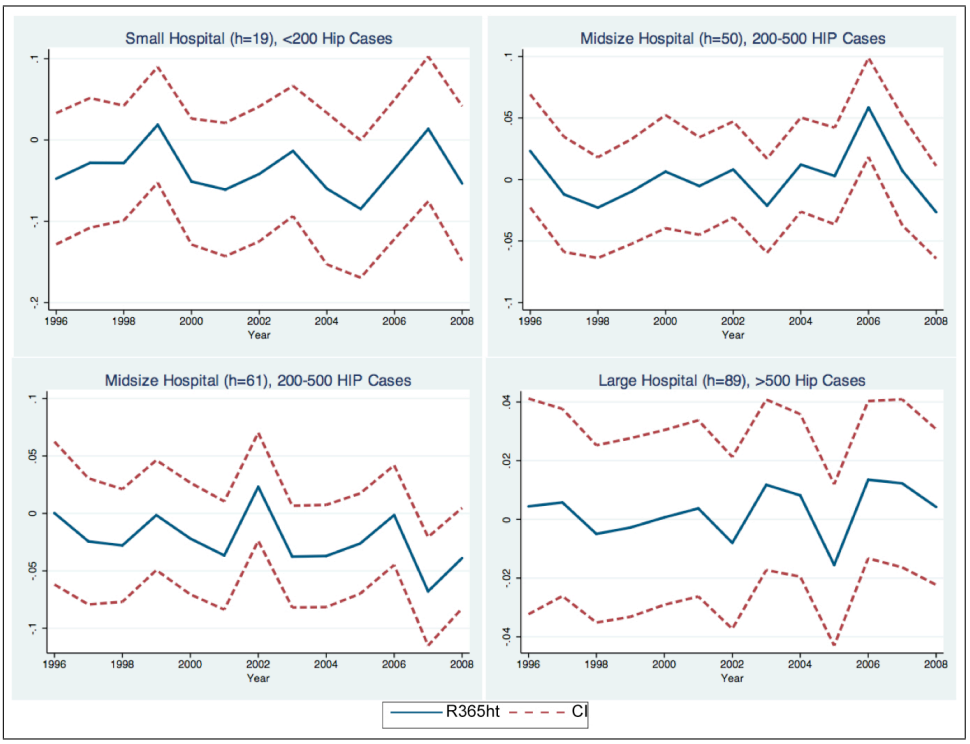


Figure 2.18: Trends across years in latent Hip 365-day readmissions for selected hospitals.



Model 2

The second model uses the aggregate outcome measures as dependent variables to test for fundamental relationships amongst outcomes and hospital characteristics. Tables 2.4 – 2.7 present the results for the regressions estimated for Model 2, one for each of the aggregate outcome measures, for each condition, run separately for each of the seven conditions. The number of instruments for each model is reported together with the regression results. As only a emergency AMI conditions were included in the sample, waiting times and lagged waiting times were not included in any of the AMI models. Additionally, different specifications of the models were run for the different conditions, and the one which best met model fit criteria is reported. For this reason lagged waiting times and length of stay are sometimes not included in selected models. Most models passed the AR(1) test for autocorrelation with over 95% confidence, aside from the readmission models for Stroke. All models also passed the Sargan test for instrument validity with over 95% confidence, rejecting the null hypothesis that the overidentifying assumptions are valid. The models indicate that few hospital characteristics are significant in influencing the change in latent outcomes over time. However, for most conditions some element of performance is dynamic – demonstrating that change does not occur instantaneously but is incremental.

Table 2.4: Model 2 regression results for latent 30-day mortality.

	AMI	MI	IHD	CCF	Stroke	TIA	Hip
L. Latent	0.961***	-0.0711	0.488**	-0.0265	0.342**	0.0166	-0.603***
Mortality	(0.209)	(0.176)	(0.247)	(0.121)	(0.153)	(0.0939)	(0.103)
L. LOS	0.00738***	0.00380	-1.92e-05	-0.000525	0.000656	-1.32e-06	-0.00147*
	(0.00113)	(0.00543)	(0.000538)	(0.00233)	(0.000768)	(0.000744)	(0.000815)
L. Cases	-0.000188	7.44e-05	3.95e-08	0.000558	-2.08e-05	-5.77e-05	6.09e-05**
	(0.000394)	(0.000195)	(1.19e-06)	(0.000447)	(5.98e-05)	(7.07e-05)	(3.01e-05)
L. Waiting	-	-	-	-	-	6.51e-07**	2.56e-05
Times	-	-	-	-	-	(2.85e-07)	(4.97e-05)
Cases	-0.000329	0.000404	-4.86e-07	-0.000966	3.94e-05	2.36e-05	-5.47e-05
	(0.000525)	(0.00117)	(3.32e-06)	(0.00120)	(0.000283)	(8.79e-05)	(3.85e-05)
$Cases^2$	5.92e-07	-9.94e-07	1.84e-10	8.63e-06	-9.43e-09	1.08e-07	-7.12e-09
	(6.63e-07)	(4.38e-06)	(9.40e-10)	(8.83e-06)	(1.96e-07)	(1.36e-07)	(2.42e-08)
LOS	-0.0111***	-0.00525	0.000815	0.00181	-	7.47e-05	0.00240**
	(0.00144)	(0.0102)	(0.00122)	(0.00307)	-	(0.00155)	(0.000940)
Waiting	-	6.53e-06***	-1.91e-06	6.80e-05	3.01e-06	4.81e-06	-2.04e-05
Times	-	(1.82e-06)	(2.44e-06)	(0.000148)	(1.24e-05)	(5.68e-06)	(5.11e-05)

	AMI	MI	IHD	CCF	Stroke	TIA	Hip
Specialist	0.0603	-0.0373	0.000995	-0.0603***	0.0384	0.00990***	0.00400
Trust	(0.0723)	(0.105)	(0.00433)	(0.0203)	(0.0443)	(0.00368)	(0.00523)
Foundation	-0.00120	0.00210	-0.000576	0.0167	0.00447	-0.000772	-0.000743
Trust	(0.00936)	(0.0154)	(0.000766)	(0.0163)	(0.00474)	(0.00227)	(0.00163)
University	-0.00149	0.0141	-0.000822	-0.0471***	-0.0133	0.000767	-0.00114
Hospital	(0.0171)	(0.0187)	(0.00257)	(0.0173)	(0.00914)	(0.00143)	(0.00205)
Constant	0.142	-0.331**	-0.0208	-0.237***	-0.179*	-0.0324***	-0.0484***
	(0.121)	(0.134)	(0.0135)	(0.0617)	(0.0930)	(0.00861)	(0.00971)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Instruments	22	28	33	33	30	34	47
N	986	341	919	352	830	509	1,047
Groups (hospitals)	132	105	129	101	125	104	121

* Significant at p $\leq$ 0.1

** Significant at p $\leq$ 0.05

*** Significant at p $\leq$ 0.01

Table 2.5: Model 2 regression results for latent 365-day mortality.

	AMI	MI	IHD	CCF	Stroke	TIA	Hip
L. Latent	-0.194	-0.130	0.686***	-0.140	0.376	0.0983	-0.450***
Mortality	(0.181)	(0.125)	(0.222)	(0.0963)	(0.428)	(0.0871)	(0.111)
L. LOS	0.00471***	0.00147	-0.00248	0.000786	0.00114**	-0.000830	-0.00421***
	(0.00100)	(0.00363)	(0.00214)	(0.00218)	(0.000542)	(0.00159)	(0.00149)
L. Cases	0.000250	6.86e-05	-1.14e-07	0.000123	1.11e-05	7.87e-05	0.000103
	(0.000185)	(0.000159)	(3.22e-06)	(0.000469)	(3.08e-05)	(5.48e-05)	(7.49e-05)
L. Waiting	-	-	-	-	-	-1.36e-07	0.000144
Times	-	-	-	-	-	(4.36e-07)	(9.69e-05)
Cases	-0.00110**	-8.26e-05	8.69e-06	-0.000267	0.000237	2.88e-05	-0.000138
	(0.000529)	(0.000761)	(6.09e-06)	(0.00115)	(0.000246)	(0.000205)	(8.79e-05)
$Cases^2$	7.83e-07*	8.61e-07	-1.59e-09	5.54e-06	-1.76e-07	-2.16e-07	3.34e-08
	(4.35e-07)	(3.07e-06)	(1.65e-09)	(8.90e-06)	(1.68e-07)	(4.80e-07)	(4.65e-08)
LOS	0.00506**	-0.00525	0.00370	0.00445	-	-0.00151	0.00780***
	(0.00231)	(0.0102)	(0.00237)	(0.00354)	-	(0.00290)	(0.00211)
Waiting	-	7.46e-06***	-6.37e-06*	-0.000129	3.08e-05	9.99e-06	-0.000146
Times	-	(1.48e-06)	(3.55e-06)	(0.000106)	(1.96e-05)	(2.42e-05)	(0.000104)
Specialist	-0.0487	-0.0504	0.00154	-0.0491	0.0650	0.0102	0.0105

	AMI	MI	IHD	CCF	Stroke	TIA	Hip
Trust	(0.0475)	(0.0411)	(0.00738)	(0.0373)	(0.0518)	(0.00900)	(0.00698)
Foundation	-0.00101	0.0129	0.000429	0.0142	0.00184	-0.00626	-0.00107
Trust	(0.00959)	(0.0144)	(0.00176)	(0.0149)	(0.00613)	(0.00540)	(0.00346)
University	-0.0332*	0.00480	-0.00301	-0.0354	-0.00872	0.000457	-0.00126
Hospital	(0.0190)	(0.0124)	(0.00374)	(0.0230)	(0.00748)	(0.00409)	(0.00423)
Constant	-0.269**	-0.347***	-0.0222*	-0.343***	-0.394*	-0.327***	-0.0809***
	(0.106)	(0.0532)	(0.0128)	(0.0651)	(0.236)	(0.0324)	(0.0185)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Instruments	22	32	29	33	24	35	47
N	986	341	919	352	830	509	1,047
Groups (hospitals)	132	105	129	101	125	104	121

* Significant at p $\leq$ 0.1

** Significant at p $\leq$ 0.05

*** Significant at p $\leq$ 0.01

Table 2.6: Model 2 regression results for latent 28-day readmissions.

	AMI	MI	IHD	CCF	Stroke	TIA	Hip
L. Latent	-0.0775	0.0596	0.0191	-0.515***	-0.164***	0.960	-0.0459
Mortality	(0.0589)	(0.138)	(0.410)	(0.183)	(0.0442)	(1.446)	(1.247)
L. Latent	0.210	-0.448***	0.323*	-0.255	0.0550	-0.390	0.0894
Readmissions	(0.313)	(0.139)	(0.178)	(0.197)	(0.105)	(0.406)	(0.205)
L. LOS	-0.000251	0.00336	-0.000531	-0.000907	0.000287	-0.00172	-0.00113
	(0.000389)	(0.00234)	(0.00123)	(0.00244)	(0.000244)	(0.00227)	(0.00131)
L. Cases	-1.52e-05	-3.38e-05	-1.88e-07	0.000263	-2.12e-06	-3.37e-05	3.66e-05
	(1.60e-05)	(0.000160)	(3.33e-06)	(0.000329)	(8.46e-06)	(0.000108)	(6.13e-05)
L. Waiting	-	-0.000138	-9.10e-06*	7.72e-05	4.59e-06	4.60e-07	-5.10e-05
Times	-	(0.000170)	(5.07e-06)	(7.95e-05)	(3.76e-06)	(1.40e-06)	(0.000133)
Cases	1.93e-05	9.12e-05	5.55e-07	1.27e-05	6.00e-06	0.000106*	5.87e-05
	(1.56e-05)	(0.000173)	(4.08e-06)	(0.000223)	(7.36e-06)	(6.24e-05)	(0.000102)
LOS	-0.000906	0.00229	0.000316	-0.00448*	-	-	0.00517**
	(0.00114)	(0.00309)	(0.00239)	(0.00235)	-	-	(0.00241)
Waiting	1.77e-05	0.000117	2.65e-07	-5.78e-05	2.65e-06	0.000109**	-3.11e-05
Times	(1.47e-05)	(0.000158)	(3.96e-06)	(0.000205)	(7.45e-06)	(5.58e-05)	(6.60e-05)
Specialist	-0.000906	0.00248	-0.000185	-0.00439*	-0.0140	-0.0118	0.00295
Trust	(0.00114)	(0.00345)	(0.00241)	(0.00246)	(0.0188)	(0.0246)	(0.00230)

	AMI	MI	IHD	CCF	Stroke	TIA	Hip
Foundation	0.00507	6.59e-06***	-9.40e-06	-3.55e-05	4.51e-06	-1.45e-05	6.80e-05
Trust	(0.00359)	(2.00e-06)	(6.44e-06)	(0.000182)	(3.75e-06)	(2.79e-05)	(0.000156)
University	-0.0264	-0.0400	-0.000396	0.0162	-0.0140	-0.0118	-0.00946
Hospital	(0.0207)	(0.0279)	(0.00433)	(0.0420)	(0.0188)	(0.0246)	(0.0108)
Constant	-0.0138	0.0466	3.58e-05	0.0945	0.0289*	0.0291	-0.0283
	(0.0215)	(0.0751)	(0.0386)	(0.0676)	(0.0157)	(0.0626)	(0.0442)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Instruments	30	36	29	31	34	26	42
N	986	182	879	200	744	509	1,047
Groups (hospitals)	132	71	125	65	122	104	121

* Significant at p $\leq$ 0.1

** Significant at p $\leq$ 0.05

*** Significant at p $\leq$ 0.01

Table 2.7: Model 2 regression results for latent 365-day readmissions.

	AMI	MI	IHD	CCF	Stroke	TIA	Hip
L. Latent	-0.448***	-0.0690	0.191	0.616	-0.227	0.0182	-0.821
Mortality	(0.149)	(0.198)	(0.944)	(0.476)	(0.267)	(0.634)	(0.527)
L. Latent	-0.220**	-0.312	1.047***	0.162	-0.161	-0.250	0.0497
Readmissions	(0.110)	(0.400)	(0.232)	(0.311)	(0.398)	(0.266)	(0.183)
L. LOS	0.00132**	0.00142	0.000196	0.00252	0.000724	0.000634	-0.00107
	(0.000560)	(0.00366)	(0.00245)	(0.00390)	(0.000860)	(0.00344)	(0.00235)
L. Cases	-5.91e-06	-0.000195	-1.31e-05**	5.72e-05	6.20e-05	-0.000306	3.14e-06
	(2.74e-05)	(0.000179)	(5.47e-06)	(0.000782)	(6.24e-05)	(0.000452)	(2.09e-05)
L. Waiting	-	4.70e-05	-5.40e-06	0.000286**	1.55e-06**	6.18e-06***	-4.33e-05
Times	-	(0.000114)	(1.98e-05)	(0.000114)	(7.26e-07)	(1.79e-06)	(0.000164)
Cases	3.56e-05	-1.39e-05	1.18e-05*	0.000361	-0.000263**	-0.000227	2.53e-05
	(2.40e-05)	(0.000452)	(6.76e-06)	(0.000480)	(0.000108)	(0.000491)	(4.06e-05)
LOS	-0.00206	-0.00604	-0.00166	-0.0141**	-	-	0.00740
	(0.00215)	(0.00614)	(0.00408)	(0.00608)	-	-	(0.00642)
Waiting	3.87e-05	-4.98e-05	1.30e-05**	0.000114	-5.58e-05	0.000319	1.20e-05
Times	(2.49e-05)	(0.000174)	(5.89e-06)	(0.000510)	(7.07e-05)	(0.000476)	(2.15e-05)
Specialist	-0.00135	0.000957	-0.000672	-0.0101*	1.48e-05	-4.58e-05	0.00484
Trust	(0.00177)	(0.00357)	(0.00472)	(0.00572)	(2.22e-05)	(4.77e-05)	(0.00346)

	AMI	MI	IHD	CCF	Stroke	TIA	Hip
Foundation	-0.000908	-3.94e-06	-6.40e-07	4.47e-05	-0.0308	-0.0273	3.45e-05
Trust	(0.00424)	(3.11e-05)	(1.07e-05)	(0.000383)	(0.0622)	(0.0229)	(0.000192)
University	-0.00519	-0.0500	-0.00752	0.0188	0.00454	-0.000680	-0.0169
Hospital	(0.0273)	(0.0597)	(0.0121)	(0.0704)	(0.00520)	(0.0125)	(0.0212)
Constant	-0.117**	0.0315	0.0231	0.476*	0.00320	-0.00228	-0.110***
Hospital	(0.0568)	(0.103)	(0.268)	(0.259)	(0.00778)	(0.00861)	(0.0350)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Instruments	25	36	29	29	36	27	41
N	986	182	879	200	744	509	1,047
Groups (hospitals)	132	71	125	65	122	104	121

* Significant at p $\leq$ 0.1

** Significant at p $\leq$ 0.05

*** Significant at p $\leq$ 0.01

AMI

Tables 2.4 – 2.7 show the results for the regressions run for Model 2, using the AMI latent outcome measures as dependent variables. The results suggest that an increase in lagged 30-day mortality is associated with higher 30-day mortality and lower 1-year readmissions. Table 2.4 indicates, that aside from lagged mortality, the only other variables found to be significantly related to 30-day mortality were length of stay and lagged length of stay. Where lagged length of stay has a significant negative association, such that higher LOS and length of stay a significant positive association. In the model for long term AMI mortality, presented in Table 2.5, length of stay and lagged length of stay were also significant, but both had a positive association, such that higher length of stay is associated with higher 1-year mortality estimates. Lagged caseload is negatively associated with year long mortality, such that more cases are associated with higher 1-year mortality. In addition, teaching hospitals were significantly related to lower 1-year mortality than acute care trusts, but with 90% significance. The results using short term readmissions as a dependent variable, presented in Table 2.6, show no significant associations. Table 2.7 indicates that year-long readmissions are influenced by lagged readmissions and lagged length of stay in addition to lagged mortality. The direction of the results suggests that higher lagged readmissions will lead to lower readmissions, and higher lagged length of stay will lead to higher 1-year readmissions.

MI

Tables 2.4 – 2.7 show the results for the MI regressions run for Model 2. The results suggest that only for 28-day readmissions is past performance a significant predictor of current performance, such that higher lagged readmissions result in lower 28-day readmissions. The only variable found to impact the latent mortality measures was waiting times, where higher waiting times were associated with higher short term and long term mortality. Table 2.6 also indicates that foundation trusts are associated with higher 28-day readmissions than acute care trusts. None of the variables for the model with 1-year readmissions as the dependent variable were significant.

IHD

The results from the Model 2 regressions, investigating what hospital characteristics influence the IHD latent outcome measures, are presented in Tables 2.4 – 2.7. Tables 2.4 and 2.5 indicate that both long term and short term mortality have a significant positive association with lagged mortality. No other variables are significant in the 30-day mortality model 2.4. The model for year-long mortality is negatively associated with waiting times at 90% significance. Tables 2.6 and 2.7 also show that both long term and short term readmissions have a significant positive association with lagged readmissions. The model for 28-day readmissions is positively associated with lagged waiting times at the 90% while the model for year-long readmissions is positively associated with waiting times, Year-long readmissions are and positively associated with caseload, such that more cases are linked to higher readmissions. However it is negatively associated with lagged caseload.

CCF

Table 2.4 presents the regression results of Model 2 for latent CCF outcomes. The 30-day mortality model indicates that there is no significant relationship between predicting lagged latent mortality and latent mortality. The only significant variables in the short term mortality model are the hospital type, where CCF latent mortality is significantly lower in teaching hospitals and in specialist trusts as compared to acute care hospitals. The results from Table 2.5 do not indicate similar results for long term latent mortality. In fact latent long term mortality is not associated with any of the explanatory variables included in the model. The regression results for latent 28-day readmissions, presented in Table 2.6, show a high negative association between lagged latent 30-day mortality and 28-day latent readmissions, while lagged latent 28-day readmissions have no significant effect on the dependent variable. Higher length of stay is associated at the 90% significance level with lower 28-day readmissions. Specialist trusts have significantly lower readmissions,

both 28-day and year long, than acute care trusts, although only at 90% significance. Table 2.7 indicates that the other explanatory variables associated with latent 365-day readmissions are lagged waiting times and length of stay. Neither past latent mortality nor past latent readmissions are significantly associated with year-long latent readmission rates. Higher length of stay is associated with lower year-long readmissions, and higher lagged waiting times are associated with higher year-long readmissions.

Stroke

The results in Table 2.4 indicate that lagged mortality is positively associated with 30-day latent mortality, such that higher rates of past mortality will lead to higher rates of current mortality. None of the other variables included in the short term mortality regression are found to be significant. Table 2.5, shows that past mortality is not significantly associated with year-long latent mortality at any level of significance. The only variable that is significantly related with long-term mortality is lagged length of stay, where an increase in length of stay is significantly associated with higher one-year mortality. Table 2.6 indicates that 28-day latent readmissions are significantly associated with lagged latent 30-day mortality, but not with lagged latent 28-day readmissions. Table 2.7 shows that year-long readmissions are not significantly associated with either of these past performance indicators. Year-long latent readmissions are negatively associated with caseload and positively associated with lagged waiting times. Higher lagged waiting times lead to higher year-long readmissions while fewer cases lead to lower year-long readmissions.

TIA

Tables 2.4 and 2.5 show the Model 2 regression results for the latent outcome models for TIA. Past mortality is not a significant determinant of 30-day mortality or 365-day mortality. Of the variables included in the regression, lagged waiting times are positively associated with higher 30-day mortality, and specialist trusts are associated with higher 30-day mortality relative to acute care trusts. None of the variables included in the regression for 365-day mortality were significant at any level. Tables 2.6 and 2.7 show the Model 2 results for the latent readmission regressions. Neither past mortality nor past readmissions are significantly associated with 28-day readmissions or 365-day readmissions. The only variable significant in the 28-day readmission regression is caseload, where higher caseload is associated with higher readmissions. The only significant variable in the 365-day readmission regression, presented in Table 2.7, is lagged waiting times. The positive association indicates that higher past waiting times are associated with higher year-long readmissions.

Hip Replacement

Tables 2.4 shows the Model 2 results for the latent 30-day mortality regressions. The results indicate that 30-day mortality has a negative and significant relationship with past mortality. Current and lagged length of stay are also significantly associated with 30-day mortality, such that higher lagged length of stay is associated with lower 30-day mortality, while higher length of stay is associated with higher 30-day mortality. In addition caseload is significant, such that it has a positive association with 30-day mortality. The results for year-long mortality indicate similar results. Lagged year-long mortality is significantly associated with year-long mortality, such that an increase in lagged mortality results in lower mortality. Length of stay and lagged length of stay are also significantly associated with year-long mortality, indicating a positive and negative association respectively. The regression results for readmissions, presented in Tables 2.6 and 2.7, indicate fewer significant explanatory variables. The only significant variable in the 28-day readmission model is length of stay where higher length of stay is associated with higher readmissions. There are no significant variables in when the year-long readmission outcome is used as the dependent variable.

2.5 Discussion

This chapter focuses on the need to develop more sensitive indicators to measure the quality of health care providers. Performance measurement is increasingly being used by stakeholders such as managers, politicians, regulators, researchers, service users and the general public to inform their decision making – whether this be individual choices or broader health policy decisions. While increasingly process and structure measures are being used in quality measurement, outcome measures remain lucrative in their ability to present a simple, all-encompassing measure of health care efforts. Hitherto, stakeholders have relied on raw or risk-adjusted outcome measures as indicators of organizational quality. Yet, these types of indicators are largely determined by exogenous factors such as patient characteristics, and often risk-adjustment techniques are highly sensitive to technical choices which can bias measures in different directions depending on the risk adjustment method selected.

The method applied in this chapter employs a systematic approach for evaluating hospital quality using outcome indicators. This approach assumes a latent hospital level variable that is unmeasured but includes all of the unobserved factors influencing hospital quality. It is captured by vectors of hospital level intercepts estimated from individual patient level equations measuring determinants of outcomes for seven different conditions

across hospitals. These intercepts, or 'latent outcome measures' will thus exclude known confounding patient level variables, but reflect factors such as unmeasured resources, environmental and organizational characteristics as well as random error and data imperfections. They can then be used to explain how much unobserved hospital effects contribute to changes in mortality and readmission in any given year. Closely observing the trend of the latent outcome measures over time allows for a better examination of the rate of change in hospital quality than the examination of raw outcome measures, that can be of practical use to individual providers or policy makers.

The average latent mortality estimates, plotted over time, show the rate of change in outcomes over time, controlling for patient characteristics. In most conditions, for most years, average mortality attributable to hospital quality is declining, where the rate of decline appears to slow from 2005 onwards. This is especially pronounced for the year-long mortality measures in all conditions, and for 30-day mortality intercepts in AMI, IHD, Stroke, TIA and Hip Replacement. The trends in readmissions vary more by condition. The average of the hospital intercepts suggest increasing readmissions for AMI, CCF and Stroke and decreasing readmissions for IHD, Hip Replacement and TIA. While this chapter has not used these measures to investigate the possible factors that could be responsible for these changes, it is likely that policy changes in this time associated with the introduction of Payment by Results and increased competition play some role, either in terms of the effects they have on quality through incentives or more likely on the effect they have on coding of mortality. Under this payment scheme, hospital revenue is very closely linked to coding. This may result in better coding, but also can lead to adverse behaviours such as miscoding or fraud. Discrepancies in coding practices have recently been reported in the literature, such as hospitals coding deaths as palliative care in order to reduce mortality rates (Hawkes, 2010b). Further in-depth investigation is necessary to draw any conclusive results. Given the concerns about mortality coding, future estimates using this methodology would benefit from considering palliative outcomes in addition to mortality and readmission rates at different levels, or controlling for palliative care by including it in the explanatory variables of the patient level regressions.

Another interesting observation for many of the average latent estimate plots is that the difference between the coefficients from one year to the next, which indicate the absolute rate of change in mortality. Thus an increase of 0.02 in the average mortality intercept from one year to the next means that the mortality attributable to hospital quality increased by 2% over that time period. This information can allow us to make important conclusions about changes over time, controlling for patient factors. Finally, the confidence intervals often become narrower for the mortality intercepts, and either wider or

narrower for the readmissions estimates. This suggests that the variation amongst relative hospital performance in any given year is also changing. Again this may be linked to policies occurring during this time period, and perhaps indicate a need to examine possible explanatory factors.

The latent outcome measures can also be examined at the individual hospital level. Investigating the outcome measures at this level of analysis provides more information on how hospitals are performing relative to each other. The results indicate that where there is sufficient sample size these estimates can be informative and quite precise in distinguishing quality trends over time, and relative to peers. However, in cases where there is a small number of patients treated annually, the estimates will be subject to wide year-to-year fluctuations in quality, and surrounded by wide confidence intervals. Both factors make it harder to draw conclusive results about the quality of provider, making it difficult to draw any conclusions with an adequate degree of certainty. This suggests that the best sample with which to conduct this type of analysis at the hospital level is for conditions where there are large number of patients, provided at many hospitals in the country.

The final part of the analysis of this chapter examined the latent outcome measures to determine how much they can be explained by known hospital characteristics. One of the main areas of interest of this section was to determine how dynamic hospital outcomes are. The regression results produced mixed findings. Each condition except TIA had at least one outcome measure that was influenced by past performance, but only for IHD were all four outcomes significantly influenced by past performance.

In general, there appear to be two ways in which past performance can influence current performance: through positive association or a negative association. The first way, a positive association, occurs when good/bad past performance is a predictor of continuing good/bad performance. This can be best explained through a notion of path dependency, where hospitals performing a certain way in one period are likely to continue to do so regardless of the characteristics of the patients admitted. The second way, as indicated by a negative association, occurs where past good/bad past performance is associated with bad/good current performance. This can be regarded as a process of change. Where bad performance precedes good performance, we can conceptualise this as a process of improvement, where poor outcomes in one period motivate internal change so that outcomes will improve in the next period. Where good performance precedes bad performance, we can consider this change an indication of decline.

In the 30-day mortality model AMI, IHD, Stroke and Hip Replacement were significantly influenced by past performance. All conditions but Hip Replacement had a positive

association suggesting path dependency. For Hip Replacement the association is negative, suggesting there has been change over time. The average trend in latent 30-day mortality indicated in Table 2.14 suggest that mortality is improving over the time period studied, thus the negative association is indicative of an improvement. This result possibly reflects a change in procedure which we find as a result of higher uptake of a newer technology in this area - this finding is discussed in detail in Chapter 6 where Hip Replacement is analyzed in more detail. The results from the year-long mortality regression model, and Figure 2.14 also indicate a dynamic effect indicating an improvement in performance. The only other condition to have a significant dynamic effect for this year-long mortality is IHD, where the sign suggests path dependency.

Unlike the mortality models, the short and long term readmission models were estimated using two lagged outcome measures to test for a dynamic effect; past readmissions and past mortality. The positive and negative associations for the lagged readmission variables can be interpreted in terms of path dependency and change as explained above. However, the sign on the past mortality variables indicates the association between mortality and readmissions. Common sense would suggest that high readmissions are an indication of poor quality, as are high mortality rates. However as noted in Chapter 1 this may not always be the case. High readmissions may also be indicative of good quality where severe patients have been saved, or even an indication of other factors such as patient lifestyle and behaviour after discharge.

Only in the AMI model for year-long readmissions were both lagged outcome measures significant for the same condition. In this case, the negative sign on the lagged readmission measure suggests change over time. The average latent trend for this AMI outcome, plotted in Figure 2.2 indicates increasing readmissions in the beginning of the sample, followed by decreasing readmissions in the final year. Thus the negative sign on the lagged measure is indicating the improvement in readmissions. The negative sign on the lagged mortality variable indicates that 365-day mortality and 365-day readmissions are negatively correlated, meaning that higher year-long readmissions may not be indicative of worse performance for this condition.

The results for the other condition's readmission models, as indicated in Tables 2.6 and 2.7, indicate a negative association between lagged 30-day mortality and 28-day readmissions for CCF and Stroke. No other conditions show a significant association between lagged mortality and readmission. As in the case of AMI, the results for CCF and Stroke are indicative of how to correctly interpret the readmission indicators for these conditions. In both conditions, the negative association between the two indicators suggests that high short term readmissions are not necessarily indicative of poor quality. These differing

associations by condition (and time period) are extremely important as they influence the interpretation of performance information. For this reason they are more closely investigated for all indicators in Chapter 3.

The results of the lagged readmissions on current readmissions also provide more information as to the dynamic associations in readmission indicators across the other conditions. For IHD there is a significant dynamic effect in both 28-day and year-long readmission rates, indicating a positive association across time. This can be explained through the path dependency framework outlined above. The model for MI shows a negative association between lagged 28-day readmissions and current readmissions, indicating change over time. Figure A.2 indicates variation in 28-day latent readmissions across the time period investigated, such that readmissions are increasing throughout most of the period, but with a period of decline in the middle of the sample. Thus, it is likely that the negative coefficient is capturing these changes in readmissions. In the remainder of the conditions investigated, lagged readmissions are not significant predictors of current readmissions in either the short or long term models.

The results from Model 2 do not suggest any conclusive findings on how hospital type influences performance. There are very few specialist trusts in the sample[1]. However, specialist trusts were associated with lower CCF 30-day mortality, 28-day and year-long readmissions but also with higher TIA 30-day mortality as compared to acute care trusts. While there are more foundation trusts and university hospitals in the sample, foundation trusts were only significant in having higher 30-day mortality for MI, and university hospitals for having lower 28-day readmission rates for CCF. There are mixed results for the directions on the significant variables for caseload, waiting times and length of stay. In most cases, where significant higher waiting times and lagged waiting times are negatively associated with all outcome measures. Although this is not true for IHD where it has a positive association. Lagged length of stay where significant was always negatively associated with all outcome measures. Long length of stay is likely to be an effect of worse quality or co-morbidity, both likely to contribute to worse outcomes. However, same period length of stay is associated with lower readmissions for CCF and 30-day mortality for AMI. Caseload also has mixed effect, this could be because good hospitals get more cases or because more too many cases may lead to lower quality.

Overall we find that it is difficult to attribute changes in performance to hospital characteristics and exogenous factors using only this analysis. However we are able to identify areas where past performance is an important predictor of future performance, as well as areas where the association between different outcome measures may be informative

[1]There were four specialist trusts in the sample and of those two are orthopaedic.

as to their correct interpretation. These findings leads us to extend the methodology to use this information to creating better quality measures which take into account the dynamic nature of performance and between indicators, and over time, as is done in Chapters 3 and 4. Moreover, our results suggest that we can distinguish between two different types of dynamic effect, path dependency and change, and identify the conditions where one occurs instead of the other. Indeed, in the areas where we see change and not path dependency there is room for further analysis to discover what factors are motivating this change, and whether it constitutes an improvement or decline in performance. In Chapter 6 we attempt to do this by investigating in detail the conditions of AMI and Hip Replacement so as to better understand the mechanisms of improvement suggested in these results.

We conclude that while the latent variable technique used in this chapter does not offer the solution to quality measurement it is an important addition to the tool box of methods with which to better understand the hospital contributions to quality of care. The latent outcome measures provide an interesting way to examine hospital performance over time and relative to one another, provided that they are used for conditions with large sample size. Furthermore, the latent outcomes generated through the latent variable approach provide useful indicators to use in further analysis of hospital performance. These indicators can be useful in correcting for methodological bias that arises from using statistical models that combine aggregate variables (such as hospital mortality rates) with individual observations to determine relationships. In conclusion, while these types of variables may only be only risk adjusted outcome measures, they provide a straightforward way to present, observe and analyse outcome data as well as to provide directions for further research.

3 Using a Vector Autoregression Framework to measure the quality of English NHS hospitals

3.1 Introduction

In interpreting health service performance, people are immediately drawn to measures of health outcome. Chapters 1 and 2 have already reviewed some of the main challenges in using these types of measures, ranging from the existence of suitable data to the methodologies used analyse it. The main motivation of Part II is to be able to create a robust measure of hospital quality using the raw outcome measures provided in administrative data. This measure can then be used, confidently, to evaluate the effectiveness of health policy.

The challenge we face is taking raw outcome data and using it to identify how much of this reflects hospital quality of care. As noted previously, outcome data will be influenced not only by quality of care but also by random error, systematic bias and patient case-mix. In order to extract the true quality signal from this data, we need to control for all other factors as much as possible. Chapter 2 used a latent variable technique to create measures of unobserved quality in English hospitals for the treatment of a range of conditions. These measures were analysed and compared with raw outcome data in order to determine their suitability for quality assessment purposes. While the latent measures did appear to be more sensitive to changes in hospital quality than the raw measures, they did also vary considerably across institutions and over time.

Theory tells us that the latent measures do address some of the methodological challenges we outlined earlier, such as the multidimensional nature of quality and the need to adjust for patient-case mix, yet also that they are limited in their ability to address others. While the latent measures are adjusted for the case-mix factors we were able to control for, namely: age, gender, co-morbidity, deprivation and type of admission, they will still contain random and systematic error in the estimates along with the unobserved quality

we are trying to measure. This noise contributes to the variation we see across hospitals and over time.

Moreover, outcomes are often the product of the inputs of previous years, and will not necessarily be a reflection of the performance of the current health system. Conversely, current inputs may contribute in part to future attainment. It is vital that when assessing the outcomes collected these are analysed in a way that takes into account these time lags, which are often more pronounced in the health care sector than in other areas of the economy (due to training of staff, or testing of treatments where inputs may take a number of years before they can be translated into outputs). In line with best practice discussed earlier, a performance measurement system should aim to be dynamic, reflecting their overarching objective of informing policy. Indeed the analysis from Chapter 2, indicates that hospital performance is dynamic. However, as the latent estimates are computed year-by-year they are not able to capture this dynamic element.

This chapter attempts to address some of these hurdles by applying technique published in 1999 and applied to US hospital data by McClellan and Staiger. We use this method to evaluate quality for English hospitals using English patient level data. Their method uses vector autoregressions (VARs) to capture dynamic interactions in the time series and across measures. This step allows information from the dynamic interactions of outcomes over time and across dimensions to be used to filter out more of the noise captured by the measures, and also use the time series and cross sectional information contained in the estimates to further adjust them. Moreover, the VAR methodology is commonly used for forecasting, and thus can be used to predict and forecast hospital quality extremely well. This chapter reviews the entire methodology and uses it to replicate the McClellan and Staiger (1999) quality measures for English hospitals.

3.2 Background

In health economics, and many other areas of applied economics, we face problems of endogeneity amongst dependent and independent variables. Endogeneity can occur in cases where there is a two-way influence between the independent and dependent variables. This influence can arise from autoregression with autocorrelated errors, omitted variable bias, simultaneity between variables as well as measurement and/or sample selection error. Different methodological techniques have been adopted to deal with this issue, such as instrumental variable (IV) methods, simultaneous equation models, non-linear techniques and GMM estimators, such as those used in Chapter 2. Yet this problem of endogeneity is not unfamiliar to economists who have come across the same problems when attempting to explain the relationships among money, interest rates, prices and output. In 1980,

Christopher Sims (1980) championed the VAR approach which took away many of the restrictions models impose and allowed the data to be modelled in an unrestricted reduced form, where all variables are treated as endogenous. Predictions of the VAR model performed well, and so the technique has become popular in economics despite critiques. The basic idea behind the model is to treat all variables symmetrically, such that variables which that we are not confident are exogenous are modelled as endogenous. This leads to an n-equation, n-variable linear model, where each variable is explained by its own lagged values, plus the current and past values of the other lagged variables. While VAR models are often used in macroeconomics to analyse the relationship between different policy tools, they have rarely been used in the area of health economics.

This chapter considers using a VAR methodology similar to the McClellan and Staiger (1999) method used to create better quality indicators that will control for these issues but also use them to inform their estimation. The simplest form of a VAR is a first-order VAR specification, VAR(1), where the longest lag length modelled is unity. Different specifications of the model however are also able to incorporate more lags. Indeed identifying the correct number of lags is important in order to specify the model correctly, and is likely to influence the results. There are various tests available that indicate how many lags are appropriate, including the Akaike information criterion (AIC) and the Schwartz criterion.

Stock and Watson (2001) also note that the VAR can come in three different varieties, each of which places different restrictions upon the data being modelled, these are: reduced form, recursive and structural. A structural VAR use theory to produce instrumental variables that can test contemporaneous links between variables (Stock and Watson, 2001). In practice structural VARs differ considerably from their reduced form and recursive counterparts, because of the restrictions placed upon the model. As we do not use this type of VAR we will not go over it in detail[1]. A reduced form VAR expresses each variable as a linear function of its own past values, the past values of all other variables being considered and a serially uncorrelated error term. In our evaluation of quality a VAR(1) model of this type would be represented by this simple system:

$$\begin{aligned}
D30_{ht} &= \alpha + \beta_1 D30_{h(t-1)} + \beta_2 D365_{h(t-1)} + \beta_3 R28_{h(t-1)} + \beta_4 R365_{h(t-1)} + \epsilon_{D30ht} \\
D365_{ht} &= \alpha + \beta_1 D365_{h(t-1)} + \beta_2 D30_{h(t-1)} + \beta R28_{h(t-1)} + \beta R365_{h(t-1)} + \epsilon_{D365ht} \\
R28_{ht} &= \alpha + \beta_1 R365_{h(t-1)} + \beta_2 D30_{h(t-1)} + \beta_3 D365_{h(t-1)} + \beta_4 R28_{h(t-1)} + \epsilon_{R28ht} \\
R365_{ht} &= \alpha + \beta_1 R365_{h(t-1)} + \beta_2 D30_{h(t-1)} + \beta_3 D365_{h(t-1)} + \beta_4 R28_{h(t-1)} + \epsilon_{R365ht} \,.
\end{aligned} \tag{3.1}$$

[1]For an in-depth discussion on structural VARs see Stock and Watson (2001); Enders (2004).

Each equation in this system is estimated by Ordinary Least Squares (OLS). The error terms represent the 'surprise' movements in the variables after the past variables have been taken into account. If the different variables are correlated with each other, than the error terms in the reduced form model will also be correlated across equations.

A recursive VAR constructs the error terms in each regression to be uncorrelated with one another by including some contemporaneous values of the variables in the regression. So our system from above, would be modified to look something like:

$$\begin{aligned}
D30_{ht} &= \alpha + \gamma_1 D365_{ht} + \gamma_2 R28_{ht} + \gamma_3 R365_{ht} + \beta_2 D365_{h(t-1)} \\
&\quad + \beta_3 R28_{h(t-1)} + \beta_4 R365_{h(t-1)} + \epsilon_{D30ht} \\
D365_{ht} &= \alpha + \gamma_1 D30_{ht} + \gamma_2 R28_{ht} + \gamma_3 R365_{ht} + \beta_1 D30_{h(t-1)} + \beta_2 D365_{h(t-1)} \\
&\quad + \beta_3 R28_{h(t-1)} + \beta_4 R365_{h(t-1)} + \epsilon_{D365h} \\
R28_{ht} &= \alpha + \gamma_1 D30_{ht} + \gamma_2 D365_{ht} + \gamma_3 R365_{ht} + \gamma_3 R365_{ht} + \beta_1 D30_{h(t-1)} + \beta_2 D365_{h(t-1)} + \\
&\quad \beta_3 R28_{h(t-1)} + \beta_4 R365_{h(t-1)} + \epsilon_{R28ht} \\
R365_{ht} &= \alpha + \gamma_1 D30_{ht} + \gamma_2 D365_{ht} + \gamma_3 R28_{ht} + \beta_1 D30_{h(t-1)} + \beta_2 D365_{h(t-1)} \\
&\quad + \beta_3 R28_{h(t-1)} + \beta_4 R365_{h(t-1)} + \epsilon_{R365ht} \,. \qquad (3.2)
\end{aligned}$$

Equations (3.2) are not reduced form equations, for example $D30_{ht}$ will have a contemporaneous effect on the other three quality variables, and they will have a contemporaneous effect on $D30_{ht}$. This system can be better represented in matrix algebra, allowing the VAR model to be represented in standard form (Enders, 2004). Again each regression can be estimated by OLS, however if the right hand variables are not identical, because some contemporaneous effects are dropped than estimation by OLS will no longer provide uncorrelated error terms. In this case a Seemingly Unrelated Regression (SUR) may prove to be more efficient (see Chapter 4 for more discussion on this).

As VARs involve current and lagged values of multivariate time series they are able to capture co-movements between variables that other models cannot. Thus, VAR models can be very useful for data description. Typically results from Granger-causality tests, impulse responses and forecast error variance decompositions are used to inform researchers about different relationships between the variables. While coefficients and R-squared values are

often not reported, as the other statistics are more informative. Granger causality tests examine whether lagged values of one variable help to predict another. Impulse responses are able to trace out the responses of current and future values of each variable to a one-unit increase in any one of the VAR errors. While forecast error decompositions indicate the percentage of the variance of the error made in forecasting a variable due to a specific shock in a given time horizon. In addition, VAR models have been shown to be very good at forecasting, especially when over-parametrization can be properly adjusted for (Stock and Watson, 2001). These statistics will be used later on, in Chapter 4, where we continue to use the VAR model for analysis.

The McClellan and Staiger (1999) methodology uses a reduced form VAR between the latent quality variables to understand the interactions between the variables which are thought to be co-determined. Indeed by closely studying the residuals and the coefficients they are able to better understand just how persistent quality is for various conditions. The relationship amongst different quality indicators and information about the variables which is important in their interpretation. Following this analysis, the authors use the output produced from the VAR model to create smoothed time-series estimates of each of the outcome variables that take into account the time-series and cross-sectional variations they have identified. The empirical steps to this process taken to replicated this process are reviewed in detail the following section before the results are presented and discussed.

3.3 Methodology

Hospital performance over the period 1996 to 2008 is evaluated by a two step process, as outlined by McClellan and Staiger (1999). The first step, undertaken in Chapter 2, derives latent outcome measures at the hospital level (h) by estimating patient level (i) regressions (in the form of equation (2.1)) replicated below. The patient level regressions include hospital fixed effects (β) and a set of patient characteristics, $\sum \phi X$, known to influence outcomes (age, gender, deprivation, co-morbidities, and elective or emergency treatment). The regressions are run separately for each year (t) and outcome measure (k), and the hospital intercepts, representing the mean value of outcomes of each hospital holding patient characteristics constant across all hospitals, are extracted and used to create a new dataset at the hospital level.

$$Y^k_{iht} = \beta q^k_{1h} + \sum \phi X_{jht} + u_{iht}\,.$$

As explained in detail in Chapter 2, the latent measures, β, describe the rate of change in outcomes as explained by risk-adjusted hospital quality. This chapter uses these latent

measures in a VAR framework to create new quality measures which describe, summarize and forecast hospital quality. The newly constructed dataset contains Q_h a $1 \times TK$ vector of the estimated latent hospital outcome for hospital h, adjusted for differences in patient characteristics, such that:

$$Q_h = q_h + \epsilon_h \,,$$

where q_h is a $1 \times TK$ vector of the true hospital effects for hospital h, and ϵ_h is the estimation error (which is mean zero and uncorrelated with q_h). The variance of ϵ_h is estimated from the patient level regressions (equation (2.1)) and is equal to the variance of the regression estimates Q_h, where Ω_{jh} represents the covariance matrix of the hospital effects estimates for hospital h in year t. Or simply:

$$E(\epsilon'_{ht}\epsilon_{ht}) = \Omega_{ht}$$
$$E(\epsilon'_{ht}\epsilon_{ht}) = 0, \text{ for } t \neq s \,.$$

Thus, the estimation problem McClellan and Staiger (1999) lay out is how to provide estimates of Q_h to predict q_h. They propose creating a linear combination of each hospital's observed measures in such a way that minimizes the mean squared error of the predictions, conceptualised as running the following hypothetical regression:

$$q^k_{ht} = Q_{ht}\beta^k_{ht} + \omega_{iht} \tag{3.3}$$

They note that equation (3.3) cannot be estimated directly, as q represents unobserved performance and the optimal β varies by hospital and year. Thus, the measurement challenge is to predict the true hospital effect, q, from its noisy estimate Q. The idea is to attenuate the coefficient of Q towards zero, such that a prediction of q can be derived that will reduce the noise without distorting the true effect. This is a similar idea to a smoothing techniques as outlined, for example, in Titterington et al. (1985).

While equation (3.3) can not be directly estimated, the parameters of the hypothetical regression can be estimated from the existing data. The minimum least squared predictor is given by:

$$W(q_h|Q_h) = Q_h\beta \,,$$

where

$$\beta = [E(Q'_hQ_h)]^{-1}E(Q'_hq_h) \,. \tag{3.4}$$

This best linear predictor can be calculated using the following estimates:

$$E(Q'_hQ_h) = E(q'_hq_h) + E(\epsilon'_h\epsilon_h) \tag{3.5}$$

$$E(Q_h'Q_h) = E(q_h'q_h)\,, \tag{3.6}$$

where $E(\epsilon_h'\epsilon_h)$ is estimated using the individual patient level estimates of the covariance matrix for the parameter estimates Q_h, which we call S_h. S_h varies among hospitals. $E(q_h'q_h)$ can be estimated by $E(Q_h'Q_h - S_h) = E(Q_h'q_h)$. Plugging these estimates into equation (3.4) allows the calculation of the desired least squares estimates, such that:

$$\hat{q}_{ht} = Q[E(Q_h'Q_h)]^{-1}E(Q_h'q_h) = Q_h[E(q_h'q_h) + E(\epsilon_h'\epsilon_h)]^{-1}E(q_h'q_h)\,. \tag{3.7}$$

Using estimates (3.5)and (3.6), the R-squared statistic can also be calculated, based on the least squared formula.

Estimation of equation (3.7) provides the basis for the second step of the methodology, undertaken in this chapter. McClellan and Staiger (1999) coin these estimates 'filtered estimates' as they optimally filter out the estimation error of the observed quality measures. They note three attractive properties of the filtered estimates. First, that allows information for many years and different indicators to be combined in a systematic manner. Second, by nature of their construction, these estimates are optimal linear predictors for mean squared error. Finally, the estimates are simple to construct using standard statistical software.

Given the time-series nature of the data, information of the performance in each hospital effect over time is used to better predict and further forecast the outcome measures. Using a VAR model, further structure is imposed on the filtered estimates, by assuming that each performance measure in a given its past performance, plus a contemporaneous shock that can be correlated across the different outcome measures. Thus a first order VAR model for $q_{ht}(1 \times K)$ is estimated, where:

$$q_{ht} = q_{h,t-1}\Phi + v_{ht}\,. \tag{3.8}$$

$Z = V(v_{ht})$ the $(K \times K)$ variance matrix of the residuals, and $\Gamma = V(q_{h(t=1)})$ the $(K \times K)$ initial variance matrix from the first year of the data sample are also estimated. Φ represents a $(K \times K)$ matrix containing the estimates of the lag coefficients. The VAR structure implies:

$$E(Q_h'Q_h) - S_h = E(q_h'q_h) = f(\Phi, Z, \Gamma)\,. \tag{3.9}$$

Using the parameters estimated from the VAR model we are able to estimate equation (3.9), using the Broyden algorithm in eViews to estimate non-stochastic predictions, or the 'filtered outcome measures'.

The above analysis is estimated using a large pooled cross section that spans over many individuals and providers. The first part of the analysis, reviewed in Chapter 2, is performed using the statistical package STATA, the remainder of the analysis is undertaken in eViews, which includes more options to perform time-series analyses, and especially the VAR model. The size and amount of information on each patient and provider allows us to avoid many of the technical and methodological challenges presented in time series analysis.

3.4 Data

The data used to calculate this model is the same data as used in Chapter 2, presented in the data section of Chapter 1. This chapter builds on the methods used in Chapter 2 which used individual patient mortality rates and readmission rates at different intervals to contract latent outcome measures at the hospital level. These latent measures are collected into a new data set at the hospital level, distinguished by hospital identifiers and variables indicating the year of the measure. The creation of the latent variables and the hospital level data set are described in detail in the data section of Chapter 2. In order to conduct the analysis described above all hospitals with missing years of data are dropped from the sample. The sample size described in terms of number of hospitals and average number of cases per hospital across all years are presented in Table 3.1.

Table 3.1: Summary statistics of the sample of hospitals included.

Condition	ICD-10/ OPCS 4.3 codes	Years Analysed	Number of Hospitals	Average Cases per Hospital per year
AMI	ICD-10: I21	2000-2008	119	331
MI	ICD-10: I22, I23	2000-2008	113	74
IHD	ICD-10: I20, I25	2000-2008	121	1295
CCF	ICD-10: I11.0, I13.0, I25.5, I50.0, I50.1, I50.9, J81X	2000-2008	120	31
Stroke	ICD-10: I60-I67	2000-2008	121	522
TIA	ICD-10: G45.0-G45.4, G45.8-G45.9, G46.0-G46.8	2000-2008	120	116
Hip	OPCS4.3: W37-W39 W46-W48 W58	1996-2008	120	332

3.5 Results

The methodology of this chapter uses VAR models to describe and summarise hospital quality. By quantifying what is known about the different dimensions of measured quality and the time trend associated with the different latent outcome measures. The results of this chapter attempt to illustrate how well the filtered estimates perform at predicting in sample hospital quality and forecasting out of sample hospital quality. This is done by comparing the filtered measures to the latent measures diagrammatically as to visualize how the methodology reduces the noise in the estimates, by measuring the signal to noise ratio of the filtered estimates, and by estimating the goodness of fit measures of the estimates. Each of these steps is explained in more detail below. This section shows that in all of these areas the filtered estimates appear to be very good predictors and forecasts of true hospital quality.

Of the seven conditions for which this analysis was conducted the results of AMI, Stroke and Hip Replacement are presented in this section, by condition, while the results for MI, IHD, CCF and TIA are presented in Appendix B. This is because of the relatively large set of results which, if presented in totality, might obscure the main objective of this chapter which was to present general operation of the methodology. Suffice to say that with all conditions the general performance is similar. For each reported condition, the first table reports the VAR parameters of interest: the lag coefficients, the variance and correlation for the residuals to each effect, and the initial variance and correlation of the effects in the first year of the sample. These are discussed separately for each condition. All VAR models were tested for stability and passed unit root tests with all roots lying inside the unit circle.

Initially the VAR parameters are estimated using the information on all five aggregated outcome measures (i.e. the three mortality and the re-admission rates for all years in the sample, separately for each condition). The VAR(1) specification is as given in equation (3.8), and other specification of the model were tested with different lag lengths, the inclusion of additional lags yielded similar scores, sometimes marginally better, using the Akaike information criterion and the Schwartz criterion. Given the small difference in scores we chose to use the VAR(1) specification for all models as it fits the data relatively well and makes the analysis more parsimonious and the models easier to interpret.

The signal variance, which measures the underlying quality signal of each outcome measure is one of the parameters which the VAR model is able to extract from the original hospital data. These estimates can be used together with the estimates of the estimation error in each measure, defined as S_h in equation (3.9) above, to estimate the signal-to-noise ratio for each of the outcome measures, as specified in equation (3.10). For each condition

a figure is therefore included which plots the estimates of the ratio of signal variance to total (signal plus noise) variance in the observed hospital outcome measures against the number of cases treated in each hospital (the cases upon which this measure is based in the first step of the analysis).

$$\text{Signal}/(\text{Signal} + \text{Noise}) = V_{ht}/(V_{ht} + S_{ht)} \tag{3.10}$$

This plot provides statistical information on the level of "true" signal in each of the quality measures relative to underlying noise and indicates which performance measures have large associated variances across the specific observed outcomes and across the relevant sample.

The methodology further uses the VAR framework to further refine the latent outcome measures estimated in Chapter 2 by creating new 'filtered' measures of quality which contain more information as, by using the underlying time-series structure of the latent variables, they filter out more noise. The figures reported in each section report the latent outcome measures used in the analysis together with the predicted (in sample) filtered and forecasted (out of sample) filtered quality indicators for each condition. The predicted filtered estimates are constructed for the entire time period using the latent measures from the entire time period, while the forecasted indicators are constructed for the entire time period using the latent measures only up to 2006. Thus, the last two filtered measures are forecasted using existing data, but can be assessed as compared to the existing measures for those years.

Each figure plots the latent and predicted filtered estimates constructed from the data in four panels for four separate hospitals: small hospital (upper left), a large hospital (lower right), and two midsize hospitals. These hospitals are not a random sample, but chosen to illustrate the results in different settings, and are the same hospitals represented in the corresponding figures in Chapter 2. Each panel plots data for a single hospital from 2000 through to 2008, apart from the figures for Hip Replacement which plot the data on the larger sample available for that condition, from 1996 through to 2008. The figures plot two lines, a solid line indicating the aggregated outcome measures, estimated from a linear model run separately by year controlling for patient characteristics (see the data section above), and a long dashed line, indicating filtered outcome measures, estimated by a multivariate VAR framework including all the outcome-based measures. The solid lines can be interpreted as absolute outcome differences, or risk-adjusted mortality rates. A value of 0.02 indicates that the hospital's mortality was 2% above the average hospital in that year, with negative values indicating lower mortality than average, controlling for patient characteristics. The dashed lines are based on a multivariate VAR model, thus incorporating all of each hospital's data from 2000-2006 (1996-2006 for Hip Replacement), and using this data to forecast the values for 2007-2008. The two short dashed lines

indicate the 95% confidence intervals of the parameter estimates (long-dashed line). These figures are discussed below, separately for each condition.

In order to assess the ability of the filtered estimates to predict variation in true hospital effects, McClellan and Staiger (1999) construct an R-squared measure that can be applied to this setting, using the standard R-squared formula:

$$R^2 = 1 - \frac{\sum_{h=1}^{N} \hat{u}_h^2}{\sum_{h=1}^{N} q_h^2} \,. \tag{3.11}$$

As the purpose of this goodness of fit measure is to estimate how well the filtered estimates minimize the mean square error of the prediction, the numerator should measure prediction error, such that:

$$\hat{u} = q - \hat{q} \,.$$

Since q is not observed, estimates must be used for both the numerator and the denominator. McClellan and Staiger (1999) propose using the estimate of $E(q_h' q_h)$ for the denominator and $E(q_h - \hat{q}_h)'(q_h - \hat{q}_h)$ for the numerator. Both of these can be estimated using estimates 3.5 and 3.6 above.

These R-squared measures are calculated for the predicted values, and presented separately for each condition. Each table reports the results for predictions using different amounts of data, similar to the McClellan and Staiger (1999) analysis. The first column reports the R-squared for predictions using all years of data for both outcomes, the second column uses data from all years but only from the outcome being considered. The following columns calculate the R-squared for predictions based on 3 years of data, and 1 year of data, for both outcomes and one outcome respectively.

A similar goodness of fit measure is constructed in order to measure the accuracy of the VAR model in forecasting outcomes. In order to compare the forecast to the actual measurement, the model was estimated using data from 2000-2006 (1996-2006 for Hip Replacement) and used to forecast outcomes for 1 and 2 years ahead (2007-2008). The R-square measure for the forecasts, was thus used to measure the fraction of the true hospital variation found in the aggregate measures that was successfully explained in the forecasts:

$$R^2 = 1 - \frac{\sum_{h=1}^{N} (\hat{u}_h^2 - S_h)}{\sum_{h=1}^{N} (Q_h^2 - S_h)} \,. \tag{3.12}$$

In this measure the forecast error is estimated as:

$$\hat{u} = Q - \hat{q}$$

and S_h measures the variance of the OLS estimate Q_h. Thus the R-squared for the forecasts estimates the amount of variance in the true hospital effects that has been forecasted. This

R-squared measure can be negative if the forecasts lie out of sample. The expected R-squared values are calculated for the forecasted values using the measure estimated for the predicted values (equation (3.11)), the actual R-squared measures, based on actual estimates (equation (3.12)) are also calculated. These R-squared measures for predictions and forecasts are presented below, separately for each condition.

The final part of the results section (3.5), ranks the hospitals in the same using three different performance measures (raw, latent and filtered measures).This allows for a better understanding of the differences between the indicators and can be useful in drawing conclusions as to their applicability to policy.

AMI

The parameter estimates of basic model coefficients in Table 3.2 indicate the effect past values of each outcome measure have on their own performance. The model suggests that one-year hospital mortality, $D365ht$, is the most persistent of all four outcome indicators, with a value of the coefficient on its own lag of approximately 0.8. $R28_{ht}$ exhibits a weak dynamic effect, with a coefficient of around 0.4, while $D30_{ht}$ and $R365_{ht}$ both show an almost negligible dynamic effect. The standard deviation of the residuals indicate about 6% variation in short term mortality rates, and long term readmission rates across hospitals, while short term readmission rates vary by nearly 4% across hospitals. Long term mortality rates however are subject to much wider variation at about 17% across hospitals. The standard deviations from the year 2000 suggest that both readmission measures and year-long mortality have an annual variation around $3-4\%$, however 30-day mortality rates fluctuate more, varying around 10% annually. The correlation between variables in the year 2000, indicates a negative association between the outcome measures 30-day mortality, $D30_{ht}$ and short term re-admissions, $R28_{ht}$. The correlation of residuals indicates a similar negative association between $D365_{ht}$ and $R365_{ht}$, and a positive association between $R28_{ht}$ and $R365_{ht}$.

Table 3.2: Estimates of AMI multivariate VAR(1) parameters for hospital specific effects.

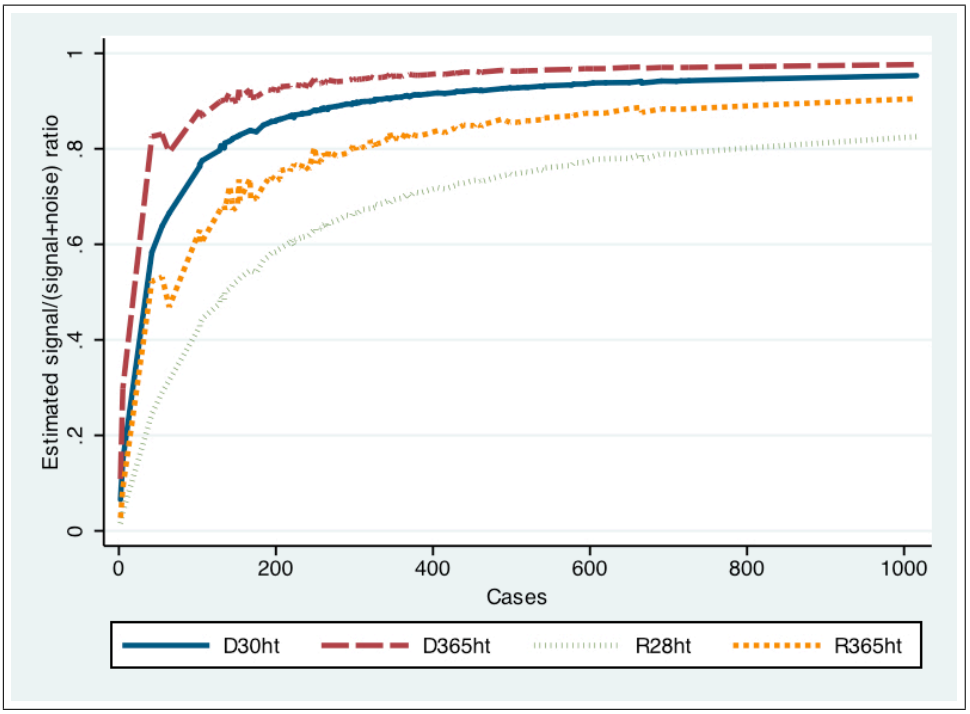
	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D30_{h(t-1)}$	0.078627	-0.023861	0.582003	-0.330667
	(0.04077)	(0.02525)	(0.07844)	(0.04201)
	[1.92840]	[-0.94497]	[7.41973]	[-7.87205]
$R28_{h(t-1)}$	-0.299568	0.404420	-1.651768	0.478057
	(0.05853)	(0.03625)	(0.11260)	(0.06030)
	[-5.11841]	[11.1577]	[-14.6699]	[7.92850]

	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D365_{h(t-1)}$	0.166596	-0.052642	0.797091	-0.044305
	(0.01356)	(0.00840)	(0.02608)	(0.01397)
	[12.2879]	[-6.26978]	[30.5604]	[-3.17204]
$R365_{h(t-1)}$	0.043576	0.012759	0.536484	-0.003055
	(0.03673)	(0.02274)	(0.07066)	(0.03784)
	[1.18648]	[0.56097]	[7.59290]	[-0.08073]
Residuals				
S.D. dependent	0.057489	0.036205	0.172179	0.058462
Correlation of residuals ($D30_{ht}$)	1.000000	-0.195636	0.281587	-0.272041
Correlation of residuals ($R28_{ht}$)	-0.195636	1.000000	-0.172637	0.478933
Correlation of residuals ($D365_{ht}$)	0.281587	-0.172637	1.000000	-0.437937
Correlation of residuals ($R365ht$)	-0.272041	0.478933	-0.437937	1.000000
Initial Conditions				
S.D. dependent in 2000	0.095917	0.029137	0.038380	0.03838
Correlation with $D30_{ht}$ in 2000	-	-0.5124	0.0335	0.0641
Correlation with $R28_{ht}$ in 2000	-0.5124	-	-0.0304	0.0334
Correlation with $D365_{ht}$ in 2000	0.0335	-0.0304	-	-0.0431
Correlation with $R365ht$ in 2000	0.0641	0.0334	-0.0431	-
Sample (adjusted): 2001 2008				
Included observations: 952 after adjustments				
Standard errors in ( ) & t-statistics in [ ]				

Figure 3.1 presents the signal to noise ratio of the four AMI outcome measures. This is calculated as specified by equation (3.10) using the signal variance estimated in the VAR equation as well as the observed measurement error from the patient level equations. The ratio estimates of the amount of signal variance to total (signal plus noise) variance in the observed hospital outcome measures, and plots this ratio against the number of cases treated in each hospital. Recall that these are the cases used to construct the measures in Chapter 2. What is immediately apparent from Figure 3.1 is the very high signal to noise ratios, especially once the number of cases rises above 200, which is indicative that the outcome measures are strong estimates of quality. Of the four measures, the two mortality measures have the strongest signal, where year-long mortality is a better predictor of performance than 30-day mortality due the higher variance across hospitals in the true effects observed in Table 3.2. However, as the sample exceeds 300 patients, the difference between the two indicators ratios begins to shrink, suggesting that both

indicators can be used to detect a large amount of the mortality-related quality difference between hospitals. While, the readmission measures also have good signal to noise ratios, and especially year-long readmissions, they are lower than the mortality measures. In the larger hospitals the indicators do have relatively strong signals, but for the small hospitals they remain, as might be expected given the smaller sample sizes, relative noisy measures of performance.

Figure 3.1: Signal to noise ratio for the four AMI outcome measures (year 2005).



Figures 3.2 – 3.5 present the filtered AMI outcome measures (black dashed line) for selected hospitals, together with their confidence intervals (red dotted lines), and the latent outcome measures derived in Chapter 2 (blue solid line). There are two features of the filtered estimates that stand out when compared to the latent measures. The first is that, as expected, the filtered estimates move smoothly from year to year, while the latent indicators are more erratic. The filtered estimates tend to be closer to zero than the aggregated estimates, indicating their tendency to approach the average. The other noticeable difference between Figures 3.2 – 3.5 and the corresponding Figures 2.3 – 2.6 in Chapter 2 are the confidence intervals which are much wider for the filtered measures than they were for the latent variables. Thus while the filtered measures seem more consistent over time, the wider confidence intervals surrounding them make it harder to interpret them with certainty as compared to the latent measures.

Figure 3.2: Filtered and latent estimates for AMI $D30_{ht}$ for selected hospitals.

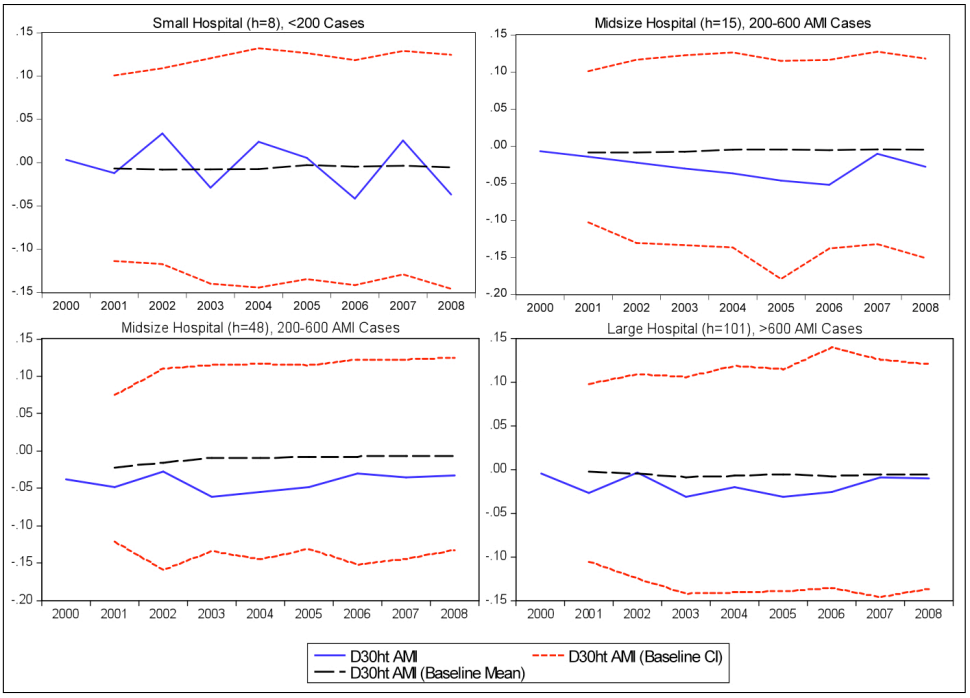


Figure 3.3: Filtered and latent estimates for AMI $D365_{ht}$ for selected hospitals.

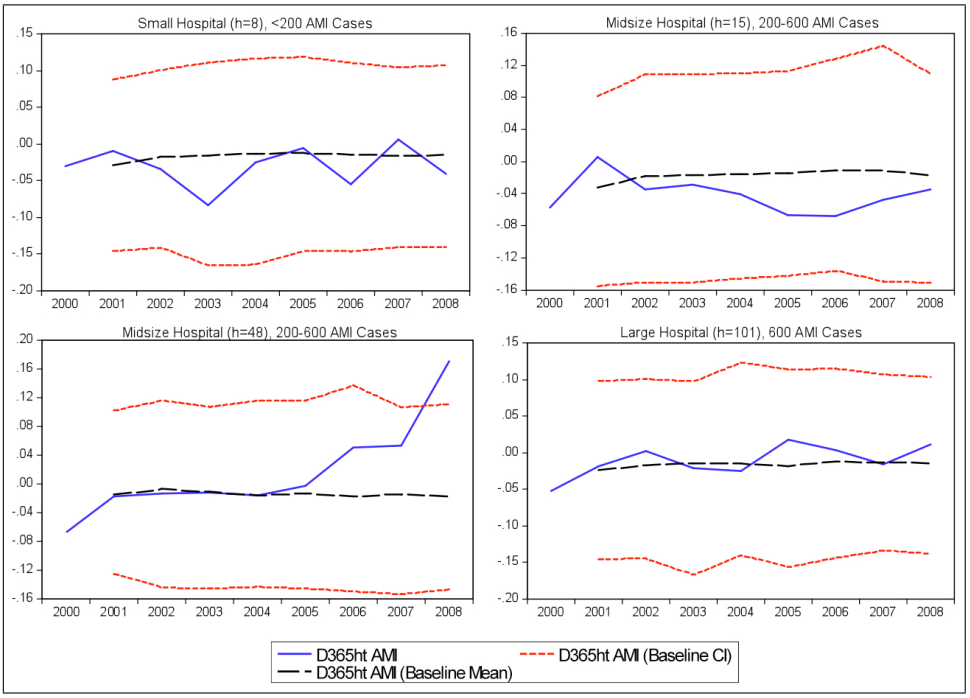


Figure 3.4: Filtered and latent estimates for AMI $R28_{ht}$ for selected hospitals.

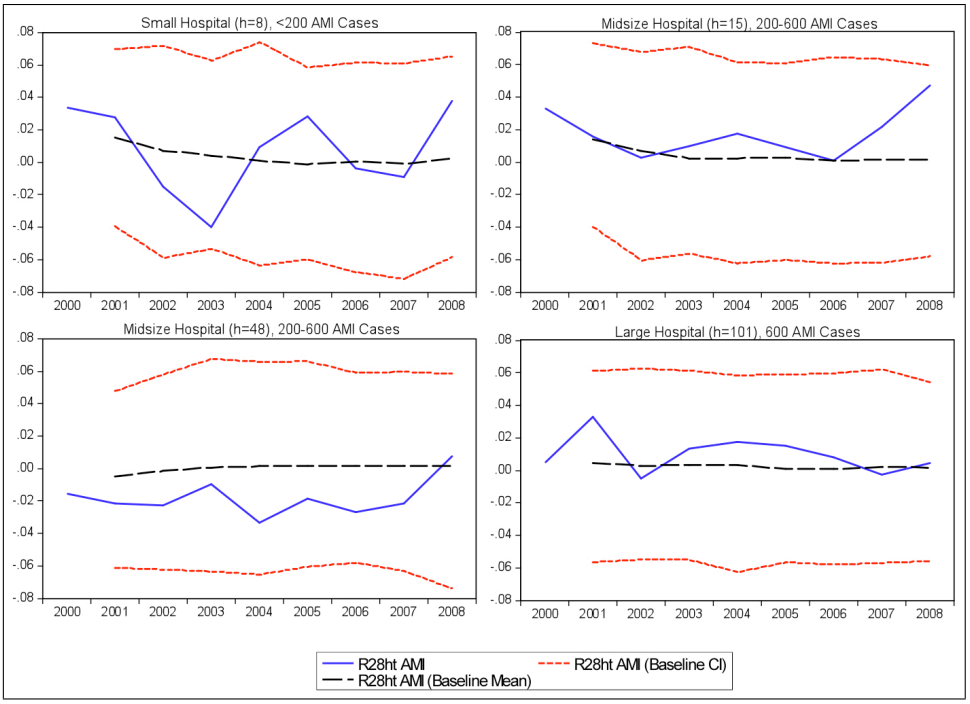


Figure 3.5: Filtered and latent estimates for AMI $R365_{ht}$ for selected hospitals.

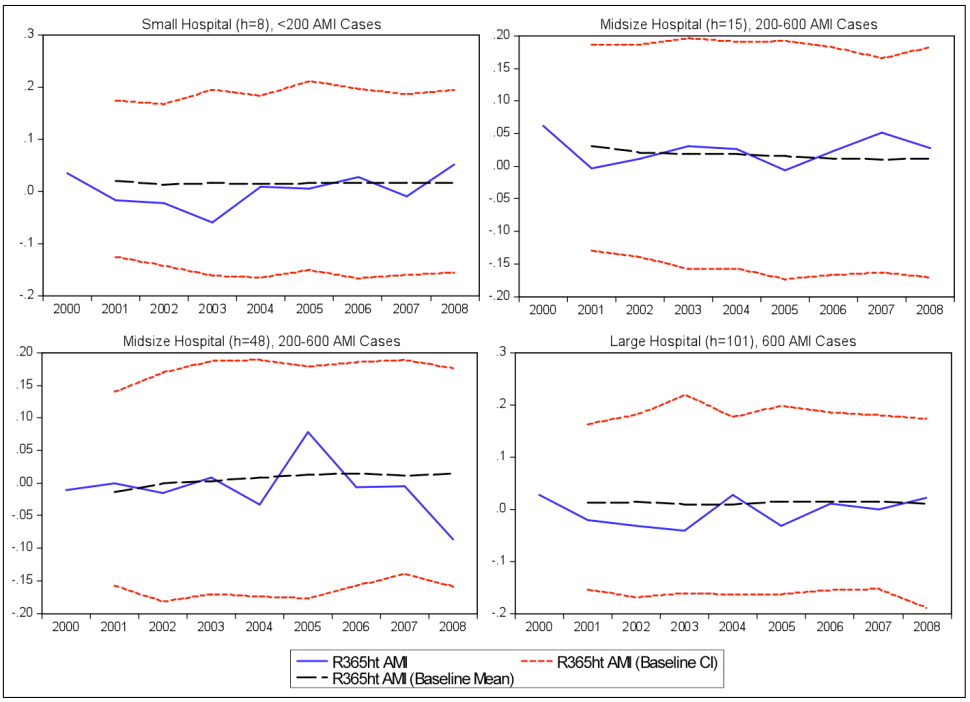


Table 3.3 indicates the R-squared estimates as calculated from equation (3.11) discussed above. These are presented for the predictions made of the different outcome measures, using different amounts of past data. The table indicates very high R-squared

values for all measures, suggesting that the filtered estimates are able to predict extremely well. In all cases the predicted R-squared values suggest that the filtered estimates capture over 90% of the true variation across hospitals in the different outcomes measures. Only for one-year mortality are the estimates a bit lower, although even then they do not fall below 79%. Table 3.3 also indicates that the filtered estimates are able to predict just as well using fewer years of data.

Table 3.3: Summary of estimated prediction accuracy using alternative methods of signal extraction. All estimates based on the VAR(1) model from Table 3.2.

Expected R^2 prediction based on:						
	All 8 years		3 most recent years		Concurrent year	
	All outcomes	Same outcome	All outcomes	Same outcome	All outcomes	Same outcome
$D30_{ht}$						
2004	0.993171	0.993224	0.993237	0.993246	0.994526	0.994452
2006	0.979275	0.979259	0.981738	0.981795	0.979818	0.979875
$D365_{ht}$						
2004	0.891798	0.892396	0.891843	0.891521	0.990980	0.990974
2006	0.981158	0.980648	0.916352	0.916693	0.796221	0.796244
$R28_{ht}$						
2004	0.996880	0.996899	0.996901	0.996891	0.997927	0.997931
2006	0.996920	0.996921	0.997074	0.997065	0.997650	0.997664
$R365_{ht}$						
2004	0.991736	0.991746	0.991792	0.991701	0.992516	0.992544
2006	0.989215	0.989353	0.989767	0.989848	0.991058	0.991133

The R-squared values for the outcome forecasts are presented in Table 3.4. The expected R-squared values are derived using equation (3.12) and represent how well the forecasts are able to predict the true values. The actual R-squared values indicate how well the predictions fit the data when using a full sample. Both the actual and the expected R-squared values are very high. While the expected R-squared values are lower than the actual R-squared values the difference is very small, and never more than14%. This indicates that the forecasts are also able to predict the true values extremely well for up to two years after the end of the data set. The results are also presented for a VAR(2) specification of the model, and are almost identical to the VAR(1) results. This indicates that the forecast performance is not sensitive to the lag choice specified for this VAR model.

Table 3.4: Summary of forecast accuracy using alternative forecasting models. Forecasting 2006-2008 values using data from 2000-2006.

	All outcomes	Same outcome	All outcomes	Same outcome
	VAR(1), forecasting with		VAR(2), forecasting with	
$D30_{ht}$				
2007(expected)	0.997908	0.997619	0.998164	0.998201
2007 (actual)	0.9939783	0.9940615	0.9927514	0.9927658
2008(expected)	0.994683	0.994478	0.997798	0.997928
2008 (actual)	0.9489663	0.9486998	0.9446982	0.9446459
$D365_{ht}$				
2007(expected)	0.973235	0.971065	0.979825	0.979843
2007 (actual)	0.9774626	0.9764693	0.9616151	0.9613662
2008(expected)	0.968023	0.96491	0.976735	0.979905
2008 (actual)	0.9759809	0.9745514	0.9708943	0.9708727
$R28_{ht}$				
2007(expected)	0.97878	0.979752	0.993951	0.992514
2007 (actual)	0.9911799	0.9912462	0.9912541	0.9912401
2008(expected)	0.924943	0.912794	0.953368	0.957072
2008 (actual)	0.993593	0.9936331	0.9943355	0.9943442
$R365_{ht}$				
2007(expected)	0.890177	0.890824	0.895657	0.867804
2007 (actual)	0.9843904	0.9845041	0.9845231	0.9842737
2008(expected)	0.846979	0.84891	0.828721	0.841011
2008 (actual)	0.980951	0.981266	0.9836124	0.9833608

Stroke

Table 3.5 presents the parameter estimates from the Stroke VAR model. The lag coefficients suggest that death is a persistent dimension of hospital quality; ranging at about 0.6 for both $D30_{ht}$ and $D365_{ht}$. The variance of the residuals indicate an annual standard deviation of 5% and 8% respectively, while the initial variance indicates a variance of around 6% to 9% across hospitals. The parameters on the readmission indicators suggest that these are much less dynamic. The lag coefficients are always less than 0.07, and the sign varies between positive and negative. The variance in the year 2000, indicates little variation across hospitals; corresponding to a standard deviation of around 3% for $R28_{ht}$ and 5% for R365ht. The variance of residuals indicates similar variation annually;

indicating a standard deviation of about 4% and 6% respectively. Similar to the other conditions, there is indication of a high positive correlation between the pairs of mortality and readmission measures. There is also a mild negative correlation between $R365_{ht}$ and both mortality measures. All other correlations are negative and weak.

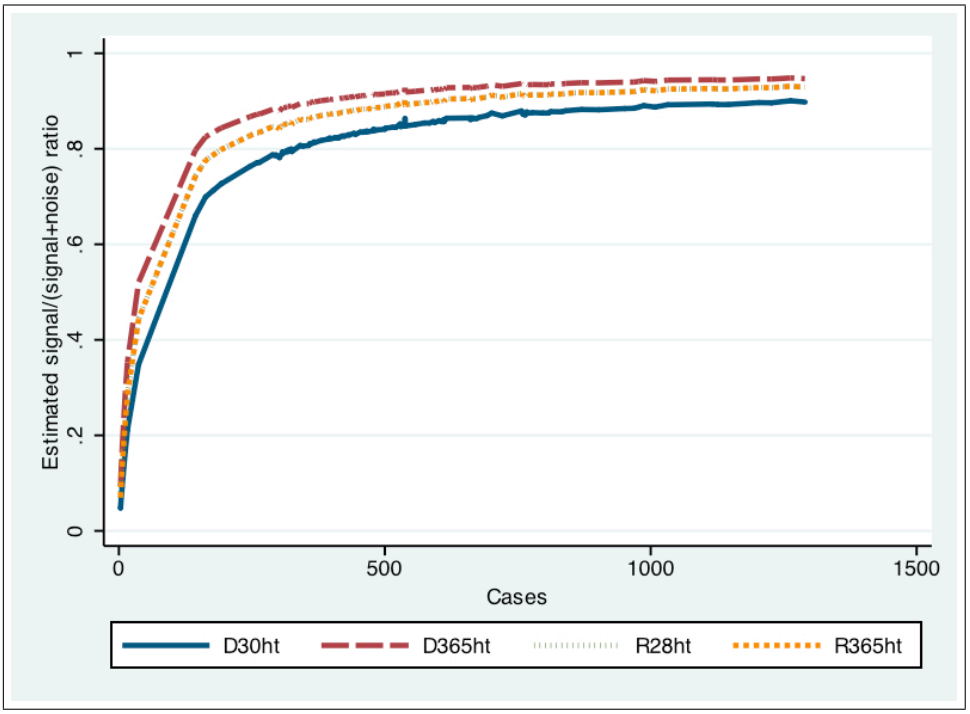
Table 3.5: Estimates of Stroke multivariate VAR(1) parameters for hospital specific effects.

	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D30_{h(t-1)}$	0.423853	-0.117846	-0.203754	0.073602
	(0.03406)	(0.02607)	(0.05454)	(0.03794)
	[12.4429]	[-4.52056]	[-3.73594]	[1.94019]
$R28_{h(t-1)}$	0.089437	-0.040873	-0.190125	-0.118715
	(0.05469)	(0.04186)	(0.08757)	(0.06091)
	[1.63524]	[-0.97649]	[-2.17116]	[-1.94903]
$D365_{h(t-1)}$	0.158739	0.160186	0.840151	0.171265
	(0.01975)	(0.01511)	(0.03161)	(0.02199)
	[8.03908]	[10.6003]	[26.5745]	[7.78821]
$R365_{h(t-1)}$	0.116206	0.056785	0.286192	0.010468
	(0.03664)	(0.02804)	(0.05867)	(0.04081)
	[3.17113]	[2.02485]	[4.87787]	[0.25651]
Residuals				
S.D. dependent	0.068801	0.042570	0.122831	0.062766
Correlation of residuals ($D30_{ht}$)	1.000000	-0.136723	0.553742	-0.418894
Correlation of residuals ($R28_{ht}$)	-0.136723	1.000000	-0.195895	0.680239
Correlation of residuals ($D365_{ht}$)	0.553742	-0.195895	1.000000	-0.399833
Correlation of residuals ($R365_{ht}$)	-0.418894	0.680239	-0.399833	1.000000
Initial Conditions				
S.D. dependent in 2000	0.056062	0.032202	0.088837	0.053572
Correlation with $D30_{ht}$ in 2000	-	-0.1007	0.6780	0.0523
Correlation with $R28_{ht}$ in 2000	-0.1007	-	0.3266	0.4023
Correlation with $D365_{ht}$ in 2000	0.6780	0.3266	-	0.2434
Correlation with $R365_{ht}$ in 2000	0.0523	0.4023	0.2434	-
Sample (adjusted): 2001 2008				
Included observations: 968 after adjustments				
Standard errors in ( ) & t-statistics in [ ]				

Figure 3.6 plots the estimates of the signal to noise ratio in the observed hospital

outcome measures for Stroke against the number of hospital admissions. Similar to AMI, all measures have a very strong signal, which improves substantially in hospitals with more cases. Of the four measures, year-long mortality has the strongest signal, again this is probably related to the high signal variance for this indicator, as observed in Table 3.5. However the other measures also perform very well. The readmission measures have almost the same signal to noise ratio and overlap in Figure 3.6, while 30-day mortality has the weakest signal but which is still quite high. As the sample exceeds 250 patients, the difference between the four indicator's ratios begins to shrink, suggesting that all indicators can be used relatively confidently to detect a large amount of the quality differences between hospitals.

Figure 3.6: Signal to noise ratio of the four Stroke outcome measures (year 2005) .



Figures 3.7 – 3.10 present the filtered Stroke outcome measures (black dashed line) for selected hospitals, together with their confidence intervals (red dotted lines), and the latent outcome measures derived in Chapter 2 (blue solid line). There same two features identified above when interpreting the AMI figures can also be observed for the Stroke figures: namely the smoothed out estimates for the filtered measures as compared to the latent indicator, and the wide confidence intervals. However, in some cases such as the small hospital in the upper left hand panel, the filtered estimates provide a clearer picture as to performance over time, as the latent figures change value significantly from year to year. The filtered measures indicate that the hospital ($h = 26$) has declining mortality over

the time period studied combined with declining readmissions. This improved performance results in the hospital moving from being worse than average on all outcomes to average by the year 2008. The other three hospital's filtered measures illustrated in the panel all indicate average performance throughout the period being investigated.

Figure 3.7: Filtered and latent estimates for Stroke $D30ht$.

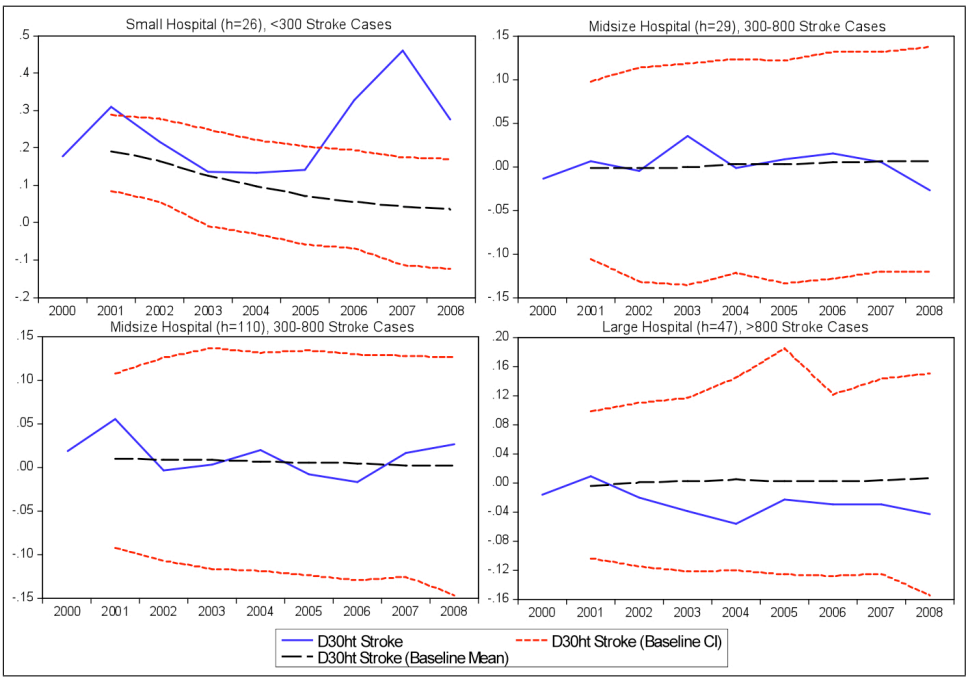


Figure 3.8: Filtered and latent estimates for Stroke $D365_{ht}$.

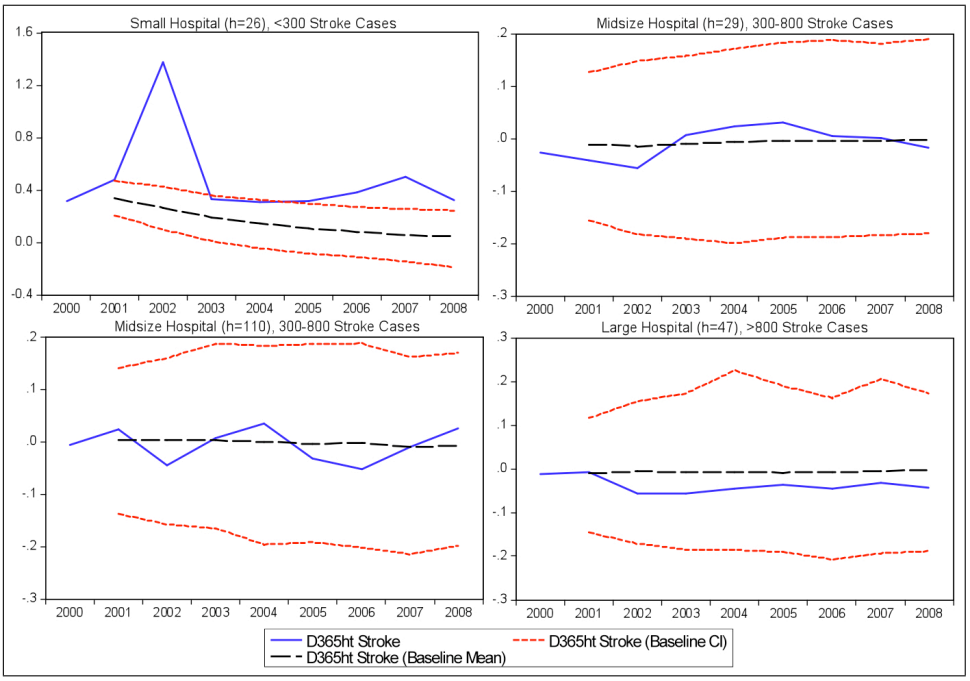


Figure 3.9: Filtered and latent estimates for Stroke $R28_{ht}$.

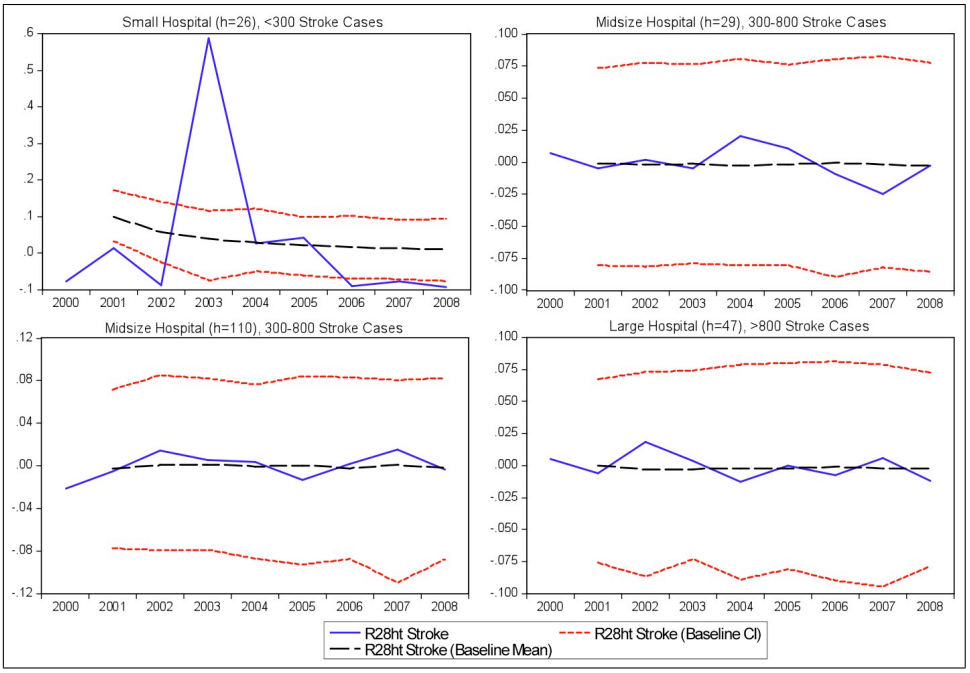


Figure 3.10: Filtered and latent estimates for Stroke $R365_{ht}$.

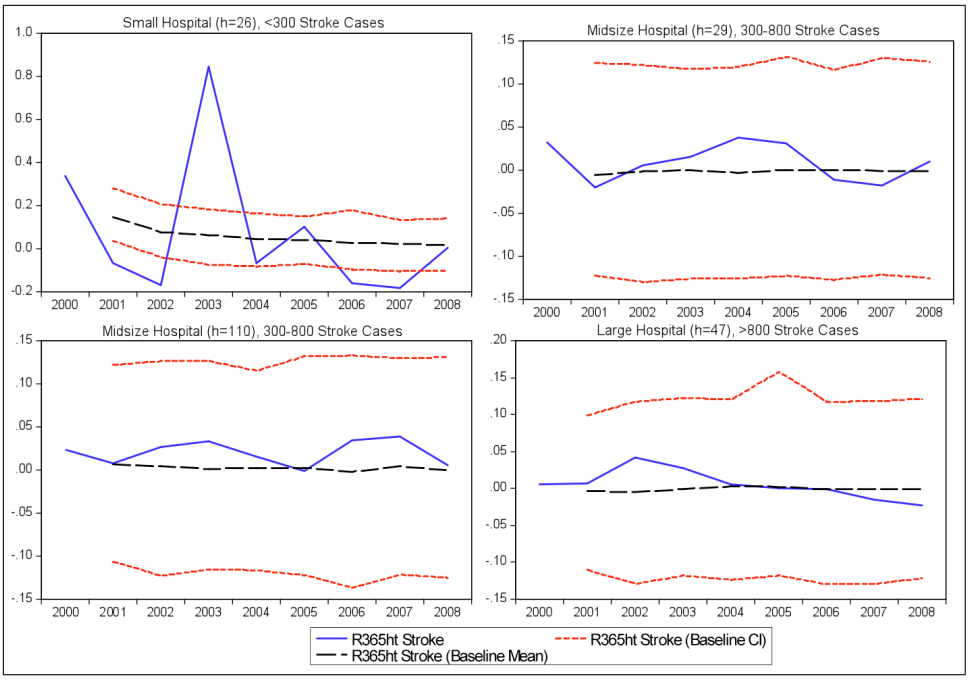


Table 3.6 indicates the R-squared estimates for the predictions made for Stroke outcomes measures, using different amounts of past data. While the R-squared values for Stroke are not as high as for AMI, they remain high for all outcomes. The mortality estimates are able to predict better than the readmission estimates, but even these are

able to capture over 77% of the true variation in hospitals. Similar to the AMI results, the filtered estimates constructed from as little as one year of data are able to predict the true variation just as well as the larger samples. Indeed, in some cases they are even better predictors than the larger samples. The R-squared values for the outcome forecasts presented in Table 3.7 indicate that the model is able to forecast estimates as well as it is able to predict them from a full set of data. Both the actual and the expected R-squared values are very high, and almost identical. The results for the different model specifications indicate the forecast performance is not sensitive to the lag choice specified in the VAR model, as the results for both the VAR(1) and the VAR(2) specifications are very high and almost the same in value.

Table 3.6: Summary of estimated prediction accuracy using alternative methods of signal extraction. All estimates based on the VAR(1) model from Table 3.5.

Expected R^2 prediction based on:						
	All 8 years		3 most recent years		Concurrent year	
	All outcomes	Same outcome	All outcomes	Same outcome	All outcomes	Same outcome
$D30_{ht}$						
2004	0.983698	0.983669	0.983509	0.983862	0.991067	0.991066
2006	0.980058	0.980046	0.982017	0.981926	0.987698	0.987725
$D365_{ht}$						
2004	0.938718	0.9384	0.938640	0.938740	0.969390	0.969351
2006	0.981663	0.981831	0.980129	0.979591	0.940582	0.939948
$R28_{ht}$						
2004	0.994965	0.994965	0.994909	0.994961	0.993634	0.993594
2006	0.833847	0.833997	0.833539	0.833902	0.833445	0.833399
$R365_{ht}$						
2004	0.772766	0.773075	0.772993	0.772996	0.807686	0.806639
2006	0.937973	0.937642	0.937648	0.937475	0.93510	0.935357

Table 3.7: Summary of forecast accuracy using alternative forecasting models. Forecasting 2006-2008 values using data from 2000-2006.

	All outcomes	Same outcome	All outcomes	Same outcome
	VAR(1), forecasting with		VAR(2), forecasting with	
$D30_{ht}$				
2007(expected)	0.9993262	0.999328	0.9994763	0.9995843
2007 (actual)	0.9893464	0.9895585	0.9897618	0.989777
2008(expected)	0.9982069	0.9982541	0.9986149	0.9987863
2008 (actual)	0.9914728	0.9915513	0.9916081	0.9917418
$D365_{ht}$				
2007(expected)	0.9925504	0.9929932	0.9930512	0.9935482
2007 (actual)	0.9859509	0.9866322	0.991607	0.9913715
2008(expected)	0.9913359	0.9915887	0.9919757	0.9928179
2008 (actual)	0.9813546	0.9818443	0.9898155	0.9895275
$R28_{ht}$				
2007(expected)	0.9915956	0.9906852	0.9906945	0.9904812
2007 (actual)	0.9973422	0.9973438	0.9961982	0.9961572
2008(expected)	0.9991251	0.99915	0.9993581	0.9992571
2008 (actual)	0.9939768	0.9939387	0.9943756	0.9943268
$R365_{ht}$				
2007(expected)	0.9982554	0.9983872	0.9980332	0.9981403
2007 (actual)	0.9909626	0.9909122	0.9905966	0.9904634
2008(expected)	0.9987442	0.9989667	0.9989013	0.9988669
2008 (actual)	0.9871836	0.9872606	0.9882355	0.9880505

Hip Replacement

The parameter estimates of the basic model run for Hip Replacement are presented in Table 3.8. The estimates suggest that $D365_{ht}$ is persistent over time, but that the other quality indicators being considered are not. The lag coefficient of $D365_{ht}$ is almost 0.6, as compared to lag coefficients of about 0.2 for $R28_{ht}$ and $R365_{ht}$, and about 0.01 for $D30_{ht}$. The variance of initial conditions indicates a standard deviation of about 2% across hospitals for $D30_{ht}$, 3% for $R28_{ht}$, 4% for $D365_{ht}$ and 5% for $R365_{ht}$. Similarly the variance of their residuals shows an annual standard deviation of 1% for $D30_{ht}$ and $D365_{ht}$, 3% for $R28_{ht}$ and 4% for $R365_{ht}$. The correlation coefficients amongst indicators, and amongst residuals, indicate a high positive correlation between $R365_{ht}$ and $R28_{ht}$,

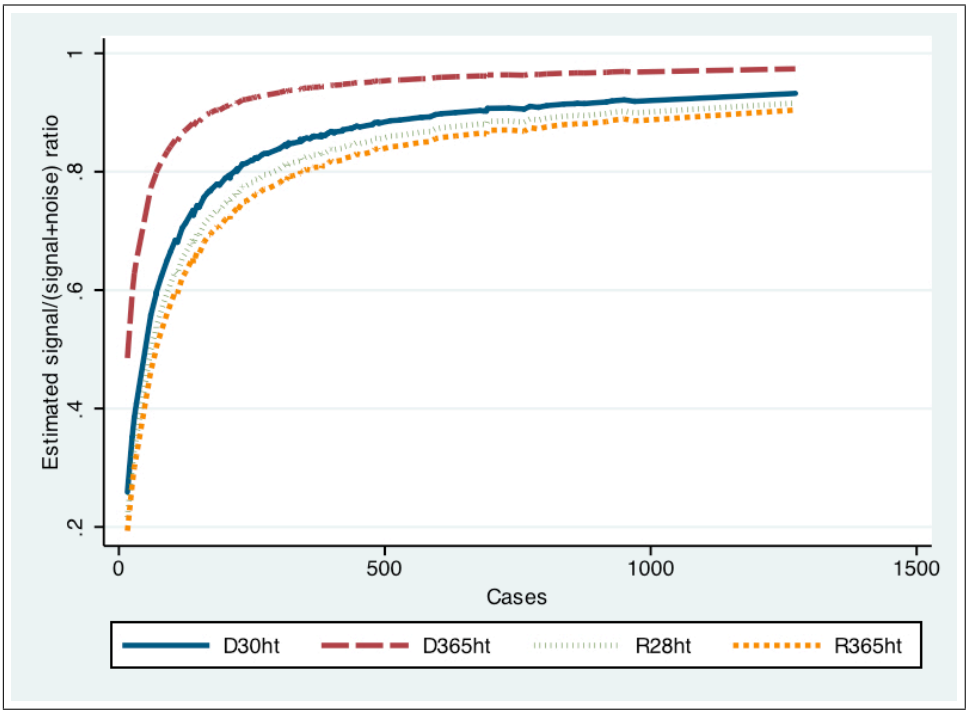
and a weak positive correlation between $D30_{ht}$ and $D365_{ht}$. There is a positive correlation between the residuals of $D365_{ht}$ and $R28_{ht}$, while the correlation coefficient amongst these two indicators in the year 2000 is low and negative. The opposite is true for the pair $D365_{ht}$ and $R365_{ht}$ which have a negative correlation in the year 2000, but a low positive correlation between their residuals. Finally there is a positive correlation between $D30_{ht}$ and $R28_{ht}$.

Table 3.8: Estimates of Hip Replacement multivariate VAR(1) parameters for hospital specific effects.

	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D30_{h(t-1)}$	-0.047351	-0.224300	-0.627994	-0.282623
	(0.02543)	(0.07851)	(0.08952)	(0.09652)
	[-1.86231]	[-2.85705]	[-7.01536]	[-2.92803]
$R28_{h(t-1)}$	-0.030140	0.312121	-0.359189	0.468140
	(0.01479)	(0.04567)	(0.05207)	(0.05615)
	[-2.03789]	[6.83480]	[-6.89816]	[8.33795]
$D365_{h(t-1)}$	0.036579	0.058774	0.633914	-0.029772
	(0.00686)	(0.02119)	(0.02417)	(0.02606)
	[5.32910]	[2.77313]	[26.2315]	[-1.14255]
$R365_{h(t-1)}$	-0.016563	-0.045723	-0.039086	0.018910
	(0.01098)	(0.03390)	(0.03865)	(0.04168)
	[-1.50871]	[-1.34884]	[-1.01124]	[0.45373]
Residuals				
S.D. dependent	0.011466	0.036723	0.049172	0.046638
Correlation of residuals ($D30_{ht}$)	1.000000	-0.197193	0.262098	-0.250718
Correlation of residuals ($R28_{ht}$)	-0.197193	1.000000	0.350683	0.790476
Correlation of residuals ($D365_{ht}$)	0.262098	0.350683	1.000000	0.149165
Correlation of residuals ($R365_{ht}$)	-0.250718	0.790476	0.149165	1.000000
Initial Conditions				
S.D. dependent in 2000	0.019079	0.033392	0.044777	0.046217
Correlation with $D30_{ht}$ in 2000	-	0.3661	0.2470	0.1459
Correlation with $R28_{ht}$ in 2000	0.3661	-	-0.1613	0.7196
Correlation with $D365_{ht}$ in 2000	0.2470	-0.1613	-	-0.4921
Correlation with $R365_{ht}$ in 2000	0.1459	0.7196	-0.4921	-
Sample (adjusted): 1997 2008				
Included observations: 1462 after adjustments				
Standard errors in ( ) & t-statistics in [ ]				

Figure 3.11 illustrates the signal to noise ratios of the observed hospital outcome measures against the number of Hip Replacement cases treated in each hospital. For Hip Replacement, the signal to noise ratios are quite high, indicating that the four outcome measures are good indicators of hospital performance. Similar to the previous conditions, the signal to noise ratio increases as more cases are included in the analysis, and the differences between the four indicators begin to shrink. Yet, year-long mortality consistently has the strongest signal of the four conditions, despite not having as high a signal variance as it did for AMI and Stroke. While year-long readmissions have a higher signal variance than year-long mortality (Table 3.8), they most probably have higher amounts in the variance of the estimation error, causing them to perform the worst of the four measures.

Figure 3.11: Signal to noise ratio for the four Hip Replacement outcome measures (year 2005) .



Figures 3.12 – 3.15 present the filtered Hip outcome measures, their 95% confidence intervals and the corresponding latent outcome measures derived in Chapter 2 for selected hospitals. The sample for Hip Replacement is longer than for AMI and Stroke, and so all figures present information back to 1996. Similar to the other two conditions, the filtered estimates are smoothed averages of the latent measures, and the confidence intervals are wider, again due to a limited number of hospitals available in the data. Also similar to Stroke, the latent measure for the small hospital, upper left hand corner, is more erratic than for the medium and large hospitals, thus making the filtered estimates useful in terms of interpreting a trend over time.

Figure 3.12: Filtered and latent estimates for Hip Replacement $D30_{ht}$.

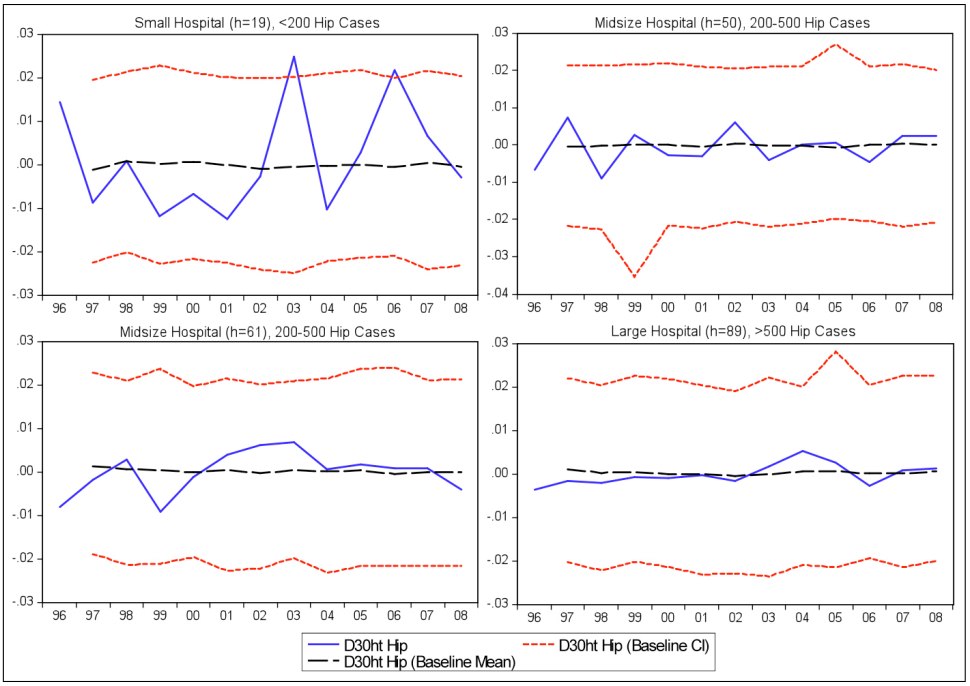


Figure 3.13: Filtered and latent estimates for Hip Replacement $D365_{ht}$.

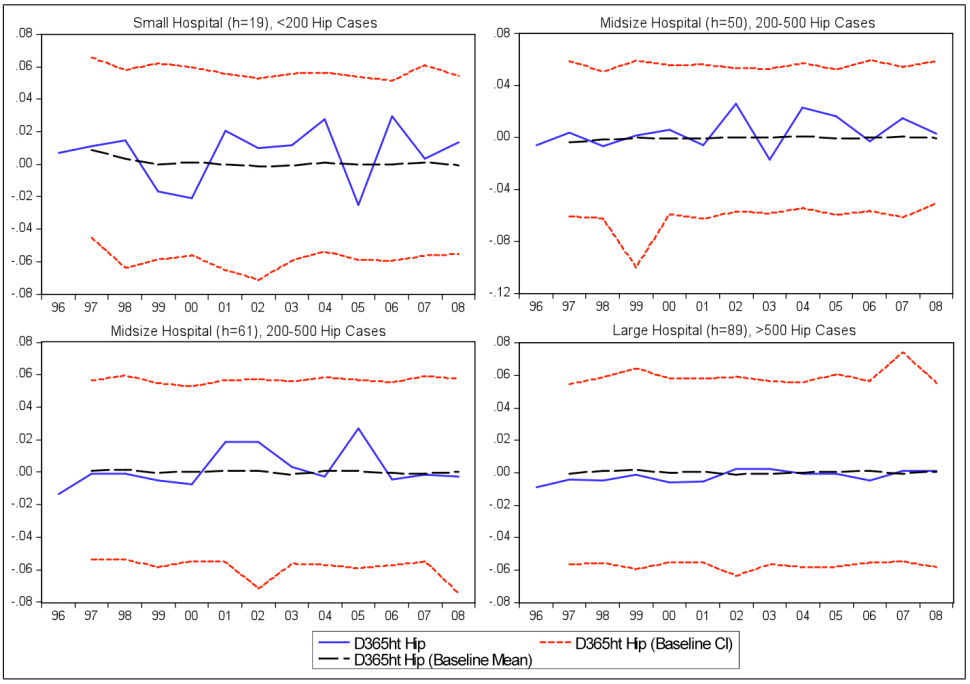


Figure 3.14: Filtered and latent estimates of Hip Replacement $R28_{ht}$.

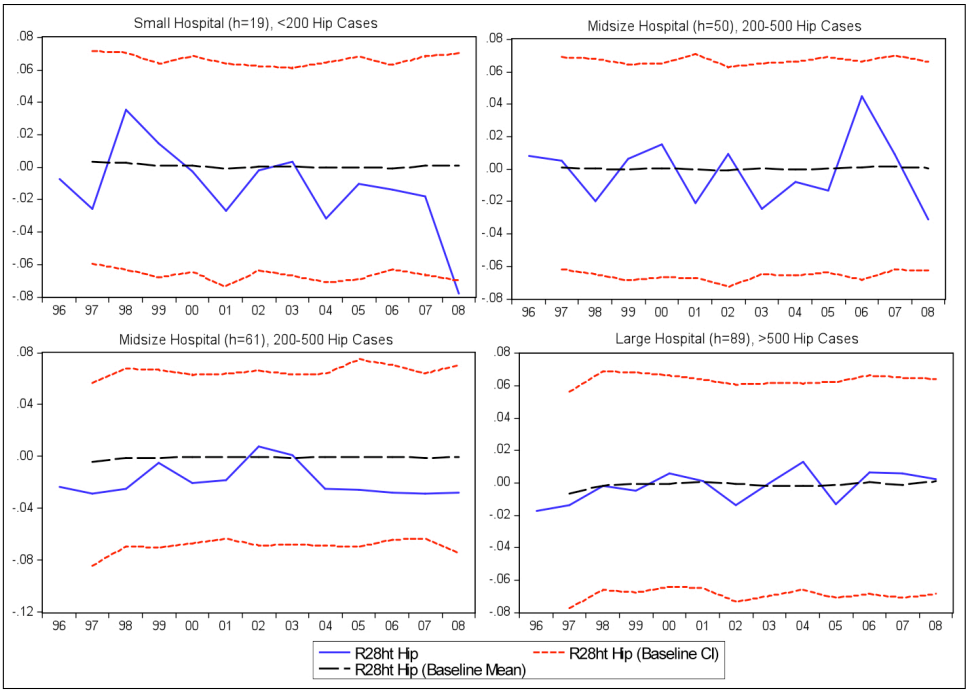


Figure 3.15: Filtered and latent estimates of Hip Replacement $R365_{ht}$.

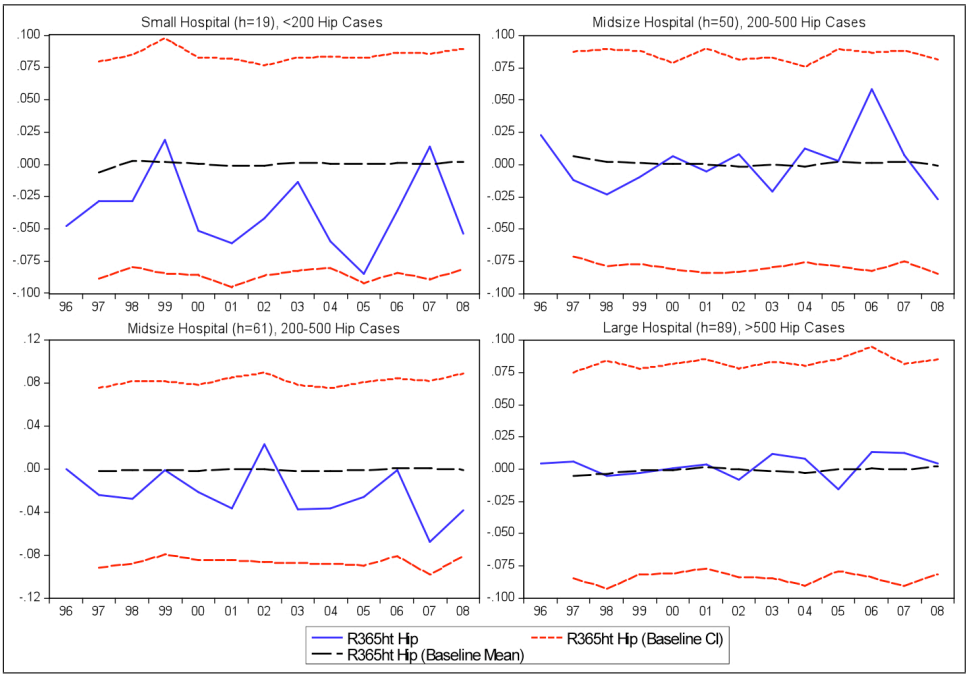


Table 3.9 indicates the R-squared estimates for the predictions made for the Hip filtered outcomes, using different amounts of past data. The R-squared values for Hip are extremely high, indicating a near perfect prediction for all measures, even when using only

one year of data. Table 3.10 indicates the R-squared values for the outcome forecasts, estimated using equation (3.12), and predictions estimated using equation (3.11). These are also near perfect for both the forecasts and predictions, and both the VAR(1) and VAR(2) specifications. This indicate that the model is able to forecast estimates as well as it is able to predict them from a full set of data, regardless of the lag choice specified in the model.

Table 3.9: Summary of estimated prediction accuracy using alternative methods of signal extraction. All estimates based on the VAR(1) model from Table 3.8.

Expected R^2 prediction based on:						
	All 11 years		3 most recent years		Concurrent year	
	All outcomes	Same outcome	All outcomes	Same outcome	All outcomes	Same outcome
$D30_{ht}$						
2004	0.999851	0.999851	0.999850	0.999852	0.999824	0.999829
2006	0.999856	0.999852	0.999860	0.999857	0.999840	0.999840
$D365_{ht}$						
2004	0.993021	0.992983	0.992833	0.992773	0.998047	0.998065
2006	0.994185	0.994248	0.991052	0.990711	0.982275	0.982161
$R28_{ht}$						
2004	0.998588	0.998589	0.998595	0.998593	0.998714	0.998706
2006	0.997845	0.997845	0.997835	0.997836	0.997967	0.997969
$R365_{ht}$						
2004	0.995829	0.995849	0.995807	0.995831	0.996284	0.996242
2006	0.993924	0.993940	0.993907	0.993959	0.995122	0.995136

Table 3.10: Summary of forecast accuracy using alternative forecasting models. Forecasting 1996-2008 values using data from 1996-2006.

	All outcomes	Same outcome	All outcomes	Same outcome
	VAR(1), forecasting with		VAR(2), forecasting with	
$D30_{ht}$				
2007(expected)	0.999837	0.9998281	0.9998208	0.9998139
2007 (actual)	0.9998575	0.9998609	0.9998577	0.9998588
2008(expected)	0.9997321	0.999688	0.9997113	0.9996896
2008 (actual)	0.9998561	0.999858	0.9998613	0.9998624

	All outcomes	Same outcome	All outcomes	Same outcome
$D365_{ht}$				
2007(expected)	0.9968599	0.9970006	0.9963497	0.9963019
2007 (actual)	0.9869273	0.9871355	0.9848145	0.9850215
2008(expected)	0.9965712	0.9964086	0.9957694	0.9954451
2008 (actual)	0.9840067	0.9841068	0.9814323	0.9818322
$R28_{ht}$				
2007(expected)	0.9985577	0.9983832	0.9987864	0.9986095
2007 (actual)	0.9980288	0.998031	0.9980153	0.9980155
2008(expected)	0.9995171	0.9995244	0.999558	0.9995869
2008 (actual)	0.9767528	0.9767398	0.9767273	0.9767253
$R365_{ht}$				
2007(expected)	0.9989753	0.9989704	0.9990094	0.9990171
2007 (actual)	0.9928861	0.9929147	0.9931077	0.993055
2008(expected)	0.999464	0.9994054	0.999514	0.9994828
2008 (actual)	0.9878172	0.9878773	0.9880453	0.9880126

Comparison of Indicators

In this subsection, we are able to relate our findings to policy by ranking the hospitals in the AMI sample using three different indicators of performance for the year 2005. The first indicator is an aggregated 30-day mortality rate as available in the raw data. The second performance indicator is the latent 30-day mortality rate derived in Chapter 2, while the third measure is the filtered 30-day mortality rate estimated in this chapter. The hospitals are also ranked by the other outcomes and these are reported in Appendix B.5 due to space constraints. The year 2005 is presented as it is in the middle of the sample and allows enough information to construct the filtered measures from, however the R-squared values in the AMI section suggest that even with less data the filtered measures are still good predictors. The outcomes are ranked only for AMI at not the other conditions, as the results are very similar and do not provide further insight.

Table 3.11: Rankings of 2005 AMI $D30_{ht}$ measures.

Ranking	Mean $D30_{ht}$	Hospital	Latent $D30_{ht}$	Hospital	Filtered $D30_{ht}$	Hospital
Top 10						
1	0.0521401	55	-8.417754	83	-2.490163	17
2	0.0532544	9	-5.088554	81	-2.113111	54
3	0.0536913	89	-5.00683	42	-1.934144	22

Ranking	Mean $D30_{ht}$	Hospital	Latent $D30_{ht}$	Hospital	Filtered $D30_{ht}$	Hospital
4	0.0594286	119	-4.887803	47	-1.651729	103
5	0.0645161	62	-4.834379	22	-1.651613	3
6	0.0681818	19	-4.648541	15	-1.608179	18
7	0.0681818	97	-4.089908	1	-1.47745	7
8	0.0684932	80	-4.078938	50	-1.438395	107
9	0.0758808	52	-3.924413	16	-1.425196	21
10	0.0774194	42	-3.834195	68	-1.343411	89
Bottom 10						
110	0.1702128	12	2.985045	3	0.2998581	33
111	0.1727941	36	3.342186	7	0.3957789	118
112	0.1759531	96	3.580219	41	0.4082001	41
113	0.1787072	17	3.738158	89	0.4182017	99
114	0.19	53	4.557611	90	0.5266839	66
115	0.1901408	71	4.750142	17	0.5433974	38
116	0.1929825	41	5.562703	53	0.5688122	35
117	0.1987578	90	5.586496	71	0.9426492	27
118	0.2	66	18.70218	43	1.04961	9
119	0.3426574	43	28.97059	66	1.091938	56

Table 3.11 presents the top and bottom 10 hospitals as ranked by the three different performance measures together with the values of each measure. Each hospital is represented by a number which has been randomly assigned to be its identifier. Figure 3.16 illustrates the different rankings for the first 15 hospitals in the sample. What is immediately apparent from both Table 3.11 and Figure 3.16 is that depending on the indicator used the ranking of hospitals changes substantially, although not always in the same direction. Some hospitals go from a very high ranking to a very low ranking. Hospital 9 went from being ranked second best to second worst when using the filtered measure to rank performance instead of the raw aggregated mortality measure. Hospital 3 on the contrary, went from a very low ranking, 96 to a very high ranking, 5. There are also cases where two measures seem to be more similar to one another, but where rankings stay relatively consistent such as hospitals 11 and 15.

Figure 3.16: Rankings of 2005 AMI quality measures for $D30_{ht}$.

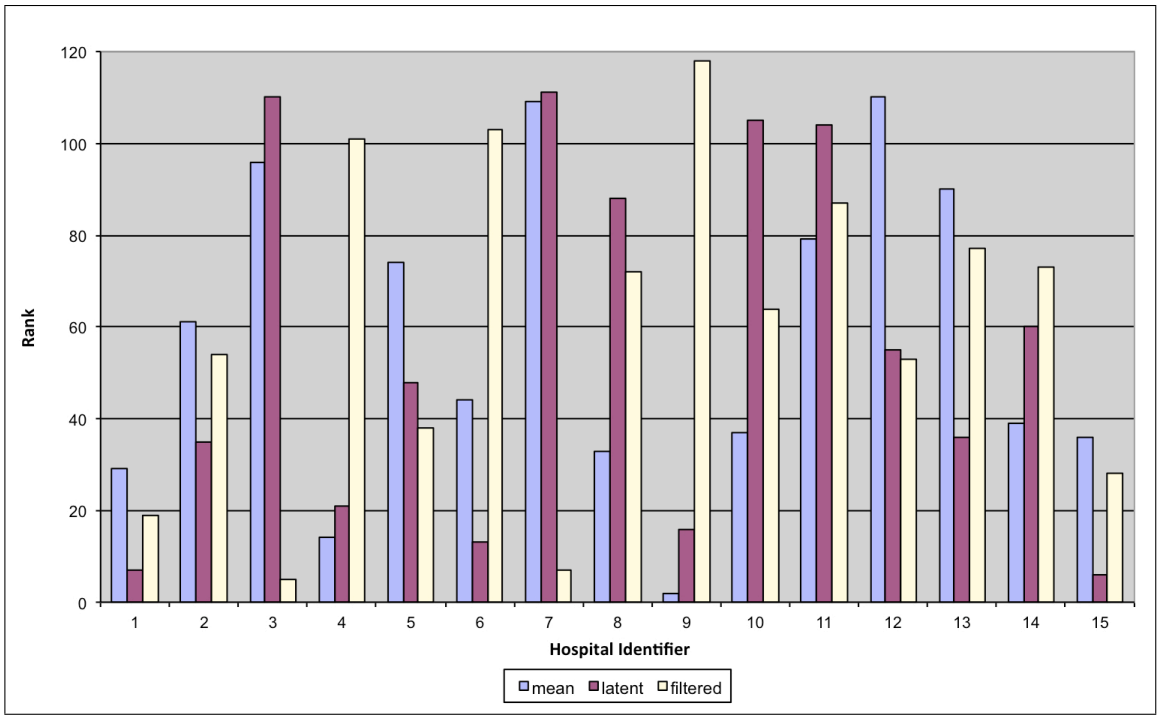
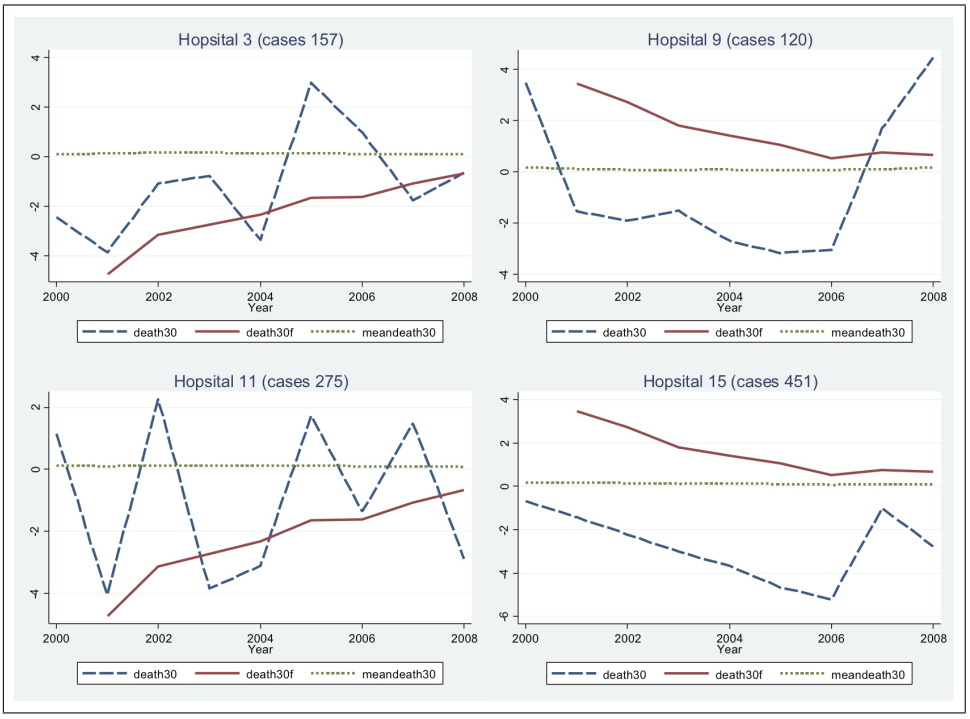


Figure 3.17 presents the full time series of the three different performance indicators for hospitals 3, 9, 11 and 15. This alternative presentation of the data can help to better understand why the rankings are different from one another. In the upper left hand corner the trajectories of hospital 3's indicators are presented. The mean raw mortality only ranges between 0 and 1, as each patient is coded as either having died or survived. When ranked according to this indicator, hospital 3 does relatively poorly coming in 96th out of 119 in 2005. This indicator does not adjust for differences in patient characteristics, such as co-morbidity or deprivation, while the latent measure does. When looking at the performance of hospital 3 as reported by latent measure there is much more variation from year to year. The year 2005 is the worst year in terms of hospital 3's performance, and the hospital is ranked 110 of 119. In all other years however, the hospital performs above average. The third indicator, the filtered measure, is constructed using the information provided throughout the time-series and from the other outcome measures. While the filtered indicator does reflect hospital 3's worsening performance over time, it smooths out the year-to-year variation allowing for a more representative overall picture when singling out one year. The performance ranking for hospital 3 using the filtered measure is 5 out of 119, which is a huge difference from the latent measure but reflects the hospital's above average performance in all the other years.

When looking at hospital 9 in the upper right hand panel, again the raw mortality

has much less variation than the other two indicators. Using this indicator hospital 9 ranks 2nd out of the 119 in 2005. The latent measure adjusts for some of the patient differences through the and shows a very different picture of performance, with much larger year to year variation. Performance as reported by this indicator starts out much worse than average in 2000, improving in the years 2001-2006, but worsening again after. In 2005 performance is still above average, but adjusting patient characteristics, the ranking falls from 2 to 16. The third indicator, the filtered measure, is constructed using the information provided throughout the time-series and from the other outcome measures. Thus, the improvement in performance is indicated, however not as sharply as by the latent measure, and never so much that it results in above average performance. This adjustment causes the ranking to drop down to 118.

Figure 3.17: AMI $D30_{ht}$ quality indicators for selected hospitals.



Hospital 11 in the bottom right hand panel ranks 79 out of 119 when using the aggregated raw mortality measure. However the latent variable indicates that when controlling for patient characteristics performance varies considerably from year to year, sometimes reaching very high levels above average, and others falling far below average. 2005 is one of the years where performance is below average, and thus when ranked according to it does poorly coming 104^{th} out of 119. The filtered indicator by definition provides a smoothed out measure of average performance across time and incorporating the performance of the other outcome measures. This is apparent from the diagram which shows less volatility

over time in the filtered indicator. Using this indicator the ranking falls down to 87^{th} out of 119, which lies between the two other measures. Finally, when looking at the performance of hospital 15 in the bottom right hand panel we see a similar result. The latent measure shows much more erratic performance from year to year once it controls for all the patient characteristics, and the filtered measure is able to summarize these into a much smoother, consistent trend.

Overall, the analysis provides support for the following: Aggregate raw measures are unable to produce a consistent performance ranking of hospitals that controls for systematic differences in patients case mix, such as deprivation or severity. The latent measures do adjust explicitly for these differences, but exhibit year-on-year variation and therefore different rankings of hospital performance depending on the year selected, making it difficult to draw conclusions on overall hospital performance over time. The filtered measures are able to summarize the information provided by the latent variable over time and consider the performance of the other indicators alongside it, thus providing a much more consistent picture of performance.

The largest difference in rankings is observed in hospitals treating fewer patients. Small caseload leads to increased volatility in the raw mortality and readmission measures across the years. While the latent measures control for systematic patient differences in hospitals, the volatility due to small numbers remains. This was obvious in Chapter 2 where the small hospitals always had the most erratic performance measures from year to year. The filtered measures are better at smoothing out the jumps from year to year as they combine all the information from the time-series and across the other variables. Thus in these cases, the filtered measure will be a better indication of performance in any one year.

3.6 Discussion

In their paper The Quality of Health Care Providers, McClellan and Staiger (1999) propose a methodology with which to evaluate health care providers. Their framework is able to tackle some of the main limitations inherent to quality measurement, allowing them to create indicators which: integrate different dimensions of quality into one measure, reflect the multifaceted nature of performance; filter out much of the noise inherent to this type of measure as a result of the small number of patients treated and the large number of factors which contribute to outcomes; and to eliminate much of the bias created from systematic differences in patient mix which may result in variations in treatment. Their paper uses US patient level data for elderly American's suffering from heart disease to create performance indicators at the hospital level. They are able to prove that the indicators they create predict and forecast quality remarkably well, better than many existing methods.

Despite its advantages over traditional methods, this analysis has not been applied to evaluate hospitals outside the US[2], or for other conditions. This chapter, together with Chapter 2, attempts to replicate their analysis using English patient level data for a wider range of conditions. The chapter is also able to address some of the limitations acknowledged by the authors, due to gaps in their data on patient co-morbidity, which can be used to create even more robust indicators. Our results indicate that this method can be applied to other countries with similar data, and when controlling for co-morbidity are able to produce indicators with high prediction accuracy. However, in our application of this method to a different setting we are also able to identify other difficulties, which arise to do a smaller sample of hospitals available in the English data as compared to the US data.

The first step of the methodology, creating latent measures of performance for each of the outcomes of interest, is presented in Chapter 2. These latent measures serve essentially as risk adjusted measures of performance, as they are able to control for exogenous patient characteristics such as age, gender, deprivation and co-morbidity. They proved to be useful for detecting trends and comparing hospital performance to their peers. When analysed more closely, to see what factors influenced performance, the results indicated that many of the indicators are dynamic, and also related to one another. This chapter replicates the second step of the methodology which uses a VAR framework that is able to incorporate the time series information, as well as the relationship to other the other outcome variables into new performance indicators. Both the VAR models, and the indicators inform us on the performance of hospitals.

The results of the VAR models indicate which dimensions of hospital performance are persistent across different conditions, indicate how much they vary across hospitals and over time, and provide insight as to their relationship with each other. The results for all conditions suggest that of the four measures included in the model, year-long mortality is the most persistent dimension of performance. While the coefficient of year-long mortality on its own lag is only around 0.2 for CCF, it is over 0.63 for all other conditions suggesting a strong dynamic presence for these conditions. For most conditions, this indicator also exhibits a high standard deviation across hospitals, ranging from 20% to 5%, and is over 10% for all conditions except Hip Replacement. The high variation associated with year-long survival most likely stems from a variety of factors outside the provider's influence, such as patient behaviour and lifestyle. Although the extent of this influence will vary by condition.

The persistence of the 30-day mortality indicator varies considerably more by condi-

[2]It has been applied to evaluate educational outcomes in the USA, for more information see Kane et al. (2002).

tion. The coefficient on its own lag is high for IHD and CCF and around 0.4 for Stroke. In most cases it is quite low, such as for the conditions of AMI, Hip Replacement, MI, and TIA. The variation of indicator across hospitals also varies considerably by condition. The standard deviation is around 10% for MI and CCF, around 6% for AMI and Stroke, and around 1% for Hip Replacement, IHD and TIA.

Unlike our results, the McClellan and Staiger (1999) paper finds that 30-day mortality is more persistent than year-long mortality for AMI, and that shorter term mortality is more persistent than year-long mortality for IHD. This difference could be explained by variations in the UK and US treatment pathways. It could also be linked to the different samples being analysed by the different investigations; their analysis focused only on the elderly while ours examined all patients. It may also be related to the fact that we were able to adjust for patient co-morbidity which they did not have the data to do.

Similarly the mortality models in Chapter 2 only identified a significant dynamic relationship between IHD and Hip Replacement for year-long mortality, and a significant dynamic relationship for AMI, IHD, Stroke and Hip Replacement for 30-day mortality. Given the performance of the filtered estimates on the different goodness of fit measures this could be related to the noise in the latent estimates which obscure the 'true' quality effect. It could also reflect the number of restrictions set in the GMM model, which the VAR model does not apply.

Moreover, we mention in the results section that for this analysis the VAR(1) specification was chosen for ease of interpretation and parsimony. However, different specifications were indicated as marginally better fits for the model by the Aikake and Swartz lag tests. Yet, when tested with alternative specifications the results did not differ substantially. Moreover, the R-squared estimates calculated for a VAR(2) specification, as reported in the results section, indicate similar results for all conditions, and in many cases do not indicate improved fit. However, investigation for each condition could benefit from the inclusion of more lags to create more robust predictions and forecasts, especially if there is a longer time-series being analysed.

The readmission indicators are by and large less persistent indicators of quality as compared to mortality. The coefficients on the lags of 28-day emergency readmissions range between 0.4 and 0.5 for AMI, IHD and Hip Replacement, while year-long readmissions are only persistent for IHD with a coefficient around 0.5. The variation in short and long term readmissions varies more considerably by condition. The standard deviation on both indicators is around 5% for AMI, Stroke and Hip Replacement but rises to around 10% for MI and CCF. Both TIA and IHD indicate little variation between hospitals in 28-day readmissions, with standard deviations around 2-3%, but high variation in year-long

readmissions, with standard deviations around 7-8%.

The AMI, Stroke and CCF VAR models indicate a strong positive correlation between 28-day readmissions and year-long readmissions, and weaker positive correlation between 30-day mortality and year-long mortality. These associations are expected as they all represent worse outcomes. However, the model also shows a negative association between mortality and readmissions present for some conditions and different time combinations, but strongest between year-long mortality and year-long readmissions. McClellan and Staiger (1999) also observe this result, for AMI, although for 30-day mortality and year-long readmissions. They note that while a positive correlation might be expected, as higher values for both indicators represent worse outcomes, the negative correlation may reflect the relatively poor heart function of 'marginal' patients who survive when treated in high quality hospitals. Thus, the hospitals which have worse mortality measures will perform better on the readmission measures, as fewer severely ill patients survive to be readmitted. Moreover if healthier patients led to low mortality rates, than complication rates for that hospital would also be lower, thus there are quality differences amongst hospitals which are not linked to patient selection.

The Hip Replacement, MI, IHD and TIA models suggest mixed association between the readmission and mortality variables; indicating a positive correlation between some of the mortality and readmission combinations and negative correlations between the others. For example, in Hip Replacement and MI 30-day mortality is negatively associated with both short and long term readmissions but year-long mortality has a positive association with both. In most cases however, all associations are weak. For no condition were all associations positive however, indicating that one should be cautious when interpreting readmission measures in isolation as they may not be indicative of higher quality. The results of the VAR models also report the correlation of the residuals for the different indicators. In all models short term and long term mortality are positively correlated with one another, although in most cases this is very weak, with Stroke having the highest association at 0.55. Short and long term readmissions have strong positive correlations with each other in the AMI, Stroke and Hip Replacement models, but very weak associations in the other conditions.

The signal variances estimated using the VAR parameters were also used together with the estimation error to construct signal to noise ratios for each outcome measure in each condition for the year 2005. The first striking result is how strong the signal is for the indicators in most conditions, for a sufficient sample of patients. TIA and CCF have the weakest signal to noise ratio, yet, for both conditions this is related to having relatively fewer cases admitted annually as compared to the other conditions, resulting in more

estimation error. Indeed, in all the conditions the signal to noise ratios are considerably worse for the hospitals with fewer cases. While the number of cases required to get a good signal to noise ratio varies by condition, in most cases it includes the medium to large volume hospitals. McClellan and Staiger (1999) also observe this finding in their paper, and note that it is generally harder to observe the true performance of smaller hospitals from patient outcome data. This is because the variation in the data will be more strongly influenced by differences in treatment, such as the presence or absence of an individual physician, which would have relatively smaller effects in a larger hospital. Moreover, If we consider the average number of cases per hospital (Table 3.1) together with the number of cases above which the signal to noise ratio because high enough, we see that only for hospitals of average size and above do the patient outcome measures for a single year provide relatively good information on performance.

The other striking result from the signal to noise ratios was that in all cases, except CCF, long-term mortality had the strongest signal. This suggests that for these conditions, the long term measure of mortality is a more useful measure of quality than the short term measure. Similarly, for most conditions year-long readmissions had a stronger signal than 28-day readmissions, aside from Stroke and Hip Replacement. Indeed 28-day readmissions in almost all cases tended to be the worst performing measure. For cases such as AMI, where treatment variations in the short term have high implications for survival, one would expect the short term mortality measure to have a stronger signal. Especially as long term outcomes add more noise. This finding was reported by McClellan and Staiger (1999) in the US analysis. It is interesting that this is not the case in the UK scenarios, and raises interesting questions as to why.

One possibility for the noise found in the short term estimates, may linked to the organization of the health system and different health policies within in the UK. In the NHS data collection and reporting has not traditionally been attached to financing as it is in a claims type system such as that of the US, this may lead more error in estimates if less effort is put into coding. On the other hand, since 2000 many health policies have focused on using measures such as 30-day in-hospital mortality and 28-day emergency readmissions to measure and reward the performance of hospitals, such as the star ratings. There has been criticism surrounding these policies and the distortionary results they had on indicators, such as manipulation of data collection (Bevan and Hamblin, 2009). In addition, the introduction of payment by results (2004/5) has now linked coding to hospital payments changing the importance of good coding. As a result, discrepancies in coding practices have been reported in the literature, such as hospitals coding deaths as palliative care in order to reduce mortality rates (Hawkes, 2010b). Thus, it is plausible that the

emphasis put on the short term indicators for policy has created more measurement error in their collection, making the longer term measures perform better despite the additional noise in them from other exogenous factors such as patient behaviours and/or lifestyles.

The McClellan and Staiger (1999) analysis replicated the VAR models for different samples of hospitals in order to better understand the differences in estimation parameters between them. We were unable to do this as the number of hospitals in our sample across each of our conditions were considerably less, at around 100 per condition as opposed to their sample of approximately 4,000.

While the results of the VAR models prove informative in themselves, they can also be used to create 'filtered measures' of each of the four indicators. These filtered estimates are able to encompass the time-series relationships within indicators, as well as the correlations between measures, allowing them to portray a more accurate description of overall performance. The results section presents these filtered measures together with the latent measures in a series of diagrams for each outcome, for each condition. These figures have three main similarities throughout all conditions. The first is that the filtered indicators are able to provide smoother estimates over time as compared to the latent measures which exhibit considerable year-to-year variation. The second is the wider confidence intervals of the filtered measures, which are about double the size of the latent measure confidence intervals (Chapter 3). In their analysis, McClellan and Staiger (1999) note that the confidence intervals for their filtered estimates are much tighter than those of the latent measures. We attribute this different finding to the smaller sample of hospitals we used to estimate the filtered estimates, resulting in higher uncertainty surrounding the estimates[3]. However, many critiques of the VAR methodology note that the standard errors of the variance decompositions are large that it is difficult to make inferences about them (Sims, 1980). In this instance as well, the wider confidence intervals make it much harder to draw conclusive interpretations from the estimates about relative hospital performance.

Finally, the third similarity across conditions in the performance of the estimates for the small hospitals. As noted in Chapter 2, the small hospitals exhibit more year-to-year fluctuation in the latent estimates. While the filtered estimates smooth out this performance, and have wide confidence intervals, the latent measure will often lie outside these bounds. This reflects observations noted earlier, about predicting performance for small hospitals, which the raw measures are very sensitive to differences in treatment.

An evaluation of the filtered estimates in prediction the variation of true hospital effects is estimated through R-squared estimates, based on the adapted formula in McClellan and Staiger (1999). The R-square estimates for all filtered measures, in all conditions,

[3]Their sample consisted of 3945 hospitals while we had data on around 120 hospitals per condition.

are very high, suggesting that the filtered estimates are able to predict true performance remarkably well. These high estimates are in line with the very high signal to noise ratios of the original data, discussed previously. Moreover, the R-squared measures also indicate that the model is also able to predict very accurately using different amounts of data, including that of only one year. The R-squared values presented in this chapter are much higher than the ones reported byMcClellan and Staiger (1999), especially when using a limited set of data to create predictions. This differs from the McClellan and Staiger results, where the R-squared estimates decline when a smaller sample is used to construct the indicators. This is most probably related to differences in the underlying data. For instance, unlike them, we had information on patient co-morbidity which allowed us to better adjust for case-mix. Also while their sample only considered the elderly we looked at the entire patient population.

As discussed previously, the VAR structure allows the model to forecast outcomes for future years. By using the data to estimate performance the final years of our sample, and compare these data to the true estimates we are able to assess how well the model forecasts data. The R-squared results using this formula (equation (3.12)) were also very high for all conditions, indicating the VAR's ability to forecast outcomes. While these estimates are again higher than McClellan and Staiger's, they also note the model's ability to forecast extremely well. The results are also presented for a VAR(2) specification of the model, and are almost identical to the VAR(1) results. This indicates that the forecast performance is not sensitive to the lag choice specified in the VAR model.

The last section of this Chapter considers how hospitals perform when ranked by the three different measures (raw, latent and filtered). The results are quite striking. Depending on the measure chosen, hospitals may go from the top of a ranking to the bottom, or the opposite. The hospitals with the fewest cases are most influenced by the type of measure as there is so more variance in the raw and latent estimates. The filtered measures are better at smoothing out the jumps from year to year as they combine all the information from the time-series and across the other variables. Thus in these cases, the filtered measure will be a better indication of performance in any one year. The latent estimates, while risk adjusted are very erratic from year-to-year, and rankings may change suddenly when looking at year snapshots. Raw measures do not control for exogenous characteristics that influence outcomes, and so are the worst measure of the three. While the filtered estimates are much better at providing a much more consistent picture of performance over time, we do not advocate the ranking of hospitals, as this exercise shows how sensitive rankings are to the method chosen.

Much of the analysis of this chapter focuses on identifying which indicators are more

useful for comparing performance across hospitals. The VAR models indicate which measures are more persistent for the different conditions, how much they vary across hospitals, how well they capture the true signal in the data and how they are correlated with the other measures being considered. The results overall suggest exercising caution when interpreting any indicator alone as it may be misleading given its relationship with the other outcome measures. However, the mortality indicators capture more of the true signal than the readmission measures for most conditions, and especially long-term mortality making it a better indicator to look at.

In conclusion, the analysis of the VAR models for the seven conditions chosen indicate considerable correlation of the outcomes across time and between measures. The degree of persistence varies by measure and across conditions, as does the extent to which measures vary across hospitals. However, in almost all cases the most persistent measure with the strongest signal was year-long mortality. Some of the other more generalizable findings are that predictions are weaker for hospitals with fewer cases, and variation in their outcomes from year to year is larger. However, measures overall are very good at identifying the true signal of good performance in different hospitals. Indeed the R-squared values indicate that the measures are extremely good predictors and forecasters of performance.

4 Examining the persistence of hospital quality across conditions

4.1 Introduction

Chapters 1–3 have reviewed many of the challenges in measuring quality of care using outcome measures. One of the major challenges they note is identifying suitable metric to capture the multidimensionality of quality. Increasingly most quality assessment exercises at the hospital level use a combination of different types of indicators, recognizing that the measurement hospital performance over time or across institutions is challenging, due to the diversity of services they provide and multiple factors which influence their performance, such as technological innovation and personal skill. Yet multidimensionality spans not only from the different structures, processes and outcomes associated with quality, but also in terms of the mix of conditions and patients that care is provided to. In dealing with the first of these issues, Chapters 2 and 3 consider how we can create stronger indicators of quality by combining the information known about patients and also about different outcome measures. These chapters dealt with the second issue by selecting particular conditions that are linked to known processes of care associated with good quality care, such as AMI.

While many studies take this approach, using one or more conditions as proxies of quality for whole institutions (Bloom et al., 2010; Dimick et al., 2004; Propper et al., 2004, 2008), another way to overcome this problem is to measure case-adjusted outcomes across all patients treated in hospitals as is done with HSMRs (Jarman et al., 1999). Recently, problems with the aggregated approach have been highlighted by Shahian et al. (2010) who note that these measures are highly dependent on the case-adjustment technique used, such that it is difficult to use them to draw any meaningful results about quality. Using individual conditions as proxies, takes an extreme approach to case-adjustment by examining only those conditions that are believed to have a strong relationship between quality and outcome. However, when using this approach it is not clear what the results suggest about overall quality in the institution; that is to say there is ambiguity about

just how well the selected conditions perform as proxies.

Commonly used outcome indicators are mortality and readmission rates of specific types of patients. As discussed in Chapter 1, 30-day mortality rates from AMI and Stroke have been used as measures of quality across OECD countries, while other conditions such CABG surgery, repair of Abdominal Aortic Aneurysm, Pancreatic Resection, Esophageal Resection, Pediatric Heart Surgery, Craniotomy and Hip Replacement, have been recommended by the Agency for Healthcare Research and Quality in the US (Dimick et al., 2004). Studies from the US (Kessler and McClellan, 1996; McClellan and Staiger, 1999) and England (Bloom et al., 2010; Propper et al., 2004, 2008) have used risk adjusted AMI 30-day mortality as proxies of quality of care. Indeed, since the 1980s, there has been considerable work done on 'avoidable mortality', that is identifying conditions where death is avoidable according to current medical knowledge, practice and public health interventions in a defined age/sex group of the population (Holland, 1988; Nolte et al., 2004). This chapter attempts to understand to what extent mortality and readmission rates of different conditions for same hospital are related, indicating how persistent the quality of health care providers is across treatments. Little work has been done to study these relationships, and existing studies indicate only modest correlations between the mortality of different conditions (Dimick et al., 2006). This will allow us to better understand if good outcome indicators are only reflective of providers doing certain procedures well or if they are consistent across different treatments and thus the result of some wider common factors in the hospital environment.

Using the VAR model, we examine whether there is an empirical relationship between the latent outcome measures, or risk adjusted outcome measures, from seven conditions (including AMI) within the same hospital. That is, if there are features of quality that are present across conditions within certain hospitals or if they are more treatment specific. Understanding the linkages of different performance indicators across conditions, will allow us to determine how generalizable different measures are about the overall performance of hospitals. After reviewing the methodology in more detail, the chapter will present the data being investigated. We will discuss which seven conditions have been selected, as well as the construction of the latent outcome measures from Chapter 2. Finally, the results of the models are presented in the results section, and the findings and their implications for performance measurement and policy are examined in the discussion section.

4.2 Methodology

The methodology of this chapter uses the VAR model to better understand the co-integrating relationship of the variables outlined above. The VAR analysis was chosen

as it is able to model two way relationships and track the dynamics of these relationships through time, giving it a strong advantage over other models in terms of understanding co-integrated relationships. The first model, Model 1, attempts to understand the relationship of the same outcome measures across different conditions in order to determine whether there are certain hospital characteristics which make them perform better overall. Model 2, analysed separately for each condition, attempts to use the VAR framework to better understand the nature of the co-integrating relationships amongst different performance indicators.

Model 1 assumes that a condition's outcome measure in a given year for a given hospital will depend on its quality measure in the past years plus a contemporaneous shock that might be correlated across the quality indicators for other conditions. The first model is interested in relating the different indicators of outcome across conditions, such that:

$$Q^k_{c1ht} = \alpha_c + A_{11}(L)Q^k_{c1ht-1} + A_{12}(L)Q^k_{c2ht-1} + ... + A_{1n}(L)Q^k_{cnht-1} + \epsilon_t \tag{4.1}$$

Where Q^k the denotes the outcome measure being used, c denotes the condition, h identifies the hospital and t the year. α_c denotes 1×7 vectors containing the constant terms, while A indicates the matrices of the coefficients to be estimated, and (L) their lag specification. ϵ_t denotes the vector of innovations that may be contemporaneously correlated with each other but are uncorrelated with their own lagged values and all the right-hand side variables. Each equation was estimated using lag lengths of $1-4$ years. The Akaike information criterion (AIC) indicated (SC) 3 lags were optimal, and so the VAR(3) specification was chosen. The results reported will be for this specification. Recall that only one lag was used in the VAR model used in Chapter 3 for ease of interpretation and comparability across the conditions. As we are using the model for a different purpose, and do not need to compare the different models to one another, we experiment with the 3 lag specification which is optimally indicated, albeit marginally.

Unrestricted VAR models often suffer from over-parameterization (Enders, 2004; Gujarati, 2003). In order to avoid this problem and still include the optimal lag length to capture the dynamic effect of performance, we first estimate a reduced form version of the model and use the Granger causality/block exogeneity tests and variance decomposition estimates to determine which variables are exogenous to the model. The information about the relationship between variables from the VAR(3) specification is used to modify the models and adjust the lag lengths included for the different conditions. In the final version of the model, we include three lags of the outcome for the dependent variable condition and the contemporaneous value and 1 lag of all other conditions indicators. As the right-had variables were no longer identical, we could not use OLS to estimated the equations and instead used SUR estimation. From the resulting coefficients of the SUR

model, we were able to estimate the effect performance in one condition had on another. Estimation with the SUR model will produce more efficient coefficients than estimation with OLS, especially when the disturbances are highly correlated, and the independent variables are not highly correlated.

4.3 Data and Key Variables

The risk adjusted outcome measures being used for this investigation have been constructed in Chapter 2 using a latent variable approach. The data from which these measures are constructed uses 4 different outcome measures collected at the individual level 30-day within hospital mortality rates, year-long overall mortality rates, 28-day emergency readmission and year-long readmission rates. Using patient level regressions, a dummy variable for the hospital in which every patient was treated, and controlling for patient characteristics, we are able to estimate the unobserved effect hospitals are having on the different outcomes. Thus, the latent measures estimated are essentially risk-adjusted outcome measures. For more detail on the construction of these variables see the methodology section of Chapter 2.

Chapter 2 and 3 examine the relationships between these risk-adjusted outcome measures separately for each condition. Some of the key characteristics that have emerged from these analyses are that different performance measures have different levels of persistence and exhibit different variation amongst hospitals. However, year-long mortality is almost always the most persistent variable across conditions. The associations amongst the different outcome measures also vary by condition. For the most part the mortality variables and readmission variables are positively correlated amongst themselves, although in most cases not very strongly. However, mortality and readmission variables tend to be negatively correlated with one another, indicating that higher readmission variables may not always be indicative of worse quality. Filtered outcome estimates are constructed in Chapter 3 also using a VAR framework. They are able to incorporate information provided throughout the time-series and from the other outcome measures used for each condition. As compared to the risk-adjusted (latent) measures they are smoother over time, but with greater confidence intervals. For more discussion of the two types of indicators see the results section of Chapter 3.

In order to run the VAR model in this chapter the risk-adjusted measures are collected into a new data set at the hospital level. Each hospital is distinguished by a unique identifier, and the sample is reduced for all conditions to the years 2000-2008 where data is available for all seven conditions. A description of this data sample are presented in Table 4.1. The same seven conditions examined in this chapter as in Chapters 2–3, as

from the disease and treatment codes indicated in the table.

Table 4.1: Descriptive statistics for the sample used in the cross-condition VAR.

Conditions	ICD-10/ OPCS 4.3 codes	Years Analysed	Number of Hospitals
AMI, MI, IHD, CCF, Stroke, TIA, Hip	ICD-10: I20, I21, I22, I23, I25, I11.0, I13.0, I25.5, I50.0, I50.1, I50.9, J81X, I60-I67, G45.0-G45.4, G45.8-G45.9, G46.0-G46.8 OPCS4.3: W37-W39 W46-W48 W58	2000-2008	130

The justification of the selection of the above conditions is explained in detail in the Data section in Chapter 1. AMI, MI, Stroke and TIA are extremely urgent health problems. Thus, patients suffering from these conditions are likely to go to nearby facilities for care, limiting the amount of selection bias that can occur. For this reason outcome of these type of indicator are often used in assessments. Patients with IHD and CCF are also readily hospitalized, usually when it becomes very severe or is in its acute form. Hip Replacement can be admitted as an elective or emergency treatment. Elective Hip Arthroplasty are extremely common and extremely successful, however as the treatment is mostly performed amongst the elderly population, where underlying medical conditions are likely to be present, complications occasionally arise during or –more commonly- after treatment. Acute Hip Replacements carry a much higher morbidity and mortality risk, partly due to the lack of preoperative preparedness of the patient but also partly because people who will undergo urgent surgery might not have been deemed as appropriate surgical candidates in an elective setting. Our risk-adjusted quality measures take into account the admission of these patients, however there is likely to be more selection bias amongst the elective Hip Replacement patients admitted.

Of the seven conditions selected, we can classify them in to similar groups: AMI, MI, IHD and CCF are all heart conditions, while Stroke and TIA are neurological Hip Replacement is an orthopaedic condition. We expect to see a stronger relationship between the risk adjusted outcome measures within the same underlying group, as they are more likely to have common factors influencing quality. The relationships between the conditions in different groups will be informative as to how generalizable risk adjusted outcome measures are for quality in other areas.

4.4 Results

The methodology employs a reduced form VAR framework to estimate the impact that exogenous factors have had on the quality indicators. Granger causality tests are used to evaluate whether the lags of any of the variables Granger cause any other variables in the VAR system. As normal pairwise Granger Causality tests do not yield reliable results in a multivariate VAR (Dufour and Renault, 1998; Lütkepohl, 2006), Granger Causality/block-exogeneity tests were used, under the null hypotheses that they do not affect any of the other variables in the system. As the causal relationship between a pair of variables in a multivariate system will be largely influenced by their relationship to other variables in the system, it is easier to think of the system in terms of an entire causal ordering, rather than as piecemeal elements of the relation. Close examination of the Granger causalities presented for each of the four outcome indicators allow us to examine how the different variables are associated with each other. These tables present the p-values from χ^2 (Wald) statistics for the joint significance of each of the other lagged endogenous variables in the equation being considered. The statistic in the final row, labelled 'All', indicate the p-values from χ^2 (Wald) statistic for the joint significance of all variables in the equation. If the null hypothesis is rejected, then the direction of causal relations for each pair of variables is determined. The Granger tests tell us nothing of the polarity or magnitude of these relationships– but together with the variance decomposition estimates we can acquire a more complete understanding of the relationships amongst the variables.

Forecast error variance decompositions can be used to ascertain the importance of the interactions between the variables in the VAR system, at it determines how much of the forecast error variance of each variable can be explained by exogenous shocks to other variables within a specific time horizon. The variance decompositions were obtained using a Choleski decomposition. As the order of the variables is likely to influence the results these were estimated using many different orderings. However, given the low correlation between the errors, the change in ordering is unlikely to make a big difference. We estimated the variance decompositions using different orderings and indeed found this to be the case. In addition the variation observed from the different orderings is reduced at longer forecasting horizons (Enders, 2004). The set of variance decompositions reported in this section are estimates at 10 lags, where there was little variation between the different orderings of the conditions.

Using the results from the Granger Causality/block-exogeneity tests and the forecast error variance decompositions, we are able to construct and run a better specified model for each of the conditions, that considers the dynamic variables where indicated. As the right hand side variables for each equation differ, we estimated this system using a SUR. The

SUR technique estimates each regression by Ordinary Least Squares (OLS) and uses the residuals to estimate the error variances both for each equality and across equations. The errors are then transformed so that they all have the same variance and are uncorrelated. The other variables then also undergo the transformation, and OLS estimation is applied to the transformed variables (Martin and Smith, 2005). The coefficients from the SUR model help us to identify the relationships between the outcome measures of the different indicators. Finally we also conduced the Spearman's rank correlation coefficients of the filtered outcome indicators. The results are all presented separately by outcome indicator.

Latent 30-Day Mortality Estimates

The results of the Granger Causality/block-exogeneity tests for risk-adjusted 30-day mortality, presented in Table 4.2, indicate whether the lags of the excluded variable affect the endogenous variable. The null hypothesis is that the lagged coefficients are significantly different than zero, while the 'All' column is a joint test to see if the lags of all other variables affect the endogenous variable. Reading off the columns of the table, we can observe some evidence of endogeneity for IHD, CCF at the 1% level and Stroke at the 10% significance level. For IHD we see that lagged values of AMI and TIA mortality have a significant effect on IHD mortality. Lagged values of AMI and Stroke mortality have a significant effect on CCF mortality, while lagged values of IHD and TIA mortality have significant effects on Stroke mortality.

Table 4.2: Pairwise Granger Causality Test/Block Exogeneity Wald Tests for $D30_{ht}$ VAR(3) specification.

30-Day Mortality VAR(3)							
Excluded Variable	Dependent Variable						
	AMI	MI	IHD	CCF	Stroke	TIA	Hip
AMI	-	0.2290	0.0018	0.0006	0.7363	0.2070	0.1059
MI	0.6145	-	0.8446	0.5487	0.5909	0.3359	0.3482
IHD	0.6819	0.1281	-	0.2534	0.0911	0.7110	0.2192
CCF	0.3981	0.4945	0.9034	-	0.1105	0.4034	0.9869
Stroke	0.3398	0.5988	0.1675	0.0251	-	0.8450	0.2573
TIA	0.9213	0.8012	0.0541	0.2915	0.0296	-	0.6568
Hip	0.5942	0.8607	0.2069	0.7999	0.8613	0.9764	-
All	0.8112	0.4808	0.0086	0.0010	0.0764	0.8020	0.3623

The forecast error variance decompositions presented in Table 4.3 indicate the percentage of the variance of the error made in forecasting a variable due to a specific shock at a

given horizon, and are informative about the interaction of the variables. Reading off the rows of the table it is immediately apparent that most of the variance of each mortality measure is explained by its own lags, this ranges from 98% for AMI to 94% for TIA. No other condition accounts for over 2% of the variance of 30-day mortality in any of the seven conditions being investigated.

Table 4.3: Variance Decomposition percentages for $D30_{ht}$ using the VAR(3) specification.

% of FEV in	Standard errors	Typical Shock in						
		AMI	MI	IHD	CCF	Stroke	TIA	Hip
AMI	0.036036	98.22361	0.069142	0.306507	0.773694	0.267182	0.230910	0.128951
MI	0.117229	0.152837	97.63818	1.546354	0.241875	0.144968	0.195015	0.080769
IHD	0.016468	0.705988	0.510289	95.84055	0.483150	1.799155	0.210501	0.450366
CCF	0.130813	1.486912	0.483108	0.720106	94.88037	1.914901	0.168820	0.345779
Stroke	0.033281	0.105104	2.133255	1.189376	1.141058	94.85801	0.315946	0.257246
TIA	0.012641	1.598555	1.333829	1.948989	0.615825	0.197573	94.29646	0.008770
Hip	0.008657	0.328762	1.334577	1.047900	0.025746	0.405228	0.165104	96.69268

Forecast Horizon at 10 years

Note: The Choleski ordering for this table was AMI, MI, IHD, CCF, Stroke, TIA, Hip

Table 4.4 presents the results from the SUR estimated for each condition and includes 3-lags of the dependent variable, and 1 lag of the risk-adjusted mortality of the other conditions. 3-lags are chosen as they were specified as optimal by the AIC in the previous model. While the variance decomposition percentages of the other conditions, at a ten-year forecast horizon, were low, 1-lag was included to draw some conclusions about the dynamic quality effects within hospitals. The R-squared estimates show that for all conditions except Hip Replacement, the model is successful in capturing about $50 - 60\%$ of the variance in risk-adjusted mortality rates, and that this is largely explained by their own past outcomes.

For all conditions, the three lags of itself are significant, indicating that outcomes are very dynamic. For all conditions, except Hip Replacement, the sign of the coefficient for the first lag of the mortality indicator for all conditions is positive, suggesting that an increase in risk-adjusted mortality in period $t-1$ will cause an increase in risk-adjusted mortality in period t. The size of this effect varies by condition, indicating a very high dynamic effect for MI at 0.7, followed by 0.5 for most of the other conditions. Hip Replacement exhibits an extremely low negative effect. While the second lag is still significant at 1% for all conditions, the coefficient is smaller, indicating that there is less of an effect on

the dependent variable. This is not true for Hip Replacement, where the coefficient has increased and become positive. The results for the third lag are still significant 5% or above for all conditions, although the coefficients are very small, and negative for Hip Replacement.

All of the risk-adjusted mortality rates are influenced by the risk-adjusted mortality of at least one of the other conditions. Contemporaneous and lagged risk adjusted AMI mortality is significant in influencing IHD, CCF and Stroke, and contemporaneous mortality alone significantly affects Hip Replacement. The contemporaneous effect is positive for all these conditions apart from Stroke, indicating that high AMI mortality in one year would be associated high mortality in these conditions, while the opposite would occur with Stroke. However the effect is small for all conditions apart from CCF. Where significant, lagged AMI mortality had the opposite sign to the contemporaneous effect, such that higher AMI mortality in year $t-1$ is associated with lower CCF mortality for year t.

Table 4.4: Seemingly Unrelated Regression for risk adjusted $D30_{ht}$ estimates.

	AMI	**MI**	**IHD**	**CCF**	**Stroke**	**TIA**	**HIP**
Explanatory Variables	**Dependent Variables**						
DV L.2	0.187352***	0.097468***	0.188667***	0.179875***	0.330422***	0.151277***	0.17563***
	(0.030487)	(0.035901)	(0.031609)	(0.033126)	(0.032770)	(0.032052)	(0.039135)
DV L.3	0.052758**	0.050641***	0.100506***	0.132121***	0.065510**	0.068117**	-0.11076***
	(0.027356)	(0.031205)	(0.029447)	(0.030825)	(0.030264)	(0.029580)	(0.039361)
AMI	-	-0.087226	0.042875***	0.776358***	-0.060751**	0.009030	0.161110***
		(0.103193)	(0.016592)	(0.126146)	(0.031501)	(0.013853)	(0.039311)
AMI L.1	0.598694***	0.161837	-0.05151***	-0.79946***	0.052723*	-0.031333	0.016209
	(0.033098)	(0.104108)	(0.016732)	(0.127075)	(0.031842)	(0.013962)	(0.013711)
MI	-0.010628	-	0.02495***	-0.009493	0.035748***	0.03989***	-0.008035
	(0.014527)		(0.006285)	(0.048857)	(0.011939)	(0.005143)	(0.013789
MI L.1	0.003800	0.727239***	-0.02164***	0.028105	-0.026248**	-0.03570***	0.014491***
	(0.015095)	(0.033746)	(0.006533)	(0.050687)	(0.012412)	(0.005364)	(0.005245)
IHD	0.169505**	0.94626***	-	1.136166***	0.062988	-0.19640***	-0.005887
	(0.087293	(0.233088)		(0.289356)	(0.071711)	(0.031133)	(0.005443)
IHD L.1	-0.111375	-1.06570***	0.574897***	-0.86949***	-0.025188	0.15864***	-0.046022
	(0.087830)	(0.233463)	(0.032420)	(0.291326)	(0.072025)	(0.031432)	(0.031588)
CCF	0.064823***	0.009531	0.019433***	-	-0.02687***	-0.01549***	0.036073
	(0.011145)	(0.030310)	(0.004868)		(0.009319)	(0.004048)	(0.002547)

	AMI	MI	IHD	CCF	Stroke	TIA	HIP
CCF L.1	-0.05880***	-0.025352	-0.011640**	0.583471***	0.03299***	0.012896***	0.004365*
	(0.011453)	(0.031079)	(0.005002)	(0.033777)	(0.009435)	(0.004160)	(0.003230)
Stroke	-0.102036**	0.287342**	0.016743	-0.330910**	-	-0.006029	-0.002102
	(0.043764)	(0.118373)	(0.019189)	(0.147598)		(0.015999)	(0.015885)
Stroke L.1	0.090545**	-0.133387	-0.007538	0.343633**	0.507251***	0.007762	0.020552
	(0.044021)	(0.119247)	(0.019240)	(0.148889)	(0.034267)	(0.016123)	(0.015939)
TIA	0.010064	2.170857***	-0.30043***	-1.40150***	0.035551	-	-0.033395**
	(0.106381)	(0.280108)	(0.045451)	(0.354007)	(0.087414)		(0.038220)
TIA L.1	-0.064007	-1.40431***	0.189141***	1.08720***	-0.100899	0.570750***	-0.018239
	(0.104526)	(0.278669)	(0.044985)	(0.348211)	(0.085746)	(0.032231)	(0.037502
Hip	0.128751	0.789460***	-0.061220	0.390161	0.089048	-0.013219	-
	(0.107485)	(0.289552)	(0.046937)	(0.360638)	(0.088640)	(0.039021)	
Hip L.1	-0.131166	-0.206313	0.066904	-0.177648	-0.005013	0.019547	-0.01243***
	(0.107259)	(0.290172)	(0.046777)	(0.361123)	(0.088847)	(0.039033)	(0.004214)
R^2	0.533735	0.629012	0.533405	0.557296	0.663021	0.488114	0.071582
N	636	636	636	636	636	636	636

* Significant at p $\leq$ 0.1

** Significant at p $\leq$ 0.05

*** Significant at p $\leq$ 0.01

When looking at the results for MI, we see that the risk-adjusted mortality for contemporaneous and lagged MI is associated with IHD, CCF and TIA. The coefficients are very small for all these variables. In all cases the sign is positive for the contemporaneous effect and negative for the lagged effect. The results for the IHD model suggest that contemporaneous and lagged IHD risk adjusted mortality is significant in influencing MI, CCF and TIA. The contemporaneous effect is positive for all variables apart from TIA, and the sign is reversed for the lagged effect. The size of the coefficients is largest for the relationship between IHD and CCF. Contemporaneous risk adjusted IHD has a significantly effect on AMI, such than an increase in IHD mortality is associated with an increase in AMI mortality.

Contemporaneous and Risk-adjusted CCF is significantly associated with AMI, IHD, Stroke and TIA. Similar to the other conditions, the sign is reversed when looking at the contemporaneous and lagged effects. The effect of contemporaneous CCF on AMI and IHD is positive, while the effect of lagged CCF on AMI and IHD is negative. CCF has the opposite dynamic relationship with Stroke and TIA. CCF only has no significant effect on MI. The size of all these associations are very small.

Contemporaneous and lagged risk-adjusted mortality for Stroke significantly influences AMI and CCF, such that they are negatively associated with the former and positively with the latter. Contemporaneous Stroke mortality negatively influences MI but lagged mortality does not. The coefficients are small on all significant variables. The coefficients indicate that contemporaneous and lagged risk-adjusted TIA is significantly associated with MI, IHD and CCF. The sign on the contemporaneous effect is positive for MI and negative for the other indicated conditions, including Hip Replacement. Similar to the other dynamic relationships discussed above, the sign is reversed for the lagged effect. The size of the coefficient is large for MI and CCF, indicating a strong association amongst these conditions. Finally, risk-adjusted mortality for Hip Replacement is only significant in influencing IHD, and only contemporaneously. The magnitude of the effect is relatively large.

Latent 365-Day Mortality Estimates

Table 4.5 indicates the Granger causality/block-exogeneity tests for the year-long risk-adjusted mortality estimates. From the 'All' column, we observe some evidence of endogeneity for IHD, TIA and Hip Replacement. This is significant at 1% for IHD and TIA, but at 10% for Hip Replacement. Looking at the individual variables, year-long risk adjusted IHD mortality is significantly affected by lagged values of risk adjusted year-long AMI and Stroke mortality with 5% and 1% significance. TIA is significantly affected by lagged values of CCF and Stroke at 1% significance and Hip Replacement with 10% significance. Hip Replacement is effected by lagged values of risk adjusted year-long CCF and TIA mortality at 5% significance.

Table 4.5: Pairwise Granger Causality Test/Block Exogeneity Wald Tests for $D365_{ht}$ VAR(3) specification.

365-Day Mortality VAR(3)							
Excluded Variable	Dependent Variable						
	AMI	MI	IHD	CCF	Stroke	TIA	Hip
AMI	-	0.0514	0.0349	0.0373	0.2070	0.1373	0.1975
MI	0.6638	-	0.8967	0.1956	0.3079	0.1642	0.4893
IHD	0.4008	0.8255	-	0.3778	0.6662	0.3690	0.7697
CCF	0.6007	0.9504	0.1407	-	0.2100	0.0048	0.0390
Stroke	0.4543	0.6799	0.0009	0.8869	-	0.0004	0.9122
TIA	0.6884	0.7996	0.1788	0.2624	0.9507	-	0.0244
Hip	0.0216	0.0316	0.8101	0.6398	0.9555	0.0606	-
All	0.3423	0.3226	0.0072	0.2788	0.7008	0.0001	0.0978

The variance decomposition percentages for the year-long risk adjusted mortality rates in Table 4.6 indicate that most of the variance of the forecast error is explained by itself. This is the highest for AMI IHD and CCF and Hip Replacement, all above 95%, but at around 90% for MI, Stroke and TIA. For AMI, none of the mortality rates for the other conditions explain over 0.5 of the variance in the forecast error. For MI, Hip Replacement explains over 6% of the variance, and AMI just over 1%. For IHD, CCF explains just over 2% of the forecast variance, while all other conditions less than 1%. For Stroke, MI accounts for nearly 6.5% of the forecast error, and CCF and AMI about 2%. IHD explains another 1% and TIA and Hip Replacement are almost negligible. 4% of variance in forecast error of TIA is explained by Hip Replacement, 3% by IHD, 2% by AMI and less than 1% by Stroke and MI. 2% of Hip Replacement's forecast error is accounted for by AMI, and the other conditions all account for less than 1%.

Table 4.6: Variance Decomposition percentages for $D365_{ht}$ using the VAR(3) specification.

% of FEV in	Standard errors	Typical Shock in						
		AMI	MI	IHD	CCF	Stroke	TIA	Hip
AMI	0.150565	97.73275	0.414054	0.377599	0.263897	0.279656	0.363818	0.568221
MI	0.231973	1.229519	91.53107	0.013705	0.046967	0.227374	0.490768	6.460593
IHD	0.107316	0.372524	0.049894	96.20926	2.258779	0.929918	0.130017	0.049607
CCF	0.157662	0.746149	1.064262	1.376480	95.13499	0.220446	1.313856	0.143821
Stroke	0.090850	1.944060	6.462655	0.791335	2.001118	88.70909	0.010900	0.080844
TIA	0.167406	2.230687	0.344484	3.213478	1.294588	0.876988	87.95642	4.083352
Hip	0.044039	1.945740	0.313211	0.348649	0.987421	0.272940	0.615216	95.51682

Forecast Horizon at 10 years

Note: The Choleski ordering for this table was: AMI, MI, IHD, CCF, Stroke, TIA, Hip

The results of the SUR models for year-long mortality are presented in Table 4.7. The R-squared values are all over 40%, and around 60% for CCF and Hip Replacement, indicating that the model is able to explain nearly half the variance in year-long mortality for the seven conditions. In all conditions, the two lags of the dependent variable are significant and positive, indicating a dynamic effect. The third lag of the dependent variable is also significant for MI, CCF, Stroke and Hip Replacement, although it is negative for Hip Replacement. The magnitude of the coefficients for the first lag of the dependent variable are all above 0.4, and highest for Hip Replacement. The coefficient on the second lag drops for all conditions, but stays the highest for CCF indicating more persistent performance over time. Where significant the value of the third lag is even lower, apart from

Hip Replacement.

Contemporaneous values of risk adjusted year long AMI are significant predictors of mortality for all other conditions, while lagged AMI mortality is only significant for MI, CCF and Hip Replacement. The sign on the coefficients indicate that contemporaneous AMI mortality is negatively correlated with CCF and TIA mortality, but positively correlated with all other conditions. Where the lags are significant, the sign switches for the lagged effect. Contemporaneous and lagged MI readmissions are a significant predictors for Stroke and TIA. Contemporaneous MI is also positively with both, as well as being positively associated with AMI, while lagged MI is negatively associated with both. Contemporaneous IHD is positively associated with CCF, and negatively associated with Stroke and TIA, lagged IHD is positively associated with TIA but not any of the other conditions. Contemporaneous CCF is positively associated with IHD, and negatively associated with AMI and Stroke. AMI and Stroke are also associated with lagged CCF mortality, but the sign is reversed. Lagged CCF mortality is also negatively associated with TIA, while contemporaneous CCF is not. The value of the coefficients for AMI, MI, IHD and CCF are low in all cases.

Table 4.7: Seemingly Unrelated Regression for risk adjusted $D365_{ht}$ estimates.

	AMI	**MI**	**IHD**	**CCF**	**Stroke**	**TIA**	**HIP**
Explanatory Variables	**Dependent Variables**						
DV L.2	0.161948***	0.181597***	0.171739***	0.300041***	0.124456***	0.186148***	0.172038***
	(0.031015)	(0.035379)	(0.032690)	(0.031442)	(0.032658)	(0.032206)	(0.026438)
DV L.3	0.021315	0.099929***	0.034530	0.132290***	0.117344***	0.036815	-0.19543***
	(0.027945)	(0.032646)	(0.030260)	(0.030359)	(0.029996)	(0.030771)	(0.030123)
AMI	-	0.175202***	0.039790**	-0.12643***	0.03499*	-0.11078***	0.02847***
		(0.050913)	(0.024988)	(0.031964)	(0.020098)	(0.040070)	(0.010203)
AMI L.1	0.523541***	-0.11664***	-0.056303	0.115093***	-0.005309	0.025578	-0.02696***
	(0.032861)	(0.049212)	(0.024053)	(0.030832)	(0.019405)	(0.038800)	(0.009840)
MI	0.095908***	-	-0.003710	-0.036364	-0.09504***	-0.10978***	6.22E-05
	(0.028862)		(0.018717)	(0.024047)	(0.014823)	(0.029959)	(0.007655)
MI L.1	-0.028046	0.539948***	-0.008871	0.056827**	0.055226***	0.108085***	0.004828
	(0.029618)	(0.034758)	(0.019139)	(0.024573)	(0.015324)	(0.030599)	(0.007794)
IHD	0.065604	-0.009880	-	0.150792***	-0.065911**	-0.35761***	0.014212
	(0.059854	(0.078939)		(0.049503)	(0.031136)	(0.061489)	(0.015722)
IHD L.1	-0.088878	0.013162	0.549683***	-0.060642	0.028807	0.153661***	-0.013623
	(0.059992)	(0.078990)	(0.033391)	(0.049685)	(0.031248)	(0.062302)	(0.015732)

	AMI	MI	IHD	CCF	Stroke	TIA	HIP
CCF	-0.19850***	-0.081064	0.096322***	-	-0.13360***	-0.025682	-0.009332
	(0.042975)	(0.057087)	(0.027744)		(0.022080)	(0.044840)	(0.011386)
CCF L.1	0.15403***	0.062583	-0.040138	0.455009***	0.10609***	0.104848**	0.004333
	(0.042185)	(0.055884)	(0.027251)	(0.032218)	(0.021719)	(0.043824)	(0.011137)
Stroke	0.117409	-0.60999***	-0.082676*	-0.33780***	-	-0.48723***	-7.93E-05
	(0.073675)	(0.096335)	(0.047669)	(0.060746)		(0.075541)	(0.019437)
Stroke L.1	0.001221	0.44688***	-0.012943	0.26165***	0.574844***	0.454457***	0.006845
	(0.073415)	(0.096519)	(0.047495)	(0.060968)	(0.033050)	(0.075216)	(0.019350)
TIA	-0.10418***	-0.16788***	-0.13765***	0.019878	-0.11352***	-	0.005231
	(0.036801)	(0.048672)	(0.023610)	(0.030844)	(0.019004)		(0.009758)
TIA L.1	0.044732	0.094912**	0.08893***	0.030258	0.07980***	0.507934***	-0.013777
	(0.036467)	(0.048338)	(0.023465)	(0.030438)	(0.018979)	(0.033376)	(0.009579)
Hip	0.567393***	0.069135	0.051744	-0.047415	-0.002148	0.011115	-
	(0.139307)	(0.185637)	(0.090656	(0.117412)	(0.072887)	(0.146540)	
Hip L.1	-0.41574***	-0.173236	-0.009076	0.026400	-0.010329	0.073965	0.720947***
	(0.144665)	(0.192635)	(0.094051)	(0.121399)	(0.075620)	(0.151944)	(0.029407)
R^2	0.437267	0.458630	0.470580	0.588279	0.485304	0.422073	0.601383
N	636	636	636	636	636	636	636

* Significant at p $\leq$ 0.1

** Significant at p $\leq$ 0.05

*** Significant at p $\leq$ 0.01

Contemporaneous Stroke readmissions are a negative predictor of MI, IHD, CCF and TIA mortality, and are relatively high for MI, CCF and TIA. Lagged values of Stroke mortally are positively associated with MI, CCF and TIA and remain relatively large. TIA mortality is significant and negative for AMI, MI, IHD and Stroke, but the values of the coefficients are low. Lagged TIA mortality is positively associated with MI, IHD and Stroke and the values indicate an even weaker association. Hip Replacement is associated with AMI, such that the values indicate a relatively strong positive association between contemporaneous hip mortality and AMI mortality, and a relatively weak negative association between lagged hip mortality and AMI.

Latent 28-Day Readmission Estimates

Table 4.8 presents the results of the Granger Causality/block-exogeneity tests for risk-adjusted 28-day readmissions. The 'All' column indicates some evidence of endogeneity

for AMI, MI, IHD and TIA with 5% significance or over. For AMI we see that lagged values of MI and IHD have a significant effect on risk adjusted readmissions. Lagged values of IHD, CCF have a significant effect on risk-adjusted MI readmission, while lagged values of AMI, MI, TIA and Hip readmissions have significant effects on IHD mortality. Finally lagged values of MI, IHD and CCF risk adjusted readmissions significantly impact TIA.

Table 4.8: Pairwise Granger Causality Test/Block Exogeneity Wald Tests for $R28_{ht}$ VAR(3) specification.

28-Day Readmissions VAR(3)							
Excluded Variable	Dependent Variable						
	AMI	MI	IHD	CCF	Stroke	TIA	Hip
AMI	-	0.5722	0.0327	0.7555	0.6180	0.2852	0.3671
MI	0.0011	-	0.0444	0.4556	0.8225	0.0250	0.6089
IHD	0.0022	0.0195	-	0.6405	0.7458	0.0359	0.2892
CCF	0.6202	0.0374	0.5672	-	0.1572	0.0468	0.1806
Stroke	0.4302	0.5713	0.5273	0.1150	-	0.3894	0.0754
TIA	0.6501	0.4177	0.0006	0.3734	0.3070	-	0.1449
Hip	0.1234	0.2008	0.0123	0.8875	0.3786	0.3271	-
All	0.0002	0.0159	0.0001	0.6673	0.6975	0.0168	0.1659

Table 4.9 presents the variance decomposition percentages for 28-day risk adjusted readmission rates. Again the largest amount of the variance of the forecast error in the conditions is explained by itself, although this is lower than for the mortality outcomes measures. This is highest for CCF, Stroke and Hip Replacement at 95%, IHD and TIA are also above 90%, but AMI is relatively low at 86%. For AMI, the risk adjusted readmission rates of IHD explain over 10% of the variance in the forecast error, and the readmission rates of MI explain nearly 2%. All other conditions have very low values. Similarly for MI readmissions, IHD explains the largest portion of the forecast error variance at over 5% and AMI just over 1%. For IHD, AMI explains just over 5%, Hip Replacement a bit over 1% and all other conditions less than that. For the remaining conditions, no other condition accounts for over 2% of the forecast error.

Table 4.9: Variance Decomposition percentages for $R28_{ht}$ using the VAR(3) specification.

% of FEV in	Standard errors	Typical Shock in						
		AMI	MI	IHD	CCF	Stroke	TIA	Hip
AMI	0.035841	86.39919	1.882459	10.38385	0.602159	0.104970	0.236702	0.390671
MI	0.085341	1.361604	91.38744	5.315513	0.524300	0.281605	0.246366	0.883176
IHD	0.022533	5.634419	0.191761	91.69390	0.051047	0.408958	0.782296	1.237622
CCF	0.094715	0.818006	1.086792	1.116505	95.03112	1.018157	0.899300	0.030118
Stroke	0.019286	1.619055	0.039725	0.213028	0.490413	95.35720	1.227889	1.052692
TIA	0.040074	0.191572	0.771842	2.109950	2.584053	0.531111	93.08755	0.723920
Hip	0.027452	0.556988	0.521928	1.152340	0.198788	1.180246	0.707539	95.68217

Forecast Horizon at 10 years

Note: The Choleski ordering for this Table was: AMI, MI, IHD, CCF, Stroke, TIA, Hip

The results of the SUR models for the risk adjusted 28-day readmission estimates are presented in Table 4.10. The R-squared estimates for the models indicated that in most cases it is able to explain about half the variation in readmissions. It performs less well for MI and Hip Replacement, where it is able to explain about 40% and 15% respectively. The first, second and third lags of the dependent variable are significant for most conditions, indicating that readmissions are very dynamic. The exception is Hip Replacement where only the first and third lags are significant. The sign on the first lag is positive for all conditions, however the value of the coefficients are not as high as they were for the mortality estimates, indicating a weaker effect. The values of the coefficients for the second and third lags are lower, indicating the dynamic effect wearing off.

Examining the other explanatory variables reveals that contemporaneous AMI is a positive predictor of IHD and Stroke, and a negative predictor of CCF and Hip Replacement, although the value of the coefficient suggests that association all but CCF is quite weak. Lagged AMI readmissions are also associated with CCF and Stroke, but the sign is reversed from the contemporaneous effect and the magnitude of the coefficient is much lower. Risk adjusted MI readmissions are negatively associated with CCF and Hip Replacement and positively associated with IHD, but again the values on the coefficient are very low. Lagged MI readmissions are weak positive predictors of AMI, CCF and Hip Replacement. Contemporaneous CCF is a weak negative predictor of AMI, MI and TIA and a weak positive predictor of IHD, lagged CCF is only significant for MI but the sign of the association is positive.

Contemporaneous Stroke readmissions are positive predictors of AMI and CCF and

negative predictors of TIA and Hip Replacement. The value of the coefficient in all cases indicates a relatively strong effect. Lagged Stroke readmissions also predict the same conditions, and while the effect remains relatively strong the direction is reversed. There is a weak positive association between contemporaneous TIA readmissions and IHD, and a negative association MI, CCF and Stroke which is relatively strong for CCF. Lagged TIA readmissions are continue to be significant for IHD, CCF and Stroke but the coefficients are lower in value. Contemporaneous Hip Replacement has a weak negative association with AMI, Stroke and TIA and a slightly stronger negative association with MI. Lagged values of hip readmissions are only significantly associated with Stroke, and the value of the coefficient is quite low.

Table 4.10: Seemingly Unrelated Regression for risk adjusted $R28_{ht}$ estimates.

	AMI	**MI**	**IHD**	**CCF**	**Stroke**	**TIA**	**HIP**
Explanatory Variables	**Dependent Variables**						
DV L.2	0.157913***	0.217126***	0.091447***	0.286443***	0.131106***	0.180445***	0.009773***
	(0.030767)	(0.032785)	(0.033130)	(0.029342)	(0.031707)	(0.031370)	(0.026020)
DV L.3	0.093962***	0.084173**	0.127575***	0.103890***	0.126648***	0.144845***	-0.003702
	(0.027977)	(0.030322)	(0.028892)	(0.027611)	(0.028084)	(0.028320)	(0.023535)
AMI	-	0.083821	0.042350**	-0.32619***	0.077571***	-0.022906	-0.082793*
		(0.112618)	(0.019702)	(0.119601)	(0.019903)	(0.044197)	(0.043747)
AMI L.1	0.538134***	-0.081332	0.004349	0.19594*	-0.046470**	0.001574	0.017708
	(0.031668)	(0.110923)	(0.019132)	(0.116559)	(0.019445)	(0.043040)	(0.042633)
MI	-0.001651	-	0.023651***	-0.19417***	-0.001229	-0.023807	-0.04721***
	(0.012872)		(0.006611)	(0.040030)	(0.006747)	(0.014907)	(0.014569)
MI L.1	0.040562***	0.422964***	-0.01994***	0.144207***	-0.002297	0.009063	0.029921**
	(0.012034)	(0.032677)	(0.006234)	(0.037848)	(0.006374)	(0.014084)	(0.013801)
IHD	0.157951**	0.782934***	-	0.525372**	0.035743	0.368363***	0.130834
	(0.074639)	(0.219396)		(0.233855)	(0.039051)	(0.086299)	(0.085934)
IHD L.1	0.026594	-0.320072	0.677362***	-0.114767	-0.043997	-0.210736**	0.010558
	(0.075932)	(0.224195)	(0.031474)	(0.238382)	(0.039754)	(0.088248)	(0.087699)
CCF	-0.026238**	-0.17810***	0.012530**	-	0.005265	-0.04114***	0.003389
	(0.011540)	(0.033705)	(0.005946)		(0.006053)	(0.013358)	(0.008272)
CCF L.1	0.014692	0.145293***	-0.008757	0.399054***	0.002493	0.010246	0.010488
	(0.010762)	(0.031461)	(0.005548)	(0.030796)	(0.005651)	(0.012472)	(0.010747)
Stroke	0.30980***	-0.038083	0.031349	0.404678*	-	-0.28643***	-0.22787***

	AMI	MI	IHD	CCF	Stroke	TIA	HIP
	(0.072249)	(0.214670)	(0.037486)	(0.228827)		(0.083946)	(0.083106)
Stroke L.1	-0.24936***	0.096421	-0.031976	-0.55247**	0.618191***	0.22790***	0.30337***
	(0.072137)	(0.213821)	(0.037309)	(0.228180)	(0.032134)	(0.083722)	(0.082382)
TIA	-0.019106	-0.166589*	0.07998***	-0.35914***	-0.05417***	-	-0.051038
	(0.032433)	(0.095637)	(0.016620)	(0.101478)	(0.016889)		(0.037082)
TIA L.1	0.033241	0.137201	-0.08366***	0.299642***	0.05618***	0.541382***	0.035751
	(0.032530)	(0.095857)	(0.016690)	(0.102159)	(0.016928)	(0.032807)	(0.037257)
Hip	-0.060652*	-0.31015***	0.028778	-0.003382	-0.04854***	-0.073073*	-
	(0.034608)	(0.101701)	(0.017823)	(0.108838)	(0.018086)	(0.040182)	
Hip L.1	-0.005289	-0.065736	0.020697	-0.041625	0.04939***	0.063768	0.365077***
	(0.035609)	(0.104740)	(0.018300)	(0.111714)	(0.018580)	(0.041237)	(0.042943)
R^2	0.554428	0.397957	0.696157	0.451002	0.607697	0.519184	0.157403
N	636	636	636	636	636	636	636

* Significant at p $\leq$ 0.1

** Significant at p $\leq$ 0.05

*** Significant at p $\leq$ 0.01

Latent 365-Day Readmission Estimates

Table 4.11 presents the results of the Granger Causality/block-exogeneity tests for risk-adjusted year long readmissions. The 'All' column indicates some evidence of endogeneity for Hip Replacement at 5% significance and IHD, Stroke and TIA at 1%. For Hip Replacement, the chi-squared test indicates that lagged values of Stroke and TIA have a significant effect on risk adjusted readmissions. For IHD and Stroke, lagged values of Hip Replacement have a significant effect, while lagged values of AMI readmissions have significant effects on TIA and Stroke.

Table 4.11: Pairwise Granger Causality Test/Block Exogeneity Wald Tests for $R365_{ht}$ VAR(3) specification.

365-Day Readmissions VAR(3)							
Excluded Variable	Dependent Variable						
	AMI	MI	IHD	CCF	Stroke	TIA	Hip
AMI	-	0.0595	0.3958	0.6906	0.0414	0.0530	0.3546
MI	0.5631	-	0.6822	0.9941	0.5638	0.6878	0.1397
IHD	0.1446	0.5421	-	0.1179	0.9093	0.1234	0.4443
CCF	0.9966	0.5395	0.0668	-	0.0961	0.3037	0.7247
Stroke	0.8594	0.4699	0.4562	0.7797	-	0.2694	0.0743
TIA	0.1046	0.4439	0.2916	0.0577	0.9837	-	0.0623
Hip	0.9722	0.2882	0.0751	0.7930	0.0966	0.5448	-
All	0.6240	0.3118	0.0909	0.4486	0.1076	0.0989	0.0495

The variance decomposition percentages for year long risk adjusted readmission rates are presented in Table 4.9. Again the largest amount of the variance of the forecast error for each condition is explained by itself. This is around 95% for AMI, MI, IHD and Stroke, and over 99% for CCF, TIA and Hip Replacement. For AMI, a negligible amount of the variance in the forecast error is explained by the risk adjusted readmission rates of the other conditions. For MI, AMI readmissions explain over 2% of the forecast error, and Stroke nearly 1.5%. 2% of IHD forecast error is accounted for by hip, while all other conditions have very low values. TIA accounts for over 4% of the forecast error of CCF readmissions, while IHD for over 2% and AMI and MI about 1% each. The variance decomposition of the forecast errors for Stroke readmissions are such that Hip Replacement accounts for over 2%, and all other conditions for less than 1%. For TIA, AMI and Stroke explain almost 3%, MI over 1% and all other conditions less than 1%. Finally the nearly 3% of the forecast error of Hip Replacement is accounted for by Stroke and TIA, and IHD explains about 1.5%, while all other conditions account for less than 1%.

Table 4.12: Variance Decomposition percentages for $R365_{ht}$ using the VAR(3) specification.

% of FEV in	Standard errors	Typical Shock in						
		AMI	MI	IHD	CCF	Stroke	TIA	Hip
AMI	0.090501	98.70889	0.254707	0.391759	0.009760	0.070692	0.472412	0.091779
MI	0.147368	2.276795	94.77693	0.470095	0.067534	1.443554	0.126491	0.838599
IHD	0.038481	0.097818	0.070058	96.05539	0.347611	0.631246	0.248768	2.549107
CCF	0.137021	0.951112	1.207000	2.553303	90.81431	0.064143	4.309704	0.100431
Stroke	0.034611	0.631405	0.380211	0.936857	0.447049	95.14921	0.173727	2.281544
TIA	0.071102	2.790180	1.032996	0.734510	0.747682	2.620303	91.91310	0.161223
Hip	0.040651	0.583124	0.925351	1.609994	0.207120	2.417708	2.634518	91.62218

Forecast Horizon at 10 years

Note: The Choleski ordering for this table was: AMI, MI, IHD, CCF, Stroke, TIA, Hip

The results for the SUR models for risk adjusted 365-day readmissions are presented in Table 4.13. The R-squared values for the individual regressions indicate that in most cases they are able to explain about half, and sometimes nearly 60% of the variation in year-long readmissions. For Hip Replacement the value of the R-squared estimate is much lower, indicating it is able to explain only about 17% of the variance. The results also indicate that two lags of the dependent variable are significant for all conditions, suggesting that the readmission outcome measures are dynamic. For many conditions the third lag is also significant. The value of the coefficients for the first lag is relatively high, however it declines for the second and third lag. In all conditions apart from Hip Replacement the sign is positive indicating a positive association across time.

Table 4.13: Seemingly Unrelated Regression for risk adjusted $R365_{ht}$ estimates.

	AMI	**MI**	**IHD**	**CCF**	**Stroke**	**TIA**	**HIP**
Explanatory Variables	**Dependent Variables**						
DV L.2	0.199535***	0.251365***	0.209415***	0.289965***	0.129079***	0.198856***	0.223241***
	(0.032807)	(0.035135)	(0.033799)	(0.029456)	(0.033797)	(0.031974)	(0.040228)
DV L.3	0.181245***	0.030915	0.138477***	0.086428***	0.099908***	0.142555***	0.052245
	(0.030576)	(0.034243)	(0.029919)	(0.028069)	(0.029364)	(0.029620)	(0.033917)
AMI	-	-0.174574**	0.004590	0.151741***	-0.06361***	-0.061671**	0.145974***
		(0.057263)	(0.012758)	(0.052254)	(0.012731)	(0.026568)	(0.029880)
AMI L.1	0.456406***	0.086576	-0.008331	-0.101208**	0.056009***	0.018620	0.017445

	AMI	MI	IHD	CCF	Stroke	TIA	HIP
	(0.035409)	(0.056306)	(0.012528)	(0.051315)	(0.012495)	(0.026069)	(0.022717)
MI	-0.07071***	-	-0.000273	-0.14169***	-0.008398	0.011503	0.006214
	(0.024189)		(0.008419)	(0.034543)	(0.008499)	(0.017558)	(0.022249)
MI L.1	0.037782	0.535474***	-0.004920	0.097650***	0.001198	0.010343	-0.023324
	(0.024672)	(0.034966)	(0.008590)	(0.035276)	(0.008659)	(0.017880)	(0.015035)
IHD	-0.107421	-0.027109	-	-0.240048	0.062793*	-0.128157*	0.033722**
	(0.107110)	(0.167736)		(0.153482)	(0.037589)	(0.077643)	(0.015282)
IHD L.1	0.170033	0.202768	0.575135***	0.445533***	-0.023260	0.064946	0.208220***
	(0.108279)	(0.169484)	(0.032381)	(0.154866)	(0.038001)	(0.078429)	(0.066109)
CCF	0.070856***	-0.13329***	-0.02335***	-	0.012842	-0.003959	-0.090814
	(0.025289)	(0.039598)	(0.008805)		(0.008909)	(0.018475)	(0.067139)
CCF L.1	-0.048273**	0.06927*	0.018451**	0.480992***	-0.004650	-0.010544	-0.018348
	(0.024549)	(0.038545)	(0.008545)	(0.031090)	(0.008646)	(0.017911)	(0.015745)
Stroke	-0.45293***	-0.199637	0.079010**	0.178707	-	0.424074***	0.016708***
	(0.108713)	(0.171749)	(0.037991)	(0.157234)		(0.078218)	(0.015271)
Stroke L.1	0.28943***	-0.020309	-0.066767*	-0.226086	0.601554***	-0.23635***	-0.135720**
	(0.108813)	(0.170398)	(0.037760)	(0.156521)	(0.033002)	(0.078302)	(0.067749)
TIA	-0.110117**	0.041914	-0.029641	0.015638	0.102043***	-	0.216961
	(0.051544)	(0.080926)	(0.018080)	(0.074005)	(0.017910)		(0.066969)
TIA L.1	0.052456	0.020536	0.016290	0.088372	-0.06967***	0.520754***	-0.017698
	(0.051605)	(0.081025)	(0.018119)	(0.074046)	(0.018066)	(0.032767)	(0.032036)
Hip	0.085811	-0.168676*	0.067343***	-0.135723	-0.042703**	-0.039790	-
	(0.062420)	(0.097932)	(0.021637)	(0.089421)	(0.021891)	(0.045364)	
Hip L.1	0.001821	0.031260	0.032483	0.072690	0.058228***	-0.017455	-0.03758***
	(0.060836)	(0.095530)	(0.021082)	(0.087120)	(0.021275)	(0.044118)	(0.009917)
R^2	0.624146	0.490366	0.661888	0.561351	0.574192	0.535415	0.168016
N	636	636	636	636	636	636	636

* Significant at p $\leq$ 0.1

** Significant at p $\leq$ 0.05

*** Significant at p $\leq$ 0.01

Contemporaneous values of AMI readmissions are negatively associated with MI, Stroke and TIA and positively associated with CCF and Hip Replacement. All of these associations are weak. Lagged values of AMI readmissions are also associated with CCF and Stroke lagged however the sign is reversed, and the association is weaker. Contemporaneous MI is weakly negatively associated with AMI and CCF, while lagged MI is weakly

positively associated with CCF. Contemporaneous IHD has a weak and positive association with Stroke and Hip Replacement and a weak negative association with TIA. Lagged IHD is also positively associated with Hip Replacement and CCF, and the value on the coefficient of the latter is relatively large. Contemporaneous values of CCF readmissions have a weak positive association with AMI and a weak negative association with IHD and MI. The lagged values of CCF are also correlated with these conditions, but the sign is reversed.

Contemporaneous values of risk adjusted year long readmissions for Stroke are positively associated with IHD, TIA and Hip Replacement and negatively associated with AMI. The value of the AMI and TIA coefficients are relatively high, at around 0.45. Lagged values of Stroke, are also significant in influencing these conditions, however the value of the coefficient is lower, and the sign is reversed. Risk adjusted TIA readmissions have a significant negative association with AMI and a positive association with Stroke, both coefficients indicate that the association is weak. Lagged TIA readmissions also significantly affect Stroke, although the relationship is negative and weaker. Readmission for Hip Replacement are positively associated with IHD and negatively associated with MI and Stroke, however all associations are weak. Lagged Hip Replacement readmissions continue to be associated with Stroke, at about the same strength as contemporaneous readmissions, but positively.

4.5 Discussion

Outcome measures are commonly used to determine whether health service providers are delivering high quality care. Mortality rates have been used to evaluate the performance of individual physicians, or hospitals in many industrialized countries, and are often used in composite scores of performance. While outcome measures are often measured more generally, such as overall mortality rates, the outcomes of selected conditions, such as AMI, are sometimes preferred. By focusing on a particular condition, users of the data can better adjust it for exogenous factors such as patient case-mix and choose conditions may have higher instances of occurrence and/or stronger linkages between outcome and treatment. However, when performance is assessed using these piecemeal outcome measures, we need to be cautious about generalizing these findings to draw conclusions about provider's performance in general.

Chapters 2 and 3 consider how to use methodological tools to create better estimates of quality, by reducing bias and systematic error as well as incorporating time series and cross section information from different variables. However, these chapters did not consider the association of quality across different conditions. Better understanding of

these relationships will allows us to determine to what extent the outcomes of one condition can be used to make generalizations about the overall performance of an institution. This chapter investigated the relationship between the latent outcome measures constructed in Chapter 2 using a VAR model, the same model used in Chapter 3 to transform these measures into filtered estimates.

The models are run separately for each of the four risk-adjusted outcome indicators, 30-day in hospital mortality, year long mortality, 28-day emergency readmissions and year long readmissions. For each model a VAR(3) specification is run and the Granger causality and forecast decompositions are reported to describe the data. The Granger causality estimates indicate, for all four outcomes, where the lagged values of one condition's outcome significantly help to predict the other condition's outcomes. The estimates for the different conditions suggest that where significant, a condition influences the performance of the other conditions. For example in the 30-day mortality estimates, indicate that TIA outcomes are significant predictors of Stroke mortality, however Stroke is not a significant predictor of TIA mortality. Thus, there is a uni-dimensional causality of TIA on Stroke.

The forecast error decompositions of all four outcome measures indicate that the dependent variable is largely determined by its own past performance. While there is variation amongst the conditions, as to which are more endogenously determined, they are all extremely dynamic indicators. No condition is ever able to forecast over 10% of the variance in forecast error of another for any of the outcome measures. Risk adjusted 30-day mortality is the most endogenous of the outcome measures, where all the indicators for all conditions are able to forecast about 95% of their own variance ten years into the future. The other three outcome measures perform similarly, with all conditions being able to forecast around 90% of their own variance.

The information about the relationship between variables from the VAR(3) specification is used to modify the models and adjust the lag lengths included for the different conditions. In the final version of the model, we include three lags of the outcome for the dependent variable condition and the contemporaneous value and 1 lag of all other conditions indicators. We include the 1 lag, as the Granger causality estimates indicate some endogeneity between the lagged outcomes of other conditions, but not more as the variance decomposition estimates indicate that a very small portion of the variance will be determined by the outcomes of the other conditions. We keep the three lags for the dependent variable as this was the appropriately defined lag length by our tests, and the variance decomposition estimates indicate that a large percent of the forecast error is autoregressive. These versions of the model are estimated by SUR, as the right hand side variables are no longer the same for every equation.

The results of the SUR models for each condition are informative as to how the outcomes of different conditions are associated with one another contemporaneously and dynamically. The interpretation of the coefficient on the contemporaneous outcomes is straightforward, if the condition is significant this means that the outcome of that condition is somehow associated with the dependent variable's outcome. If the association is positive it suggests that good quality in that condition is associated with good quality in the other condition, for example good CCF outcomes are associated with good AMI outcomes. This may be because the cardiology department is good in this particular hospital because of the skill of physicians or equipment available. If the association is negative it indicates that good quality in one condition is associated with poor quality in another. For example risk adjusted 30-day AMI mortality is negatively correlated with contemporaneous 30-day risk adjusted Stroke mortality. This could be a result of competing resources amongst departments within a hospital, resulting in a trade-off in equipment or staff that impacts quality. If all contemporaneous coefficients were positive this would indicate that there is some overarching driver that encourages good quality throughout the hospital, such as management. Instead, the results of our SUR models indicate that for all conditions the outcomes of some conditions have a positive contemporaneous effect on the dependent variable, while others have a negative effect.

One of the most interesting results of from the SUR models is that in almost all cases when one condition is significantly associated with another the sign on the contemporaneous effect is opposite from the sign on the lagged effect. For example 30-day risk adjusted CCF mortality is positively associated with 30-day risk adjusted AMI mortality, however lagged CCF mortality is negatively associated with 30-day AMI mortality. Similarly, while contemporaneous Stroke and AMI 30-day mortality is negatively associated, lagged Stroke mortality is positively associated with AMI mortality. What is striking is that the sign switch is present in all cases, but one, where the contemporaneous and lagged effect are significant. This suggests that when the conditions are dynamically associated with one another there is a particular type of relationship.

Through the examination of the contemporaneous effects we know which conditions are reinforcing and which are competing. Reinforcing conditions are characterized by a positive contemporaneous relationship, where good quality in one is associated with good quality in the other. In the example used above, CCF and AMI would be reinforcing conditions. Competing conditions are characterized by a negative contemporaneous relationship, where good quality in one is associated with bad quality in the other. In the example used above, Stroke and AMI would be competing conditions. The results of the lagged indicators suggest that in cases where the reinforcing relationship is dynamic, that

is both the contemporaneous and lagged condition are significantly associated with the dependent variable, the sign switches. Going back to our example, a decline in risk-adjusted CCF mortality is associated with a decline in AMI mortality, but in the dynamic reinforcing relationship the decline in lagged CCF mortality is associated with an increase in AMI mortality. Similarly in the dynamic competing relationship, an increase in contemporaneous Stroke 30-day mortality is associated with a decrease in AMI mortality, but in the dynamic competing relationship a decline in lagged Stroke outcomes is also associated with a decrease in current AMI mortality.

This sign switch, or the nature of the dynamic reinforcing and dynamic competing relationships are more difficult to explain, and are more specific to the pairs of variables which are correlated. For example, the association between CCF and AMI is more complicated to disentangle because the outcomes for either of these conditions will impact the outcomes of the other. AMI may appear as a complication of CCF, or heart failure may appear as a sequence of AMI in patients who had no disturbance of cardiac function prior to the formation of the infarct (Gerbode and Selzer, 1948). Indeed the Granger causality estimates for 30-day mortality suggest that lagged AMI outcomes are significant predictors of CCF mortality, but that lagged CCF outcomes are not significant predictors of AMI mortality. Similarly, the results from Chapter 2 tell us that 30-day AMI mortality is negatively correlated with emergency readmissions, such that a 'good' hospital will be able to successfully save the very severe AMI patients who then return with complications later on. It is possible, then that some of these patients develop heart failure as a result of a severe MI, such that lower lagged AMI mortality is associated with higher CCF mortality. This tells us that CCF mortality may be a less good overall performance indicator of the provider due to its relationship with AMI. Thus, in order to understand the dynamic reinforcing and dynamic competing relationships we need to consider these in relation to the related conditions. Below we consider in detail the results for AMI only, while the other results are discussed in Appendix C. We chose to focus on AMI because it has been used more so than other conditions as an indicator of hospital quality (Kessler and McClellan, 1996; McClellan and Staiger, 1999; Propper et al., 2004) and recently in the UK, AMI outcomes have been used as a proxy of hospital quality to assess management practices (Bloom et al., 2010), as well as to inform publicly available hospital rankings such as the Dr. Foster 'Good Hospital Guide'.

Another interesting result that appears, is that while two conditions may have a dynamic reinforcing relationship with regards to one of the outcome indicators, they may have a dynamic competing relationship with regards to another. For example, TIA and Stroke have a dynamic competing effect with regards to year-long mortality and 28-day

readmissions, but a dynamic reinforcing effect when looking at year-long readmissions. This may initially seem counter-intuitive, as we associate both lower mortality and lower readmissions as indications of higher quality. However, in Chapters 2 and 3 we already came across a negative correlation between mortality and readmission indicators for many of the conditions, indicating that in some instances higher readmissions may be mistaken for lower quality, when in fact they may be a result of the lower mortality of 'marginal' patients. A similar explanation may be applied to these cases. For example, when considering the relationship between TIA and Stroke, we know that when a patient is admitted with TIA it initially appears exactly the same as a Stroke, the only difference being that the patient will recover in 24 hours. After having a TIA, patients are at a great risk of going on to develop a Stroke in the next few weeks, so it is important that they receive appropriate treatment shortly after. In the case of TIA thus, we can assume that high 28 day readmissions are an indicator of poor quality. Stroke patients will spend a longer time in hospital than TIA patients, on average, and thus low short-term readmissions in this case may very well reflect better quality.

Indeed, looking back at Chapter 3 we see that the in the Stroke model, 28-day readmissions were negatively correlated with 30-day mortality, while year-long readmissions were not. The TIA model from Chapter 3 also tells us that TIA 28-day readmissions are positively correlated with 30-day mortality as are year-long readmissions. Thus, the competing effect we see between the two conditions in the short term could very well reflect this. That is to say that hospitals with low readmissions for Stroke (better quality) are negatively correlated with hospitals that have high 28-day readmissions for TIA (better quality). Thus it appears that TIA patients benefit from being treated at institutions with lower long-term Stroke readmissions. This is confirmed by the long run readmissions model which indicates that there is a reinforcing effect between the two conditions. The relationship between these two conditions highlights the importance of the correct interpretation of readmissions indicators. Not being aware of what exactly high readmissions indicate for a particular condition can lead to misinterpretation of outcome indicators.

AMI

The forecast error variance decompositions for AMI indicate that its performance is highly dynamic. For all mortality outcomes, and year-long readmissions AMI is able to predict over 97% of its own variance in a 10 year horizon. The results from the AMI SUR models confirm that performance is very highly dynamic, with all three lags being significant predictors of current outcomes, such that high outcomes in a previous period are highly associated with high outcomes in the current period. The Granger causality results for AMI

suggest that the lagged values of other conditions do not significantly help to predict any AMI outcomes aside from 28-day readmissions, where lags of MI and IHD are significantly associated with AMI readmissions. Moreover, the forecast error variance decomposition indicates that these conditions are able to explain 2% and 10% of the forecast error in a ten-year period, respectively.

AMI & MI

The results of all tables suggest that 30-day mortality AMI and MI outcomes are not significantly associated. In the case of year-long mortality the Granger causality estimates indicate that lagged AMI is a significant predictor of MI year-long mortality, but not the reverse. Indeed, the SUR model for year-long mortality shows a contemporaneous reinforcing effect, for both AMI and MI models. However, a dynamic reinforcing effect appears only in the MI model, such that higher lagged AMI mortality is associated with lower MI mortality. Thus, good mortality outcomes in one are indicative of good mortality outcomes in the other, but ultimately higher lagged AMI mortality will lead to lower MI mortality, presumably because this results in less MI patients.

The Granger causality estimates for both short and long term readmissions suggest unidirectional causality, such that lagged AMI readmissions significantly influence MI. The SUR results confirm this showing that lagged MI is a positive predictor of AMI 28-day readmissions, that the two conditions exhibit a contemporaneous competing relationship, such that higher long term readmissions in MI will lead to lower long term readmissions in AMI. This allows us to interpret causality and suggests that higher AMI readmissions lead to higher MI readmissions, possibly because readmitted AMI patients are having a MI. In all cases where there is a significant effect the value of the coefficient is very low, indicating that when this effect is occurring it affects only a small percentage of the outcomes being considered. Indeed, this is confirmed by the forecast error variance decomposition estimates, which show that one condition will never explain more than 2% of the variance in the other.

Overall, in the short term if an AMI patient survives and is readmitted it is likely that this will increase MI readmissions in the long term. If the AMI patient dies it will lead to lower MI mortality later on, as there are fewer patients. Thus, in this case using AMI to measure the quality of providers may be more informative than using MI, as what seems to be 'worse' outcomes for MI may be the result of better AMI treatment for the relatively poor heart function of 'marginal' patients who survive when treated in high quality hospitals. This is similar to the finding from Chapter 3, where AMI readmissions and mortality were negatively correlated, indicating that higher readmissions in some cases

is an indication of better quality, as they indicate 'marginal' patients that were able to survive due to good treatment.

AMI & IHD

The 30-day mortality SUR model suggests that AMI and IHD have a dynamic reinforcing contemporaneous effect on each other. Thus, low mortality in one condition is associated with low mortality in the other contemporaneously, but the lagged mortality of one is negatively associated with the contemporaneous mortality of the other. The Granger causality estimates indicate a unidirectional casual effect between lagged outcomes of AMI and IHD mortality, such that lagged values of AMI influence IHD but not the opposite. Both the variance decomposition estimates, and the values of the coefficients suggest that the magnitude of the effect is small. The same casual effect is suggested by the Granger causality estimates for year-long mortality. However, in the SUR model lagged AMI is not a significant predictor of IHD mortality, and IHD is not a significant predictor of AMI mortality. The only significant effect is the positive effect between contemporaneous AMI and IHD.

The SUR model for 28-day readmissions also suggests there only a contemporaneous effect between AMI and IHD, such that higher readmissions in one are associated with higher readmissions in the other. The Granger causality estimates suggest significant bi-directional causality. While, this effect is present for both conditions, the IHD coefficient in the AMI model is higher, as is the forecast error variance decomposition. This suggests that IHD readmissions explain more of the variation in AMI readmissions. However, when looking at year-long readmissions there is no association between the two conditions.

Thus it appears that lower lagged AMI mortality will lead to higher IHD mortality, yet the same factors will influence the short term readmissions for both conditions. This suggests that AMI mortality is a better indicator of overall performance than IHD mortality, as the 'worse' outcomes for IHD may be the result of better AMI treatment for the relatively poor heart function of 'marginal' patients who survive when treated in high quality hospitals. However, treatment of the two are clearly related, and poor IHD treatment will result in higher readmissions for AMI.

AMI & CCF

The dynamic reinforcing relationship between AMI and CCF mortality outcomes has already been discussed briefly above. The Granger causality estimates in Tables 1 and 6 suggest that there is unidirectional causality between AMI and CCF, such that lagged AMI short and long term mortality Granger causes CCF short and long term mortality. Yet the

variance decomposition percentages, for both long and short term measures suggest that this applies to less than 1% of the variance in mortality of either condition. Thus, the SUR dynamic reinforcing relationships in the short and long term models can be interpreted as AMI and CCF being contemporaneously positively correlated, possibly because good outcomes of both reflect a strong cardiology unit. However, lower lagged AMI mortality will cause worse CCF outcomes, and worse CCF outcomes are associated with higher AMI mortality. In the case of short-term mortality, the coefficient of the former effect is very small, while the coefficient of the latter indicates that it explains a substantial amount of the variance. For long term mortality both explain a small amount of the variance.

The effect is slightly different for the readmission measures. Neither for long term nor short term readmissions are the Granger causalities significant. Moreover, the forecast error variance decompositions indicate that the conditions explain less than 1% of each others variance. The SUR model for year-long readmissions do however, indicate the same dynamic reinforcing relationship between AMI and CCF discussed above. Although because of the insignificant Granger causalities we can only interpret this as association between the conditions, yet again the coefficients suggest the effect CCF has on AMI is weaker than AMI's effect on CCF. The short-term readmissions model suggests only a positive contemporaneous association between the two conditions.

Overall, the results suggest that it is difficult to disentangle the effects between the two conditions. It appears while contemporaneously either indicator will be indicative of good performance, better AMI performance will cause worse CCF performance later on. Thus, again AMI is a more reliable indicator of provider quality, as poor CCF outcomes can be a result of successful AMI treatment.

AMI & Stroke

The Granger causality tests for short and long term mortality do not suggest any significant relationship between the lags of one condition on the other's performance. Moreover the variance decomposition percentages between AMI and Stroke are also very small, indicating they explain less than 0.5% of each others forecast error variance. The SUR results for short-term mortality show a weak dynamic competing relationship between the two conditions, such that contemporaneously they are negatively associated, while lower lagged mortality of one is associated with lower mortality in the other. As the Granger causality estimates are not significant, the dynamic effect may be a result of the decreasing trend in latent short term mortality that both conditions experienced during this time period (see Figures 2.2 and 2.8, Chapter 2). Thus, while they have a competing contemporaneous effect, they are both improving over time.

The results of the long-term mortality SUR model are quite different, indicating a significant relationship of contemporaneous AMI mortality on Stroke mortality only. Moreover the association is only significant at 10%, and of very low value. However, the association between the two is positive, unlike the association of contemporaneous short term mortality. The figures from Chapter 2 do indicate a steep decline in long-term mortality for both these conditions over the period being investigated which may account for this association. Interestingly, AMI is significant in the Stroke model, but Stroke is not significant in the AMI model. It is possible that this is because AMI explains more of the variance in Stroke outcomes than the reverse, as demonstrated by the variance decompositions in Table 4.6. The Granger causality tests for readmissions indicate that lagged values AMI year-long readmissions are significant in influencing Stroke readmissions, while the variance decomposition percentages between the short and long term outcomes of the two conditions are very low. The long-term SUR model indicates a dynamic competing relationship similar to the one observed for short-term mortality.

The SUR model for short term readmissions indicates a dynamic reinforcing effect, such that the contemporaneous effect between the two conditions is positive, while a decline in the lagged readmissions of one condition will lead to an increase in the readmissions of the other. The Granger causalities suggest no significant causality in any direction, while the variance decomposition suggest that AMI readmissions explain more of the variance in Stroke than the other way around. However, the value of the Stroke coefficients in the SUR model are higher than the AMI coefficients for the Stroke model.

These results are very difficult to explain. There is some literature on the association between ischemic Stroke and AMI. The risk of Ischemic Stroke in patients presenting with AMI has declined from 2.4% to 3.5% in earlier reports to about 0.6% to 1.8% in more recent studies incorporating thrombolytic or anticoagulant therapy in the acute phase (Suarez, 2006). However, the Granger causalities between the conditions do not suggest that this is what is driving the association. Indeed, while is observed that over the past $40 - 50$ years Stroke case fatality rates have decreased in high-income countries (Feigin et al., 2009) it is difficult to know what to attribute this to. There has been evidence to suggest that hospitals with Stroke units are able to achieve better outcomes, however it remains uncertain what about these units is responsible for them (Langhorne et al., 2000). Perhaps the relationship between Stroke and AMI is linked to the resource allocation between departments within a hospital. More in-depth research into the relationship between treatment of the two conditions would be needed to help disentangle the factors driving this relationship.

AMI & TIA

AMI and TIA have no significant relationship in any of the short-term outcome models. In both of the long term models, they have a weak competing contemporaneous effect which is not dynamic, indicating that contemporaneous long term outcomes of TIA are negatively associated with long term AMI outcomes. The Granger causality estimates do not suggest a significant causal link in any direction for mortality. However, they do suggest unidirectional causality, such that lagged AMI readmissions cause TIA readmissions. Again it is not clear what is driving the negative contemporaneous effect between the two conditions.

AMI & Hip Replacement

The short term mortality SUR model indicates that AMI outcomes have a small positive effect on hip outcomes. The Granger causality and variance decomposition estimates do not help in clarifying the relationship between the two variables. The long term Granger causality estimates however do suggest that lagged hip mortality is a significant predictor of AMI, and the variance decomposition estimates indicate that it explains nearly 2% of AMI's forecast error. Indeed the SUR results show a dynamic reinforcing effect, such that the two conditions have a positive contemporaneous association, and a negative lagged association. Moreover the value of the hip coefficients in the AMI model are relatively high. Indeed, a high number of patients suffer from AMI in the year following their surgery. The results of the year-long readmission model show that AMI has a positive contemporaneous effect on Hip year-long replacements, but not the other way around. While, short term readmissions between the two conditions are negatively associated. The explanation of these effects is difficult, but may be related to the associations indicated for mortality.

Through the investigation of the relationships between AMI to the other conditions reported by our models, we believe there is evidence to select risk-adjusted AMI indicators as good quality measures of the cardiovascular departments of hospitals The different causal relationships between AMI and the other cardiovascular conditions make it difficult to interpret some of the other indicators, where rates of higher readmissions or mortality may not always be indicative of poor quality. However, the negative contemporaneous correlation between AMI and Stroke and AMI and TIA suggest that it may perform less well as an indicator of overall hospital performance. Indeed, looking at this indicator alone may draw attention away from other conditions in the hospital which compete for resources.

Our discussion of the relationship between AMI and other conditions was limited due

to uncertainty as to what exactly these associations were reflecting. In the case of AMI and Stroke it was difficult to disentangle the complex relationship as there are a number of possible explanations to explain the associations between the different indicators. This highlights some of the limitations with trying to draw conclusions based on outcome indicators alone, and the need to have complementary process and structure indicators. For example, in the case of AMI and Stroke it would be informative to have evidence on the administration of anticoagulants, and whether the patient was treated in a Stroke unit. While this study is limited to the extent it can make conclusions about the relationships between the pairs of conditions studied, it does suggest that the methodology and risk-adjusted outcome measures are sensitive enough to be able to detect some of the subtle causal relationships suggested in the literature.

Many policy makers consider hospitals the nucleus of the health system, possibly because they are responsible for a substantial proportion of health care spending. Yet, endeavours to measure hospital performance over time or across institutions are challenging, due to the diversity of services they provide, and multiple factors which influence their performance, such as technological innovation and personal skill. In practice many hospital indicator frameworks use a combination of indicators to assess performance. Often mortality and readmission rates are included, these may be overall rates or ones that express the outcomes of the specific services that are provided in the hospital (e.g orthopaedic surgery) of specific types of patients (AMI, Stroke). In addition, measures of throughput, such as waiting times and the average length of hospital stay are commonly used to measure responsiveness and efficiency respectively. However, when used alone these indicators they may ignore the ultimate effectiveness or appropriateness of the intervention, especially in situations where health services are dealing with patients with multi-morbidities, and from different socio-economic backgrounds. Overall, we find that it is difficult to make generalizations about quality of providers based on outcome indicators alone, and for separate conditions. While AMI is associated with many of the other conditions we are interested in the relationships between them are often more complex than they initially appear. Moreover, in many cases where associations are identified they apply only to a small number of patients. However, our findings suggest that the VAR model is well suited for understanding health provider performance as allows us to model the endogeneity inherent in the provision of health care, and use this in order to determine casual effects and relationships between performance indicators and across time.

Part III

Evaluating Quality

5 The effect of Payment by Results on quality of care

5.1 Introduction

As part of the Blair Government's health reforms, in 2003/04 the English NHS moved from a bulk contract system of funding hospital episodes to a DRG type system known as "Payment by Results" (PbR) (Audit Commission, 2004). Under this system Primary Care Trusts (PCTs) – the commissioning agencies in the English NHS - reimburse hospitals for each procedure they perform though a national tariff based on Healthcare Resource Groups (HRGs), the English version of DRGs. This type of provider payment, typically referred to as case payment or activity payment, has been increasingly adopted in many health care systems because of the positive effects it can have on cost containment and transparency. Indeed in Chapter 1 we review the theoretical underpinnings this type of payment system and consider how is able to incentivize providers to deliver more efficient, transparent care within a reasonable cost. However, in Chapter 1 we also note that both theory and experience indicate a number of other unwanted effects that can also arise from this type of system, depending on various design, organizational and system factors such as how cases are defined, how the tariff is set, the organizational setting in which it is applied and the structure of the health system to which it is applied. It is important to understand how and when these unwanted effects emerge, especially when they can have adverse quality implications for patients.

The US Medicare program was the first to introduce DRG payments, under the Prospective Payment System (PPS) in 1983, where hospitals were reimbursed a fixed amount per patient based on reported diagnosis. Since its adoption there has been considerable literature documenting the effects of this payment mechanism on different areas of performance. Various studies reported a decrease in activity (Davis and Rhodes, 1988; Guterman et al., 1988; Kahn et al., 1990; Rosenberg, 2001). With regards to quality of care, studies report mixed findings. Kahn et al. (1990) identify improved processes of care being applied for CCF, MI, Pneumonia, Hip Replacement and Cerebrovascular accidents following the adoption of PPS, that result in decreases in mortality. However, in future work Kahn et al. (1992) report evidence of patients being admitted to hospital sicker and released less stable than before the adoption of DRGs, while again noting improvement in

processes of care, declines in in-hospital mortality as well as no change in post-discharge mortality. Wells et al. (1993) report mixed quality results for depressed elderly patients in the clinical setting, with moderate improvements since the introduction of DRGs, but poor quality for one third of patients at discharge. Cutler (1995) and Shen (2003) also reported a compression of deaths in short term hospital discharges, yet no change in the mortality of patients surviving past one year after discharge.

Some instances of selection were also reported from the US experience. Newhouse (1989) found that patients in unprofitable DRGs were more likely to be found in 'hospitals of last resort', suggesting patient selection by profitability. Similarly, there was a short term noticeable shift of treatment from DRG financed inpatient settings to outpatient clinics which were otherwise financed (Cutler and Reber, 1998; Ellis and Vidal-Fernandez, 2007; Newhouse and Byrne, 1988). Other evidence of patient selection was presented by Meltzer et al. (2002) who found greater cost decreases for high cost patients than low cost patients, mirrored by a pattern of reductions in more expensive DRGs. Similarly, Ellis and McGuire (1996) identified evidence of selection, under Medicaid's mental health services in New Hampshire where expenditures for the sickest patients were reduced under prospective payment.

Since Medicare's adoption of DRGs as the mechanism for paying hospitals in 1983, case payment systems have been increasingly adopted amongst industrialized countries. While the basic principles of the system remain intact, there is large variation amongst the motivation, system of design, and implementation of DRG systems in different countries (Busse et al., 2006). At present many countries complement their case-based funding by other forms of payment (such as fixed budgets), yet for some this is only until they fully implement case-based funding for all hospital costs (France, Germany and the Netherlands). The objectives for introducing case-based funding also vary: Sweden and Australia have introduced DRG payments as a method of reducing waiting times (Duckett, 1995; Rauner et al., 2003); Austria, Germany, France and Australia aim to use them to increase efficiency and reduce costs (Duckett, 1995; Rauner et al., 2003); and Sweden, France and the Netherlands mention increasing transparency as motivation.

In these countries, in-depth studies have also identified possible instances of adverse quality effects. In Australia, Duckett (1995) found a decrease in waiting times for urgent procedures and an increase in activity levels combined with a decrease in expenditures in Victoria hospitals in the year following DRG implementation. Patient selection was ruled out, as the overall average case weight had risen, yet concern was expressed about incentives for gaming and selection in the future. In 2000, Duckett and Jackson expressed concerns about patients being discharged 'quicker but sicker' from hospitals employing

case-based funding. In Austria, Sommersguter-Reichmann (2000) showed that ownership of hospitals influences the response to case-based funding. Upcoding was identified as was cost-shifting between inpatient and outpatient care. Rauner et al. (2003), also identified premature discharges of patients in case-funded hospitals with higher readmission rates, as well as cost-shifting and gaming by transferring patients to other hospitals before treatment was finished, while claiming the full reimbursement fee.

As case-based payment systems are increasingly adopted as the main mechanisms for paying hospitals it is important to assess the effects this policy has in different settings. These type of analyses will create a wider pool of evidence from which policy makers are able to draw better conclusions as to what organizational, system or design features are best suited to bring out more of the desirable features of the policy. This chapter attempts to understand what the effect the introduction of case-based payments in England, under PbR, have had on the quality of providers. The English case is interesting for many reasons. Since the full implementation of PbR, hospital incomes are exclusively determined by this payment mechanism. Apart from Germany, which is also in the process of applying a case payment system, no other country has experience with this. Moreover, in England, the tariff reflects average costs of cases alone. Aside from France, this is not the case of any other country. Most countries apply a more complex pricing system to provide all hospitals with an incentive to improve their performance (Street and Maynard, 2007). Finally, England adopted a case payment system after paying hospitals through bulk contracts and budgets, implementing a fast and large transition. These three factors combined make the English case unique and interesting to study both from a policy and research perspective.

While there is little dispute as to the case payment's ability to increase transparency and curb costs, there is ambiguity as to the exact effect it has on quality. Indeed much seems to depend on how widely the system is adopted, how the reimbursement is set and how strictly it is enforced, making it even more difficult to generalize findings from one context to another. The PbR system in England is one of the main pillars of the Blair Government's 'New NHS', created with the goal of providing "prompt, convenient, high quality services" (Secretary of State for Health, 2002). It is important to determine to what extent this change in payment has succeeded in meeting these objectives, not only from a policy perspective, but to gain insight as researchers as to what works and what does not. Part II of this thesis is concerned with creating robust and sensitive quality indicators that can be used for evaluating policy. This chapter allows us to apply the quality indicators we have created to a specific policy area, PbR, in order to draw conclusions as to its overall effectiveness. In the process we are able to demonstrate the indicator's value added in

drawing policy conclusions.

5.2 Background

The PbR policy is structured around the HRG, which is the case classification system and the basis of the national tariff. HRGs were developed in the early 1990s. At that time they were not used to reimburse providers, but primarily for benchmarking exercises and to set targets to encourage unit cost reductions (Street and Dawson, 2002). HRGs are designed to measure health care activity in a way that takes into account the diagnosis, mix and complexity of patients that will be receiving care. The basis of the national tariff is an average of all hospital HRG costs for the procedure in question. Separate tariffs exist for elective and emergency care, as well as for short-stay patients, while specialist work is excluded. Hospitals also received a separate payment, the Market Forces Factor, which is based on the geographical price indices for land, labor and building costs.

PbR's implementation began in 2003/4 where the tariff was first applied to marginal changes in output for 15 HRGs. The tariff was extended to a further 33 HRGs in 2004/5 (Farrar et al., 2009). For NHS foundation trusts – a new NHS organization introduced in April 2004 for high performing hospital trusts – PbR was fully introduced to all spells of care in 2004/5[1]. The policy was extended to included elective activity for all other NHS trusts 2005/6, and to non-elective and outpatient care from 2006/7 (Audit Commission, 2004, 2005). Errors in the 2006/7 tariff, published on 31 January 2006 resulted in a greater average increase of the tariff than initially designed. This led to the tariff being withdrawn and reissued on 17 March 2006 (Boyle, 2009). From this date onwards it has covered all elective and emergency activity in English hospitals.

As noted previously, case payment systems create a variety of incentives, including: increased activity for non complex patients, unit cost reductions, expenditure control, cream skimming, quality skimping and upcoding. From early on studies have being undertaken to examine the effects of the English PbR policy on length of stay, readmissions and volumes of inpatient and emergency activity. While case-based reforms are usually focused on improving the efficiency of health care delivery, they do raise concerns about their effects on quality of care, which may be adversely effected as evidence from the US experience has suggested (Cutler, 1995; Kahn et al., 1992; Shen, 2003). Moreover, the Audit Commission (2008) notes that 53% of doctors were wary of the quality effects the PbR policy would have, and while they also report increases in readmission rates between

[1]There were 29 foundation trusts by then end of 2004/5 and 34 by the end of 2005/6. Foundation trusts have more managerial and financial freedom and a different accountability regime than other trusts. For a more in depth discussion see Ham (2009).

2003/04 – 2006/07, yet do not find these directly attributable to the PbR policy. Farrar et al. (2009) also find evidence to suggest a mild increase in acute hospital activity along with a reduction in unit costs.

Research also suggests changes in the recording of inpatient activity of hospitals. After the first year of PbR, the Audit Commission (2005) reported little difference in activity growth or efficiency in foundation trusts apart from a small increase in length of stay. They report no evidence of gaming amongst the early implementers although do mention cases of perceived gaming having been reported by some PCTs, including resubmission of patients using old referrals, artificial discharge of payments and coding and/or undertaking multiple interventions that are unnecessary in order to increase revenues. Research has also indicated a change in year to year activity among cases (Farrar et al., 2009; Rogers et al., 2005; Sussex and Farrar, 2008) although in all authors note that it is unclear whether this represents a genuine change in activity or a change in the way activity is recorded. No links have been established between PbR and quality of care, even in cases where this has been tested (Farrar et al., 2009), although in all studies to date, authors have noted that the quality variables used (in-hospital mortality, 30-day post surgical mortality and emergency readmission after hip fracture) may not be sensitive enough to detect change.

In the previous section latent and filtered estimates of hospital quality were created and analysed over the period 1996 to 2008. In these chapters our findings suggest that the filtered estimates are more sensitive predictors of quality of care. Chapters 2 and 3 investigated the trajectory of average performance using the latent and filtered indicators over the period 1996-2008 and the variation of performance amongst hospitals at any single point in time using the different indicators. While both of these characteristics differ by indicator and condition, some trends were observed that would benefit from further investigation, namely a change in the variation of performance in the later years of the sample and the change in performance from year to year. We believe part of the changes we have observed earlier on may be related to the introduction of the PbR policy during this period.

5.3 Methodology

This chapter attempts to examine whether quality of care has been influenced by PbR through a two step process involving the quality indicators constructed in Part II. The first step of the analysis investigates the effect PbR has had on levels quality over time, this step will indicate how quality has changed since the implementation of PbR. However, it is also of interest to understand whether the system design has influenced quality. It has been suggested that as the tariff is set according to average cost it incentivizes 'average'

performance. The second step considers this issue by investigating what has occurred to relative hospital performance since the implementation of the policy.

In order to undertake step one it is first necessary to adjust the quality metric that will be used for the analysis over the time period. In Chapter 2 the latent estimates were created, these estimates indicate the marginal effect a hospital has on each outcome measure controlling for patient characteristics. Thus, each point estimated for every year represents the slope of the risk-adjusted quality curve. We use this information to create this curve, for each of the four outcome indicators for every hospital, spanning throughout the years in our sample. We use the value of zero as our starting point, however, one could easily substitute zero for the mortality rate of that hospital in the same year to get the true estimate. The filtered estimates created in Chapter 2, have also been used as measures of quality. Indeed we argue that they are better estimates of true quality as they are able to reduce more of the noise in each estimate. For this reason we apply the same technique to create similar filtered quality curves for all four indicators in every hospital.

Using these latent and filtered quality measures, we are then able to examine the effect PbR has had on each of these outcomes using the following model:

$$Q_{ht}^{k} = \alpha + \beta_1 T_{ht} + \beta_2 \sum X_{ht} + \beta_3 PbR + \beta_4 \sum C_{ht} + \epsilon_{ht} \tag{5.1}$$

$$q_{ht}^{k} = \alpha + \beta_1 T_{ht} + \beta_2 \sum X_{ht} + \beta_3 PbR + \beta_4 \sum C_{ht} + \epsilon_{ht} \tag{5.2}$$

The two models are estimated separately for each of the seven conditions in the sample. In the first model, represented by equation (5.1), Q_{ht} represents the filtered quality measures for each of the four quality variables, k, each hospital h, and each year, t, for each of the seven conditions being investigated. Similarly, the second model, represented by equation (5.2), q_{ht} represents the latent quality measures for each condition. The variable T_{ht} represents the average tariff received by each hospital for the patients admitted with the particular condition in question for every year in the sample. The control variables, $\sum X_{ht}$, indicate the total caseload and average deprivation, co-morbidity, age, length of stay of each hospital for every year, again for the particular condition being investigated. PbR represents the dummy variable included for PbR which takes a value of 1 after 2006, as that is when the policy became effective for all emergency and elective conditions in all hospitals. For AMI, which only looks at non-elective procedures, it is constructed to measure a change from 2005. Finally, $\sum C_{ht}$ indicates the hospital type, such as foundation trust, independent specialist treatment centre (ISTC) or teaching hospital. Because foundation trusts and ISTCs were only introduced in from 2004, and hospitals transitioned into foundation trusts during the period under investigation, these variables are not time invariant. However the teaching hospital variable will be static, as this did not change during the period under investigation.

The second step investigates the effect PbR has had on the variation in quality between hospitals in each year of our sample. The relative performance of hospitals cane be measured through the normalized latent estimates for each hospital (see Chapter 2). Recall, that for these estimates the mean latent value has been set to zero, such that any negative value for any single hospital represents its absolute level below average mortality, while any positive value represents how high it lies above average mortality. Moreover the estimates are normalized such that the mean value of every year is equal to zero. Thus the spread of the data indicates the relative variation from year to year. The equivalent filtered measures were also constructed (as done in Chapter 3) using the normalized latent measures.

Using these latent and filtered measures of variation in quality the following models are able to examine the effect PbR has had on the spread of the performance across time using the same model:

$$V_{ht}^{k} = \alpha + \beta_1 T_{ht} + \beta_2 \sum X_{ht} + \beta_3 PbR + \beta_4 \sum C_{ht} + \epsilon_{ht} \tag{5.3}$$

$$v_{ht}^{k} = \alpha + \beta_1 T_{ht} + \beta_2 \sum X_{ht} + \beta_3 PbR + \beta_4 \sum C_{ht} + \epsilon_{ht} \tag{5.4}$$

The explanatory variables indicated in equations (5.3) and (5.4) are the same as in equations (5.1) and (5.4), while the dependent variable V_{ht} in equation (5.3) represents the normalized latent quality estimates, for each of the four outcome measures k. Similarly, in equation (5.4), v_{ht} denotes the normalized filtered quality estimates. Both equations are estimated separately for each of the seven conditions. In the regressions for all four models above, year dummy variables were included in the analysis. All models were run using fixed effects, as a result all time-invariant characteristics (such as teaching hospital) were differenced out of the equation. The sensitivity analysis section reports the results of the random effects models, and any differences that were observed.

5.4 Data

The basic data used to conduct this analysis is the same as in Part II, and is reviewed in detail in the data section of Chapter 1. The dependent variables used in the analysis are the latent and filtered quality indicators constructed in Part II for the four quality indictors: 30-day in-hospital mortality, 365-day mortality, 28-day readmissions and 365-day readmissions. For equations (5.3) and (5.4) the latent and filtered measures have been normalized, such that the mean value of each year is set equal to zero. Thus, any positive value is indicative of above average performance and any negative value is indicative of below average performance. For more information on the construction and normalization of the filtered indicators see Chapter 3.

Some of the key characteristics that have emerged from the previous analyses are that different performance measures have different levels of persistence and exhibit different variation amongst hospitals. However, year-long mortality is almost always the most persistent variable across conditions. The associations amongst the different outcome measures also vary by condition for the most part the mortality variables and readmission variables are positively correlated amongst themselves, although in most cases not very strongly. However, mortality and readmission variables tend to be negatively correlated with one another, indicating that higher readmission variables may not always be indicative of worse quality. The independent variables used in this analysis are discussed below.

Tariff & PbR

Since their implementation in the US in 1983 many countries have begun to use DRGs in their health systems. Most countries using DRGs, have found it necessary to modify their design features in order to better suit the institutional, political, and socio-demographic features of their health care systems (Busse et al., 2006). In this vein, English HRGs are the English modified version of DRGs, adopted by the Department of Health and Social Security in the late 1990s (The British Medical Association, 2008). HRGs are an English measure of case-mix which allow a clinically meaningful grouping where resource use can be expected to be roughly the same, and thus can have a particular cost ascribed to it. The HRG case-mix is constructed using the ICD-10 diagnostic codes and the Office of Population Censuses and Surveys Surgical Operations and Procedures (OPCS) version 4.3 procedural classification codes, while HRG costs are derived from national reference cost exercises. Street and Dawson (2002) note that due to(NHS, 2005) the organizational structure of the NHS and its lack of a substantial private insurance sector that would require detailed billing data, there is no history of routine patient level cost data collection. Moreover efforts in the mid-eighties to encourage such activities towards this end failed to spark an interest and were abandoned. As a result most hospitals cost activity on the basis of top-down allocations.

In the Department of Health jargon, the term 'episode' refers to any single procedure/condition being treated in the hospital. Any patient may undergo more than one episode from the time they are admitted to the hospital until they are discharged. The total episodes a patient undergoes in this period is referred to as the 'Hospital Provider Spell' (HPS), and is the main unit of currency under PbR. Within an HPS, there may be episodes that are excluded from the national tariff (such as rehabilitation). The term 'PbR spell' is used to identify the episodes which are covered by the tariff, and is derived from the HPS by the Department of Health. The national tariff is based on the dominant

HRG for the HPS. The dominant HRG is determined by considering a Department of Health provided ranking of episode level HRGs[2].

In order to calculate the cost for the key HRGs within each point of delivery, comparative costing data is collected based on a nationally agreed upon format specified in the NHS costing manual[3] (Department of Health, 2008). This manual specifies how hospitals should collect their annual statements of hospital cost structures and providing a breakdown of expenditure by specialities, services or programmes within the hospitals. Using this share of total hospital expenditure per 'patient treatment service' and dividing by the total number of bed days occupied by patients, the average cost per bed day is estimated. In order to calculated the HRGs within that speciality the number of bed days for each HRG is multiplied by the speciality cost per bed day. This bottom-up costing is discussed in detail in Street and Dawson (2002).

Each NHS provider is then required to select the HRGs that cover at least 80% of cost and activity at each point of delivery allowing the costing method to focus on the small number of HRGs that represent a high proportion the cost (Department of Health, 2008). This selection is performed using a so called 'grouper' that is provided by the NHS Information Authority. Data with excess bed days beyond a defined trim-point – calculated based on the national distribution of length of stay for each HRG – are excluded from these costs and reported separately in order to prevent outliers from skewing the data. Over time HRG groupings have been refined and different versions have been released which are able to explain more the variation in length of stay. The current version being used is HRG 3.5, in place since the financial year 2003/4. Previous versions include HRG 3.0 and HRG 3.1, constructed and collected by the Department of Health since May 1997. The national tariff is based 2-year retrospective returns. The newer HRG4 grouper has been in use for reference costs since April 2007 (for financial year 2006/7 onwards) and for Payment by Results (PbR) since April 2009 (for financial year 2009 onwards). However this lies beyond our sample of data.

The average HRG costs averaged across all NHS Trusts form the basis for the calculation of the national tariff for each HRG. For any patient whose length of stay exceeds the trim point a per-diem payment will be added to the payment. Per-diem payments are HRG specific and differ for elective and non-elective activity. The national tariff also considers local cost differences and adds what is called a Market Forces Factor (MFF).

[2]In a multi-episode spell where there is more than one occurrence of the highest ranked HRG, the first episode is taken as dominant. Since only spell level information or HRG related data is derived from the dominant episode, no issues arise from this (National Health Service, 2005).

[3]Since 1994, when efforts to encourage the use of HRGs in the care contracting process were first made official (NHS executive, 1994), multiple guidance documents have been produced to assist hospitals in apportioning costs to HRGs in this manner.

The MFF is specific to the each hospital and considers the unavoidable cost differences associated with the provision of services in different parts of the country. It is constructed by combining three separate indices which reflect labour and capital differences: the staff index, the building index and the land index.

The academic literature on case payment systems discusses the different incentives associated with alternative tariff setting mechanisms (Ellis and McGuire, 1996; Schreyögg et al., 2006; Street and Maynard, 2007). As discussed previously the English tariff reflects average costs alone, and is calculated from cost data gathered from all hospitals. This will encourage hospitals to become 'average' rather than to strive to become considerably more efficient (Street and Maynard, 2007). The English case reports the tariff in monetary units, unlike other countries which separate price and underlying cost by using a points system where policy makers decide how much to pay per point (Schreyögg et al., 2006).

At the time of their development in the early 1990s, HRGs were not used to reimburse providers, but primarily for benchmarking exercises and to set targets to encourage unit cost reductions (Street and Dawson, 2002). Currently the PbR tariff is payable for admitted patient care (elective, non-elective and emergency), outpatient attendances and accident and emergency admissions. The individual level data provided by Dr Foster in our dataset contains information on the HRG tariff throughout the entire period of the sample. Note that while this tariff does reflect approximate unit costs of the patient it is not the paid to each hospital for the entire period, but only from the years PbR was phased in. The variable included in our regression reflects the average tariff received by each hospital for the patients treated for the conditions specified in each year of the sample.

Caseload

The activity within a hospital, in terms of the number of cases treated, is likely to be related to many of the performance variables being investigated. The relationship between cases and outcomes is not clear. Increased caseload may result in lower quality due to overcrowding, or it can result in higher quality as doctors become more experienced. Moreover higher quality may lead to more cases as demand increases, or it can be the result of selecting fewer cases. Activity has also been linked to DRG type systems, although it is not always clear in what way caseload will be affected. The US experience with DRGs resulted in a decrease in hospital activity (Davis and Rhodes, 1988; Guterman et al., 1988; Kahn et al., 1990; Manton et al., 1993; Muller, 1993; Rosenberg, 2001), many other countries have reported increases in activity, such as Australia (Duckett, 1995; Healy et al., 2006), Denmark (Street and Maynard, 2007), Germany (Böcking et al., 2005; Hensen et al., 2008; Schreyögg et al., 2005) and England (Farrar et al., 2009; Rogers et al., 2005; Sussex

and Farrar, 2008). In the case of England, all authors note that the activity increase occurring in the period since PbR was implemented is hard to interpret as it may represent a genuine change in activity or a change in the way activity is recorded. The methodology used later on in this Part (in Chapter 6) allows us to explore the association between caseload and the quality, moreover it allows us to control for the change in cases when investigating the effects of PbR on quality. In this chapter, the variable used to measure caseload was estimated by the number of admissions recorded for each condition in each hospital every year, hospitals with less than 10 cases per year were dropped from the sample.

Length of stay

Length of stay (LOS) is an important indicator of performance, commonly used to measure utilization, efficiency and/or hospital management. It is unclear whether low LOS is indicative of better or worse quality and efficiency, high LOS may be a reflection of complexity of case mix (demand) or indeed poor discharge planning (supply). While some studies have indicated there is little, if any, relation between length of stay and outcome (Clarke, 1996; O'Brien, 2002), interest in the area remains high. Cost-containment efforts and payment mechanisms often use length of stay as a proxy for efficient use of resources, however it is unclear if lower LOS in these cases indicates increased effectiveness and better planning or inadequate discharge planning or premature discharges. Results from many countries indicate that a DRG type system is associated with a fall in average length of stay (Böcking et al., 2005; Davis and Rhodes, 1988; Guterman et al., 1988; Hensen et al., 2008; Kastberg and Siverbo, 2007; Kahn et al., 1990; Manton et al., 1993; Muller, 1993; Rosenberg, 2001; Schreyögg et al., 2005), including England (Audit Commission, 2008; Sussex and Farrar, 2008; Farrar et al., 2009). However, its effect on quality are often mixed, depending on the factors within the different systems that are driving the change in LOS. By including LOS in our model we hope to better understand its relation with quality, as well as to control for changes in LOS when testing for the effects of PbR on quality. The length of stay variable used in this dataset represents the average length of stay of patients per hospital for each condition. This average is estimated from the original patient level data, where there is information on the number of days each patient spend in the hospital from admission to discharge.

Average severity and deprivation of the patient population

Other hospital characteristics are likely to affect outcomes, such as the characteristics of the patients that they treat. A hospital treating older, sicker or more deprived patients,

for example, may have more patients with complications, longer length of stay and worse outcomes. However, in the construction of the latent estimates, which form the basis for the filtered estimates, we have already adjusted for co-morbidity, deprivation, age, gender and type of admission at the patient level. Thus by including these variables in our regression at this level, we are able to control for systematic biases that could still be present in the data set and not 'filtered' out correctly through the construction of the quality metrics in Part II. This means that we are controlling for these variables to see if treating a more severe or more deprived patient population has any overarching effects on hospital quality. For example, if the hospitals treating more deprived patient populations suffer from understaffing.

Hospital type

During the period of investigation the hospitals included in the sample studied could be classified into four types: acute trusts, teaching hospitals and foundation trusts and ISTCs. An NHS trust provides services on behalf of the English NHS. Each trust is headed by a board of directors consisting of non executive and executive directors. Trusts are split into commissioning and non-commissioning trusts. Secondary care services, such as the ones investigated in this paper, are provided by NHS hospital trusts, otherwise known as acute trusts. All of the hospitals in our sample are known as acute trusts. A teaching hospital is a hospital that provides clinical education and training to future and current doctors, nurses and other health professionals as well as investing in research and technology, in addition to delivering medical care to patients. Due to these different functions, this type of hospital is often singled out as having different objectives which can also contribute to quality improvement. However, under a system of price competition, teaching hospitals may be disadvantaged by these activities as they may be unable to drive costs down as easily as their competitors. Presumably in an attempt to avoid such problems, the English government chose not to take any measures to consolidate subsidies for research and technology with patient care till after the initial transition period of PbR's implementation (Boyle, 2009). In order to test for any differential effects on quality since PbR in our sample, a dummy variable was created to represent all teaching hospitals.

The idea of NHS foundation trusts developed in part as a way to place more emphasis on patient choice and provider competition that the Blair Government emphasized in their reform agenda published in 'The NHS Plan'(Secretary of State for Health, 1997) and 'Delivering the NHS Plan'(Secretary of State for Health, 2002). The original intention was to give high performing NHS trusts the opportunity to manage their services with less inference from the Department of Health and with greater involvement from local

communities, staff and other stakeholders. Moreover, NHS foundation trusts were established as public benefit corporations overseen by a new regulator, Monitor, as opposed to the Department of Health. NHS foundation trusts have freedoms available to them which acute trusts do not. These include the ability to retain operating surpluses and borrow money from public and private sectors; to establish private companies and joint ventures; to vary staff pay from nationally determined terms and conditions. Foundation trusts are expected to provide services in accordance to the specifications set out in the service agreements negotiated with PCTs, and are regulated by the Care Quality Commission, similar to acute trusts, but they are also required to comply to the terms of their authorization as determined by Monitor (Ham, 2009). The government expects all trusts to become foundation trusts once they have proved their ability to run services as a public benefit corporation. The first NHS foundation trusts were established in 2004 and increased year by year.

There is little evidence on the performance of foundation trusts, especially with regards to PbR. Early evidence suggested that foundation status was a weak signal for strong financial management (Mannion et al., 2008) With regards to PbR evidence suggests that that foundation trusts have seen a greater increase in short stay inpatient admissions from accidents and emergencies than non-foundation trusts Rogers et al. (2005), but also shorter lengths of stay (Farrar et al., 2009). However, as well performing trusts are awarded with foundation trust status and given the different regulatory and financial structures of foundation trusts, which could influence quality. A dummy variable was constructed to examine control for foundation trusts in all equations. Each trust coded as a foundation trust from the year that the trust became one, if it became one during the years investigated.

The last type of hospital considered in our investigation is an ISTC. While hospitals are made up of many departments that share labour and capital resources and treat both emergency and elective patients, treatment centres are dedicated to the provision of elective care. They are not required to provide emergency treatment and are designed to specialize in one or two high-volume procedures and avoid taking on complex operations. In 2003 the Department of Health decided to establish treatment centres throughout England and to prioritize areas with high waiting times. Shortly after this decision was made, it was decided that private, or "independent sector", providers would be allowed to establish treatment centers to treat NHS patients. By 2008 there were about 100 treatment centers operating, with about half opened by the private sector (**?**). Evidence suggests that treatment centres are treating patients from less deprived areas, who have less diagnoses and undergo less procedures than patients treated in hospitals (Mason et al., 2010). As our sample deals with non-elective admissions, there is a very small amount of ISTCs but

a dummy variable is created to control for hospitals with this status, coded as an ISTC from the year that it was opened. In most cases there were not enough ISTCs in the sample, and so they are not included in the analysis.

5.5 Results

The results section is separated by condition, in this section the methodology as applied to AMI, Stroke and Hip Replacement is presented, while the results for MI, IHD, CCF and TIA are presented in Appendix D. The first part of the results section for each condition plots of the average latent and filtered quality measures over time. Using the latent outcome measures constructed in Chapter 2, which measure the marginal effect a hospital has on each outcome measure for every year, and the filtered measures created in Chapter 3 from these latent indicators we are able to construct the latent and filtered outcomes over time. For each condition, the first figure presents these measures, normalized to zero for the first year of the sample. These plots are of interest as they are the dependent variables used in models 1 (equation (5.1)) and 2 (equation (5.2)), but also in order to observe the differences between the latent and filtered estimates. This is followed by a table with the results of models 1 and 2, which examine the effect of PbR on the eight latent and filtered quality measures. The results of this section will allow us to determine whether quality has improved, stayed the same, or worsened since the introduction of PbR, separately for each condition.

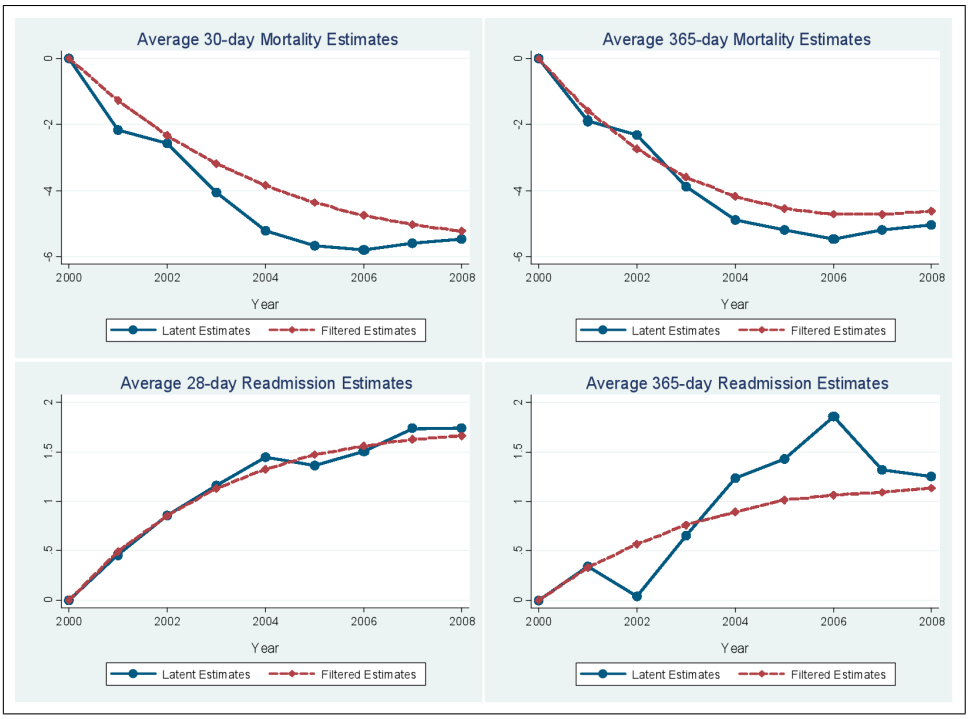
The second part of the results for each condition provide evidence on the change in relative performance of hospitals in each year. This is of interest because tariffs have been set at average cost, and thus in theory might be incentivizing average performance. The results for this section are also separated by condition, presented together with the results for the first two models. Two more plots are presented which illustrate the ranked normalized latent and filtered outcome indicators for each hospital respectively, as bound by 95% confidence intervals for each estimate, for three different years: 2002, 2005 and 2008. These years are selected as the reflect a year before the introduction of PbR (2002), a year in the middle of the rolling out of the policy (2005) and the last year of our sample (2008). The x-axis of each plot indicates the hospital rank for that year, while the y-axis shows the actual value of the outcome measure (latent or filtered). Each point on the graph represents the hospitals value, and is extended to show the values within the 95% confidence interval. Recall that the latent measures have been normalized, and the filtered estimates are constructed from the normalized latent measures. Thus, a value of zero represents the average values, while any point below zero indicates outcomes below average outcomes and any value above zero illustrates outcomes above average. These

figures are useful for visualizing the change in relative performance of hospitals over time, which is used as a dependent variable in models 3 (equation (5.3)) and 4 (equation (5.4)), but also to observe the difference between the normalized latent and filtered outcome measures. Finally the last table presents the results of models 3 and 4, which examine the effect of PbR on the variation of performance amongst hospitals.

AMI

As mentioned previously, the methodology uses the latent and filtered estimates of each hospital for each year, as calculated in Chapters 2 and 3 to construct hospital quality estimates over the entire time period available. Figure 5.1 illustrates these quality estimates over time. The latent and filtered curves plotted indicate the same overall trend in mortality and readmissions over the time period investigated. The filtered curves are able to filter out much of the noise in the latent estimates and thus show a smoother trajectory across time. The top left hand panel shows the latent and filtered plots for 30-day mortality. Both curves indicate a decline of about 5% in the period 2000-2008. The latent curve (solid line), indicates different rates of decline throughout the period, which are much more pronounced from 2000-2005, and indeed slightly increasing from 2006-2008. As the filtered curve (dashed line) smoothes out the time series it does not pick this up, and shows a much smoother decline across all years. These characteristics can also be observed for the year-long mortality plots illustrated in the top right hand panel. In this plot mortality also falls by around 5%, and the filtered curve smoothes out much of the spikes in the latent curve. Both bottom panels, illustrating the trend in 28-day and year-long readmissions, show an increase in readmissions. 28-day readmissions increase by about 1.5%, while year long readmissions increase by just over 1%. Similarly to the mortality panels, the filtered estimates provide much smoother curves which smooth out the variation amongst the different years. This is most pronounced for year-long readmissions, where the filtered curve is noticeably smoother and does not show a large interim increase in readmissions of 1% in the period 2003-2005.

Figure 5.1: Average hospital quality over time for AMI.



Models 1 and 2 explore how the hospital level indicators, used to create the hospital average plotted above, are influenced by the introduction of PbR. The results of these models are presented in Table 5.1. The R-squared values vary considerably between the models, ranging from 31% for year-long mortality to 8% for year-long latent readmissions. The PbR dummy is significant in all models except the latent 28-day readmission model. In all mortality models it indicates that since PbR has been implemented, mortality has fallen. The magnitude of the coefficients indicate that PbR is responsible for a $4-5\%$ decline in 30-day mortality and year long mortality. The coefficients on the readmission models, suggest that since PbR there has been an increase in short and long term readmissions of around $1-2\%$.

The tariff variable is only significant for the latent 28-day readmission model, and indicates that a higher tariff is associated with lower readmissions. The average LOS of AMI patients in each hospital is significant for latent 30-day mortality, latent year-long mortality and filtered year-long readmissions, such that an increase in LOS is associated with higher mortality and lower readmissions. Caseload is significant for filtered 30-day in hospital and year long mortality as well as for latent and filtered 28-day readmissions. In the filtered models the sign of the caseload variable is positive suggesting that more cases is associated with higher mortality and lower readmissions. However, the sign on the latent 28-day readmission variable is positive, suggesting that more cases are associated

with higher readmissions.

Average co-morbidity and deprivation of AMI patients admitted to each hospital are both significant in many of the models. Average co-morbidity is significant in all filtered models, as well as the latent 30-day in hospital mortality model. The sign on the latter coefficient is negative indicating that higher average co-morbidity is associated with lower 30-day mortality, while the signs on all filtered mortality models are positive suggesting that higher average co-morbidity is associated with higher mortality. Average deprivation is positively associated with the dependent variable in all latent and filtered mortality models, indicating that hospitals with higher numbers of deprived patients have higher mortality. Similarly, the deprivation coefficients are negative for the readmissions models indicating that hospitals with more deprived patients have higher readmissions. The age and foundation trust status variables are not significant in any of the models. Teaching status was dropped from the model as it is time invariant, and the model was run with fixed effects.

Table 5.1: AMI Models 1 & 2.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.00100	0.000370	0.00237	0.000203	-0.00272**	-0.000150	-0.000699	2.20e-05
	(0.00107)	(0.000604)	(0.00236)	(0.000744)	(0.00137)	(0.000190)	(0.00185)	(0.000133)
Age	-0.127	-0.0223	-0.351	-0.0531	-0.198	0.0183	0.0164	0.0292
	(0.498)	(0.139)	(0.454)	(0.204)	(0.236)	(0.0514)	(0.434)	(0.0371)
LOS	0.284**	0.0450	0.900***	0.125	-0.0819	-0.0183	-0.317	-0.0459***
	(0.134)	(0.0786)	(0.156)	(0.0952)	(0.0952)	(0.0250)	(0.344)	(0.0169)
Cases	-0.00468	0.00372*	-0.0103	0.00616*	0.00670*	-0.00150*	0.00908	-0.000991
	(0.00538)	(0.00220)	(0.00621)	(0.00326)	(0.00402)	(0.000814)	(0.00594)	(0.000667)
Co-morbidity	-7.868**	2.274**	0.204	3.623**	-5.061	-0.778*	1.490	-0.910**
	(3.168)	(1.062)	(4.242)	(1.547)	(3.853)	(0.421)	(2.421)	(0.416)
Deprivation	8.200*	0.652*	12.47**	0.987**	-3.930*	-0.233*	-3.467	-0.193*
	(4.477)	(0.335)	(5.614)	(0.483)	(2.335)	(0.125)	(2.625)	(0.111)
FT	-1.082	-0.318	-1.329	-0.353	0.524	0.116	1.176	0.151
	(1.225)	(0.664)	(1.570)	(0.974)	(1.153)	(0.246)	(1.773)	(0.186)
PbR(05)	-4.032***	-4.957***	-4.931***	-5.487***	2.308***	1.680***	0.821	1.197***
	(0.921)	(0.595)	(1.232)	(0.906)	(0.749)	(0.244)	(1.194)	(0.216)
Constant	14.67	-4.606	8.663	-5.391	29.62**	1.015	-0.0579	0.00295
	(34.68)	(9.923)	(26.78)	(14.27)	(11.43)	(3.669)	(31.03)	(2.599)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
N	1,071	1,071	1,071	1,071	1,071	1,071	1,071	1,071
$\boldsymbol{R}^2$	0.270	0.305	0.312	0.166	0.146	0.236	0.080	0.191
Hospitals	119	119	119	119	119	119	119	119

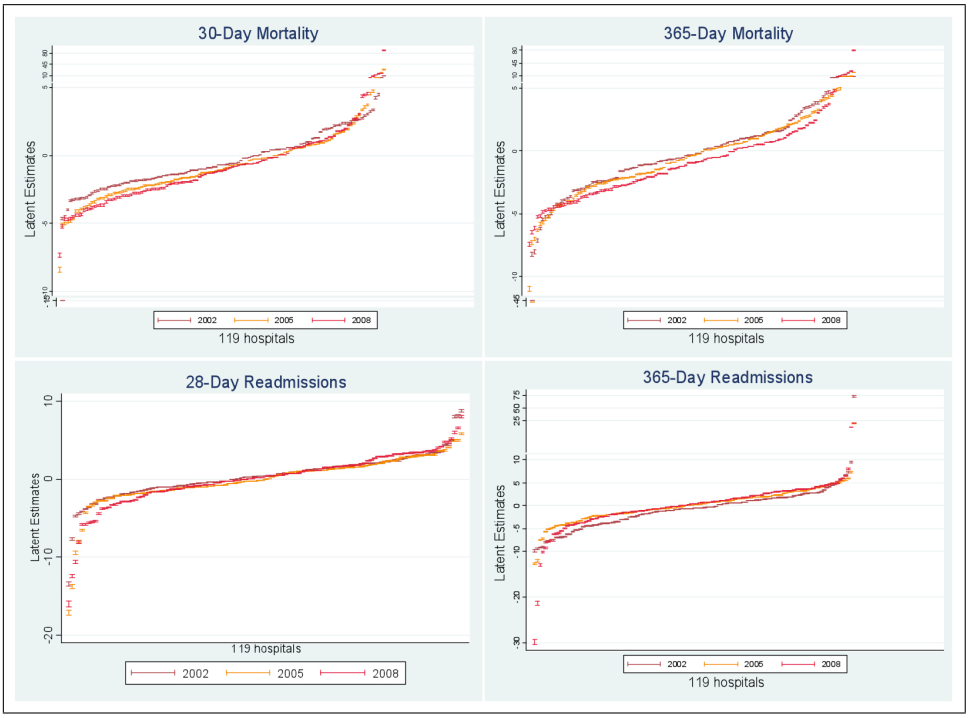
Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

The second part of the methodology considers how PbR has influenced the relative performance amongst the hospitals in our sample over time. By ranking the hospitals according to the different outcome indicators and plotting them for different years, we are able to observe whether any overall changes in performance have occurred. Figures 5.2 and 5.3 illustrate these plots using the latent and filtered outcomes measures for AMI, plotted for the years 2002, 2005 and 2008. In Figure 5.2, we are able to see the relative performance of hospitals in the different years with regards to the four latent outcomes. As the latent estimates are subject to large volatility (see Chapters 2 and 3), they have very large outliers. This makes the figures difficult to interpret. In addition, while there are some differences amongst hospitals between the years, these are very small. However, the plots with the filtered indicators (Figure 3) indicate a very different pattern. First off, we notice that the filtered indicators have fewer, and much lower outliers. Secondly the confidence intervals on each hospital estimate are much larger than those on the latent indicators. Both of these characteristics were also observed in Chapter 3.

By examining each of the four panels in Figures 5.2 and 5.3 we are able to observe the trends in relative performance between the years plotted. The top left hand corner of both figures plots the 119 hospitals according to their ranking on the latent 30-day mortality indicator in each year. In Figure 5.2, we see that from 2002 to 2008 there has been a gradual drop in mortality among the top performers such that the line has shifted to the right, indicating better performance overall. However, the curve also appears to have shifted across the axis, such that all points are further from zero, indicating that there is more variation in performance across hospitals later on in time. The same trend can also be observed in the top right hand panel for the latent year-long mortality variables, although there is more overlap at the tail ends of each curve. It is more difficult to observe any noticeable change in the readmission curves as they are largely overlapping for the years plotted. The plots in Figure 5.3, show a much more distinct change in relative performance across the years plotted. All of the plots the filtered outcome plots indicate a convergence to the mean from 2002-2008. In both the short and long term mortality plots, the convergence to the mean, is most noticeable amongst the best performers, who level out gradually from 2002 to 2005 and then from 2005 to 2008. We see a similar convergence

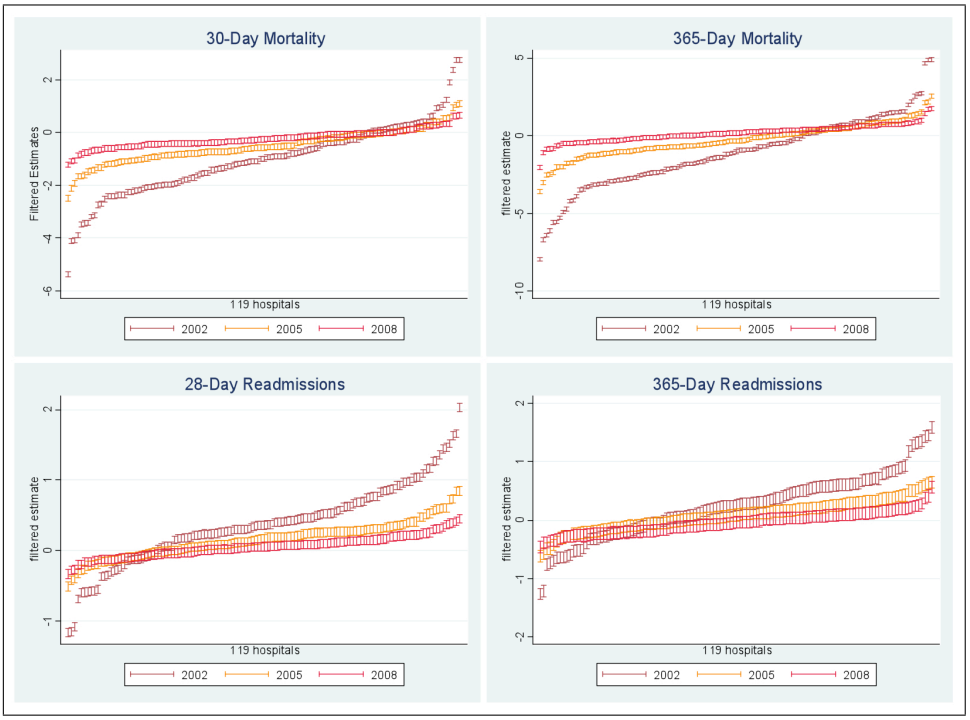
for short and long term readmissions, only for these indicators a higher number of hospitals were achieving readmissions above the average rate, as a result of the convergence, readmissions in the worst performers are levelling out towards the mean value.

Figure 5.2: Relative hospital performance over time for AMI (normalized latent outcome indicators).



Using the normalized latent and filtered outcome indicators for all years in the sample as dependent variables, we are able to evaluate the impact of PbR on the relative performance of hospitals over time. The results of these models, Models 3 and 4, are presented in Table 5.2 for all eight outcome indicators. The R-squared values indicate how much of the overall variance in outcomes the different models are able to predict. In most cases this value indicates that the model performs reasonably well. Generally, there is not that much difference in the R-squared values between the latent and filtered models, although the filtered models perform slightly better in most cases. There is one instance of large variation between the two models in the 28-day readmissions model, where the latent model predicts nearly 2% of the variance while the filtered model is able to predict about 20%.

Figure 5.3: Relative hospital performance over time for AMI (normalized filtered outcome indicators).



PbR is significantly associated with the relative performance of hospitals in all models except filtered year-long readmissions. The coefficient of the PbR dummy is positive for all mortality models, indicating that since PbR normalized mortality has been increasing relative to the mean. Recall that the latent estimates (and thus by extension the filtered estimates) have been normalized such that the mean in every year is equal to zero. Thus the positive coefficient is explaining the increase in mortality over time as illustrated in Figure 5.3, where the hospitals with negative latent and filtered values are approaching zero. However, this only tells us that relative performance in each year is approaching the mean and not how the mean has been changing over time. The PbR coefficient is negative for all short term readmission models, and latent year long readmissions. Again due to the normalization of estimates within each year, this result only tells us about the relative distribution of performance across time. The negative result indicates that since the adoption of PbR readmissions in every year are falling towards the mean.

The tariff variable is significant for all of the latent mortality measures, but not for the filtered measures. The tariff variable is constructed as the average tariff received by each hospital in each year. Part of the association between tariff and quality for the latent measures will thus be picking up information on the amount of more 'expensive' patients admitted each year. As the filtered measures incorporate information from different parts

of the time series and smooth out the values, they will be less sensitive to this information. The negative sign on the tariff coefficient indicates that hospitals with higher average tariffs are associated with better mortality outcomes within each year. If we consider that higher average tariffs are truly a result of higher numbers of expensive patients, than we can interpret this result as hospitals treating more severe patients have better outcomes. Interestingly, our variable for co-morbidity is only significant for the filtered measure, and also negative. As a higher score on the Charlson Co-morbidity score is indicative of more co-morbidity, this indicates that hospitals treating a more severe patient population on average have lower mortality outcomes. Co-morbidity is also significant for the filtered readmission measures, only in these cases the sign is positive, indicating that the higher the co-morbidity in the patient population being treated, the lower the readmissions for that hospital.

The average deprivation level of the hospital caseload also proves to be a significant predictor of hospital quality. While it is significant for both the latent and filtered outcomes, the signs are different. Deprivation is positively correlated with the latent outcome indicators and negatively correlated with the filtered indicators. The deprivation indicator measures the average Carstairs score of the patients treated in each hospital each year. A higher Carstairs score is indicative of higher levels of deprivation. Thus, the latent model find that hospitals treating a higher number of deprived patients have lower mortalities, with the filtered model suggests that these hospitals have higher mortalities. For the readmission models, only the filtered models are significant, and they suggest that the hospitals treating a higher number of deprived patients have lower readmissions. Age is not significant in either of the models, as would be expected given that this has been controlled for in the construction of both indicators.

The number of cases is significant at the 10% level for the latent short term mortality and readmission indicators, such that fewer cases are associated with lower mortality. This variable is not significant for the filtered outcome indicators. Caseload is also positively associated with short and long term latent readmissions, such that an increase in cases leads to higher readmission rates. This result is also indicated for the filtered long-term readmissions model. LOS is significantly associated with latent year-long mortality, such that longer length of stay is associated with higher mortality. Latent year-long readmissions are also correlated with LOS at the 10% level. The positive sign indicates that higher LOS is linked to lower year-long readmissions. As the model was run using fixed effects, all the time invariant hospital characteristics were dropped from the model (such as teaching hospital status). Because foundation trusts were only introduced from 2005 onwards, they were not dropped from the model. However, the dummy for foundation

trusts is only significant for the latent year-long mortality model, where it is associated with lower mortality rates.

Table 5.2: AMI Models 3 & 4.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	-0.00099***	2.73e-05	-0.00174*	-0.000271	-1.72e-05	6.99e-05	0.00279	0.000327
	(0.000359)	(9.21e-05)	(0.000972)	(0.000287)	(0.000347)	(8.57e-05)	(0.00196)	(0.000262)
Age	-0.433	0.0294	-0.174	0.0319	0.114	-0.0160	0.739	-0.00597
	(0.359)	(0.0230)	(0.413)	(0.0449)	(0.0759)	(0.0131)	(0.493)	(0.0211)
LOS	0.0231	-0.0184	0.735***	0.0174	-0.00286	-0.00501	-0.553*	-0.0440
	(0.0820)	(0.0115)	(0.201)	(0.0382)	(0.0508)	(0.00992)	(0.299)	(0.0337)
Cases	-0.00453*	-0.000211	-0.0094***	-0.00105	0.00250*	0.000141	0.00686**	0.000624**
	(0.00265)	(0.000488)	(0.00255)	(0.000759)	(0.00129)	(0.000223)	(0.00274)	(0.000300)
Co-morbidity	1.373	-0.399***	-3.462*	-0.733***	-0.782	0.172*	-2.291	0.232**
	(2.362)	(0.152)	(2.031)	(0.269)	(0.555)	(0.100)	(1.434)	(0.112)
Deprivation	6.087**	-0.136***	1.979**	-0.195***	0.245	0.0603**	0.386	0.0517*
	(2.457)	(0.0460)	(0.942)	(0.0724)	(0.395)	(0.0264)	(0.559)	(0.0283)
FT	-0.464	0.00513	-1.150**	0.0349	0.155	-0.0401	0.310	-0.0717
	(0.578)	(0.121)	(0.555)	(0.203)	(0.366)	(0.0611)	(0.540)	(0.0723)
PbR(05)	2.140***	0.121***	3.227***	0.213***	-0.944**	-0.0603***	-1.341***	-0.0377
	(0.481)	(0.0351)	(0.557)	(0.0575)	(0.367)	(0.0178)	(0.483)	(0.0364)
Constant	30.10	-1.873	15.73	-0.394	-6.629	0.807	-52.71	-0.746
	(20.51)	(1.585)	(26.53)	(3.258)	(5.204)	(0.976)	(35.22)	(1.894)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,071	952	1,071	952	1,071	952	1,071	952
R^2	0.306	0.310	0.274	0.335	0.018	0.198	0.104	0.108
Hospitals	119	119	119	119	119	119	119	119

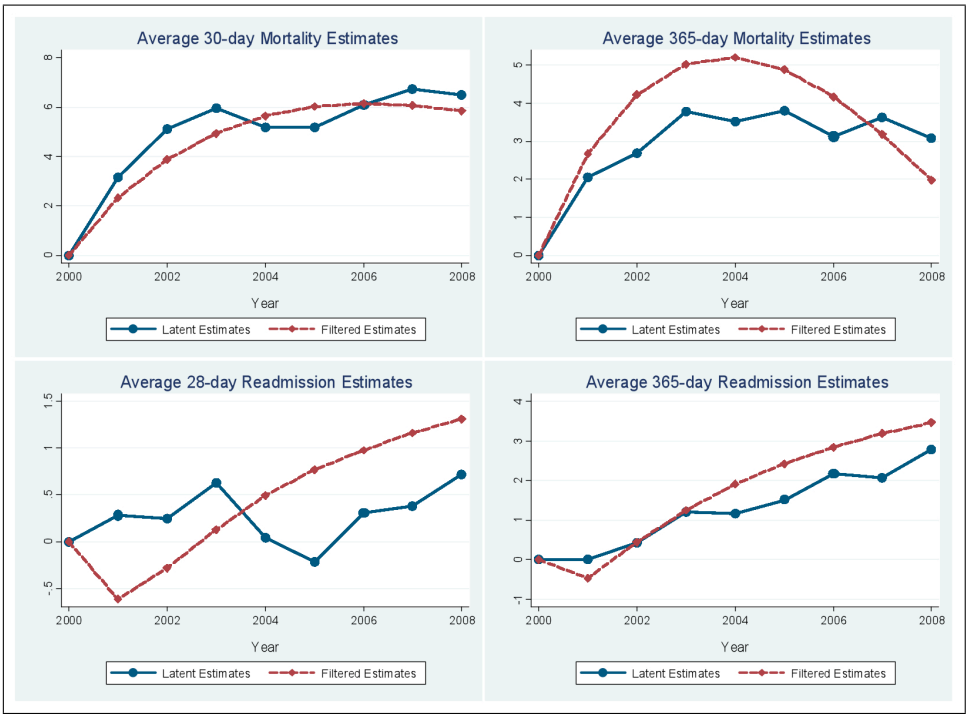
Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Stroke

The latent and filtered measures used as measures of Stroke quality, calculated for the different outcomes in models 1 and 2, are plotted in Figure 5.4. Unlike the corresponding figure for AMI (Figure 5.1), the filtered and latent estimates often show quite a different trajectory, indicating more noise in the latent Stroke measures than in the latent AMI measures. Looking at the top left hand panel of Figure 5.4, we are able to observe that 30-

day in hospital mortality for Stroke has increased, on average, by about 6% during the time period studied. The top right hand panel also indicates an increase in year long mortality, although there is a larger difference between the filtered and latent estimates, with the later reporting a 3% increase by the end of the sample, and the former a 2% increase. Also the filtered estimate shows a different average mortality trajectory, indicating a much larger initial rise in mortality from 2000 to 2004, which then falls in the remaining period, although not down to its initial 2000 level.

Figure 5.4: Average hospital quality over time for Stroke.



The bottom left hand panel indicates the change in 28-day readmissions. While both latent and filtered estimates indicate a rise over the 2000-2008 period, again the amount differs. The filtered estimates indicate a rise of about 1.25%, while the latent estimates indicate a rise of around 0.7%. Again the trajectory demonstrated in average performance of this indicator is very different for the filtered and latent curve. Finally the bottom right hand panel indicates the change over time in year-long readmissions. This too shows an increase in readmissions of around 3% for both latent and filtered estimates. As with the filtered estimates in the short term readmissions panel, the filtered curve indicates a larger increase in readmissions than the latent curve.

The latent and filtered quality measures, whose averages are plotted in Figure 5.4, are used as the dependent variable for models 1 and 2. The results of these models are presented in Table 5.3, and indicate how the outcomes are influenced by the adoption

of PbR and other explanatory variables. The R-squared values indicate that the models do not explain a large amount of the variance in the dependent variables, but range considerably from 5% to 41%. In all cases the filtered models perform better than the latent measures, which is to be expected given the properties of the filtered measures as noted in Chapter 3.

The PbR dummy is significant in most models indicating that the policy is having an effect on outcomes. PbR is not significant for latent 30-day mortality, and only significant at 10% for filtered 30-day mortality. The coefficient on the filtered mortality indicator is positive suggesting that since the policy mortality has increased by nearly 2%. The coefficient on the latent and filtered year long mortality suggests that since PbR long term mortality has fallen. The filtered model indicates that this decline is around 5%, where as the latent model shows a different magnitudes, of about 9% for year long mortality. The PbR dummy is also significant for latent and filtered 28-day readmissions as well as filtered year-long readmissions. The coefficient on the latent variable indicates a 2% fall in short term readmissions since PbR, however the coefficients on the two filtered readmission variables indicate a rise in readmissions, of around 0.5% for short term readmissions and over 2% in long term readmissions.

Few of the other explanatory variables are significant in the models. Average tariff, age, caseload, deprivation and foundation trust status are not significant in any of the models. Average co-morbidity of patients treated in hospitals is significant in all of the latent mortality models, and for the latent year-long readmission model. It indicates that more severe patients have lower mortality and higher readmissions. Average length of stay is significant at the 10% level for latent 30-day mortality, as well as for year-long readmissions and significant at the 5% level for latent short term readmissions. Its coefficient is negative in all these models, indicating that higher average LOS is associated with lower mortality and readmissions.

Table 5.3: Stroke Models 1 & 2.

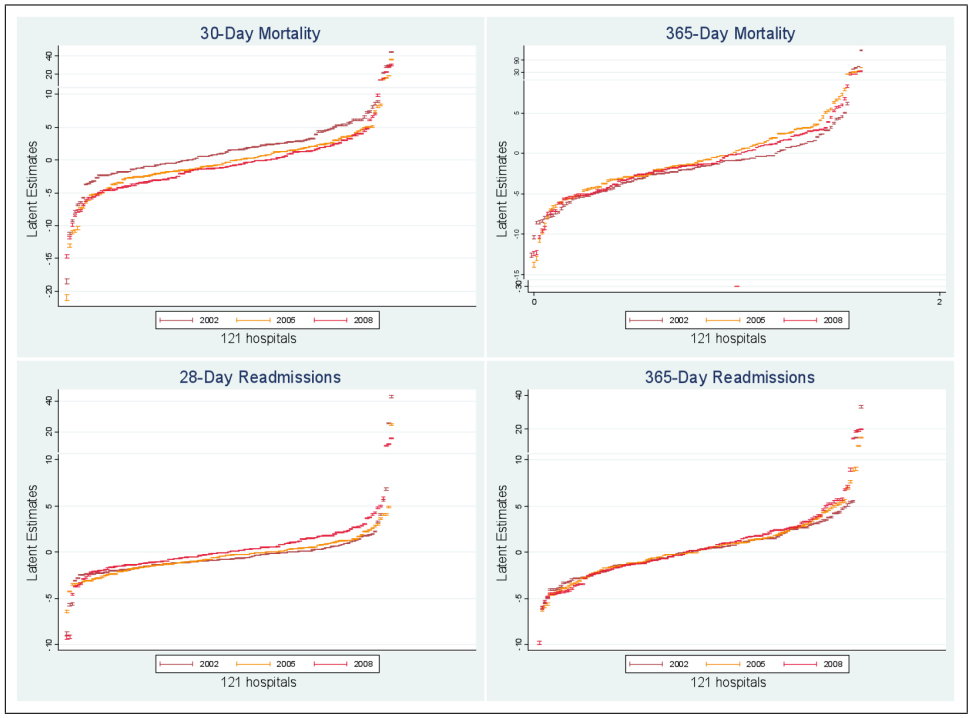
	Latent	Filtered	Latent	Filtered	Latent	Filtered	Latent	Filtered
	$D30_{ht}$	$D30_{ht}$	$D365_{ht}$	$D365_{ht}$	$R28_{ht}$	$R28_{ht}$	$R365_{ht}$	$R365_{ht}$
Tariff	0.00275	-7.73e-05	0.00108	-0.000154	0.000611	2.55e-05	0.000800	2.22e-05
	(0.00266)	(0.000472)	(0.00238)	(0.000695)	(0.000702)	(0.000102)	(0.00125)	(0.000235)
Age	0.308	-0.0231	0.151	-0.0451	-0.182	0.00119	-0.168	0.00946
	(0.298)	(0.0983)	(0.303)	(0.152)	(0.146)	(0.0247)	(0.239)	(0.0460)
LOS	-0.481*	0.0381	-0.282	0.0370	-0.162**	0.00206	-0.333*	0.0247
	(0.281)	(0.0750)	(0.274)	(0.109)	(0.0757)	(0.0145)	(0.179)	(0.0345)

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Cases	0.00636	-4.84e-05	0.00313	-0.000566	0.00116	-0.000417	0.00235	9.44e-05
	(0.00846)	(0.00170)	(0.00767)	(0.00262)	(0.00226)	(0.000412)	(0.00412)	(0.000844)
Co-morbidity	-13.21***	0.825	-11.76***	2.065	1.998	0.206	4.531*	-0.221
	(4.073)	(0.958)	(4.225)	(1.425)	(1.691)	(0.245)	(2.545)	(0.512)
Deprivation	0.821	-0.185	0.594	-0.255	-0.825	0.0319	-0.819	-0.000105
	(1.779)	(0.561)	(2.700)	(0.878)	(0.668)	(0.129)	(1.443)	(0.236)
FT	0.911	0.657	-0.0790	1.228	-0.462	0.114	1.611	0.217
	(2.176)	(0.565)	(2.454)	(0.866)	(0.701)	(0.128)	(1.367)	(0.246)
PbR(06)	-1.279	1.742*	-9.362**	-4.926***	-2.313***	0.564***	-2.971	2.336***
	(4.931)	(0.961)	(4.662)	(1.449)	(0.775)	(0.207)	(1.914)	(0.434)
Constant	-5.429	-0.183	8.024	0.128	11.10	-0.413	9.326	-1.148
	(22.16)	(7.390)	(23.66)	(11.36)	(11.30)	(1.865)	(17.96)	(3.657)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,035	1,035	1,035	1,035	1,035	1,035	1,035	1,035
$\boldsymbol{R}^2$	0.085	0.128	0.159	0.223	0.097	0.411	0.052	0.432
Hospitals	115	115	115	115	115	115	115	115

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

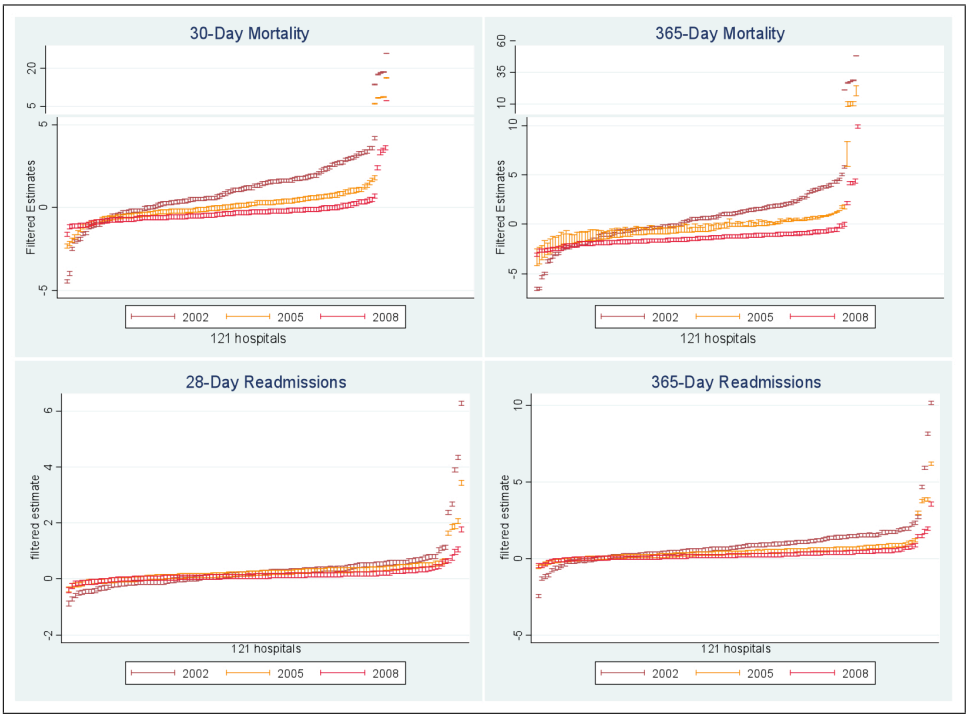
Figures 5.5 and 5.6 present the relative performance over time for Stroke, as measured by the latent and filtered outcome indicators respectively. Similar to AMI, the latent outcome indicators have more extreme outliers than the filtered measures and much smaller confidence intervals around each hospitals outcome. However, the trends in relative hospital performance in the different years are different. From 2002 to 2008, the latent figures show a gradual decline in short term mortality, and a gradual increase in short term readmissions. It is harder to discern a pattern for the long term performance indicators as there is considerable overlap in the lines, however, it does appear that long term mortality is lower in 2005 and 2008 as compared to 2002.

Figure 5.5: Relative hospital performance over time for Stroke (normalized latent outcome indicators).



The filtered indicators in Figure 5.6 are easier to read, partly because there are fewer outliers and performance is smoothed out. The confidence intervals for each of the hospitals are larger, indicating the higher levels of uncertainty associated with the filtered estimates. Similar to AMI there is a convergence to the mean for all filtered outcome measures. However unlike AMI where the mortality of the below average performers approached zero in each year, in Stroke the hospitals with above average mortality approach zero in the later years of the sample. This same pattern can be observed for the readmission estimates, although to a smaller degree. Moreover a closer look at the year-long mortality panel reveals that estimates in each year, and especially in 2008 are not converging towards the mean (zero) but towards a value below it. Thus indicating that hospitals are converging to slightly below average mortality in 2008.

Figure 5.6: Relative hospital performance over time for Stroke (normalized filtered outcome indicators).



The normalized latent and filtered outcome indicators are used in models 3 and 4 to understand the change in yearly relative performance of hospitals since PbR. The R-squared values indicated that for each outcome, the filtered models are much better at explaining the variance in the filtered outcomes. The R-squared values range from just over 40% to just over 70% for the filtered models, compared to a range of nearly 5% to 40%. The PbR dummy is significant for all mortality models. Where significant, the coefficient is always negative. This indicates that since PbR, relative mortality rates have been declining. The PbR coefficient is only significant for filtered long term readmissions, and negative indicating the same effect for relative performance with regards to readmissions.

The tariff variable is only significant for the filtered mortality models, and suggests that tariff is positively associated with mortality, such that an increase in tariff is associated with worse mortality rates. co-morbidity is only significant for latent year-long mortality and readmissions, where higher co-morbidity is associated with higher readmissions. Length of stay is only significant in the latent mortality models, where longer length of stay is associated with lower mortality. Similarly caseload is associated with lower latent mortality, and higher latent year-long readmissions. Age and deprivation are not significant in any of the models. Foundation trusts are associated with higher latent short-term mortality and lower filtered short-term readmissions.

Table 5.4: Stroke Models 3 & 4.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.000791**	-0.000199	0.000982**	-9.76e-05	1.25e-05	2.43e-05	-0.000326	-7.61e-05
	(0.000365)	(0.000142)	(0.000489)	(0.000165)	(0.000239)	(3.35e-05)	(0.000355)	(9.31e-05)
Age	0.0488	0.0261	0.243	0.0593	-0.159	0.00193	-0.00620	-0.0140
	(0.111)	(0.0338)	(0.167)	(0.0460)	(0.156)	(0.00675)	(0.119)	(0.0161)
LOS	-0.168***	0.00990	-0.139*	0.00677	-0.0449	0.00263	-0.0272	0.00431
	(0.0508)	(0.0175)	(0.0777)	(0.0224)	(0.0278)	(0.00466)	(0.0673)	(0.00992)
Cases	-0.00486***	-5.18e-05	-0.00536***	0.000374	0.00111	-0.000102	0.00500**	-0.000447
	(0.00142)	(0.000575)	(0.00168)	(0.000695)	(0.00114)	(0.000134)	(0.00196)	(0.000388)
Co-morbidity	-0.722	0.169	-3.266***	-0.247	-0.190	-0.114	-1.047	0.303
	(1.070)	(0.303)	(0.834)	(0.385)	(0.424)	(0.0853)	(0.848)	(0.209)
Deprivation	-0.562	-0.0136	0.605	-0.0283	0.746	-0.0251	1.154	-0.0733
	(0.735)	(0.175)	(0.891)	(0.230)	(0.658)	(0.0709)	(1.092)	(0.104)
FT	0.703*	-0.107	0.643	-0.289	-0.283	-0.0790**	-0.0253	-0.0387
	(0.369)	(0.142)	(0.417)	(0.204)	(0.243)	(0.0325)	(0.489)	(0.0820)
PbR(06)	-3.364***	-0.389***	-1.285*	-0.439***	0.285	-0.0302	-0.302	-0.172***
	(0.568)	(0.0879)	(0.700)	(0.115)	(0.405)	(0.0282)	(0.720)	(0.0593)
Constant	2.427	-1.490	-11.21	-4.786	12.35	0.123	1.911	1.406
	(9.091)	(2.431)	(12.55)	(3.243)	(11.56)	(0.546)	(9.565)	(1.277)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,035	920	1,035	920	1,035	920	1,035	920
R^2	0.224	0.485	0.129	0.507	0.080	0.704	0.048	0.404
Hospitals	115	115	115	115	115	115	115	115

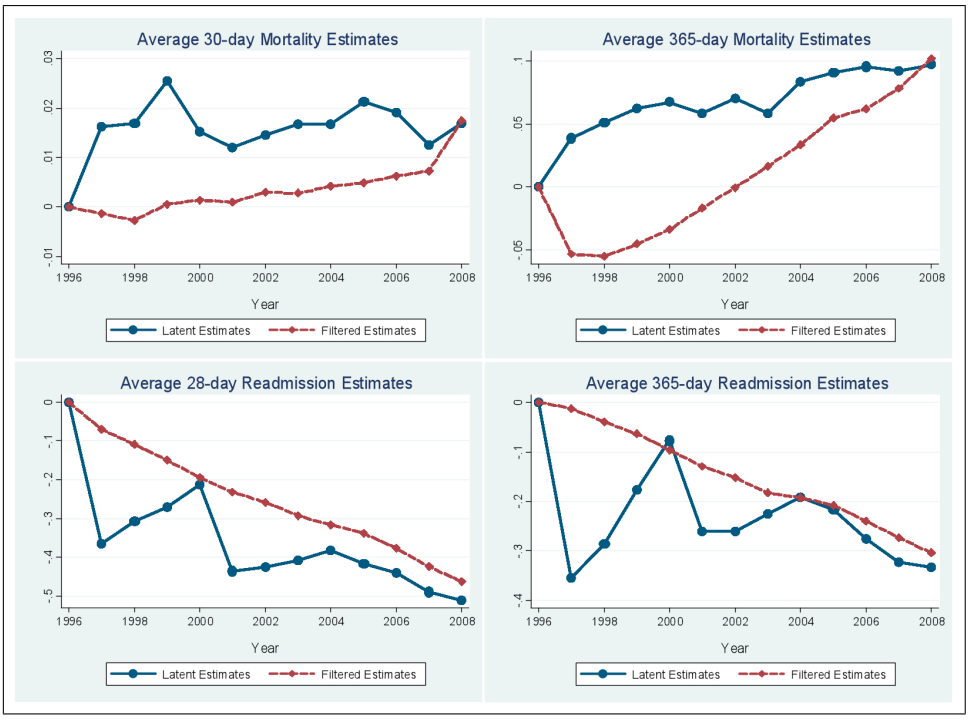
Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Hip Replacement

The sample of data for Hip Replacement is longer for the other conditions, ranging from 1996-2008. The latent and filtered quality indicators for this time period are plotted in Figure 5.7, using the slopes calculated from the latent measures in Chapter 2. The figures indicate a very small increase in short and long term mortality, over the entire time period, of 0.1% and 1% respectively. Moreover, they indicate a decline in readmissions of about 5% for 28-day readmissions and 3% for year-long readmissions. In all four panels the trajectory of the latent and filtered estimates are very different, with the filtered estimates

being much smoother, but also in the case of mortality, showing a different trend over time.

Figure 5.7: Average hospital quality over time for Hip Replacement.



The results for models 1 and 2 are presented in Table 5.5. These models use the latent and filtered outcome measures presented in Figure 5.7 to understand what factors have contributed to the change in outcomes over time. The R-squared estimates indicate that most models explain less than 6% of the variance in the dependent variables. The exceptions to this is the filtered year-long mortality model which is able to explain nearly 13% of the variance in the dependent variable. Across all conditions, the the filtered model has a higher R-squared value than the latent model. The PbR dummy is not significant for most models, aside from the filtered short and long term readmission models, where it is highly significant. The coefficients for these models indicate that since PbR short and long term readmissions have fallen by around 0.05%.

Many of the other explanatory variables are also not significant across the different conditions. Average tariff is not significant for any of the models, while average patient age is only significant at 10% for filtered 30-day mortality. The coefficient indicates that an increase in age leads to and increase in mortality. Average LOS is significant for latent 30-day mortality at the 10% level, showing a negative association. Caseload is positively associated with filtered 30-day mortality, but negatively associated with latent year-long mortality. Average co-morbidity is positively associated with latent and filtered 30-day and

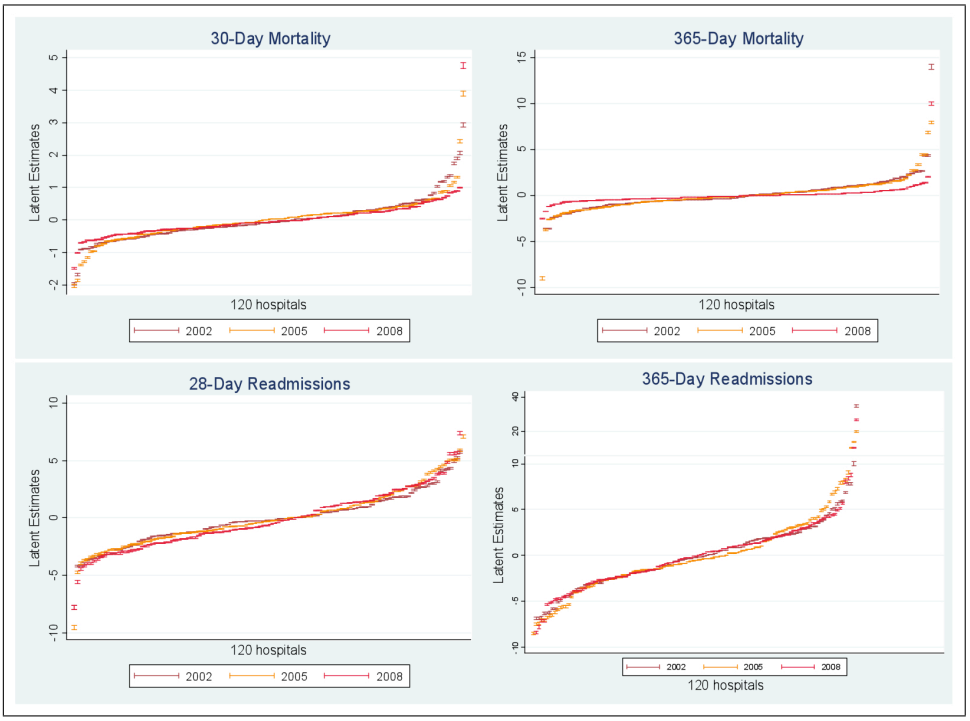
year-long mortality, such that hospitals with more severe patients have higher mortality rates. Average deprivation is also significant, for latent 30-day mortality and filtered year-long mortality, such that hospitals with more deprived patients have lower latent short term mortality, but higher filtered year-long mortality. Finally, hospitals with foundation status are significantly associated with higher filtered year long mortality, and filtered short and long term readmissions.

Table 5.5: Hip Models 1 & 2.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.000218	-1.82e-07	-0.000918	-2.76e-05	-0.000365	2.54e-05	-0.00121	6.87e-05
	(0.000270)	(9.04e-06)	(0.000621)	(2.17e-05)	(0.00114)	(3.35e-05)	(0.00169)	(4.69e-05)
Age	0.0363	0.00261*	0.0387	0.000375	0.137	-0.00525	-0.0315	-0.0110
	(0.0462)	(0.00136)	(0.139)	(0.00336)	(0.246)	(0.00588)	(0.290)	(0.00831)
LOS	-0.115*	0.00253	-0.0646	0.00507	0.127	-0.00905	0.419	-0.0141
	(0.0597)	(0.00184)	(0.142)	(0.00410)	(0.274)	(0.00730)	(0.423)	(0.0103)
Cases	-0.000717	6.83e-05**	-0.0037**	7.61e-05	0.000207	-0.000158	-0.00116	-4.03e-05
	(0.000615)	(3.21e-05)	(0.00152)	(9.56e-05)	(0.00277)	(0.000155)	(0.00417)	(0.000205)
Co-morbidity	1.282*	0.0568*	3.486**	0.152*	0.886	-0.0995	0.451	-0.204
	(0.745)	(0.0327)	(1.439)	(0.0795)	(3.219)	(0.129)	(4.834)	(0.158)
Deprivation	-0.748**	0.0186	-1.057	0.0421*	-0.0675	-0.0121	-0.653	-0.0554
	(0.371)	(0.0127)	(1.110)	(0.0251)	(0.907)	(0.0486)	(1.556)	(0.0576)
FT	-0.118	-0.00229	-0.233	0.0349*	1.506	0.0666**	2.020	0.0806*
	(0.165)	(0.00868)	(0.326)	(0.0205)	(0.966)	(0.0318)	(1.405)	(0.0410)
PbR(06)	-0.122	0.00324	-0.137	0.00632	-0.0970	-0.0495***	-0.00600	-0.0463***
	(0.0890)	(0.00405)	(0.151)	(0.00940)	(0.339)	(0.0125)	(0.480)	(0.0156)
Constant	-2.931	-0.237**	3.751	0.0806	-9.625	0.0439	4.633	0.334
	(4.037)	(0.110)	(7.350)	(0.277)	(16.68)	(0.470)	(20.24)	(0.640)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	826	826	826	826	826	826	826	826
$\boldsymbol{R}^2$	0.051	0.057	0.060	0.128	0.015	0.226	0.019	0.109
Hospitals	118	118	118	118	118	118	118	118

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Figure 5.8: Relative hospital performance over time for Hip Replacement (normalized latent outcome indicators).

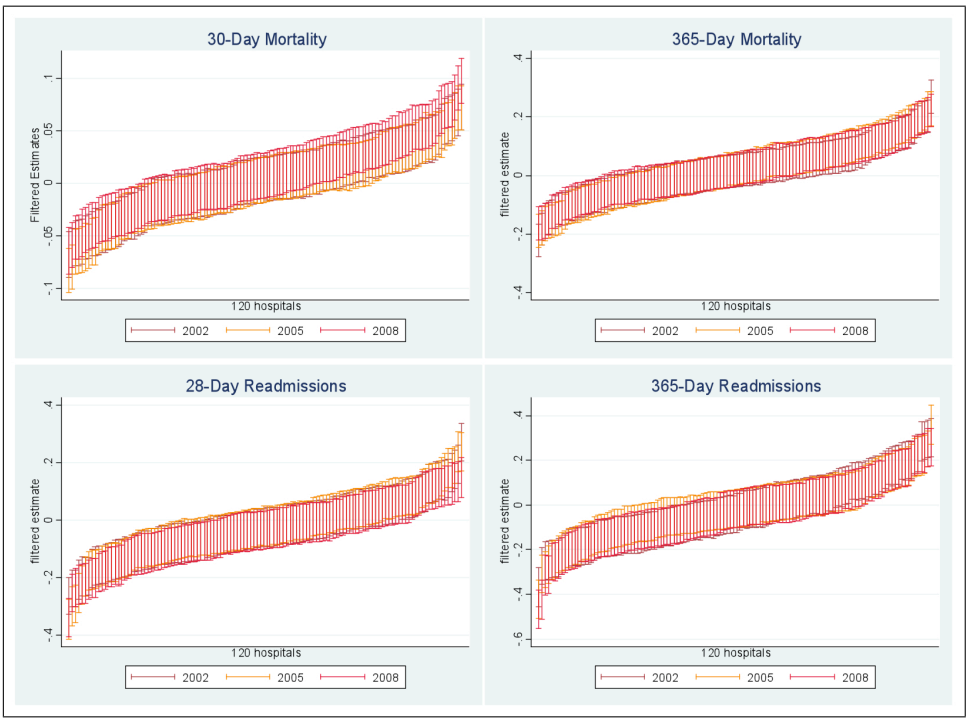


Relative hospital performance over selected years in the sample are illustrated in Figures 5.8 and 5.9. Figure 5.8 plots the hospitals ranked by the different latent outcome measures in 2002, 2005 and 2008. While this plot shows fewer outliers than the corresponding figures for AMI and Stroke, they are still apparent at the far ends of the distribution. Moreover, the range of the latent estimates, plotted on the y-axis, is much smaller than those of other conditions. This indicates relatively little variation between hospitals in the outcomes. The only exception to this is for year-long readmissions, where there is more variation than other Hip outcomes. Similar to the other plots of this type, for the other conditions, the confidence intervals for each hospital estimate are very small. Using the curves to try to make a comparison of relative performance over time is very difficult as the lines tend to overlap considerably. Only for year-long mortality is it possible to see that the indicators at the ends of the distribution have converged closer to the mean in 2008, as compared to the other two years.

The same plots, using the filtered indicators, are constructed and presented in Figure 5.9. The confidence intervals are larger for the filtered measures as compared to the latent measures. Also, the range of the indicators is also much smaller, such that there are no outliers for any of the conditions. Moreover, when comparing the relative performance of hospitals across the different years using the filtered indicators, it is still difficult to

detect any apparent change. In fact the three curves for each year overlap almost entirely. Interestingly, there does not appear to be any convergence to the mean for any of the Hip Replacement outcome indicators, as we saw for AMI and Stroke. Indeed, looking closely at the upper left hand panel, showing the trend in 30-day mortality we see the opposite. The outcomes for 2008 are moving away from the mean as compared to 2005 and 2002. However this change is very small and does not seem to be occurring in the other three panels, where it is difficult to detect any substantive change.

Figure 5.9: Relative hospital performance over time for Hip Replacement (normalized filtered outcome indicators).



Models 3 and 4, specified in equations (5.3) and (5.4), use the filtered and latent outcome indicators plotted in Figures 5.8 and 5.9 as the dependent variable. The results from these models are presented in Table 5.6. Similar to the results in Table 5.4, the R-squared values are quite low suggesting that the latent models only explain between 1-8% of the variance in most outcomes. Moreover the PbR dummy is only significant for 28-day filtered mortality, where it indicates a decrease in readmissions. Tariff, co-morbidity, and foundation trust status are all insignificant for all conditions. Caseload is significant in both filtered readmission models, suggesting that increased cases are associated with declining readmissions. Similarly, LOS is significantly positively associated with filtered short term readmissions. Finally deprivation is negatively associated with both latent readmissions such that an increase in deprivation is associated with an increase in readmissions.

Table 5.6: Hip Replacement Models 3 & 4.

	Latent $D30_{ht}$	Filtered $D30ht$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	-0.000199	7.08e-06	-0.00101	7.34e-06	0.000334	-1.07e-05	0.00172	-2.34e-05
	(0.000193)	(5.01e-06)	(0.000658)	(1.31e-05)	(0.000650)	(1.36e-05)	(0.00115)	(2.16e-05)
Age	0.150***	0.000532	0.277**	0.000836	0.299**	0.00151	0.302	-0.00455
	(0.0391)	(0.000860)	(0.114)	(0.00256)	(0.143)	(0.00279)	(0.209)	(0.00353)
LOS	0.0408	-0.000403	0.152*	-0.00378	0.0325	-0.00606**	-0.0463	-0.00189
	(0.0312)	(0.00100)	(0.0850)	(0.00256)	(0.164)	(0.00266)	(0.149)	(0.00404)
Cases	0.000558	2.94e-05	-5.43e-05	-7.47e-05	-0.000684	-0.000194***	-0.00196	-0.000154**
	(0.000344)	(2.28e-05)	(0.000830)	(5.52e-05)	(0.00167)	(5.41e-05)	(0.00237)	(6.33e-05)
Co-morbidity	0.204	0.0245	2.762	0.0497	-1.026	0.0582	-3.144	0.0508
	(0.596)	(0.0189)	(2.843)	(0.0495)	(3.001)	(0.0586)	(5.065)	(0.0751)
Deprivation	0.241	-0.00965*	0.0460	-0.0190	-1.254**	0.00748	-2.244***	0.0152
	(0.307)	(0.00556)	(0.433)	(0.0189)	(0.585)	(0.0153)	(0.631)	(0.0179)
FT	-0.0153	-0.00211	-0.153	-0.00723	-0.253	0.0131	0.127	0.0406**
	(0.0984)	(0.00463)	(0.239)	(0.0124)	(0.399)	(0.0144)	(0.465)	(0.0195)
PbR(06)	0.0132	-0.000219	0.0277	-0.0179	-0.00520	-0.0242*	-0.163	-0.0214
	(0.0972)	(0.00447)	(0.211)	(0.0125)	(0.337)	(0.0129)	(0.427)	(0.0156)
Constant	-9.888***	-0.0968	-15.60**	-0.0343	-23.32**	0.0619	-30.51*	0.509*
	(2.580)	(0.0653)	(6.103)	(0.211)	(11.68)	(0.214)	(17.77)	(0.274)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	826	826	826	826	826	826	826	826
$\boldsymbol{R}^2$	0.077	0.023	0.084	0.011	0.057	0.044	0.077	0.033
Hospitals	118	118	118	118	118	118	118	118

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Sensitivity Analysis

In order to make sure the results were consistent, but also to test if the effect of time invariant characteristics, such as teaching status of hospitals, the models were also run with random effects. The coefficients for these regressions are presented in Appendix D.5. Indeed the results are consistent, such that the PbR dummies and other explanatory

variables are significant for the same models indicating similar effects. Teaching status like foundation trust status in not significant in all models, and varies considerably by condition. It is significantly associated with higher AMI and Hip readmissions and lower MI, CCF and Stroke mortality. The coefficients on the latent Stroke mortality outcomes suggest that teaching hospitals have nearly 10% lower latent mortality than acute care trusts. However, only in the MI model is it associated with the filtered outcome measures.

As patient selection is a commonly cited effect of case-payment systems we also ran a model with interaction variables between the PbR dummy and the co-morbidity variable, and between the PbR dummy and the deprivation variable. The coefficients for all models are also presented in Appendix D.5. For many of the conditions the interaction effects show that indeed the quality changes indicated in the results section have do to with changes in the quality of severe or deprived patients. For AMI, the results of the interaction model suggest that while the average deprivation of patients admitted to each hospital is not a significant determinant of filtered quality before the policy, it is after. Indeed the interaction terms show that since PbR hospitals treating more deprived patients experience an increase in all mortalities and a decrease in short term readmissions. The interaction terms for IHD, like AMI, show that prior to the policy hospitals with more severe patients had higher filtered mortality outcomes overall, and lower year-long filtered readmissions. Since PbR, hospitals with higher amounts of severe patients have lower mortalities and higher year long readmissions. Also hospitals treating more deprived patients have higher mortalities since PbR, but there is not a significant difference between quality, as indicated by the mortality models, in these hospitals prior to the policy.

The model for Stroke also indicates that prior to PbR, hospitals with a more severe patient population had higher mortalities. Since PbR, the coefficient on the interaction term suggests that hospitals treating more severe patients have lower mortalities, and lower readmissions. Yet the effect of PbR alone, as indicated by the PbR dummy, is to increase 30-day mortality as well as short and long term readmissions. This appears to indicate some sort of specialization, where certain hospitals known to be better at treating Stroke patients get more severe patients and have better results. More in depth research would be necessary to make any substantive conclusions on this.

The Hip interaction model drops the PbR coefficient due to collinearity, and so we are unable to interpret the interaction effect. While, the interaction models for the remaining conditions, MI, CCF and TIA, indicate marginal or no effects of a significant interaction effect. In the case of MI the interaction terms indicate no difference in quality between the hospitals with different amounts of severe and deprived patients since PbR. The CCF and TIA models indicate the interaction effect is only significant for readmissions, such

that year-long readmissions for hospitals treating more deprived patients are higher since PbR for CCF and short term readmissions for hospitals treating more deprived patients are lower since PbR for TIA.

5.6 Discussion

A payment systems is concerned with how, and how much, health care providers are paid. The mode of payment can create powerful incentives affecting a provider's behaviour and these changes in turn will affect quality, quantity, costs, choice of alternative medical recommendations to various patients, and hence the allocative efficiency, of health interventions. While payment systems date back to the foundations of medicine itself, new payment systems emerge as a result of a need to address topical issues, such as rising costs or demands for increased transparency and accountability. There are various positive and negative incentives attached to any payment system. Different modes of payment are thus important to consider not only on a theoretical basis, but also in practice as they will be shaped by wider system and organizational factors.

When looking at the monetary incentives within any single health system, we may find different incentives, or different effects of policies, on different conditions due to the varying degrees of ease associated with reducing costs and/or improving quality for diverse medical conditions. Indeed this is discussed in much more detail in Chapter 4, where we see that the treatment quality of hospitals for different conditions may not always be linked to the same factors. Thus the interpretation of quality change in one area is not always straightforward to interpret. This chapter attempts to evaluate the effect PbR has had on the quality of hospitals by assessing the change in our latent and filtered outcomes since the policy's implementation in the seven conditions studied throughout this Thesis. Our results confirm that the policy has had differential effects across the seven conditions.

Models 1 (equation (5.1)) and 2 (equation (5.2)) investigate the latent and filtered measures of all hospitals using a fixed effects model in order to determine if and how the quality indicators have changed since the implementation of the PbR policy. The dependent variables for Models 3 (equation (5.3)) and 4 (equation (5.4)) are slightly different, in that they represent each hospitals relative performance in every year. However, all dependent variables are calculated using the same underlying methodology. The first interesting result in all four models was that the filtered and latent outcome models did not always indicate the same effect. In almost all cases the coefficients had a different magnitude indicating that the policy explained a different amount of change in quality. This is to be expected as the filtered estimates, by construction, smooth out the latent estimates and have a much smaller ranges. Thus the coefficients on the PbR dummy in these models

will, in most cases, indicate a smaller change in quality.

There were also instances where the PbR dummy was significant in the latent model but not the filtered model or the other way around. Given that the filtered variables have been show to have a much stronger signal to noise ratio (Chapter 3), we can assume that where the latent variables are significant and the filtered measures are not the model is picking up noise in the estimates, however in the reverse scenario the noise is obstructing the true effect. The later effect seems to be especially pronounced for the readmission models, where in most cases the latent models are not significant while the filtered models are.

The construction of the filtered estimates takes into account the variable's time series information as well as its relationship with the other indicators. This adjustment eliminates considerable measurement error, still present in the latent variables, and smooths out large amounts of their variation. Of the four outcome variables studied, our analysis in Chapter 3 found that year-long readmissions tend to be very noisy while the year-long mortality estimates almost always have the strongest signal. Indeed, in models 1 and 2, the PbR dummy is significant in the filtered year-long readmissions models and not the latent models for all seven conditions. However, when looking at the year-long mortality models, this is only the case for CCF, which of the seven conditions is the only one where the signal to ratio estimate for year-long mortality is relatively weak.

The third difference between the PbR effect on the filtered and latent models was in instances where both models were significant but their coefficients had a different sign, indicating that the policy had different effect on quality. This only occurred in one instance, regarding Stroke 28-day readmissions. In this case, because of the filtered measure's ability to smooth out variations over time, and take into account relationships between the different outcome measures, we chose the interpretation of the filtered model as the 'true' effect on quality. Indeed the R-squared value of the filtered model, which is about 40%, suggests a better fit of the data compared to the latent model where it is about 10%. Moreover, in Models 1 and 2, for all of the outcome measures the R-squared estimates are higher for the filtered models than the latent models, indicating that it is a better fit to the data.

The differences in significance, magnitude and direction between the coefficients in the latent and filtered models do not only apply to the coefficient on the PbR dummy, but also to those of the other explanatory variables. We interpret these results with the same approach as described above for the PbR coefficient, such that when there are differences in significance, magnitude and direction the interpretation of the filtered model is preferred. Yet, when examining the other explanatory variables there are some other

interesting differences that appear. Recall that the latent measures, and by extension the filtered measures, have already been adjusted for some of the individual patient level characteristics such as patient age, co-morbidity and deprivation (see Chapter 2). However in many instances across the four models and various conditions, we observed that the average of these indicators for each hospital were significant. In these cases the models are suggesting that there are differences in quality between hospitals that have different types of patients, regardless of the differences this has on the individual patient. For example, for most of the outcome indicators and most conditions, hospitals that had a more deprived population, on average, had higher mortality outcomes. While more deprived patients are expected to have worse outcomes, for most conditions there appear to be some wider systematic factors influencing the quality of hospitals in deprived areas.

The average co-morbidity of population is also often significant, however not always in a manner that suggests that hospitals treating more severe patients on average have worse outcomes. Indeed there are a few instances suggesting the opposite, most of these instances appear in the latent models, and only in one case in a filtered model (for IHD mortality). If we were to interpret only the latent models, the results would suggest that in some cases hospitals treating more severe populations on average have better outcomes, possibly indicating selection. However, the interpretation of the filtered model suggests that in the majority of cases this effect is noise that was not completely eliminated by the risk adjustment in the patient level regressions. This highlights the importance in the selection of quality metric used to analyze quality changes, a finding highlighted in most of the previous chapters, as the type of indicator used will influence findings substantially.

Indeed, some of the findings from the previous chapters need to be considered in the interpretation of the models in this chapter. For instance, when trying to understand the overall implications of the PbR policy on quality, we need to consider the relationship of the different indicators to each other. In Chapters 2 and 3 we noted that mortality and readmissions are often negatively associated, such that high readmissions are not always an indicator of bad quality. For instance when very severe patients are treated at a good hospital they are more likely to survive, but are also more likely to have a difficult recovery and be readmitted. In these cases readmissions will be higher for good quality hospitals, as the patients would have likely died in the poor quality hospital. Taking these findings into account it is easier to interpret the findings generally and by condition. Overall we find that PbR does have significant effects on quality, as measures by our eight outcome indicators.

Aside from influencing the levels of quality we find the that policy also influences the distribution of quality as well as the relative performance of hospitals with one another,

which is discussed later on. Yet, the effects the policy has had on outcome indicators vary, such that it is more beneficial for some conditions than others. In almost all instances PbR is associated with declining mortality. The is the most pronounced for AMI, where all quality indicators in Models 1 and 2 are significant and negative, and least pronounced for MI and Hip Replacement, where none of the quality variables are negative. The amount of change also varies by condition, ranging from highs of $5-6\%$ declines in AMI and IHD, to mid range changes around $2-3\%$ for Stroke, MI, TIA and CCF and small changes around 0.03% for Hip Replacement. Moreover, while Stroke year-long mortality is declining since PbR's introduction, filtered 30-day mortality is rising. This suggests that while the policy has benefited longer-term outcomes and those who survive their initial Stroke, the cost containment may have had adverse effects on initial rates of survival.

The results on readmissions are more mixed, for AMI and IHD, short and long term readmissions seem to be unequivocally rising. Although in both cases this is only around $1-2\%$. The filtered models suggest that this is also the case for Stroke readmissions, albeit with a smaller increase in short term readmissions of about 0.5%. Short term readmissions are also rising for MI, again around 2%. For the remaining conditions, CCF, TIA and Hip Replacement – the models suggest falling readmissions. For Hip Replacement this is by a very small amount, only 0.05%, but in both short and long term readmissions. For CCF, this is experienced only in long term readmissions, by about 2%, and for TIA only in short term readmissions by around 0.5%. Yet as noted previously, extra caution needs to be taken in the interpretation of the effects of the policy on readmissions, as higher readmissions are not always an indication of poor quality. The effect on readmissions needs to be interpreted alongside the effect on the other explanatory variables, such as co-morbidity and deprivation, as well as the mortality outcomes with which they are correlated. For example, we know from Chapter 3 that mortality and readmissions are negatively associated for the conditions of Stroke and MI, so the increases in readmissions could in fact be indicative of quality improvements.

In order to understand the entire effect on quality however it is not enough to compare the change in levels of quality. It is also of interest to understand what occurs to the relative performance of hospitals. In the English scenario we expect there to be some effect in this area because of the way the national tariff is set. As the national tariff is determined by average costs across hospitals it is likely to incentivize average performance. Models 3 and 4 examine changes in relative hospital performance within each year, and the effect PbR has had on this. For MI and CCF there appears to be no significant effect of PbR on relative performance, and for Hip Replacement and TIA it is only noticeable in a small subset of the models. Yet in AMI, IHD and Stroke there appears to be a

strong effect on relative performance, such the policy is encouraging hospitals to converge to average performance. In practice the interpretation of this effect varies depending on what is happening to mean performance during this period. For example, Models 3 and 4 suggest that hospitals treating AMI performance are converging to the mean for both mortality and readmissions. The coefficients on the PbR coefficient suggest that from year to year there are fewer hospitals with below average mortality, and fewer hospitals with above average readmissions. However, from Models 1 and 2 we know that mortality during this period is falling, and readmissions are rising. Thus while hospitals are converging to the mean, it is the hospitals with above average mortality that are improving and the hospitals with below average readmissions that are worsening, since the mean is changing.

Using the four models together, we can gather a more complete picture of the changes occurring in quality after the implementation of the PbR policy for each condition. As already noted above, Models 1 and 2 indicate that mortality across hospitals is declining by about 5%. Moreover the explanatory variables indicate that hospitals with more severe and more deprived patients on average will have worse outcomes. The results from Models 3 and 4, when interpreted alongside the results in Models 1 and 2, tell us that the major improvements in mortality are driven by the low performing hospitals which are able to improve mortality gradually to meet the average national level. However, this improvement is most difficult for hospitals with higher amounts of severe and deprived patients. The models are also able to tell us about the performance of readmissions during this period. As already noted Models 1 and 2 indicate that readmissions for AMI are rising, moreover the same models indicate a negative association between readmissions and average co-morbidity and deprivation of hospitals. This suggests that the rise in readmissions may indeed be a reflection of worse quality as the more deprived and severe patients, which are more likely to be 'marginal patients', are correlated with lower readmissions. Indeed, the results from Models 3 and 4 also suggest that readmissions are falling from hospitals with above average readmissions. Moreover the models with the interaction terms indicate that since PbR hospitals treating more deprived patients experience and increase in all mortalities and a decrease in short term readmissions. As we have already controlled for these variables at the patient level in the construction of the quality estimates, this indicates some sort of systematic bias. That is in AMI, PbR is contributing to some other behaviour that may may be resulting in improved quality for less complex patients, but worse outcomes for the most severe and deprived. Theory suggests that patient selection may have this effect, however the cases we are investigating for AMI are non-elective making selection very difficult. We look at AMI in depth in the next chapter to try and understand in more detail what factors are at play.

Interestingly, the policy appears to have had a much smaller effect on MI. Indeed models 1 and 2 indicate that the policy had no effect on mortality, but did increase short and long term readmissions by 2-4%. However, we know for this condition that readmissions and mortality are negatively correlated. Moreover, we find that co-morbidity is not significantly associated with readmissions but deprivation is, such that more deprived patients have lower readmissions. MI models 3 and 4 suggest that PbR has had no effect on relative performance of hospitals, and neither has co-morbidity and deprivation. The interaction model, also does not show any significant interaction between the policy and hospitals treating more severe or more deprived patients who are receiving poorer quality care. Taken together the results indicate that on the whole the policy has had limited effects for MI patients, although it appears to have disadvantaged the treatment of the most deprived patients. The policy also appears to have had a limited effect on Hip Replacement outcome, where the results from Model 2, indicate that PbR was only significant in influencing readmissions, such that they fell but only by about 0.05%. Moreover, the PbR dummy was not significant at all in models 3 and 4 indicating that relative performance did not change.

While the effects for AMI and MI appear to be mixed, suggesting that the policy has only benefited the less severe or less deprived patients, we find more positive results for IHD and CCF where there appear to be genuine quality improvements all around. The PbR coefficient on IHD models 1 and 2 suggests that mortality has been declining by about 3-6%. Long term mortality outcomes are worse for hospitals with higher numbers of sever patients, but deprivation is insignificant in all the mortality models. IHD readmissions have been rising by 2-3%, However, in all cases hospitals with more deprived patients have lower readmissions. The results for models 3 and 4 indicate that mortality is rising from year to year relative to the mean. As the mean is falling, this indicates that the improvements in mortality are driven by the hospitals with above average mortality whose mortality is decreasing from year to year since the implementation of the policy. In models 3 and 4 the PbR dummy is not significant for either of the filtered readmission models. The average co-morbidity variable is also insignificant. Although, the average deprivation variable indicates that more deprived patients have higher readmissions. Thus, unlike AMI, the improvement in mortality does not appear to be at the expense of severe patient groups but a genuine quality improvement. Yet, quality remains worse for hospitals in deprived areas. The results from interaction models confirm this interpretation as they indicate that since PbR's implementation, hospitals with higher amounts of severe patients have lower mortalities and higher year long readmissions.

Similarly, the results from CCF Models 1 and 2 indicate a decline in mortality since

the implementation of PbR at around 0.5% for 30-day mortality and nearly 3% in year-long mortality. The results also indicate a decline in year long readmissions of about 1.5%. Average co-morbidity and deprivation of patients for the different hospitals are not significant in any of these models. Moreover, Models 3 and 4 suggest that none of these variables are significant in influencing hospitals relative performance as measured by the filtered indicators. These results taken together suggest that the policy has resulted in genuine quality improvements across hospitals.

The results for TIA also suggest an overall improvement in quality. Models 1 and 2 indicate declines in mortality of about 0.6% for 30-day in hospital mortality and nearly 3% for year-long mortality. Moreover, the filtered normalized indicators in model 4 indicate that there is no significant change in relative performance. In neither models 2 or 4 is average co-morbidity significant. While higher average deprivation is associated with higher relative readmissions in Model 4, PbR is not significant in that model.

However, the results for Stroke are more mixed. Model 2 suggests that since PbR 30-day mortality is rising by nearly 2%, however year-long mortality is falling. While this is coupled with increases in 28-day and year-long readmissions of 0.5–2%. While the coefficient on the PbR dummy is negative for all three filtered mortality indicators in Model 3, the interpretation differs because of the different coefficients in Model 2. As 30-day in-hospital mortality was increasing, the negative coefficient in Model 4 indicates that the hospitals with below average mortality are worsening from year to year. However, other two mortality indicators were decreasing since PbR, thus the negative coefficient for year-long mortality indicates a decline in the mortality of the worse performing hospitals. The PbR dummy is insignificant in Model 4 for short term readmissions, suggesting that the small increase indicated in Model 2 is felt across hospitals. However, as mortality and readmission are negatively correlated this can be an indication of improved quality. The PbR dummy for year-long readmissions in Model 4 is significant and negative. As year-long readmissions have been increasing since PbR, this tells us that they are increasing for the hospitals with below average readmissions.

While average co-morbidity and deprivation are not significant in any of the filtered models. The Stroke interaction models indicate that after PbR, hospitals treating more severe patients have lower mortalities, and lower readmissions. Yet the effect of PbR alone, as indicated by the PbR dummy, is to increase 30-day mortality and increase readmissions. As mentioned previously this could be an indication of specialization, where hospitals that are known to be good providers of a certain treatment are taking on more cases in this area. Theory indeed suggests that under a case-payment system, specialization will be incentivized, as it allows providers to make profits in areas where they are already efficient.

However, more research is needed to determine if this is what is going on in the case of Stroke.

Yet, in interpreting the effect of these models it is important to point out that the way we have classified these conditions, by ICD-10 groups is much broader than individual HRGs. Indeed within any one condition, we may be looking at the performance of many HRG groups, which will have varying performance. While it would be desirable to investigate the effect on each HRG group instead, as this is the unit of payment and more likely to be associated with quality, the small numbers within each group make this more difficult and less methodologically sound. The following chapter will look in more detail at four specific HRG groups to try and understand differences in performance since the implementation of PbR, however for the purposes of this chapter we would like to acknowledge that in our interpretation of the change in quality of treating conditions we will face limitation in understanding how the true quality effects vary by HRG group.

To conclude, we find when applying the hospital quality measures developed in previous chapters to assess a policy question, that the filtered measures do indeed perform better than their latent counterparts. While the filtered and latent models indicate the same trends in readmissions and mortality outcomes in most cases, there differences in the magnitude of the effect they indicate. Moreover, while the latent models show more of the other explanatory variables to be significant, they have lower R-squared values. These differences reflect the filtered measure's ability to better eliminate noise from the latent measures and better capture the true quality signal. As the filtered measures are able to filter out noise and adjust for time trends and associations between outcome indicators they are able to capture more of the true relationships occurring. For these reasons we believe in this type of analysis the filtered measures are a more accurate measure of quality.

The other main finding of this chapter is that we are able to link the implementation of PbR and to quality of hospital care. We find that PbR has a stronger effect on some conditions, such as AMI, Stroke and IHD, than others, such as TIA, MI and CCF. Moreover, while the PbR policy has reduced mortality across most conditions, it has had all around positive effects for only a few of the conditions studied, such as MI, IHD, CCF and TIA – yet for CCF and TIA these are very small improvements. In AMI, where there have been large changes in mortality and readmissions, we find that the changes are not beneficial for all hospitals, and it appears that those treating the most deprived and severe patients have had to skimp on quality. Finally the results for Stroke are most difficult to disentangle. It appears that the most severe patients have benefited the most as has fallen, yet the less severe have had worsening quality in terms of rising mortality. While readmissions are rising - this cannot be unequivocally interpreted as a decline in quality

due to the indicator's association with mortality. This results could be explained by some sort of adverse behaviour such as quality skimping, but more evidence is required to tease out the true underlying effect. Overall we find that the PbR policy in England has had a variable impact on health system performance.

6 The effect of Payment by Results on the activity and quality of AMI and Hip Replacement

6.1 Introduction

Health care providers are economic agents. While there are competing theories as to what drives a provider's behaviour, be it utility, income, profit or quantity, most researchers believe that money is an important variable. In addition, asymmetries of information between providers and users, and providers and third-party payers, coupled with often incompatible utility functions, mean the provider may not always be the perfect agent. Driven by their own economic motives, providers will act in their own best interests. However, money can be used to alter providers behaviour towards the interests of the health system and/or their patients. For these reasons the consideration of different payment systems and the incentives attached to them offer interesting possibilities to policy makers.

England's PbR policy was introduced in 2003/04 radically changing the way in which hospitals are paid. Moving from a system where hospitals received block contracts for activity, hospitals payment is now determined by the activity they undertake which is reimbursed at a national tariff determined by national average NHS provider costs. This type of payment system, known as a case payment system, has become increasingly popular in the last two decades and is being adopted throughout the world. Evidence from the experiences of different countries using this type of payment system suggests mixed results as to the impact it has on quality and efficiency. In part the extent to which these benefits are accrued seem to be closely related to overall system factors and organization of health services, but also largely associated with the design of the case-payment itself.

There are many desirable benefits attributed to a case-based payment system and help to account for its recent popularity, such as increased transparency and efficiency. These benefits result as the system fosters hospital investment the collection of cost information that is used to calculate future payments. Moreover, as hospitals are paid according

to the activities they perform, they are encouraged to respond to patient preferences and demands. However, for a system to accrue these benefits the policy needs to be carefully designed to ensure that costs are collected accurately and payments are set in a way that successfully takes into account patient heterogeneity. Failure to do so can have serious impacts on behavioural incentives, and in the worst case can lead to providers turning severe patients away, discharging patients too early, skimping on quality of care or even reporting false activity. To avoid such unwanted behaviours, the payment made to hospitals must be able to distinguish between costs that occur because of patient heterogeneity and the costs that arise because of inefficient service provision.

Heterogeneity in a case payment can be accounted for in different ways, ultimately all of which are important. One important feature is the number of 'case groups' that are defined in a system. There can be few broad case categories, or much finer groupings based on more detailed clinical information. Underlying the decision of how detailed to make a payment system lies a trade-off: a system with fewer and broader payment categories will be easier and cheaper to monitor and administer, at the expense of increased fairness and the increased possibility of creating adverse selection incentives, or vice versa. Another important feature is how to set the rate for the different cases, the rate can reflect average costs across a subset of providers, or the whole set of providers; marginal costs and even normative costs. Indeed there is a growing literature on the implications of setting the case payment rate, noting the different options and the advantages and disadvantages of each (Ellis and McGuire, 1996; Schreyögg et al., 2006; Street and Maynard, 2007).

Since the implementation of PbR, some work has been done to assess the impact of the policy on quality, including our work presented in Chapter 5. In the English setting, Farrar et al. (2009) find evidence to suggest a mild increase in acute hospital activity along with a reduction in unit costs. Using in-hospital mortality, 30-day post surgical mortality and emergency readmission after hip fracture as measures of quality, they find no evidence to suggest quality has declined after the implementation of PbR. The Audit Commission, 2005 reported little difference in activity growth or efficiency in foundation trusts, after the first year of PbR, apart from a small increase in length of stay. They report no evidence of gaming amongst the early implementers although do mention cases of perceived gaming having been reported by some Primary Care Trusts (PCTs), including resubmission of patients using old referrals, artificial discharge of payments and coding and/or undertaking multiple interventions that are unnecessary in order to increase revenues. Research has also indicated a change in year to year activity among cases (Farrar et al., 2009; Sussex and Farrar, 2008; Rogers et al., 2005) although in all authors note that it is unclear whether this represents a genuine change in activity or a change in the way activity is recorded.

Evidence from other countries on the impact of DRG pricing has considered various aspects. Keeler et al. (1990) find an increase in sickness at admission following the introduction of the Prospective Payment System (PPS) as well as increased expected mortality. Other studies of the US Medicare PPS system, indicate reductions in average hospital length of stay (Feder et al., 1987; Newhouse and Byrne, 1988; Shen, 2003) and reductions in costs (Cutler, 1995; Shen, 2003), yet with apparent increases in the length of stay of long-stay patients (Newhouse and Byrne, 1988). There was also a short term noticeable shift of treatment from DRG financed inpatient settings to outpatient clinics which were financed differently (Cutler and Reber, 1998; Ellis and Vidal-Fernandez, 2007; Newhouse and Byrne, 1988). (Newhouse et al., 1989) found that patients in unprofitable DRGs were more likely to be found in 'hospitals of last resort', also suggesting patient selection by profitability. Other evidence of patient selection was presented by Meltzer et al. (2002) who found greater cost decreases for high cost patients than low cost patients, mirrored by a pattern of reductions in more expensive DRGs. Kahn et al. (1992) note increased readmissions after patients are discharged quicker from hospital, yet do note improved processes of care for specific conditions (Kahn et al., 1990, 1992). Similarly, Ellis and McGuire (1996) identified evidence of selection, under Medicaid's mental health services in New Hampshire where expenditures for the sickest patients were reduced under prospective payment. Carter et al. (1990) investigated changes in Medicare's Case Mix Index between 1986-1987 to identify any instances in DRG creep or upcoding, yet found no evidence to support such behaviour.

Evidence from other countries on the effects of activity-based financing show that while 'upcoding' and gaming do exist in the system, these are a rather marginal phenomenon. There are also reports of down coding, especially in Sweden (HOPE, 2006). There is limited research on the effect case-based funding has had on the efficiency and quality of care on these health care systems due to lack of reliable data, and/or a limited time frame since introduction. Indeed, instances where gaming or upcoding that have been found to be detrimental to quality of care have been few. In the Australian case, Ellis and Vidal-Fernandez (2007) reported the gradual appearance of diagnostic coding creep over ten years of DRGs. While in Austria most evidence of gaming and upcoding took the form of cost shifting (Sommersguter-Reichmann, 2000; Rauner et al., 2003). In the Swedish case, Diderichsen (1995) finds an increase in hospitalization rates across diagnostic categories along with a decrease in length of stay for the age group 80+ after the introduction of DRGs.

The aim of this chapter is to explicitly consider, in a manner not so far undertaken, whether the introduction of the PbR has changed the recording of activity within the

English NHS system. The chapter will focus on the outcomes of four case groups, two in AMI and two in Hip Replacement. The first condition was chosen as a largely emergency treatment for heart attack, where the choice of hospital is argued to be of secondary importance, while the latter is a common elective procedure. In the previous chapter we identified some interesting quality effects the introduction of this policy had on these conditions. Indeed, careful investigation of the levels of activity and diagnoses revealed interesting patterns that would suggest upcoding. However, the literature notes that during this time period substantial changes have been made in the way data is recorded. It is thus likely that what appears at a glance to be upcoding, may simply be a manifestation of better coding. For lack of better terminology we continue to use the term 'upcoding' in this chapter to refer to an increase in activity in a similar more expensive case group at the expense of another due to coding. In this chapter we are concerned with determining first if the suspicious activity changes we observe are indeed upcoding, second what the driver behind this upcoding is, and thirdly what, if any, effect has this had on quality.

In order to analyze the underlying reasons for the changes in hospital activity, and the possible implications this has had for quality of care, a three step process was undertaken for both HRG pairs. In the first step a series of regressions are used to examine whether there substitution occurring between the two HRG groups, and how this is associated with PbR and the HRG tariff. The second step examines the rate of change in each of the HRG groups as compared to the change of other factors, to determine what is driving the apparent substitution effect. The third step aims to investigate what effect the activity change is having on quality of care. This is done by using the latent and filtered quality measures constructed in the previous section in a series of regressions to determine what effect the change in activity has had on quality. If the PbR system has led hospitals to be more systematic and careful in their coding practices as incentivised by the change in the payment system this indicates an efficiency gain. However, if upcoding is taking place as a profit enhancing activity this is the sign of an inefficient system. Our findings are important to better understand some of the incentives created from the implementation of the PbR policy.

6.2 Model Specification and Estimation

The literature from DRG experiences in most industrialised countries indicates an increase in activity following the adoption of a case payment system. Moreover theory suggests that this type of system provides the incentive to increase activity in higher paying DRGs at the expense of ones reimbursed at a lower rate. Using data corresponding to the diagnostic and procedural codes linked to AMI, MI, IHD, Stroke, TIA and Hip Replacement, we

studied the trend in cases of all HRG groups in our sample over time. Where HRG codes had overlapping diagnostic or procedural codes we were particularly interested to see how activity between the cases changed. In 2008 the English system had approximately 540 HRG groups. As presented in Table 6.1, we identified certain HRG pairs corresponding to the conditions in our data where there could be theoretical possibility of upcoding. However in most of these groups the number of cases per year did not exceed 100, making any robust statistical analysis difficult. In most cases where the numbers were large enough activity change in the two groups moved in the same direction. Two HRG pairs, with large sample sizes one for AMI cases and one for Hip Replacement cases, presented an interesting relationship in which activity increased for the higher paying HRG groups while decreasing for the lower paying groups. These are of further interest as one is predominately an emergency admission, while the other is a common predominately elective procedure.

Table 6.1: HRG groups in data sample.

Condition	ICD-10/ OPCS 4.3 codes	HGR pairs investigated
AMI	ICD-10: I21	• E11 (AMI without complications) & E12 (AMI with complications)
MI	ICD-10: I22, I23	• E11 (AMI without complications) & E12 (AMI with complications)
IHD	ICD-10: I20, I25	• E22 (Ischemic heart disease w/out intervention >69 or with cc) & E23 (Ischemic heart disease with/out intervention <70 or without cc); • E24 (Hypertension w/out intervention >69 or with cc) & E25 (Hypertension with/out intervention <70 or without cc); • E35 (Chest Pain w/out intervention >69 or with cc) & E36 (Chest Pain with/out intervention <70 or without cc); • E40 (Other cardiothoracic or circulatory procedures w/out intervention >18) & E41 (Other cardiothoracic or circulatory procedures <19).
CCF	ICD-10: I11.0, I13.0, I25.5, I50.0, I50.1, I50.9, J81X	• E18 (Heart failure or shock w/out intervention >69 or with cc) & E19 (Heart failure or shock w/out intervention <70 or without cc)

Condition	ICD-10/ OPCS 4.3 codes	HGR pairs investigated
Stroke	ICD-10: I60-I67	• A22 (Non-Transient Stroke or Cerebrovascular Accident >69 or with cc) & A23 (Non-Transient Stroke or Cerebrovascular Accident <70 or without cc)
TIA	ICD-10: G45.0-G45.4, G45.8-G45.9, G46.0-G46.8	• A20 (Transient Ischemic Attack >69 or with cc) & A21 (Transient Ischemic Attack <70 or without cc)
Hip	OPCS4.3: W37-W39 W46-W48 W58	• H80 (Primary Hip Replacement Cemented) & H81 (Primary Hip Replacement Uncemented); • H82 (Extracapsular Neck of Femur Fracture with Fixation with cc) & H83 (Extracapsular Neck of Femur Fracture with Fixation w/out cc); • H84 (Intracapsular Neck of Femur Fracture with Fixation with cc) & H85 (Intracapsular Neck of Femur Fracture with Fixation w/out cc); • H86 (Neck of Femur Fracture with Hip Replacement with cc) & H87 (Neck of Femur Fracture with Hip Replacement w/out cc); • H88 (Other Neck of Femur Fracture with cc) & H89 (Other Neck of Femur Fracture w/out cc).

More specifically, as presented in Figure 6.1, we found the activity for the AMI HRG code E11 (AMI with complications) to be rising over time, while the activity for HRG code E12 (AMI without complications) was falling. This trend was consistent across the whole sample, and within hospitals, such that activity in E11 increased by about 6000 cases, almost doubling over the period 2000-2008, and activity in E12 fell to 5000 cases below its initial level in the same period, with the majority of the decline concentrated in the period 2004-2008 (Figure 6.1). A similar trend was observed for HRG pair H80 (cemented Hip Replacement) and H81 (uncemented Hip Replacement), where activity increased for the more expensive H81 and declined for H80 from 2004 onwards (Figure 6.2). While cemented Hip Replacement and uncemented Hip Replacement are two different procedures, both are used by doctors to treat patients who require a hip replacement. Cemented replacement is the older technique (developed about 40 years ago) while uncemented replacement was developed 20 years ago to avoid the possibility of loosening parts and the breaking off of cement particles which can occur in the cemented procedure. However, depending on the

patient's condition and age a cemented procedure may be preferable as there is a shorter and easier recovery period. For our purposes we are specifically interested in the sudden change in activity between these two conditions, as no major medical breakthrough was made during the time period we are investigating we can assume that the change in activity is attributable to some other factor.

Figure 6.1: Average AMI cases

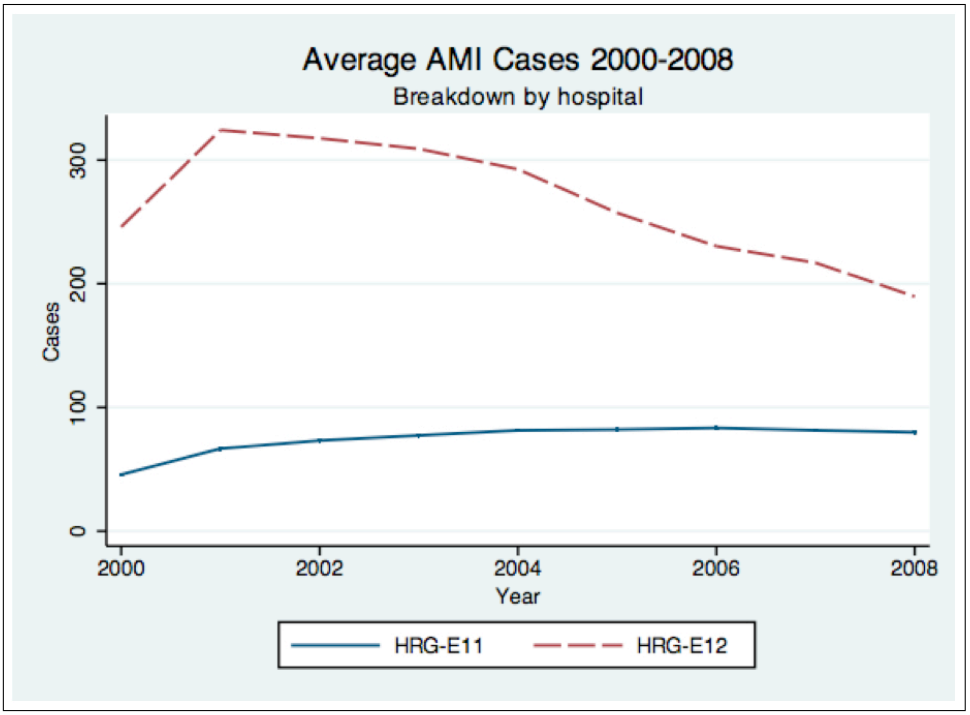
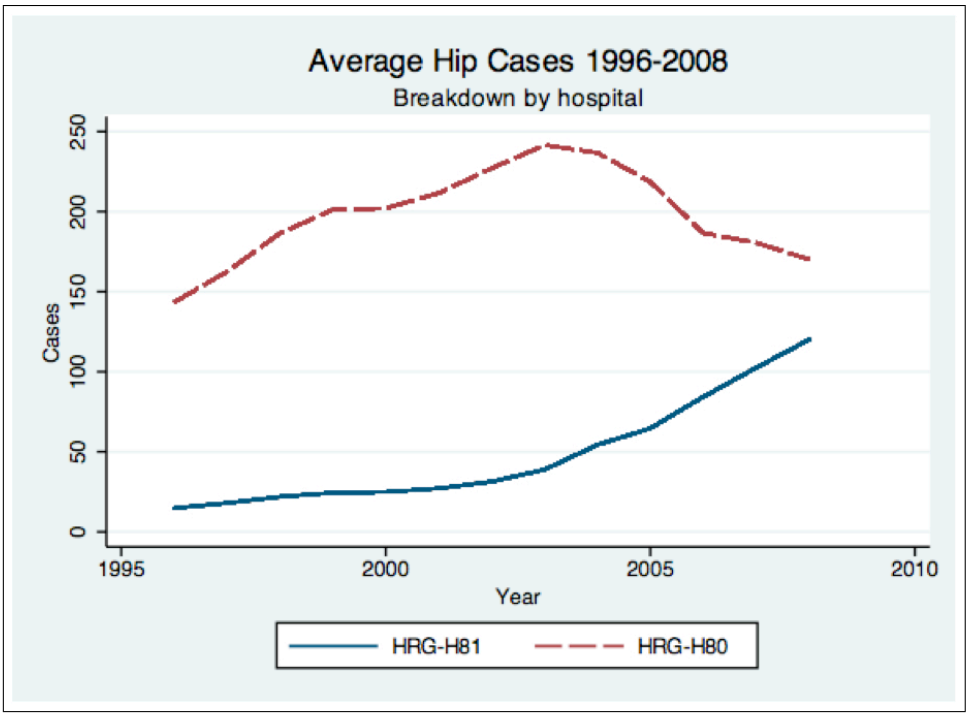


Figure 6.2: Average Hip cases



Detecting Activity Change

The first step of the analysis uses panel data to analyse the activity changes observed in the HRG groups over the period 2000-2008 for AMI and 1996-2008 for Hip Replacement. By modelling the activity in the higher paying HRG group against the activity in the lower paying group, and controlling from other hospital factors, we can detect whether there is a substitution effect taking place. A fixed effects model is estimated as it allows us to control for the individual time invariant characteristics of the hospital in order to assess the predictor's net effect. Time dummies are also included to control for within year variation.

$$E11 = \alpha + \beta_1 E12_{ht} + \beta_2 T_{ht} + \beta_3 \sum X_{ht} + \beta_4 PbR + \beta_5 \sum C_{ht} + \epsilon_{ht} \tag{6.1}$$

$$H81 = \alpha + \beta_1 H80_{ht} + \beta_2 T_{ht} + \beta_3 \sum X_{ht} + \beta_4 PbR + \beta_5 \sum C_{ht} + \epsilon_{ht} \tag{6.2}$$

In equations (6.1) and (6.2), the variables $E11$, $E12$, $H81$ and $H80$ denote the number of cases admitted to each hospital, h, each year, t, for those HRG groups. T_{ht} denotes the average tariff received by each hospital for the group of conditions being investigated. In both models a number of hospital characteristics, X_{ht}, are controlled for including average age, average deprivation, average co-morbidity and average length of stay of patients admitted for the condition being investigated. In model (6.2), for Hip Replacement where

both elective and emergency procedures are treated, an additional variable considering how many elective patients were admitted is also included. PbR is a dummy variable indicating the implementation of the PbR policy, this is set at 2005 for the AMI model and at 2006 for the Hip Replacement model, indicating the different introduction of payment for elective and non-elective procedures. The variable C_{ht} indicates a group of dummy variables that indicate whether the hospital is a foundation trust, teaching hospital or independent sector treatment centre (ISTC)[1]. Some of these variables represent time-invariant characteristics and are thus differenced out of the equation under fixed effects. However, they are included in the random effects estimations conducted in the sensitivity analysis.

Following the estimation of equations (6.1) and (6.2), we modify the model to investigate how the rate of change in activity with the higher paying HRG groups is related to the rate of change in activity in lower paying groups, and the rate of change of other independent variables. This is done taking the value of the year-to-year difference in the different variables, such that the value for time period $t+1$ is subtracted from time period t. Note in models (6.1) and (6.2) neither the dummy variable controlling for the introduction of PbR is differenced nor the time dummies.

$$\Delta E11 = \alpha + \beta_1 \Delta E12_{ht} + \beta_2 \Delta T_{ht} + \beta_3 \Delta \sum X_{ht} + \beta_4 PbR + \beta_5 \sum C_{ht} + \epsilon_{ht} \tag{6.3}$$

$$\Delta H81 = \alpha + \beta_1 \Delta H80_{ht} + \beta_2 \Delta T_{ht} + \beta_3 \Delta \sum X_{ht} + \beta_4 PbR + \beta_5 \sum C_{ht} + \epsilon_{ht} \tag{6.4}$$

The variables included in equations (6.3) and (6.4) are the same as those specified in equations (6.1) and (6.2), only that they indicate differences rather than levels as indicated by the symbol Δ. Thus the explanatory variables are as outlined above. In all models we are interested in the significance, sign and magnitude of the coefficient β_1 which denotes the marginal elasticity of substitution between the two HRG groups. A negative sign will indicate that there is indeed a substitution effect, while the magnitude will indicate the rate of substitution. Again both models are run with fixed effects and year dummies.

Analysis of Quality Change

The third step of the analysis is concerned with determining what effect the substitution between cases has on quality. In order to assess this, we use the quality metrics constructed in Part II of the thesis, that is the latent and filtered estimates. These metrics are estimated separately for all HRG groups being considered and used as dependent variables to determine how the change in activity has influenced quality of care. For this analysis we use the latent and filtered measures, which are constructed for the AMI and Hip sub-samples using the methodology outlined in Chapters 2 and 3. Due to limitations

[1] ISTCs are only included in the Hip Replacement model (Model 1.2).

in the numbers of patients classified in the H80 group, per hospital, across years, there were too many gaps in the panel to construct the filtered indicators. For this reason we only use the latent measures to analyse the quality change for Hip Replacement, but use both latent and filtered measures to assess AMI quality.

$$QE11_{ht} = \alpha + \beta_1 E12_{ht} + \beta_2 T_{ht} + \beta_3 \sum X_{ht} + \beta_4 PbR + \beta_5 \sum C_{ht} + \epsilon_{ht} \tag{6.5}$$

$$QE12_{ht} = \alpha + \beta_1 E11_{ht} + \beta_2 T_{ht} + \beta_3 \sum X_{ht} + \beta_4 PbR + \beta_5 \sum C_{ht} + \epsilon_{ht} \tag{6.6}$$

$$QH80_{ht} = \alpha + \beta_1 H81_{ht} + \beta_2 T_{ht} + \beta_3 \sum X_{ht} + \beta_4 PbR + \beta_5 \sum C_{ht} + \epsilon_{ht} \tag{6.7}$$

$$QH81_{ht} = \alpha + \beta_1 H80_{ht} + \beta_2 T_{ht} + \beta_3 \sum X_{ht} + \beta_4 PbR + \beta_5 \sum C_{ht} + \epsilon_{ht} \tag{6.8}$$

Each of the quality models (equations (6.5) – (6.8)) estimates how much quality has changed since the implementation of PbR, controlling for tariff, average deprivation, average co-morbidity, average length of stay, average age and type of hospital. Again all models are estimated with fixed effects and year dummies.

6.3 Data Description and Variable Construction

The Sample

The data used to conduct this analysis is the same data used for the other chapters in this thesis. A detailed description of the data can be found in the Data section of Chapter 1. Our initial investigation began with data selected for all AMI, MI, IHD, Stroke, TIA and Hip Replacement (see Table 1.2). We chose these conditions as they require prompt medical attention, are common in the population and thus provides a large annual sample size to be studied, and most importantly the quality of care provided by the hospital is known to have a significant impact on patient health outcomes. Of these conditions we only found pairs of HRGs that indicated a possible substitution of activity in AMI and Hip Replacement. While Dr. Foster provided data for both conditions over the financial years 1996-2008, there were problems with the sample sizes of some of the years before 2000 for AMI, and so these years were not included in the analysis.

From this data we extracted all individual cases that were coded under HRG codes E11 and E12 for AMI and HRG codes H80 and H81 for Hip Replacement. In the AMI sample only emergency admissions were examined, and only for patients with a length of stay over two days. For the Hip Replacement sample both elective and emergency admissions were investigated, but any day cases were dropped. This was an attempt to drop any cases initially misdiagnosed, for example a patient admitted as AMI when suffering general chest pains. In addition, any hospital trust that had less than 5 admissions in any of the years

investigated was dropped from the analysis as were any primary care trusts, private trusts acting as NHS providers and social care trusts. The sample size for this selection included approximately $33,500$ patients per year, and 121 hospitals for AMI and about 330 patients per year, and 126 hospitals for Hip Replacement.

Hospital Outcomes and Quality

The outcome measures from both samples were used as dependent variables to estimate the latent and filtered measures. The outcome measures used were 30-day in-hospital mortality rates, 365-day overall mortality rates as well as 28-day readmission and 365 day readmission rates. A trust code was used to distinguish each acute trust in the data. As noted in Chapters 2 and 4, the latent estimates constructed indicate the marginal effect a hospital has on each outcome measure controlling for patient characteristics. Thus, each point estimated for every year represents the slope of the risk-adjusted quality curve. We use this information to create this curve, for each of the five outcome indicators for every hospital, spanning throughout the years in our sample. We use the value of zero as our starting point, however, one could easily substitute zero for the mortality rate of that hospital in the same year to get the true estimate. The filtered estimates created in Chapter 3, have also been used as measures of quality. Indeed we argue that they are better estimates of true quality as they are able to reduce more of the noise in each estimate. For the H81 hip adjustment group the sample size was very limited, resulting in problems creating the filtered estimates, and gaps in some of the latent estimates.

Hospital Activity

Models 1 and 2 are interested in determining whether there has been a substitution of cases from HRG E12 to E11 and HRG H81 to H80, given the decline in activity in the former group and increase in the later. This type of substitution has been reported after the introduction of a case payment system in the US (Chulis, 1991; Ginsburg and Carter, 1986; Sloan et al., 1988; Carter et al., 1990). Indeed, in their commentary on PbR (Street and Maynard, 2007) note the possibility of an increase in increased cases, and the coding of patients with multiple co-morbidities to higher priced HRGs in the first years of implementation. They predict that this effect will be short lived while hospitals adjust to counting and coding of activity, until future revisions of the price incorporate such behaviours. Caution is required in analysing changes in caseload across similar case groups, indeed, further examination of the changes in caseload identified in the US by Ginsburg and Carter (1986) could be justified by the additional complexity of patients hospitalized (Carter et al., 1990).

In order to evaluate the factors influencing activity within the two HRG groups and it is necessary to construct variables that measure the annual activity for each of these classifications at the hospital level. Going back to the individual sample constructed for the two HRG pairs, the number of cases for each of these HRG groups was aggregated for each hospital, separately for each year of the sample, and exported into the newly constructed panel.

HRGs and the National Tariff

PbR is a case-based hospital funding system, paying hospitals based on their activity. Under this system PCTs reimburse hospitals for each procedure they perform though a nationally set tariff. The national tariff is based on HRGs which are designed to measure health care activity in a way that takes into account the diagnosis, mix and complexity of patients that will be receiving care. The basis of the HRG tariff is an average of all hospital costs for the procedure in question. Separate tariffs exist for elective and emergency care, as well as for short-stay patients, while specialist work is excluded. Hospitals also received a separate payment, the MFF, which is based on the geographical price indices for land, labour and building costs. PbR started being implemented in April 2004 to NHS foundation trusts, being extended to elective activity for all other NHS trusts in April 2005, and to non-elective and outpatient care from April 2007 (Audit Commission, 2004, 2005).

Table 6.2: National tariffs for AMI and Hip HRGs being investigated

Condition	Year	HRG Code	Non-elective spell tariff (£)	Elective spell tariff (£)	Non-elective long stay trimpoint (days)	Elective long stay trimpoint (days)	Per day long stay payment* (£)
AMI	2006/7	E11	4,747	4,527	27	59	155
		E12	3,111	2,089	16	32	169
Hip		H80	7,529	5,176	16	80	213
		H81	8,286	4,967	15	94	217
AMI	2007/8	E11	4,866	4,640	30	59	183
		E12	3,189	2,141	14	32	191
Hip		H80	7,717	5,305	16	98	218
		H81	8,493	5,091	15	80	222
AMI	2008/9	E11	4,787	5,006	27	43	159
		E12	3,017	2,908	16	19	173
Hip		H80	7,308	5,220	67	12	248
		H81	7,816	5,587	71	13	213

*for days exceeding trimpoint

Source: Department of Health National Reference Costs (2006/7; 2007/08; 2008/09)

At the time of their development in the early 1990s, HRGs were not used to reimburse providers, but primarily for benchmarking exercises and to set targets to encourage unit cost reductions (Street and Dawson, 2002). Currently the PbR tariff is payable for admitted patient care (elective, non-elective and emergency), outpatient attendances and accident and emergency admissions (Table 6.2). A detailed timeline of the implementation of the PbR policy and HRGs is available in Chapter 1 (Figure 1.1). The individual level data provided by Dr Foster in our dataset contains information on the HRG tariff throughout the entire period of the sample. Note that while this tariff does reflect approximate unit costs of the patient it is not the paid to each hospital for the entire period, but only from the years PbR was phased in. For this reason, two variables are included in the regression: a variable which reflects the average E11 and E12 tariffs for each hospital for every year, and a dummy variable which controls for the year that PbR was implemented for non-elective procedures in all hospitals.

Hospital Characteristics

The organization of hospitals with regards to management, finances and autonomy plays an important role in the discussion of quality of care and the behavioural incentives of providers. International literature has shown different quality and behavioural differences between for profit and not-for profit hospitals. Indeed findings from the US suggest that depending on the type of hospital there were more or less likely to engage in upcoding behaviour. More specifically, not-for-profit hospitals were least likely to upcode, for profit-hospitals more likely and the most likely were hospitals converting to for-profit status (Silverman and Skinner, 2004). While assuming that some of the findings from other countries can be applied cautiously to the case of England, it is important to control for specific organizational differences that are relevant to English hospitals in our analysis. During the period of investigation the hospitals included in the sample studied could be classified into four types: acute trusts, teaching hospitals, foundation trusts and independent sector treatment centres (ISTCs). For a detailed discussion on the differences between these types of institutions see Chapter 5. We include a dummy variable to control for each of these different institutions.

Other hospital characteristics are likely to affect outcomes and/or behaviour, such as the characteristics of the patients that they treat. A hospital treating older, sicker or more deprived patients for example may have more patients with complications, longer length of stay and worse outcomes. For this reason control variables were included in both regressions to take this characteristics into account. Developed from the individual level data that was used to construct the quality variable, four control variables were constructed which measured the average age, length of stay, deprivation and co-morbidity of patients treated for the selected sample for each year. Similarly as the sample investigating Hip Replacement deals with both elective and non-elective treatments, we also constructed a variable to measure the number of elective treatments to see if these increased since the implementation of PbR. While these indicators have already been controlled for at the patient level, including them at the hospital level will identify any systematic biases associated with them that could not be removed by a simple case-adjustment. For this reason we include them as control variables in our analysis.

6.4 Results

AMI

In the first step of the analysis, Model 1.1 is used to determine whether there is substitution occurring between AMI HRG groups E11 and E12 (Table 6.3). The positive sign on the

variable E12 suggests there is no substitution between the two groups, but that they both have been increasing over the time period studied. Tariff is positivity and significant at 1% indicating a strong association between tariff and E11 activity. As the model cannot determine causality in this relationship this can be interpreted either as the higher tariff is driving more E11 activity, or that more E11 patients are associated with higher tariffs. Age and co-morbidity are both positive and significant at over 5% indicating that hospitals with older and more severe patients have higher numbers of E11 cases. The sign on the PbR dummy is highly significant and indicates that since PbR has been introduced the number of E11 cases has risen. The average length of stay of each hospital is also significant but negative suggesting that hospitals with more E11 cases have lower average length of stay amongst all patients. The only hospital type variable not dropped in the fixed effects model was the dummy for foundation trusts, as these where phased in during the time period under investigation, but at different points for the different hospitals. However, foundation trust is not significantly associated with number of E11 cases. Year dummies were included in the model, they are highly significant and positive for most years suggesting an increase in E11 cases throughout the period studied.

The second model, Model 2.1, analyzes how the change in E11 cases year-to-year is influenced by the rate of change of the other independent variables. This is a stricter test of the substitution effect, yet again in Model 2.1 there is no evidence of substitution (Table 6.3). The year-to-year change in E12 cases is positive and highly significant, indicating that activity for both cases has been increasing over the time period studied. Similar to the levels model, tariff is also positive and highly significant suggesting that a year-to-year increase in the tariff is associated with a year-to-year increase in E11 cases. No conclusions can be made concerning the direction of the casual effect. Year to year increases in average age and co-morbidity are associated with year-to-year increases in E11 cases, and an increase in year-to-year average length of stay with a decrease in E11 cases. Once again foundation trust status had no effect on the change in activity. The only difference between the two models is that while the PbR dummy is positive in the rate of change model, it is only significant at10%. Moreover, the year dummies, while also included are never significant.

Table 6.3: Results for Models 1 & 2.

	Model 1.1 ($E11$)	Model 1.2 ($\Delta E11$)	Model 2.1 ($H81$)	Model 2.2 ($\Delta H81$)
E12	0.175***			
	(0.0269)			
Δ E12		0.158***		
		(0.0150)		
H80			-0.285***	
			(0.0694)	
ΔH80				-0.116***
				(0.0349)
tariff	0.0167***		0.0169**	
	(0.00476)		(0.00800)	
Δtariff		0.0155***		0.00340**
		(0.00535)		(0.00152)
Age	1.807**		0.0934	
	(0.730)		(0.599)	
ΔAge		1.583***		-1.485***
		(0.513)		(0.436)
LOS	-1.031**		-6.870***	
	(0.472)		(1.803)	
ΔLOS		-1.135**		-2.133***
		(0.517)		(0.452)
Deprivation	0.198		-1.481	
	(0.805)		(9.743)	
ΔDeprivation		0.974		0.968
		(0.637)		(2.829)
Co-morbidity	49.88***		2.240	
	(11.68)		(10.02)	
PbR	12.77**	5.992*	77.63***	11.37***
	(5.116)	(3.380)	(11.37)	(3.535)
FT	-2.344	0.622	-20.21**	4.320
	(3.650)	(2.806)	(9.832)	(3.556)
ISTC			102.7**	36.31**
			(50.53)	(14.69)

$H81$elective			7.490***	
			(1.564)	
$\Delta H81$ elective				1.538*
				(0.851)
Year Dummies	Yes	Yes	Yes	Yes
Constant	-241.7***	-4.571**	40.37	-5.723***
	(52.31)	(1.954)	(39.16)	(1.284)
Observations	1,071	952	1,570	1,440
R^2	0.408	0.273	0.502	0.144
Number of h	119	119	126	126

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

The third and fourth models, presented in Table 6.4, consider what effect the changes in activity have had on hospital quality using latent measures of quality in each HRG group as the dependent variable. These latent variables are based upon mortality and readmissions for each HRG group at different intervals, for the period 2000-2008. The results below indicate what factors influence hospital quality for patients in the E11 HRG group as well as the E12 groups. The R-squared values on all mortality models are over 90%, indicating that they are able to predict remarkably well for the E11 and E12 mortality outcomes. They also perform well for the readmission models, ranging between 40-90% for both groups.

The results indicate that both E11 and E12 groups have been influenced by the implementation of the PbR policy, in a similar way. Mortality at all intervals for both groups has declined since the policy's introduction, while short term readmissions have risen. There is a difference in the effect the policy has had on year-long readmission rates, where they are significantly associated with an increase in the readmissions of the E11 and a decrease in the readmissions of the E12 group. Changes in E11 caseload only influenced E11 quality in terms of their effect on 30-day in-hospital mortality, which falls as cases increase. Changes in E12 caseload only influence year-long E12 readmissions, where an increase in cases associated with a fall in readmissions. Tariff is not significant in any of the E12 models, but is significant at 10% in the E11 30-day mortality and 28-day readmissions models. The sign on the tariff coefficients indicate that an increase in tariff is associated with a decline in mortality and an increase in readmissions. Average co-morbidity is only significant in year-long mortality models for the E12 group, where hospitals with more severe patients have higher mortality. Deprivation is significant for both E11 and E12 groups, such that 30-day in hospital mortality is higher in hospitals that have more

deprived patients. Higher deprivation is also associated with lower 28-day readmissions, but for the E12 group only. Average length of stay, and foundation trust status are not significant for either of the E11 or E12 models.

Table 6.4: Quality effects on AMI patients (latent outcome indicators).

	$D30_{ht}$	$D30_{ht}$	$D365_{ht}$	$D365_{ht}$	$R28_{ht}$	$R28_{ht}$	$R365_{ht}$	$R365_{ht}$
	$E11$	$E12$	$E11$	$E12$	$E11$	$E12$	$E11$	$E12$
E11	-0.00102***		-0.000147		-3.27e-06		-6.39e-05	
	(0.000218)		(0.000419)		(0.000169)		(0.000270)	
E12		-4.05e-05		-1.07e-05		6.49e-06		-0.00018***
		(4.11e-05)		(5.39e-05)		(3.93e-05)		(6.49e-05)
tariffE11	-2.29e-05*		1.29e-06		1.23e-05*		1.05e-06	
	(1.33e-05)		(1.52e-05)		(6.59e-06)		(1.06e-05)	
tariffE12		2.73e-06		-5.66e-06		-4.70e-06		3.41e-06
		(1.07e-05)		(1.04e-05)		(7.97e-06)		(1.52e-05)
LOS	0.00333	0.00158	0.00344	0.00181	-0.00241	0.000487	-0.00251	0.00218
	(0.00286)	(0.00145)	(0.00369)	(0.00226)	(0.00172)	(0.00116)	(0.00332)	(0.00209)
Co-morbidity	0.0181	0.0108	-0.0715	0.0608**	0.0212	0.0103	0.0229	0.0307
	(0.0496)	(0.0295)	(0.0655)	(0.0237)	(0.0379)	(0.0213)	(0.0564)	(0.0368)
Deprivation	0.0197***	0.00816***	0.000649	-0.00361	-0.00206	-0.00613*	0.00765	-0.000543
	(0.00677)	(0.00283)	(0.00986)	(0.00362)	(0.00449)	(0.00350)	(0.00875)	(0.00571)
Age	0.00146	-0.00129	-0.00513	-0.00413*	-0.00106	-0.000615	-0.00208	-0.00124
	(0.00481)	(0.00238)	(0.00675)	(0.00242)	(0.00242)	(0.00157)	(0.00365)	(0.00286)
PbR	-0.815***	-1.028***	-2.779***	-2.511***	0.460***	0.0709***	0.470***	-0.149***
	(0.0234)	(0.00908)	(0.0327)	(0.0129)	(0.0132)	(0.0107)	(0.0274)	(0.0189)
FT	-0.00779	0.00323	0.00856	-0.000777	-0.00179	-0.0132	-0.0270	-0.0176
	(0.0284)	(0.0129)	(0.0330)	(0.0133)	(0.0172)	(0.00974)	(0.0224)	(0.0165)
Constant	0.00207	0.0582	0.420	0.199	0.00498	0.0363	0.133	0.0533
	(0.295)	(0.144)	(0.437)	(0.181)	(0.162)	(0.121)	(0.273)	(0.223)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,035	1,035	1,035	1,035	1,035	1,035	1,035	1,035
R^2	0.936	0.991	0.988	0.997	0.892	0.407	0.711	0.817
Number of h	115	115	115	115	115	115	115	115

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table 6.5 presents the results for the filtered quality models for all outcomes in HRG

groups E11 and E12. The R-squared value for all models is very high, and better in all cases than the R-squared for the equivalent latent models. The results in Table 6.5 indicate that for the filtered models the PbR dummy is also significant. In all models the sign indicates the same effect on the rate of change in quality that was indicated by the latent model. For both E11 and E12 groups, since the adoption of PbR there is a decrease in mortality. The short term readmission model for E11, suggests that after PbR readmissions have increased, while the E12 model does not indicate a significant effect. For year-long readmissions, the E11 model again indicates an increase in readmissions, with the E12 model indicates that readmissions for that group have fallen.

Of the explanatory variables, very few are significant. E12 cases are significant in all the E12 models, with a negative sign. As we know E12 cases have been declining throughout this period, this most likely indicates that a decrease in caseload is associated with a decrease in mortality and an increase in readmissions. Caseload is not significant for any of the E11 models. co-morbidity is significant, but again only for the E12 models and not for any of the E11 models. It suggests that hospitals with higher levels of severe patients will have higher mortality rates and lower readmission rates. Average deprivation is only significant in the year long readmission model for the E12 cases, where it indicates that hospitals with more deprived patients will have lower readmissions. Average length of stay, average age and foundation trust status are not significant for any of the models.

Table 6.5: Quality effects on AMI patients (filtered outcome indicators).

	$D30_{ht}$	$D30_{ht}$	$D365_{ht}$	$D365_{ht}$	$R28_{ht}$	$R28_{ht}$	$R365_{ht}$	$R365_{ht}$
	$E11$	$E12$	$E11$	$E12$	$E11$	$E12$	$E11$	$E12$
E11	-4.13e-06		6.11e-05		-1.13e-05		-1.66e-05	
	(2.46e-05)		(0.000291)		(3.46e-05)		(6.61e-05)	
E12		6.23e-05*		0.000245*		-4.2e-05**		-7.27e-05*
		(3.33e-05)		(0.000134)		(2.07e-05)		(3.68e-05)
tariffE11	-1.71e-06		-1.68e-05*		1.56e-06		3.45e-06	
	(1.45e-06)		(9.85e-06)		(1.16e-06)		(2.22e-06)	
tariffE12		9.50e-06		4.71e-05		-7.31e-06		-1.36e-05
		(9.31e-06)		(3.62e-05)		(5.76e-06)		(1.01e-05)
LOS	0.000744	-0.000798	0.000716	-0.00388	-9.67e-07	0.000704	-1.74e-05	0.00154
	(0.000611)	(0.00124)	(0.00265)	(0.00481)	(0.000312)	(0.000821)	(0.000640)	(0.00144)
Co-morbidity	0.00964	0.0485**	0.0134	0.188**	-0.00444	-0.0294**	-0.00626	-0.0517**
	(0.00692)	(0.0214)	(0.0632)	(0.0845)	(0.00700)	(0.0131)	(0.0138)	(0.0235)
Deprivation	0.00211	0.00425	-0.00314	0.0138	-0.000581	-0.00283	-0.00223	-0.00592*

	$D30_{ht}$ $E11$	$D30_{ht}$ $E12$	$D365_{ht}$ $E11$	$D365_{ht}$ $E12$	$R28_{ht}$ $E11$	$R28_{ht}$ $E12$	$R365_{ht}$ $E11$	$R365_{ht}$ $E12$
	(0.00142)	(0.00295)	(0.00982)	(0.00994)	(0.00117)	(0.00183)	(0.00242)	(0.00336)
Age	-0.000760	-0.00148	-0.00606	-0.00459	0.000703	0.000698	0.00146	0.00117
	(0.000460)	(0.00248)	(0.00437)	(0.00915)	(0.000463)	(0.00140)	(0.000938)	(0.00237)
PbR	-0.809***	-0.933***	-2.092***	-2.001***	0.386***	0.00453	0.315***	-0.235***
	(0.00342)	(0.0113)	(0.0255)	(0.0394)	(0.00284)	(0.00641)	(0.00567)	(0.0108)
FT	0.00181	2.26e-05	0.0240	0.00236	-0.00219	0.000598	-0.00329	0.00138
	(0.00223)	(0.0116)	(0.0250)	(0.0464)	(0.00268)	(0.00714)	(0.00531)	(0.0127)
Constant	0.0394	-0.00658	0.471	-0.130	-0.0491	0.0215	-0.107*	0.0410
	(0.0288)	(0.187)	(0.295)	(0.682)	(0.0308)	(0.105)	(0.0621)	(0.176)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,035	1,035	1,035	1,035	1,035	1,035	1,035	1,035
R^2	0.999	0.992	0.987	0.966	0.996	0.794	0.965	0.957
Number of h	115	115	115	115	115	115	115	115

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Hip Replacement

Model 2.1 investigates the change in activity between HRGs H80 and H81.The negative sign on the H80 variable in Model 1.2 indicates that there is a substitution effect between the HRG groups H80 and H81, which is significant at 1% (Table 6.3). Average tariff is significant and positive such that and increase in activity is associated with an increase in the average tariff received by each hospital, while an increase in average length of stay is significantly associated with fewer H81 cases. The PbR dummy is highly significant and positive, indicating that since the implementation of PbR the number of H81 cases have increased. Similarly the elective variable indicates that more elective cases result in more H81 cases. In the Hip Replacement models the hospital type variables come out significant, such that foundation trusts are more likely to treat fewer H81 cases as compared to acute trusts, while ISTCs are more likely to treat more H81 cases than acute trusts. Year dummies are included and are highly significant for all years.

Model 2.2 investigates the year-to-year change in cases and how that is influenced by the same dependent variables. The results in Table 6.3 show that the effect is still present and highly significant, such that a yearly increase in H81 is associated with a yearly decline in H80. Similar to the results for Model 2.1, a yearly increase in tariff is associated with a yearly increase in H81 activity, while a yearly increase in average length of stay is significantly associated with a decline in activity. Unlike Model 2.1, change in

age is highly significant such that an increase in average age from year to year is associated with a decline in H81 cases. The PbR dummy is highly significant, indicating that since the implementation of PbR, the year to year change in H81 activity has been increasing. Similarly the positive sign on the ISTC variable indicates that ISTCs take more H81 cases year to year as compared to acute care trusts, while the dummy for foundation trusts is no longer significant. The elective variable, indicating the number of elective H81 operations each year and the change in number of elective operations for year to year in models 2.1 and 2.2 respectively, indicate that the number of H81 procedures is increasing as an elective option. Year dummies were included in the analysis and sometimes significant.

Table 6.6: Quality effects on Hip Replacement patients (filtered outcome indicators).

	$D30_{ht}$	$D30_{ht}$	$D365_{ht}$	$D365_{ht}$	$R28_{ht}$	$R28_{ht}$	$R365_{ht}$	$R365_{ht}$
	$H80$	$H81$	$H80$	$H81$	$H80$	$H81$	$H80$	$H81$
H80	-2.11e-05		-2.28e-05		1.58e-05		-3.66e-05	
	(1.93e-05)		(4.48e-05)		(5.32e-05)		(7.57e-05)	
H81		1.41e-07		-4.06e-08		3.79e-06		1.40e-05
		(1.86e-06)		(5.41e-06)		(7.19e-06)		(9.09e-06)
tariffH80		-5.45e-05		-0.000170*		-0.000140		-6.80e-05
		(3.95e-05)		(9.91e-05)		(8.93e-05)		(0.000132)
tariffH81	-4.79e-06		-2.88e-05		1.33e-05**		3.04e-05***	
	(6.36e-06)		(2.31e-05)		(5.57e-06)		(9.96e-06)	
LOS	0.00250	-0.000773	0.0148	-0.000619	-0.00396**	0.000357	-0.00147	0.000321
	(0.00247)	(0.00108)	(0.00932)	(0.00246)	(0.00183)	(0.00329)	(0.00338)	(0.00398)
Co-morbidity	-0.0301	0.0162	-0.0861	0.0322	0.00571	-0.0148	-0.0456**	0.0696
	(0.0191)	(0.0234)	(0.0803)	(0.0543)	(0.0198)	(0.0904)	(0.0204)	(0.122)
Deprivation	-0.00142	-0.00878	-0.0284	-0.00934	-0.0199	0.0274	-0.0305	0.00581
	(0.00828)	(0.00695)	(0.0296)	(0.0187)	(0.0175)	(0.0258)	(0.0245)	(0.0315)
Age	0.000766	-0.000522	0.00340	-0.000605	-0.00486*	0.00249	-0.00621**	-0.00314
	(0.00127)	(0.00181)	(0.00614)	(0.00207)	(0.00253)	(0.00381)	(0.00288)	(0.00511)
PbR	-0.135***	-0.0965***	-0.362***	-0.253***	0.00963	0.0222	-0.250***	-0.0206
	(0.0142)	(0.0100)	(0.0573)	(0.0199)	(0.0150)	(0.0453)	(0.0262)	(0.0483)
FT	-0.0107*	-0.0152	-0.0350	0.0106	0.0218	0.0207	0.0189	0.00809
	(0.00562)	(0.0103)	(0.0218)	(0.0199)	(0.0158)	(0.0252)	(0.0230)	(0.0354)
ISTC	-0.0261***	0.0518***	-0.0366	0.117*	-0.0294	-0.111***	-0.108***	-0.190
	(0.00750)	(0.00787)	(0.0415)	(0.0621)	(0.0267)	(0.0272)	(0.0285)	(0.152)
Constant	-0.0552	0.0472	-0.286	0.0479	0.309*	-0.194	0.278	0.118

	$D30_{ht}$	$D30_{ht}$	$D365_{ht}$	$D365_{ht}$	$R28_{ht}$	$R28_{ht}$	$R365_{ht}$	$R365_{ht}$
	$H80$	$H81$	$H80$	$H81$	$H80$	$H81$	$H80$	$H81$
	(0.103)	(0.116)	(0.490)	(0.145)	(0.179)	(0.278)	(0.180)	(0.376)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,510	1,366	1,510	1,366	1,505	1,273	1,510	1,366
R^2	0.744	0.475	0.761	0.609	0.285	0.094	0.681	0.079
Number of h	120	120	120	120	119	119	120	120

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

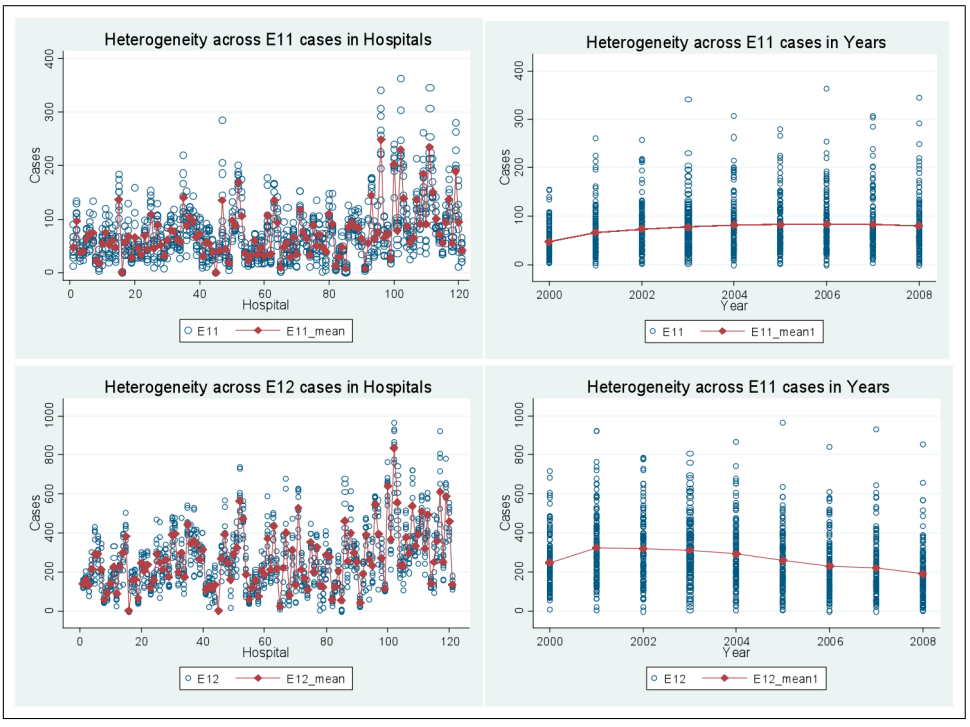
Models 7 and 8 look investigate what effect PbR and the changes in caseload had on quality o f care, using the latent variables constructed for each group. The R-squared values of the models indicate that the mortality models are able to explain around more of the variance than the readmission models, and of the two groups the H80 models have higher R-squared values than the H81 groups. The R-squared values of the H80 mortality models range between 75-80%, while the H81 groups mortality models explain about 50-60% of the variance. The R-squared values of the H80 readmission models are nearly 30% for the 28-day readmission model and nearly 70% for the year-long readmission model, the respective values for the H81 readmission models are considerably lower at 9% and 8%. The reason for this difference is most probably due to the large amount of missing values in the H81 group, which were also prohibitive in creating the filtered indicators.

The PbR dummy is significant for most models, and indicates a decline in mortality, at all intervals, for both the H80 and H81 conditions. Yet, the magnitude of the coefficient indicates that the decline in mortality is very small. The PbR dummy only indicates a decline in year-long readmissions for the H80 group, and is insignificant in all the other readmission models. Average age is negatively associated with short and long term H80 readmissions, such that hospitals with a higher age group have lower readmissions for H80. The tariff for H80 is also associated with H80 readmissions, such that an increase in the tariff leads to more readmissions for this group, while tariff for H81 is negatively associated with year-long H81 mortality. Finally, foundation trusts status and ISTC status are significant in many of the models, such that foundation trusts have lower H80 30 in-hospital mortality as compared to other hospitals, while ISTCs have higher mortalities at every level for the H81 group, and lower 30-day mortality for the H80 group. ISTCs also have lower 28-day readmissions for the H81 group, and lower year-long readmissions for the H80 group.

Sensitivity Analysis

A fixed effects model was used to estimate all models to control for the individual time invariant characteristics of the hospitals due to the observed heterogeneity across providers (Figure 6.3), similarly year dummies were used to control for the observed heterogeneity across years (Figure 6.3). However, the models were also estimated with random effects, and the key results remain unchanged. A Hausman test was used to determine which model is preferred for each case, and the fixed effect model is always indicated for Model 1 and the quality models (Models 3-6), while either estimator can be used for Model 2[2]. In the random effects model dummy variable could be included for teaching hospitals, which due to its time-invariant status was dropped from the fixed effects model. However, the results from the random effects model indicate that teaching status had no significant effect for any of the conditions. Models 1 and 2 were run using the less expensive case as the dependent variable to check the consistency of the results.

Figure 6.3: Heterogeneity across cases over provider and time



[2]When the Hausman test has a significant P-value it suggests that the coefficients estimated by the efficient random effects estimator are not the same as the consistent fixed effects estimators, and so fixed effects is preferred. When the P-value is insignificant it is safe to use either random or fixed effects.

6.5 Policy Implications

The English PbR reimbursement policy is predicated on a single tariff for predicted levels of activity adjusted by case-mix, regardless of where it is performed. The main rationale for paying providers in this way is to drive down the costs of those who are providing services in excess cost of the tariff, thus enhancing the efficiency in provision of health services. Yet, evidence from the adoption of DRG payments in the US has indicated that professional discretion plays an important part in determining hospitalization for DRGs, and that losses in hospital revenues resulting from this type of payment system can be offset if physicians modify their admission policies to produce more profit, even within the limits of medical appropriateness (Wennberg et al., 1984). This section looks at data on tariff to understand how hospital revenues for these two conditions have changed since the PbR policy.

As noted in the data section, our tariff variable is calculated by adding together all costs a patient accumulates under their spell of care, including any readmission costs, or extra length of stay. However, the tariff only represents true costs reimbursed after the adoption of PbR policy, prior to PbR the variable is only indicative of estimated patient costs, and so is less reliable. Figure 6.4 and 6.5 consider the change in spending in the four HRG groups across the time periods studied. In each of the figures, the tariff for each HRG represents the sum of costs accumulated by all patients in that HRG group for each hospital in each year. Figure 6.6 depicts the total spending across hospitals over the time period studied for treating patients in HRGs E11/E12 and H80/H81.

While overall AMI spending has decreased over time, and more noticeably since 2004, spending on Hip Replacement has been rising throughout the period with no change in the upward trend during the time PbR was implemented. Figure 6.6 illustrates the variation in spending across hospitals for each of the four HRG groups in relation to the mean spending of all hospitals in that year. While it is difficult to discern any change in variation for most of the HRG groups, there is a clear rise in heterogeneity of spending across hospitals for HRG H81, especially after 2004. The increased variation in spending across hospitals is coupled with increased average spending from year to year.

Figure 6.4: Total AMI spending 2000-2008.

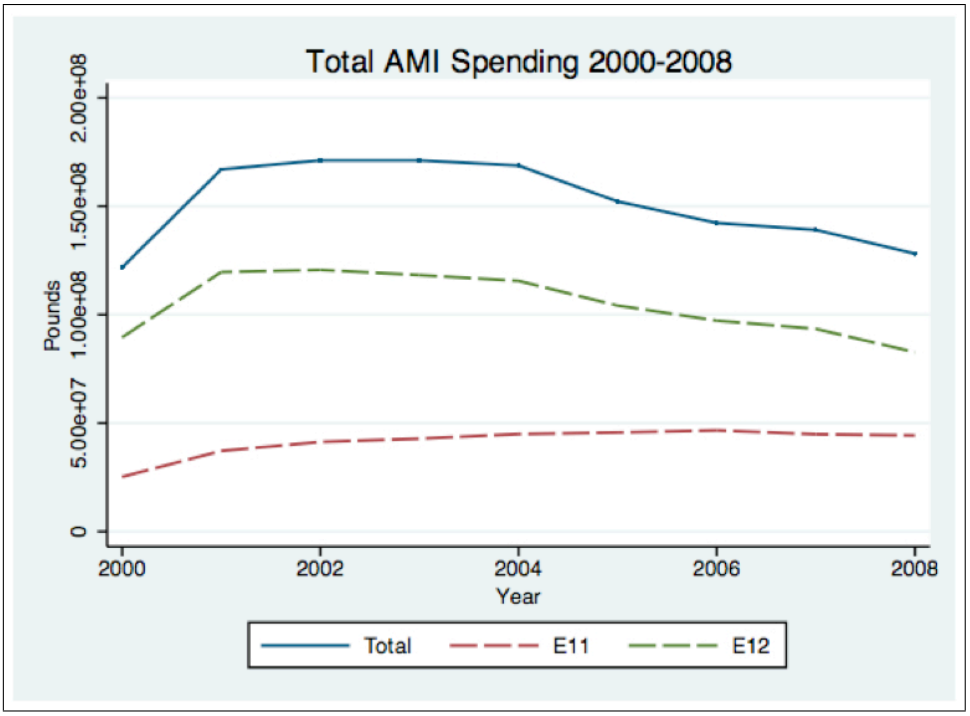


Figure 6.5: Total Hip Replacement spending 1996-2008

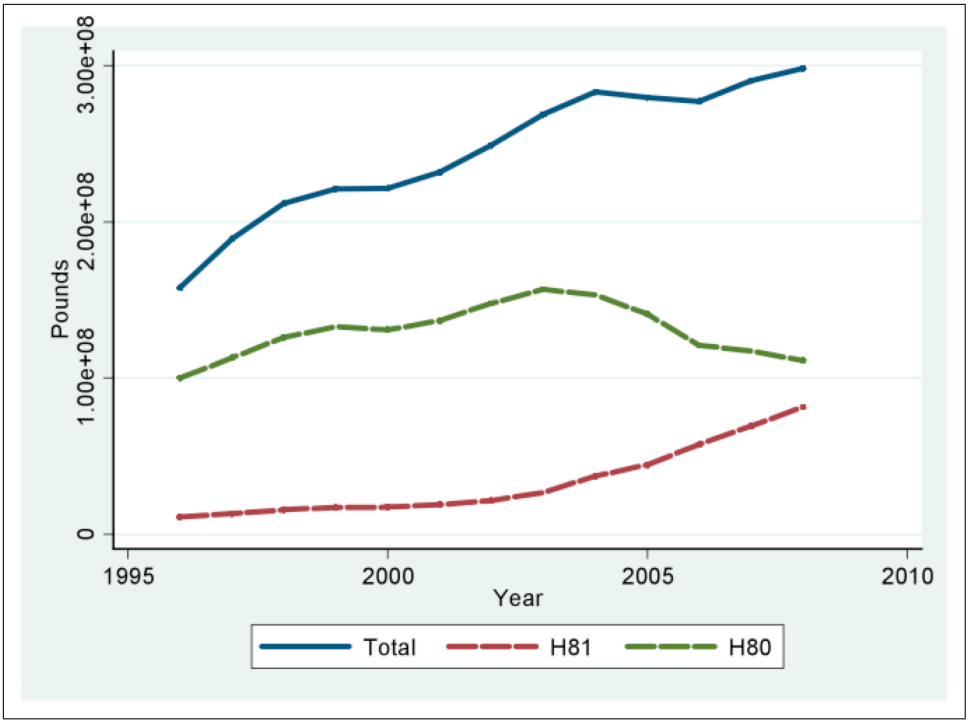
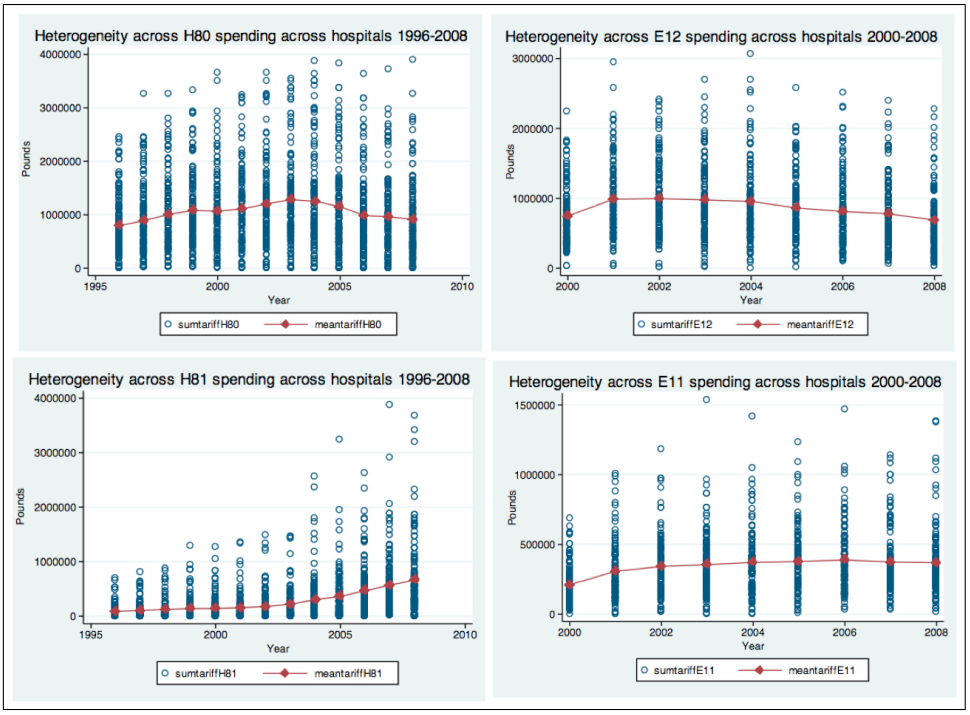


Figure 6.6: Heterogeneity in spending across hospitals over time



6.6 Discussion

The literature on case payment systems identifies many possible positive and negative incentives attached to this type of provider payment. The positive incentives, such as increased efficiency and transparency, are particularly desirable for most health systems. However, this type of payment system also runs the risk of incentivising adverse behaviours such as gaming, upcoding, and quality skimping. To date there have been no reports of such behaviours in the English PbR system, however some authors have noted increases in activity. This chapter considers two particular pairs of HRGs, where there appears to be an increase in activity for the more highly reimbursed HRG at the expense of its cheaper counterpart suggesting the possibility of upcoding.

In the case of AMI HRGs E11 (with complications) and E12 (without complications), Models 1 and 2 indicate an increase E11 cases since the implementation of the PbR policy. However, this increase is not the result of 'substitution' between the two HRG groups. In theory we would expect substitution to occur in order for providers to receive the higher tariff for less complicated cases. While tariff is also positively associated with the increase in activity for the E11 group, this is at least partly attributable to the fact that more cases lead to higher revenues rather than tariff driving more cases. Models 1 and 2 do indicate

that the activity in the E11 group is associated with more co-morbid, older patients who are expected to have an increased likelihood of complications. Indeed, the indicator used to measure co-morbidity, the Charlson Co-morbidity index, is constructed ex post using the patient's clinical data and so it is unlikely that such an indicator would also be manipulated by providers should they be upcoding. The sensitivity analysis shows that the findings are consistent when using E12 activity as the dependent variable for models 1 and 2. E12 activity is significantly associated with younger, less severe patients. However, activity only begins to decline in 2007 and since PbR, indicating that some aspect of the policy is associated with the activity change.

The results from the Hip Replacement group, on the other hand, do suggest substitution between the H80 and H81 HRG cases. Models 1 and 2 indicate that uncemented Hip Replacements are performed instead of cemented Hip Replacements, and their increase is significantly associated with the introduction of the PbR policy. While in theory upcoding is performed in order for providers to profit from the reimbursement difference between the two cases, it is difficult to support an 'upcoding' hypothesis for two different procedures such as cemented and uncemented Hip Replacement as the tariff is based on procedural costs. While the tariff variable is positively associated with the increase in the more expensive case, this is probably due to the increased revenues resulting from more activity. In Model 2, we see that uncemented procedures are associated with younger patients. However, this is expected as the uncemented procedure has a longer recovery period than the cemented alternative. The sensitivity analysis confirms the findings when using the cemented group as the dependent variable, and the decline in cemented cases is also positively associated with PbR.

It is difficult to support an argument of 'upcoding' for either of the two cases being investigated. The lack of substitution in the AMI cases leads us to believe that some other incentive of the policy is responsible for the change in activity. Conversely, in the case of Hip Replacement, where substitution is occurring, the different costs of the two procedures (reflected in the tariffs) would make 'upcoding' theoretically impossible. This leads us to believe that some other incentive of the policy is responsible for the change in hip activity as well. Re-considering the possible incentives attached to such a financing system, we conclude that the mostly likely explanation for the change in activity is a result of increased transparency, efficiency and specialization. Prior to the introduction of HRGs, England had no history of routine cost collection at the patient level. This is due to the organizational structure of the NHS and lack of a substantial private insurance sector that would require detailed billing data at this level (Street and Dawson, 2002). The early introduction of HRGs in the late nineties changed the recording of cost information, as

did the refinement of HRG groupers. Indeed the large increase in recorded activity for all four HRGs from 2000 to 2001 is probably the result of a move from HRG grouper 3.1 to 3.5. With the introduction of PbR, coding became even more important as the entirety of hospital revenues became contingent on the correct recording of information. Thus, instead of prices leading to mis-coding, upcoding or gaming, we support that the change in AMI activity is in fact a result of better coding.

Another incentive created by a case-based payment system is for providers to specialize in areas where they can make the biggest gain due to their competitive edge (Shleifer, 1985). The national tariff in England is set according to the average cost of treatment across all hospitals, thus any hospital who can perform a procedure at a lower cost will make efficiency gains. In a case such as Hip Replacement, where two relatively substitutable treatments are available, one of which is newer and less utilized, it is more likely that the average cost of the newer treatment across all hospitals will be high and that more efficiency gains can be made by specializing in this procedure. Moreover, while uncemented Hip Replacements require a longer recovery period they are less likely to result in complications such as breaking off of cement particles, that will result in readmission. Thus, in the case of Hip Replacement we believe the substitution effect is not driven by upcoding but instead by an attempt of providers to increase their efficiency.

When considering the results of Models 1 and 2 with these hypotheses in mind, the findings are much more intuitive. If the change in AMI activity is the result of better coding, we would expect to see older and more severe patients coded as with complications. There is no substitution between the cases because in all likelihood the decrease of cases in AMI without complications may not be the result of increased E11 cases but better coding overall (resulting in perhaps a different HRG grouping altogether). In the case of Hip Replacement, the substitution between the two conditions begins once PbR is implemented, despite both procedures being available for over twenty years. The controls indicated that younger patients are more likely to receive the uncemented treatment. This is most likely because the type of surgery yields better results when performed on younger, healthier patients due to the longer recovery period. However, it may also indicate an element of selection. Moreover, the choice of uncemented surgery as an elective option has been increasing in the period investigated. The results from models 3-8, which consider the effect the activity changes have had on quality, support this conclusion.

In the case of AMI, the effects of PbR are associated with a decrease in 30-day mortality for both E11 and E12 groups, and a decrease in long term mortality for E12. The change in cases over time is also associated with a decrease in 30-day mortality. If upcoding were occurring we would expect to find an improvement in the outcomes of the E11 case, but

possibly at the expense (or with no change) in the E12 category. Improved mortality in both groups could occur under increased transparency and better coding as it would result in better appropriateness of care. In both cases there is no evidence of the policy having any significant effect on the change in readmissions. However, the model does suggest that the change in cases is associated with an increase in 28-day emergency readmissions for both E11 and E12 cases. While the increase in E11 cases may partly account for this we also need to exercise caution in the interpretation of higher readmissions, which as noted in previous chapters, may not be an indicator of poor quality.

While we were able to run the AMI model using both the latent and the filtered estimates, we were limited in our ability to do the same for the hip model due to issues with sample size across the years. When attempting to construct quality indicators using this methodology at the level of the HRG group this is likely to be a problem for other conditions as well given the finer degree of classification. Moreover, that latent measures constructed will also be subject to wide variation from year to year as a result of the small sample size. Models 7 and 8 which analyse the change in quality for cemented and uncemented Hip Replacement indicate that the PbR policy has had a positive effect on quality overall. Mortality rates at all intervals have fallen for both groups, albeit by very small amounts. There has been no significant effect on readmissions, apart from the cemented group where they have fallen. However, it is possible that this is an indication of worse quality in the H80 case, as we have established that there is a negative relationship between readmissions and mortality. Moreover, the decline in readmissions is associated with more severe patients, and is also positively correlated with the change in average tariff received over time. A decline in quality could possibly be attributed to the fact that younger patients are being substituted into the uncemented group as surgery is less likely to be effective on more frail patients. However, it might also suggest an element of selection – whereby the healthier patients are transferred to the procedure that has the highest payoff.

Some of the other interesting results concern the type of trust providing treatment. While the type of trust has no effect on the number of cases or quality in AMI, it is important in the case of Hip Replacement. Part of the explanation for this difference most likely stems from the type of procedures being investigated; the treatment for AMI being considered is all non-elective, whereas the Hip Replacement conditions are a mix of elective and non-elective care. In the case of Hip Replacement foundation trusts are found to perform fewer uncemented Hip Replacements, while independent sector treatment centres perform more. However, the only effect this has on quality is that there is a higher 28-day emergency readmission rate for the ISTCs as compared to acute trusts.

The final part of our analysis considers what these efficiency gains are costing the system. This part of the analysis is considerably limited however as it only considers spending within the HRG groups and not the conditions overall, or the wider setting. However, as expected the increase in activity in the Hip Replacement cases are resulting increased spending overall, while the large decline in E12 cases, possibly as a result of better coding, has decreased spending for the pair of AMI HRGs. More interesting is the change in variations of spending across hospitals. While there appears to be no noticeable change in HRGs E11, E12 and H80, there is an obvious increase in the H81 group. Earlier results indicate an increase in uncemented surgery, H81, as an elective treatment. This could be a result of the differential pricing between the groups which unintentionally incentivizes providers to switch into the new technology. In term of costs, it appears that hospitals are indeed offsetting reductions in revenues resulting from this type of payment system by produce more profit in this area within the limits of medical appropriateness.

Considering the two case studies, and the relationships we are able to establish between PbR, activity, quality and spending we see that the payment system is providing strong behavioural incentives, however we do not find evidence of these incentives resulting in upcoding behaviour. In the case of AMI, despite the increase in E11 cases and decrease in E12 cases, the substitution effect is not significant. Instead it appears that this change in activity occurs because more attention is being paid to coding since its attachment to a monetary incentive. In the case of Hip Replacement, where there is evidence of a substitution effect between cemented and uncemented cases, again we do not attribute this to upcoding but instead to an attempt for providers to make efficiency gains.

Thus at this stage, we find that the PbR policy in England so far yields relatively positive gains, in terms of increased transparency and efficiency. While spending at the hospital level seems to be still adjusting to the new policy – we expect that with time when providers become more familiar with coding practices and costs for all areas will be adjusted to take into account activity changes such as the one investigated in this chapter – it will stabilize.

Part IV

Conclusions & Policy Recommendations

7 | Conclusions & Policy Recommendations

The ultimate drive to define and collect quality information has been from stakeholders in order to achieve broadly two purposes: the improvement of quality services, and/or the holding services to account for the quality of care they provide. The users of quality information are key stakeholders in the health system, such as clinicians, health service managers, policy makers, patients, tax-payers, the media, researchers and even industry. The uses of quality information differ according to the stakeholder's needs but can be broadly classified into a spectrum of activities, where one end is focused solely on learning from data and the other on making judgments from this data. In an extension of Freeman's work (2002), Davies (2005) considers the differences between these two approaches, coined the improvement and accountability approaches respectively.

Table 7.1: Differences between accountability and improvement approaches.

	Accountability Approaches	**Improvement Approaches**
Emphasis:	Measurement oriented; favouring verification and assurance	Insight and change oriented; favouring learning to promote continual improvement.
Rationale:	To provide external accountability and ensure/renew legitimacy.	To promote internal change and continuous quality improvement.
Culture	Comparisons drawn in order to make summative judgements on quality.	Comparisons drawn with a formative emphasis aimed at learning from difference and diversity.
Data Presentation:	Data presented as league Tables inviting naming, blaming and shaming.	Data presentation emphasises informal benchmarking and acknowledges 'casual ambiguity'.
Precision Required:	High precision needed.	Lower precisions acceptable.
Epistemology:	Empirical. Validity and reliability important alongside statistical assessments of difference.	Interpretive. Use of other data sources and local information acceptable to provide contextual and qualitative understandings.

Source: as adapted in Davies (2005) from Freeman (2002).

This body of work has been concerned primarily with the accountability approach: focusing on creating precise measures of quality to make comparisons across providers and across time. This work would be less suitable for an improvement approach as it only considers the final outcomes, and does not consider the processes and structures that contribute to their attainment. When creating quality indicators for either of these approaches however, it is crucial to be sure that the performance measures being used exhibit the characteristics of acceptability, feasibility, reliability, sensitivity to change and validity (Table 7.2). With regards to measuring quality *per se* we have noted throughout the chapters that there are numerous specific methodological challenges that need to be taken into account in order to create good quality measures that meet the above criteria. We decided to apply the McClellan and Staiger (1999) methodology to create good quality indicators, as we believe is able to overcome many of the challenges associated with quality measurement and create indicators that meet these criteria.

Table 7.2: Qualities of Good Performance Measures:

Development of Indicators:
• Face/content validity: the extent to which the indicator accurately measures what it purports to measure.
• Reproducibility: the extent to which the indicator would be the same if the method by which it was produced was repeated.
Application of Indicators:
• Acceptability: the extent to which the indicator is acceptable to those being assessed and those undertaking the assessment.
• Feasibility: the extent to which valid, reliable and consistent data is available for collection.
• Reliability: the extent to which there is minimal measurement error; that the extent to which findings are reproducible should they be collected again by another organization.
• Sensitivity to change: the extent to which the indicator has the capacity to detect changes in the unit of measurement.
• Predictive validity: the extent to which the indicator has the ability to accurately predict.

This body of work has attempted to evaluate the impact a change in payment mechanism has had on quality of providers. In order to be able to assess this impact we first faced the challenge of creating adequate measures of quality (Chapters 2 and 3) and un-

derstanding the relationships between them (Chapters 3 and 4). Finally, we were able to use them to evaluate the PbR policy (Chapters 5 and 6). In this chapter we draw out the most important findings from each of the chapters, and consider them as a whole. This exercise allows us not only to assess the findings in their entirety but also to consider what main lessons and policy recommendations emerge. We begin this exercise by briefly reviewing some of the most important findings from each chapter, before combining them to draw overall conclusions. We then go on to consider some of the limitations in data and methodology which need to be acknowledged in order to correctly interpret the findings. Finally, we conclude by considering what policy recommendations we can make from this information, and what areas would benefit from further research.

7.1 Key Findings

The fundamental requirement for any quality measurement technique is to find a suitable way to filter out the inherent noise that is present in outcome measures. This noise will be the result of measurement error, confounding variables, systematic error and chance. Moreover, quality is a multidimensional notion. While outcome measures represent the final outcome of the health system, and thus are often considered more meaningful that other measures, they are still fragmented as they consider only one area. Finding a method that is able to adequately deal with these challenges is difficult, however lacking to do so will result in inappropriate measures. The first step to applying or even developing an adequate methodology to measure quality is identifying the key challenges and methods available to deal with them, as outlined in the introductory chapter.

Measuring Quality

Part II, 'Measuring Quality', considers the application of two of the techniques discussed in the Introductory Chapter to individual patient level data from England. Chapter 2 uses a latent variable technique to create risk-adjusted measures of quality for each hospital for the treatment of seven different conditions. These measures are calculated separately for each year, and provide information about relative hospital performance in that year controlling for patient characteristics. The results indicate that where there is sufficient sample size, the latent quality measures are good risk-adjusted indicators with relatively small confidence intervals. The indicators can be used to examine the relative quality of each hospital over time, as well as to understand the rate of change in average quality across hospitals.

When used at the individual hospital level, the indicators can be subject to large

volatility from one year to the next, especially in hospitals with smaller sample size. When aggregated and used to examine trends in the rate of change in average quality across hospitals, the latent measures are able to indicate a different picture from the raw outcome measures, suggesting that when controlling for case-mix the rate of change in quality was greater than what appeared to be the case from the raw data. Indeed, this is consistent with some of the changes reported in standardized mortality rates for hospitals. There has been some discussion about the pattern of falling risk-adjusted mortality rates in the literature as crude data is relatively unchanged. Possible explanations include recent changes in coding, as well as increasing severity of conditions treated by hospitals as milder cases are increasingly treated in the community or as day cases (Hawkes, 2010b).

While the methodology appears to create risk-adjusted indicators which are consistent with other risk adjusted measures, is unable to address all the issues identified earlier. The latent indicators will not be able to separate measurement error and systematic error from unobserved quality and thus may be a poor measures of true quality in both cases. Moreover, they are not able to incorporate other dimensions of quality and thus are also limited as to the extent they provide a complete picture. Indeed, these limitations are also true of other, more common, risk-adjusted measures, such as standardized mortality rates.

In order to address these limitations, Chapter 3 uses the latent measures created for the seven conditions for each year in the sample, and applies a methodology introduced by McClellan and Staiger (1999). The methodology essentially smoothes out the variation from the indicators using the times series and cross sectional information from all indicators. This allows the indicators to pick up more of the quality signal present in the indicators and filter out more of the systematic bias and noise. The method is easily applicable to the English data, and indeed the R-squared measures used to evaluate their predictive and forecast validity indicate that it performs better than it did for the US data. This methodology allows us to predict future outcomes with very high levels of certainty. Unlike the US data however, we find the indicator with the strongest signal for almost all of the seven conditions to be year-long mortality, instead of 30-day mortality. However, 30-day mortality also has a strong signal, and in most cases is stronger than the two readmission indicators.

As the methodology uses the cross-sectional information between the different outcome measures to create the new indicators, we are also able to draw conclusions about the relationships between them. One notable finding is the relationship between readmissions and mortality. In many instances these indicators are negatively correlated – suggesting that higher readmissions may not always be indicative of poor quality, but indeed might

suggest the opposite. Similar to Chapter 2, the indicators produced can be used to examine the relative quality of each hospital over time. While the variation from year to year is smoothed out, the confidence intervals for the quality measures are larger, thus when used to assess relative performance it is difficult to draw conclusions for any one hospital.

Overall, we find evidence to support the use of this methodology for the construction of more sophisticated measures of quality that can be used with more confidence to evaluate the quality of providers.

7.2 Evaluating Quality

General Findings

While the section entitled 'Evaluating Quality' uses the indicators developed in Chapters 2 and 3 to assess the impact of the PbR policy, we also use the indicators to evaluate other areas of quality in Chapters 2 – 4. Chapter 2 uses the latent measures to assess the key determinants of quality in the period being investigated, while Chapter 3 ranks hospitals according to the different indicators for a random year, in order to compare the results. Finally, Chapter 4 considers how the different latent and filtered indicators are related across the seven conditions for which they are constructed. We briefly review the key findings from these analyses before going on to review the results from Chapters 5 and 6, summarized in the following sub-section.

In Chapter 2, the latent quality measures are regressed against lags of themselves as well as other possible determinants of quality, in order to gain insight as to what factors influence quality. The results suggest that quality is dynamic for most conditions, although not for all indicators. Moreover, while many of the lagged quality measures are significant, their influence on current quality is not always positive as might be expected. Indeed for some conditions the lagged variable is negative suggesting that poor/good quality in the past is associated with good/poor current quality. We term the latter effect 'change', and the former 'path dependency', however the analysis in Chapter 2 does not allow us to draw conclusions as to why we see one effect for some conditions and not for others. Moreover, the results from Chapter 2 do not suggest any conclusive findings on how hospital type influences quality. For some conditions foundation trusts have higher 30-day mortality, and specialist trusts lower 30-day mortality, while university hospitals have lower readmissions for CCF. Of the other exogenous variables, waiting times, caseload and length of stay have mixed results for the different conditions.

Chapter 4 considers how the different latent and filtered indicators are related across the seven conditions for which they are constructed. Increasingly we use outcome measures

of a particular condition to represent 'quality'. This is common for conditions where evidence suggests their outcome is highly linked to hospital quality, such as AMI (McClellan and Staiger, 1999). However, it is important to be aware of the relationship between these indicators and their counterparts for other conditions, to ensure that this generalization is valid. The results from Chapter 4 indicate that most conditions are very dynamic and endogenous, such that their own past performance is very highly correlated with current performance and not highly influenced by performance in other conditions. For most of the outcome measures studied in the seven conditions, we found that its own past performance from 3-years prior was often significant in determining its current performance.

In order to test the relationships between the outcomes of different conditions, we used a VAR model. This type of methodology allows us not only to observe associations between variables, but also to infer causality through Granger endogeneity tests. Indeed careful analysis of the Granger causality tests and the regression models allows us to detect relationships between the different conditions. In many cases this suggests that performance in one condition is influenced by the performance in another, although for a very small amount of patients. We identified two types of relationships between conditions, which we call reinforcing and competing. We define a reinforcing relationship is one where good quality in the treatment of one condition is positively associated with the treatment of another. This type of relationship may be found in two conditions that are treated in a similar unit, where you would expect that quality is linked because of shared resources. We define a competing relationship as existing where the quality of treatment is negatively associated between two conditions. We find that this type of relationship exists for conditions in different units, which may be competing for funding. Moreover, we find that these relationships appear to be consistent with medical literature and associations between conditions. This suggests that the quality indicators, and the model, are sensitive enough to pick up areas where poor treatment in one area will influence the success of treatment in another.

Given that each condition exhibits reinforcing and competing relationships with other conditions, we find it difficult to recommend one condition over others as a good quality indicator for a hospital as a whole. Indeed, our results suggest that focusing on any one condition alone will not give a well rounded assessment over quality. However, of the seven conditions chosen, AMI is probably best at assessing hospital quality - at least for the cardiology department, as it is least influenced by poor outcomes in other areas. Yet we would caution against using any single condition as a proxy for quality of the hospital as a whole. In fact, in our evaluation of quality in other chapters we find differences as to the factors and behaviours that influence quality for each of the seven conditions.

In Chapter 3 we rank our sample of hospitals according to the raw outcome data, the latent quality measures and the filtered quality measures for AMI in the year 2005. We find is that the three methods yield entirely different rankings. As discussed previously, one of the main motivations for quality measurement is to use indicators to hold providers to account, often by making information publicly available, attaching incentives to it or using in to inform providers as to their relative performance. The two methods we use to create measures of quality, are methodologically sound risk-adjustment techniques. Indeed, the McClellan and Staiger (1999) method builds upon the latent measures created in Chapter 2 so we would expect some association between them. However, it appears that applying the McClellan and Staiger (1999) method to the data completely changes the relative rankings of the hospitals. This implies that the methodological accuracy of the different methods varies considerably, most likely because the latent variables still contain large amounts of noise. Moreover, these results suggest that policy makers are very cautious about how they use risk-adjusted mortality indicators as they may not be accurate.

PbR

Finally, Chapter 5 and 6 use the latent and filtered indicators constructed in Chapters 2 and 3, and the information about the different indicators from Chapters 2–4, to undertake an in-depth analysis of how PbR has influenced quality in English hospitals. In Chapter 5 we investigate the effect PbR has had on the level quality of care for all seven conditions, as well as how it has influence the relative performance of hospitals to one and other. Each analysis is conducted using both the latent and the filtered indicators. Consistent with our findings in Chapter 3, where the rankings using the latent and filtered indicators were very different, we find the results of the models to differ. As the filtered indicators are able to reduce more of the noise and systematic bias from the latent indicators, we base our findings on those results.

Nevertheless, in the majority of cases, regardless of the indicator used, we find a consistent effect on quality since the introduction of PbR. Yet, the effect on quality differs for the different conditions. Our models indicate that PbR has had more of an impact on some conditions, such as AMI, Stroke and IHD than others, namely TIA, MI and CCF. Moreover, while the effects of PbR are associated with reduced mortality for most conditions, it has varying effects on readmissions. Yet, as we know from Chapter 3, increases in readmissions are not always indicative of worse quality. The results of all indicators suggest that for AMI, hospitals treating more deprived and severe patients have had to skimp on quality, while for Stroke patients the less severe patients have exhibited higher mortality and readmissions.

In Chapter 6, we look at more detail at two pairs of HRG groups, one for AMI and one for Hip Replacement where there appears to be 'upcoding'. Upcoding refers to the transfer of less severe patients to a more expensive HRG category in order to make more profits. In both cases we find that 'upcoding' *per se* is not occurring. Instead, in AMI it appears that due to the policy's emphasis on coding, which determines hospital reimbursement, coding of patients has improved. So while there appear to be more cases in the expensive HRG group, these patients are more severe patients who are properly accounted for. This in turn has had a positive impact on quality, presumably because severe patients receive more appropriate treatment.

In the case of Hip Replacement we see the substitution of cases from one group to the other between two different operations, cemented and uncemented Hip Replacement. Cemented Hip Replacement is an older technology, while uncemented Hip Replacement is newer, yet, up until the introduction of PbR it had relatively low uptake. While there are reasons to select one over the other, it is commonly accepted that for less complicated patients the newer technology has considerable benefits in the long term. Since PbR, the use of this technology has increased dramatically, while use of the cemented technique is falling. Our models do suggest there is a substitution effect occurring between the two conditions, motivated by the introduction of the policy. We believe the incentive for this switch is provided by the different rates of reimbursement for the different surgeries. Since there was a low uptake of the uncemented technique it is likely that the average cost used as the basis of the tariff was not reflective of true cost of implementing this technique. Thus, hospitals felt there was more to make an efficiency gains by performing this procedure over the cemented procedure.

In both cases we find that PbR has had unintended effects on quality, yet in both cases these are positive effects. By tying reimbursement to coding, hospitals are encouraged to improve the coding of patients, and thus quality of care, as we see in the case of AMI. For Hip Replacement, due to the low adoption of the new technology, the tariff encouraged hospitals to become efficient in its adoption in order to benefit from efficiency gains. Both these results demonstrate the power financial incentives can have on performance.

7.3 Limitations

Data limitations

The saying goes: "your results are only as good as your data"; and indeed we acknowledge the necessity to review the known limitations in the data used to conduct this analysis. While some of the limitations are difficult to overcome, we believe that the McClellan and

Staiger (1999) method is of the most capable to deal with data variations and inaccuracies. Moreover, by rigorously studying the data we are able to learn from mistakes and build upon it to create better data in the future. Nevertheless, it is important to be aware of these limitations, especially when it comes to drawing policy conclusions from any body of work that is based upon them.

The underlying data used to conduct the analyses contained in this body of work was hospital episode statistics (HES data) accessed through Dr. Foster. This is the same data that is used to create standardized hospital mortality rates, as well as the Dr Foster Good Hospital Guide, both of which are widespread indicators of quality in England. As the publication of performance information has become more prevalent internationally, but also in England, more studies have focused on understanding how 'good' the underlying data used to measure performance are. Indeed the validity and completeness of Dr. Foster and HES data has been questioned in numerous publications (Hawkes, 2010b,a; McKee and James, 1997; McKee et al., 1999; Mohammed et al., 2004, 2009; Westaby et al., 2007; Williams and Mann, 2002).

Williams and Mann (2002) note that many of the NHS data definitions are convoluted and thus open to misinterpretation, such as the definition used to describe an 'episode' or a 'spell'. They are also concerned with the definition of primary diagnosis as 'main condition treated or investigated during the episode of health care' and the low recording of secondary diagnosis, at around 10%. This is outlined in more detail in (McKee and James, 1997; McKee et al., 1999), who notes the serious implications for comparative research that these omissions may have. However, given our findings in Chapter 6 it is likely that coding has improved overall since the introduction of payment by results. Hawkes (2010b) also suggests that improved coding is occurring in hospitals accounting for what appear to be hospital improvements without any change in underlying performance.

Some authors caution that even mortality may not always be recorded accurately (Mohammed et al., 2004; Westaby et al., 2007), although not to such a degree that it would influence standardized hospital mortality rates (Mohammed et al., 2004). Another issue in the data is the increase use of the code for palliative care. In the past five years there has been a big increase in this code, from 7% in some hospitals to 50% in others Hawkes (2010a). Patients coded this way are assumed to have come to the hospital to die and are coded this way to prevent putting the blame of their death on the hospital. However, this change in coding will influence quality as is measured by risk-adjusted mortality.

With regards to the Dr. Foster data in particular, Mohammed et al. (2009) notes that there are systematic differences in the associations between hospital mortality and the

factors we have used to adjust for patient case-mix, such as age, emergency admissions and co-morbidity. While these differences will play a role in influencing standardized mortality ratios (Wright and Shojania, 2009) they will be accounted for by the McClellan and Staiger (1999) methodology adopted in Chapter 3. In addition, some of the clinical audit literature from the US (Hsia et al., 1988) and England (Cox and Koutroumanos, 2010) suggests that there will also be error in the coding of patients, which will vary from hospital to hospital and by condition. Again, the McClellan and Staiger (1999) technique adjusts for measurement error and so is arguably best suited for data facing this sort of variation and inaccuracy. Finally, the Audit Commission (2008) notes that the average HRG error rate is high, at 9.4% in 2007/8, with considerable variation across hospitals. Our analysis identified patients using ICD-10 and OPCS 4.3 codes, and only uses the HRG variable in Chapter 6 where we limit the sample to investigate four specific HRG groups. Given the limited use of the HRG variable to identify cases, we do not expect this error rate to influence our results greatly.

Method limitations

While we strongly support that the method used performs very well, we would like to acknowledge some limitations to our methodology. In all chapters we do not control for the large amounts of money injected into the English NHS over the period of investigation. It is highly likely that may of the positive quality effects we observe are related to this factor, and not to the change in funding policy alone. Moreover, there were significant changes implemented in other parts of the system which may have contributed to the changes in quality we observed and to the effects we attribute to PbR, such as changes in the payment of primary care physicians. Indeed, as we note in the introduction, the effectiveness of this type of payment-mechanism will be influenced by the organizational structure within which it operates. However, as all the changes are made in the same period it is difficult to isolate the effect of one policy from all others.

In the chapters which analyse the effect of PbR on quality the PbR dummy is set at the year 2005 for AMI and 2006 for the remaining procedures. However, PbR was implemented over the years 2003-2006 staggered over different conditions, different types of hospital, and patients admitted through different routes. Given the small sample of hospitals being examined we did not create staggered variables to control for this. However, in the sensitivity analyses we did run the model varying the year PbR was introduced. As it is likely that many of the positive quality effects observed are also a result of the expectation of PbR, resulting in better coding of patients, this difference may not be so important. Moreover, the descriptive illustrations of quality over the time period indicate

that there are significant changes around the period of implementation.

7.4 Policy Recommendations

Using the key findings that have emerged from this study and tacking into account the limitations of our data and methods, we are able to draw out the most important policy lessons. This section highlights the main policy recommendations indicated from our results:

1. *Improve data collection techniques at the patient level.*
   We begin by emphasizing the need for improved data collection that can inform the development of methodological performance measurement techniques. While England has large amounts of individual level data that is collected annually, they would benefit from ensuring that the coding is consistent across different providers. Other countries, who do not have individual level data, or where it is not made available, can not benefit from advanced methodological approaches of this sort which, as highlighted in this theses, are able to perform better than simple risk adjustment techniques.

2. *Acknowledge limitations to simple risk-adjustment techniques.*
   We believe that aside from improvement in the data used to create quality measures, all quality measurement techniques should recognize the difficulties associated with quality measurement. Chapters 1 and 2 note the many difficulties associated with measuring the quality of providers. Raw measures are inherently noisy, likely to suffer from systematic bias, measurement error and chance, and are also unidimensional. Moreover, adequate sample size is necessary to create confident measures of quality. Most current quality measurement technique are not able to address all these challenges, and thus may not be accurate indicators of quality. Users of quality information should be made aware of the difficulties involved in measurement and the uncertainty surrounding these estimates.

3. *Exercise caution when using risk adjusted measures in accountability approaches.*
   Indeed Chapter 3 indicates the extreme differences in rankings that can result from the use of risk-adjusted quality indicators (the latent measures) as compared to more accurate measures (filtered measures). This suggests that when using quality indicators to inform policy, or in any of the accountability approaches, policy makers need to be very careful of the data they are using. Use of noisy data to incen-

tivize providers, reward or penalize providers or inform decision making can lead to unwanted effects.

4. *Discourage static interpretations of quality.*
   Chapters 2–4 indicate that the latent and filtered quality indicators are highly dynamic. Chapter 4 tells us that in most cases, current outcomes are influenced from outcomes dating back 3 years. This means that the use of static indicators of quality should be discouraged, and modelling of quality should be careful to include lags. Neglecting to do so is likely to lead to incorrect results or interpretations. Indeed there are many documented relationships in economic theory where the inclusion of lags is necessary to avoid model misspecification and misinterpretation of theory, such as the relationship between prices and money (Gujarati, 2003). Drawing from the methods used to model these types of relationships may prove useful in providing a more sound methodological platform from which to analyse quality.

5. *Avoid generalizing quality effects from one condition to others.*
   It is often tempting to generalize quality effects apparent from one clinical area to an entire group of conditions, providers, hospital or system. While there may be certain factors that do influence quality at a macro level, such as management practices, financing or structural changes, there are also others which are very specific such as specialty training and reputation. Chapter 4 indicates that while treatment quality for different conditions is related, the relation is very small. Moreover, improvement in the treatment of quality of one condition does not always translate to quality improvements in another. Thus making generalized conclusions using only one medical condition or treatment may be misleading.

6. *Exercise caution in the interpretation of readmissions data.*
   Another indicator which can be misleading is readmission rates. Chapter 3 suggests that while we usually associate readmissions with lower levels of quality, this may not always be correct. We find that risk adjusted readmissions are often negatively correlated with risk adjusted mortality measures. Thus taking action to reduce readmissions, or penalizing providers for higher readmissions, may be counter-productive. Policy makers should avoid making decisions based solely on readmissions data, but instead should use them together with mortality data to better understand the information they are providing.

7. *Use DRGs to improve coding and transparency.*
   In the use of our quality data to evaluate the English PbR policy, we found that HRGs (English DRGs) were effective in improving coding and transparency. Chapter

6 indicated that this improvement had a tangible impact on quality, especially in the provision of treatment for AMI. Part of the reason this type of system has such a large effect on coding and transparency in England was due to the nature of the system. Prior to case-based payment there was little existing culture of recording cost information, and no incentive to do so carefully. When contrasted with insurance based systems, such as the US, this leads to poorer quality cost information. From a policy perspective, good information of this sort is paramount to assessing efficiency and declining where and how to target policies. We recommend this type of payment system as a way to improve transparency and the collection of cost information.

8. *Incentive payments(bonuses) can be very effective in policy making.*
   While we recommend a case-based payment system to increase transparency and improve coding, we also find that the way in which the tariff is set will be crucial in determining the effectiveness of this payment system. Chapter 6, has illustrated that even unwillingly small differences in the reimbursement of different conditions can have large effects on performance. Indeed, in the case of Hip Replacement we saw that a bonus can lead to large incentives for efficiency, the adoption of new technologies and changes in behaviour. On the one hand this suggests that the incentive payments can be a very useful tool for policy makers to incentivize providers. On the other it suggests that failure to consider all possible incentives created by a tariff may result in unwanted behaviours, which may also lead to adverse behaviours.

9. *Exercise caution when deciding how to set the DRG tariff.*
   One way to vary the incentives created by a DRG type system is to vary the way the reimbursement tariff is set. There is some literature on the existing and theoretical alternatives to this (Ellis and McGuire, 1996; Schreyögg et al., 2006; Street and Maynard, 2007). While the English tariff is set at average costs across all hospitals, other alternatives such as normative pricing, or average pricing of top performers have also been suggested. Given the results observed in Chapter 6 for Hip Replacement and AMI, there may be scope to introduce an introductory coding incentive, as well as differential pricing of new technologies to encourage uptake.

10. *Ensure a good audit system is in place when introducing a financial incentive.*
    Whatever the tariff setting process chosen, any case-based payment system needs to have a good, reliable audit system in place. Our work, as well as many other articles on case-based systems, highlight the scope for the occurrence of unexpected behaviours in reaction to the policy. In order to learn and benefit from the positive

behaviours and to catch or prevent the negative behaviours a good audit system is crucial.

11. *Further research on drivers of quality change for each condition.*
    The findings reviewed in this chapter indicate areas where further research can be beneficial. In order to better understand why quality has changed in each condition, over the time period studied, it would be interesting to do more in depth analysis of each of the conditions. Stroke and TIA in particular exhibit some interesting changes in quality, and interesting relationship to one another. It would be of interested to use these quality variable to better understand the factors that influence quality of Stroke treatment in different providers. For these two conditions, but also for other five, it would be best to look at the outcome measures together with process and structural measures. This would give a more complete idea as to what changes are responsible for influencing quality.

12. *Compare with a control group.*
    These findings would also be more robust if we were able to compare with a control group. Thus it might be interesting to compare the change in quality in English hospitals to the change in quality over the same time period in Scottish hospitals. Since UK devolution in 1998, England and Scotland have diverged substantially in the reforms they have implemented to their health systems. Prior to 1997, England and Scotland funded inpatient care in broadly the same way: health care purchasers and providers negotiated the services that would be provided through bulk contracts (Ham, 2004). Scotland changed this funding system starting in 1997 by unifying purchasers and providers under local health boards, and in 2004 started funding inpatient care through global budgets (Scottish Executive, 2004). In 2004 England moved from the bulk contract system of funding hospital episodes to PbR, a fix-priced activity based payment system. Thus, UK devolution in 1998 has unexpectedly provided one of the best natural tests of how to optimally provide health care through a NHS system, providing a rare and exciting opportunity for research.

7.5 Closing Remarks

While the tools presented in this thesis have been originally applied to the US, their application to the English setting yield interesting results for policy. In the past decade England has invested effort and money into performance measurement and management techniques, ranging from the creation of league tables that rank providers according to a combination of indicators (such as the former Star Ratings and the existing 'Dr. Foster

Good Hospital Guide') to the introduction of new approaches to the evaluation of services (such as the collection of information on Patient Reported Outcome Measures (PROMs) for selected secondary procedures). Efforts outside the public sector have also contributed to the increasing amount of performance information by provider or even health system that is now available in the public domain. This phenomenon is not unique to England, it is far more prevalent in the US system, and growing in other European countries such as the Netherlands, Spain and Germany. While the growing desire to measure, and often report, provider performance is often driven by laudable intentions such as to improve accountability and transparency, there is still uncertainty surrounding many of the methods employed to this end.

Perhaps the most popular indicator of performance, indeed one that providers are often ranked by, is mortality. Indeed mortality is the predominant indicator that contributes to most of the performance measurement initiatives reported above. Death is an obvious measure to chose as it is easily measured, common across settings and most importantly a meaningful indicator to users and providers. However, even when case-adjusted there are many problems with using mortality as a proxy for quality (Lilford and Pronovost, 2010). Some of the main problems are that it has a low signal to noise ratio, meaning that there are along a few number of deaths that are preventable by high quality care as compared to the number of deaths that stem from other causes. Even if we isolate those conditions, common risk adjustment techniques can only adjust for factors that can be identified and measured accurately. Even then, depending on the adjustment technique used results, and thus rankings, are highly variable (Shahian et al., 2010). Not to mention that in some cases, where systematic bias is present and the variable which is being adjusted for varies across the units being compared, risk adjustment can exaggerate the bias is is attempting to reduce. Moreover, most risk-adjusted mortality measures vary considerably across providers (Lilford and Pronovost, 2010). Our own latent variables, an example of this type of indicator, varied considerably from year to year and across providers, when used to rank providers in Chapter 2 we found they were so variable that it was difficult to draw reliable conclusions. These problems with mortality indicators question the appropriateness of their use to judge providers.

In this chapter, and indeed in this entire thesis, we focus on the importance of quality measurement. We argue that in the analysis of outcome measures, such as mortality, sophisticated methods are available that allow for more information to be derived from them. Indeed we find that the marginal difficulty associated with the extra steps in analysing these indicators yields large benefits as to their ability to capture the true quality signal that lies within them. This work demonstrates that much information can be derived

from outcome measures, and their improved performance when used to analyse changes in quality as compared to simple risk-adjusted measures. However, we question the suitability of using only outcomes to assess provider quality overall. A complete picture of quality requires more detailed information as to the processes and structures of care that lead towards them. Outcome measures become more meaningful when they are considered together with the processes and structures that enabled them. Moreover, other outcomes, such as PROMs are being increasingly recognized as important in being able to capture factors such as the patient experience which have been long neglected. As we move to an era where large scale performance measurement is technologically feasible, and performance information is used to inform policy, create incentives and justify reforms, policy makers are in a position to change the way decisions are made. Despite methodological advances made in risk adjustment techniques and the analysis of mortality rates it is important the policy makers keep a wider perspective and try to measure the multidimensional aspects of quality. It is thus imperative that good data collection techniques are adopted, sound methodological tools are applied, and policy makers move towards making evidence the basis for policy reforms and initiatives. With good measures of quality we are able to better assess providers, inform policy and learn from experience. While methodological techniques for quality measurement are improving, it is important that data collection efforts continue to improve alongside them.

Part V

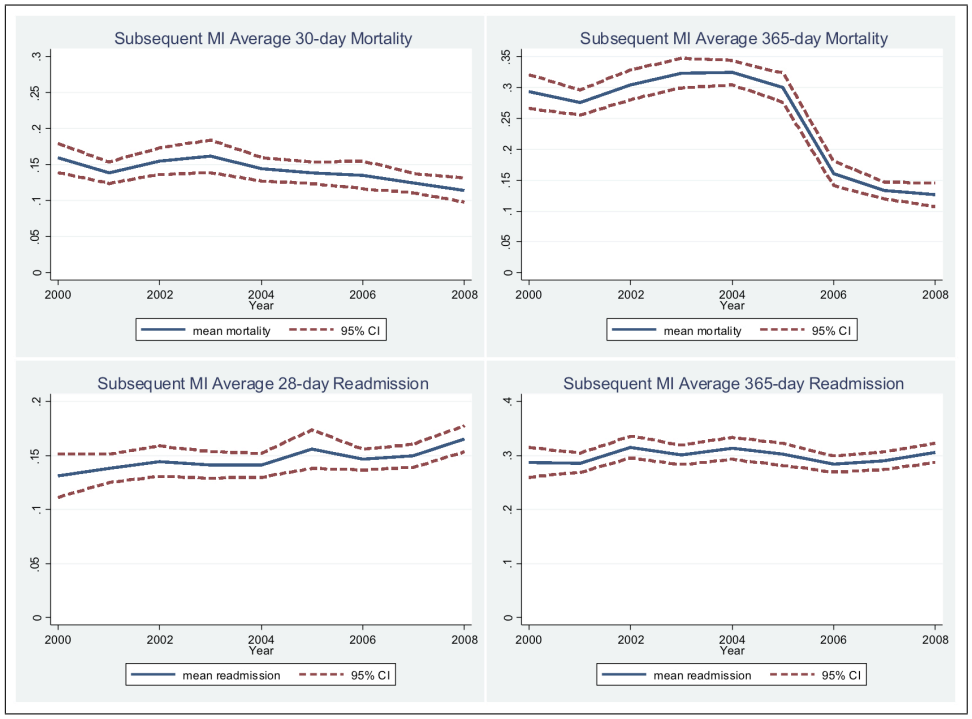
Appendices

A | Results for Chapter 2

A.1 MI

The trends in average mortality and readmission rates for MI over the 2000-2008 time period are presented in Figure A.1. The figure indicates a very small and gradual decline in average 30-day mortality from 2000 onwards, as well as a steady gradual increase in average 28-day readmission over time. The average long term readmissions are constant over the time period. While, average 365-day mortality is constant across time until it undergoes a large sharp drop in the 2005 – 2006 year.

Figure A.1: Trends across years in average MI outcome measures across hospitals.



The patient characteristics influencing the four MI outcome measures, can be determined from the Model 1 regression results presented in Table A.1. Age, and co-morbidities are significantly associated with 30-day and 365-day mortality, where higher age and co-morbidity is related to higher mortality. Gender, deprivation, and type of admission not significant for most of the years run. The results indicate that all patient characteristics included in the regressions are significant determinants of 28-day and 365-day readmissions. Higher age, deprivation and co-morbidity are linked to increased readmissions, women also

have statistically higher readmission rates than men, and in some years patients admitted with elective admission have lower readmissions than those admitted as non-elective. The trust dummies included for each hospital are highly significant for all four outcome measures.

Table A.1: Regression results for MI Model 1.

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
30-Day Mortality							
2000	5473	0.006***	0.009	0.003	0.046***	-0.088*	yes
		(0.000)	(0.011)	(0.002)	(0.005)	(0.05)	
2001	7621	0.006***	0.003	0.003*	0.037***	-0.060	yes
		(0.000)	(0.009)	(0.002)	(0.004)	(0.048)	
2002	7922	0.005***	0.012	-4.14e-04	0.034***	-0.030	yes
		(0.000)	(0.009)	(0.002)	(0.004)	(0.038)	
2003	8492	0.006***	0.005	-0.001	0.037***	-0.013	yes
		(0.000)	(0.008)	(0.002)	(0.003)	(0.040)	
2004	8982	0.005***	0.007	7.39e-04	0.036***	-0.003	yes
		(0.000)	(0.008)	(0.001)	(0.003)	(0.036)	
2005	8597	0.004***	-0.003	-0.001	0.037***	-0.002	yes
		(0.000)	(0.008)	(0.001)	(0.003)	(0.037)	
2006	8830	0.005***	-0.003	-0.002	0.030***	0.040	yes
		(0.000)	(0.008)	(0.001)	(0.003)	(0.030)	
2007	9403	0.004***	0.001	-0.001	0.032***	0.030	yes
		(0.000)	(0.007)	(0.001)	(0.002)	(0.029)	
2008	9799	0.004***	0.004	-4.83e-04	0.030***	0.027	yes
		(0.000)	(0.007)	(0.001)	(0.003)	(0.025)	
365-Day Mortality							
2000	5473	0.011***	0.009	0.003	0.099***	-0.017	yes
		(0.001)	(0.013)	(0.002)	(0.006)	(0.062)	
2001	7621	0.011***	-0.001	0.006***	0.079***	-0.008	yes
		(0.000)	(0.011)	(0.002)	(0.005)	(0.060)	
2002	7922	0.011***	0.015	-2.74e-04	0.082***	-0.037	yes
		(0.000)	(0.011)	(0.002)	(0.005)	(0.047)	

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
2003	8492	0.011***	0.008	0.001	0.082***	0.082*	yes
		(0.000)	(0.011)	(0.002)	(0.004)	(0.050)	
2004	8982	0.011***	-0.009	0.006***	0.081***	-0.023	yes
		(0.000)	(0.010)	(0.002)	(0.004)	(0.047)	
2005	8597	0.011***	-0.021**	0.002	0.076***	-0.024	yes
		(0.000)	(0.010)	(0.002)	(0.004)	(0.047)	
2006	8830	0.006***	-2.53e-0.4	-0.002*	0.036***	0.031	yes
		(0.000)	(0.008)	(0.001)	(0.003)	(0.032)	
2007	9403	0.005***	-0.002	-7.31e-04	0.034***	0.034	yes
		(0.000)	(0.007)	(0.001)	(0.003)	(0.029)	
2008	9799	0.004***	0.007	-9.35e-04	0.037***	0.023	yes
		(0.000)	(0.007)	(0.001)	(0.003)	(0.026)	
28-Day Readmission							
2000	5473	0.001***	-0.001	0.003*	0.018***	0.092*	yes
		(0.000)	(0.011)	(0.002)	(0.005)	(0.052)	
2001	7621	2.91e-05	0.025***	0.003*	0.006	0.052	yes
		(0.000)	(0.009)	(0.002)	(0.004)	(0.050)	
2002	7922	4.06e-04	-0.004	0.002	0.006	0.048	yes
		(0.000)	(0.009)	(0.002)	(0.004)	(0.039)	
2003	8492	7.87e-04**	0.014*	0.004***	0.009***	0.120***	yes
		(0.000)	(0.009)	(0.001)	(0.003)	(0.041)	
2004	8982	9.91e-04***	0.007	0.004***	0.011***	0.080**	yes
		(0.000)	(0.008)	(0.001)	(0.003)	(0.039)	
2005	8597	0.001***	-0.001	0.002	0.008***	0.023	yes
		(0.000)	(0.008)	(0.002)	(0.003)	(0.040)	
2006	8830	7.50e-04**	0.021**	0.005***	0.014***	0.014	yes
		(0.000)	(0.008)	(0.001)	(0.003)	(0.033)	
2007	9403	0.001***	0.007	0.003**	0.010***	-7.14e-04	yes
		(0.000)	(0.008)	(0.001)	(0.002)	(0.033)	
2008	9799	0.002***	0.021**	0.003*	0.011***	-0.017	yes
		(0.000)	(0.008)	(0.001)	(0.003)	(0.032)	
365-Day Readmission							
2000	5473	0.001	0.038***	0.009***	0.028***	0.164**	yes

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
		(0.001)	(0.014)	(0.002)	(0.007)	(0.066)	
2001	7621	0.001**	0.043***	0.004**	0.018***	0.153**	yes
		(0.000)	(0.011)	(0.002)	(0.005)	(0.065)	
2002	7922	0.001*	0.031***	0.008***	0.018***	0.131**	yes
		(0.000)	(0.011)	(0.002)	(0.005)	(0.051)	
2003	8492	0.002***	0.058***	0.008***	0.018***	0.198***	yes
		(0.000)	(0.011)	(0.002)	(0.004)	(0.053)	
2004	8982	0.002***	0.031***	0.007***	0.016***	0.082	yes
		(0.000)	(0.010)	(0.002)	(0.004)	(0.050)	
2005	8597	0.002***	0.024**	0.006***	0.014***	0.079	yes
		(0.000)	(0.011)	(0.002)	(0.004)	(0.050)	
2006	8830	0.002***	0.049***	0.008***	0.023***	0.055	yes
		(0.000)	(0.011)	(0.002)	(0.004)	(0.041)	
2007	9403	0.002***	0.047***	0.010***	0.012***	0.073*	yes
		(0.000)	(0.010)	(0.002)	(0.004)	(0.042)	
2008	9799	0.004***	0.047***	0.010***	0.022***	0.017	yes
		(0.000)	(0.01)	(0.002)	(0.004)	(0.039)	

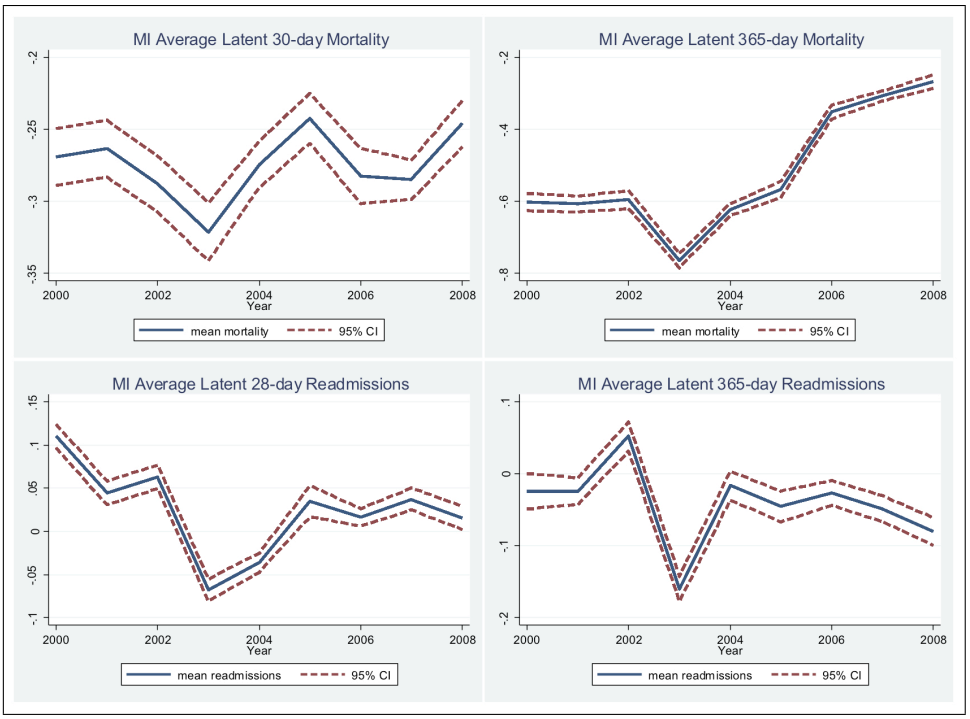
* Significant at $p \leq 0.1$

** Significant at $p \leq 0.05$

*** Significant at $p \leq 0.01$

Figure A.2 shows the average rate of change attributable to hospital quality in all four outcome measures for MI, as measured by the average hospital intercept for each year. Both mortality panels indicate a negative average intercept, meaning mortality, controlling for patient factors, is falling from year to year. For 30-day mortality the average intercept fluctuates, increase at some times and decreasing at others. This means that sometimes the rate of change is decreasing at an increasing rate and others at a decreasing rate. However, for year-long mortality the intercepts are decreasing at an increasing rate for most of the sample. The average latent short and long term readmissions both range around zero. Short term readmissions are above zero before 2003, suggesting that they were increasing during this period, albeit at a decreasing rate. After 2003, they are negative until 2005, suggesting that they were decreasing during this time, but that after 2005 are increasing again. Long term readmissions are negative for all years apart from 2002, indicating that they are decreasing throughout the sample, with a brief increase in that year.

Figure A.2: Trends across years in average latent MI outcome measures across hospitals.



Figures A.3–A.4, show the trend in the latent 30-day and 365-day mortality rates for four selected hospitals treating patients with MI. The confidence intervals for both figures, show considerable variation in latent mortality within hospitals. For both short term and long term mortality, estimates range from about 25% above average to just under 20% below average. This variation is much larger than that observed in the averages in Figure A.3. There is also large year-to-year variation, with jumps of 10% in either direction commonly observed, and in some cases outcomes change by as much as 30% in one year. Figures A.5 and A.6 show the MI latent readmission measures for the same four hospitals. The confidence intervals also indicate wide variation, although the variation among latent readmissions within hospitals is a bit less than it is for latent mortality, at about $15-20\%$. There is also year-to-year variation but of a much smaller magnitude than for mortality, with common yearly fluctuations of about 5%.

Figure A.3: Trends across years in latent MI 30-day mortality for selected hospitals.

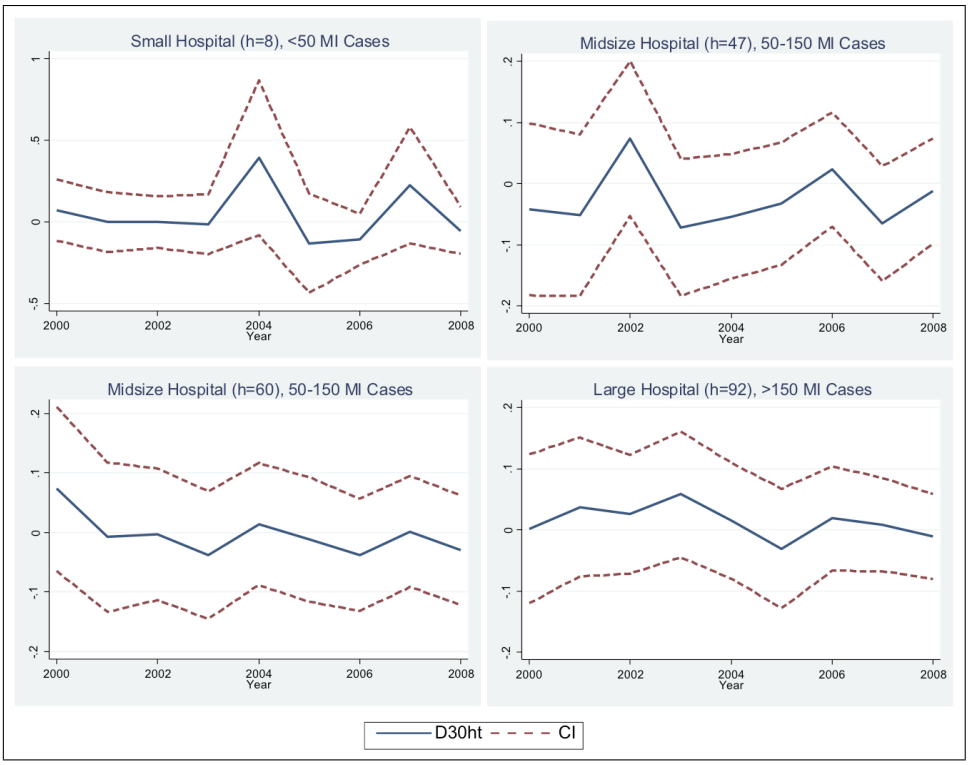


Figure A.4: Trends across years in Latent MI 365-day mortality for selected hospitals.

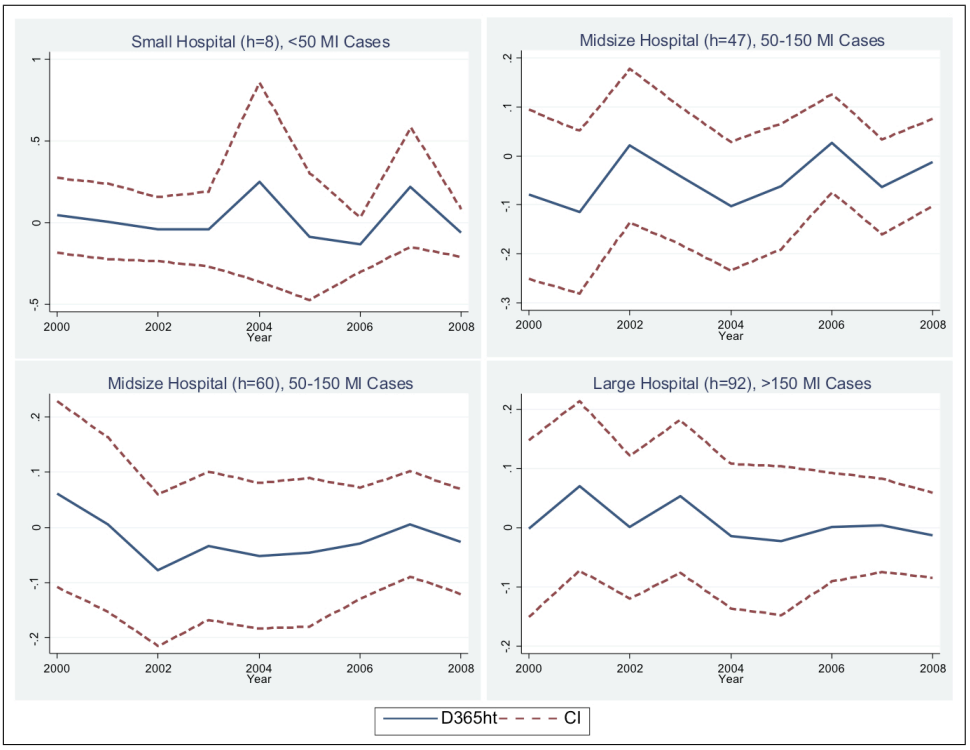


Figure A.5: Trends across years in Latent MI 28-day readmissions for selected hospitals.

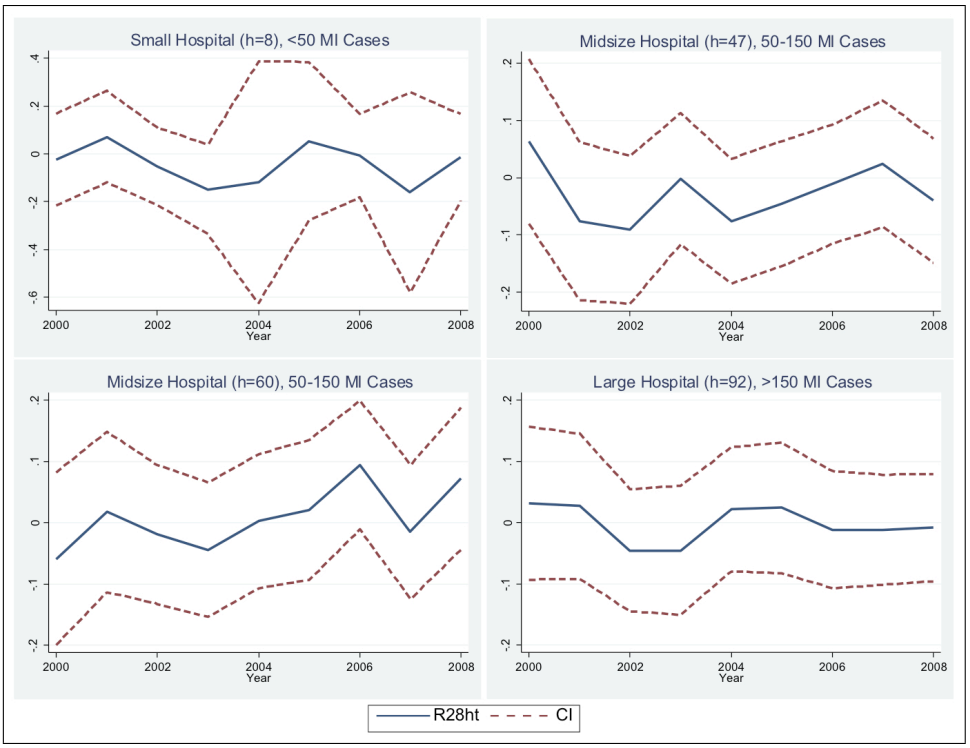
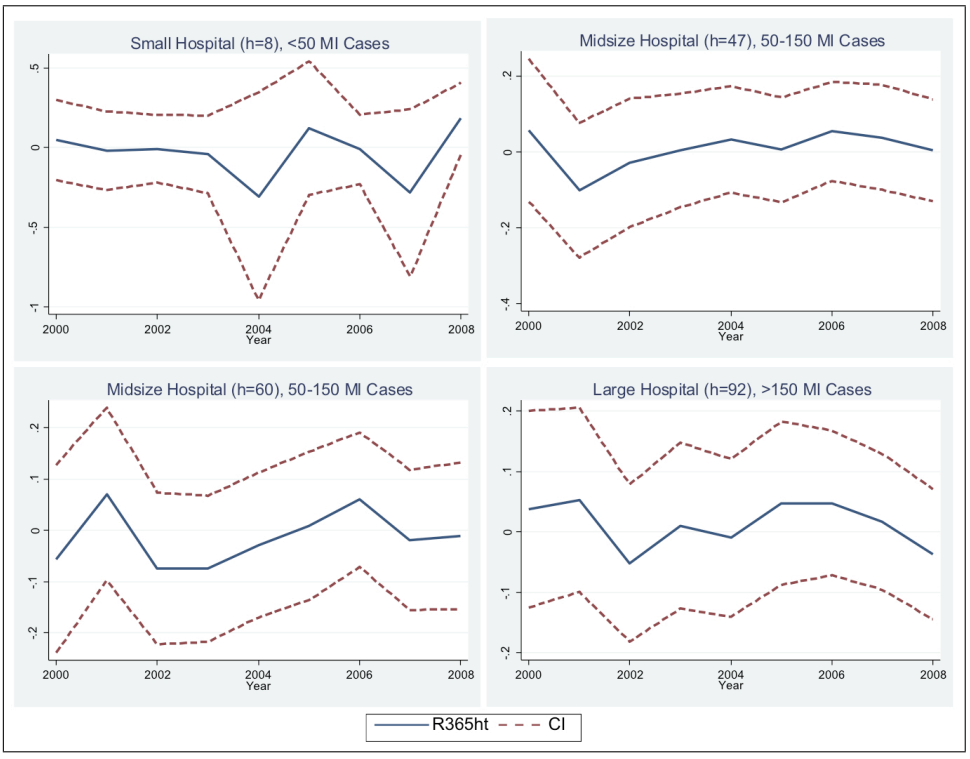


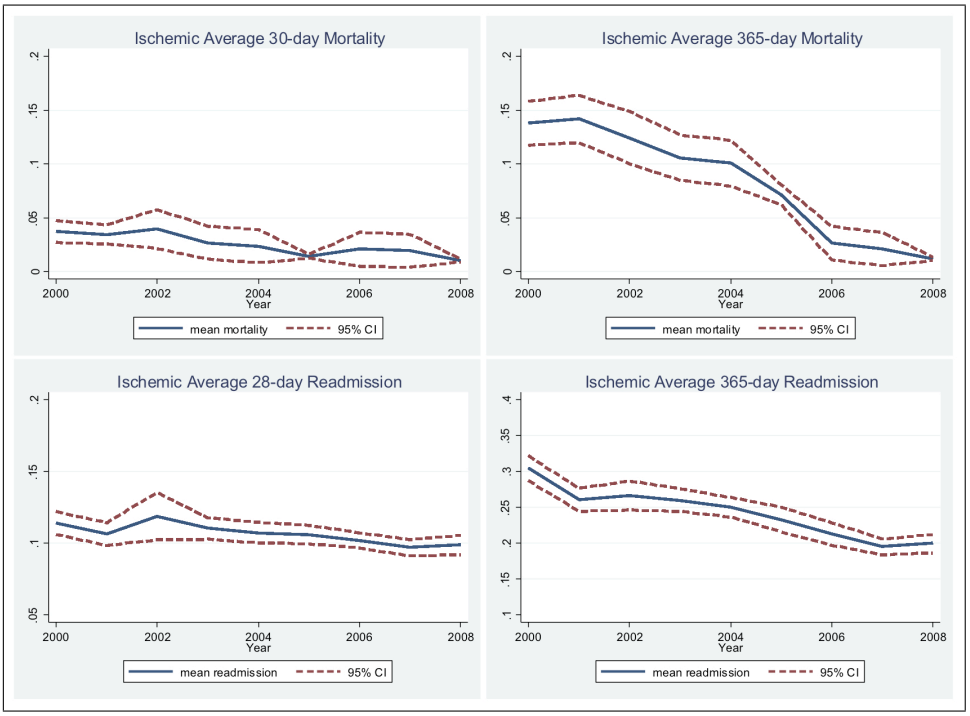
Figure A.6: Trends across years in Latent MI 365-day readmissions for selected hospitals.



A.2 IHD

Figure A.7 indicates a declining in average short and long term mortality for IHD. The downwards trend is barely discernible for 30-day mortality, but clearly pronounced for 365-day mortality. While 365-day mortality is declining for most of the years in the sample, the decline is particularly pronounced in the year 2005-2006. The trends in average readmissions indicate constant average 28-day readmissions over the period studied, and steady declines in the 365-day readmissions. The confidence intervals for all four figures indicate little variation among the hospitals in the sample.

Figure A.7: Trends across years in average IHD outcome measures across hospitals.



The results from the regressions of the first model of the analysis, where the four outcome measures are used as dependent variables, are presented in Table A.2. Age, co-morbidities and admission type are significant predictors of all four outcomes. In all cases higher age and co-morbidity is associated with worse outcomes, and elective admissions were significantly associated with better outcomes as compared to emergency admissions. Gender and deprivation are sometimes significant for mortality and readmissions, where women have significantly higher mortality and readmission, and patients with higher deprivation also have worse outcomes. Patients admitted for elective procedures are significantly linked to better outcomes than those admitted for non-elective procedures. The trust dummies included for each hospital are highly significant for all four outcome measures.

Table A.2: Regression results for IHD Model 1.

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
30-Day Mortality							
2000	118219	0.001***	-6.56 e-04	-1.15e-04	0.017***	0.014***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	
2001	156680	0.001***	-0.002***	-2.17e-04*	0.017***	0.014***	yes
		(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	
2002	158862	0.001***	-2.04e-04	-2.20e-04**	0.015***	0.014***	yes
		(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	
2003	164946	0.001***	-7.49e-04	-1.81e-04*	0.016***	0.015***	yes
		(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	
2004	170556	0.001***	-6.79e-04	-1.09e-04	0.012***	0.014***	yes
		(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	
2005	169619	0.001***	-6.79e-04	-9.55e-05	0.011***	0.015***	yes
		(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	
2006	168015	0.001***	4.65e-04	-1.91e-04**	0.010***	0.014***	yes
		(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	
2007	169918	0.001***	-0.001**	-8.38e-05	0.0093***	0.015***	yes
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	
2008	166115	0.001***	-4.22 e-04	-2.35e-04***	0.0083***	0.012***	yes
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	
365-Day Mortality							
2000	118219	0.005***	-0.012***	0.001***	0.056***	0.041***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	
2001	156680	0.005***	-0.015***	0.001***	0.056***	0.046***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	
2002	158862	0.005***	-0.011***	0.002***	0.053***	0.048***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	
2003	164946	0.005***	-0.011***	0.001***	0.052***	0.052***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	
2004	170556	0.004***	-0.0096***	0.001***	0.047***	0.050***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.00142)	

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
2005	169619	0.004***	-0.011***	0.001***	0.040***	0.046***	yes
		(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	
2006	168015	0.001***	-0.001	-0.000	0.014***	0.019***	yes
		(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	
2007	169918	0.001***	-0.001**	-9.25e-05	0.010***	0.016***	yes
		(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	
2008	166115	0.001***	-4.58e-04	-2.26e-04**	0.010***	0.013***	yes
		(0.000)	(0.000)	(0.00)	(0.000)	(0.001)	
28-Day Readmission							
2000	118219	0.001***	-0.001***	0.003***	0.014***	0.088***	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.000)	
2001	156680	0.001***	-0.007***	0.003***	0.019***	0.093***	yes
		(0.000)	(0.002)	(0.000)	(0.000)	(0.002)	
2002	158862	0.001***	-0.006***	0.003***	0.017***	0.090***	yes
		(0.000)	(0.002)	(0.000)	(0.000)	(0.002)	
2003	164946	0.001***	-0.005***	0.003***	0.016***	0.091***	yes
		(0.000)	(0.002)	(0.000)	(0.000)	(0.002)	
2004	170556	0.001***	-0.006***	0.003***	0.018***	0.090***	yes
		(0.000)	(0.002)	(0.000)	(0.000)	(0.002)	
2005	169619	0.001***	-0.004***	0.002***	0.017***	0.093***	yes
		(0.000)	(0.001)	(0.000)	(0.000)	(0.002)	
2006	168015	0.001***	-0.001	0.003***	0.016***	0.090***	yes
		(0.000)	(0.001)	(0.000)	(0.000)	(0.002)	
2007	169918	0.001***	-0.001	0.002***	0.0016***	0.088***	yes
		(0.000)	(0.001)	(0.000)	(0.000)	(0.002)	
2008	166115	0.001***	-0.002	0.003***	0.016***	0.082***	yes
		(0.000)	(0.001)	(0.000)	(0.000)	(0.002)	
365-Day Readmission							
2000	118219	0.003***	-0.011***	0.010***	0.041***	0.190***	yes
		(0.000)	(0.002)	(0.000)	(0.002)	(0.003)	
2001	156680	0.002***	-0.002	0.009***	0.036***	0.182***	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.003)	
2002	158862	0.002***	0.004**	0.009***	0.037***	0.180***	yes

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
		(0.000)	(0.002)	(0.000)	(0.001)	(0.003)	
2003	164946	0.003***	0.001	0.009***	0.034***	0.181***	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.003)	
2004	170556	0.002***	-0.001	0.008***	0.037***	0.179***	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.002)	
2005	169619	0.002***	0.003	0.007***	0.034***	0.175***	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.002)	
2006	168015	0.002***	0.003	0.007***	0.031***	0.159***	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.002)	
2007	169918	0.002***	0.006***	0.007***	0.031***	0.157***	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.002)	
2008	166115	0.00245***	0.00334*	0.007***	0.033***	0.152***	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.002)	

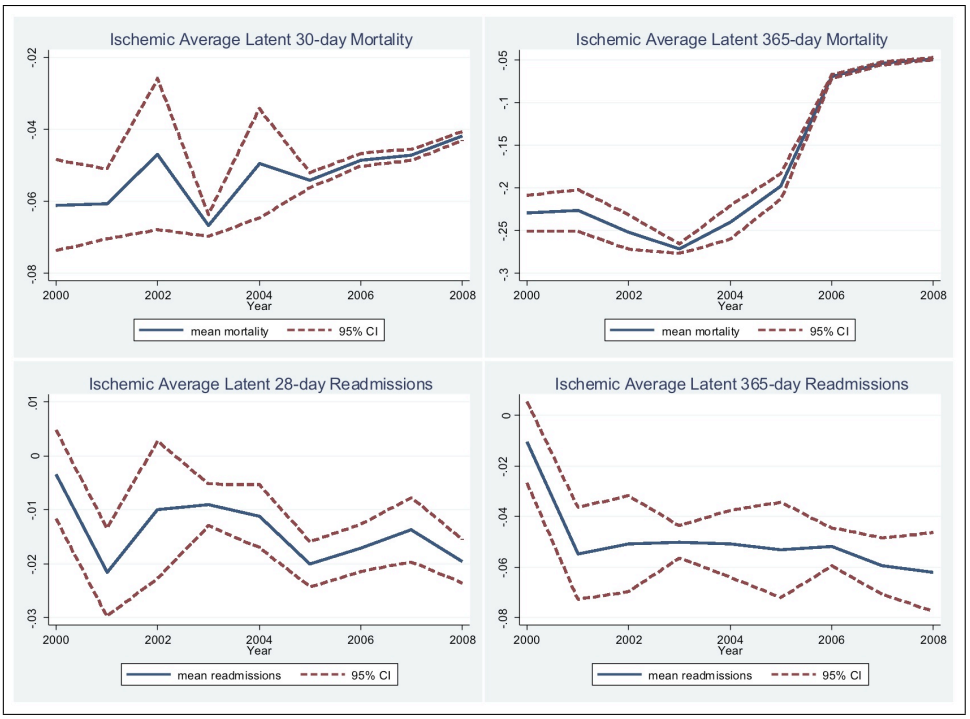
* Significant at $p \leq 0.1$

** Significant at $p \leq 0.05$

*** Significant at $p \leq 0.01$

Figure A.8 shows the trends in the averages of the four latent outcome measures estimated for IHD. Each point represents the average intercept value calculated for each year, which shows the average effect hospitals are having on each outcome controlling for patient characteristics. The average intercepts for both short and long term mortality are negative throughout the time period being investigated. This suggests that during this period mortality is decreasing, the values are becoming less negative over time, suggesting that they are decreasing at and increasing rate. Moreover, in both panels the 95% confidence intervals for the estimates become much smaller towards the end of the sample, suggesting much smaller variation between hospitals in their intercepts. The bottom two panels present the average hospital intercepts as estimated for short and long term readmissions. For both, the average values for each year are less then zero suggesting a decreasing rate of readmissions over time. For many years the averages stay constant, indicating a constant decline in the readmission rate over those periods. Unlike the mortality estimates, there does not appear to be a narrowing of the confidence interval, indeed for year-long readmissions the 95% interval is widening. This indicates an increase in the variation of hospitals from the mean values in the later years of the sample

Figure A.8: Trends across years in average latent IHD outcome measures across hospitals.



Figures A.9–A.12 show the latent mortality and readmission estimates for four selected hospitals treating patients with IHD. For all four hospitals and all four outcomes the wide confidence intervals suggest large within hospital variations in mortality. The 30-day mortality estimates range from nearly 20% below average to almost 20% above average, and the 365-day estimates range from about 15% below average to nearly 15% above average. The variation in both readmission measures, ranges around 20% below and above average. There is also year-to-year variation in all four figures, ranging around $5 - 10\%$ in either direction

Figure A.9: Trends across years in latent IHD 30-day mortality for selected hospitals.

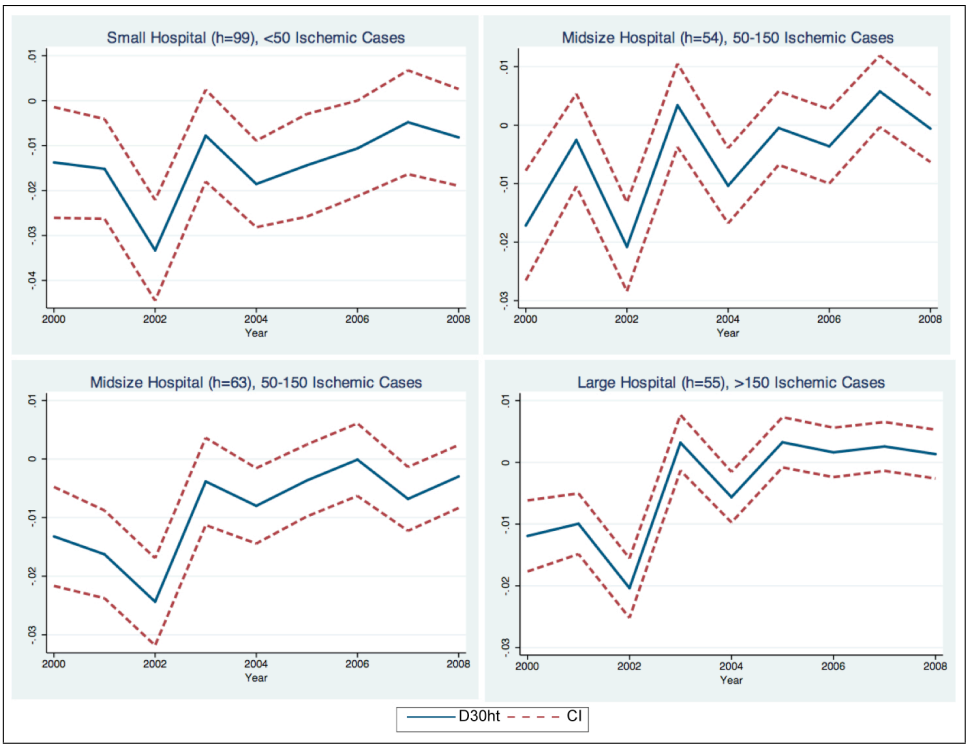


Figure A.10: Trends across years in Latent IHD 365-day mortality for selected hospitals.

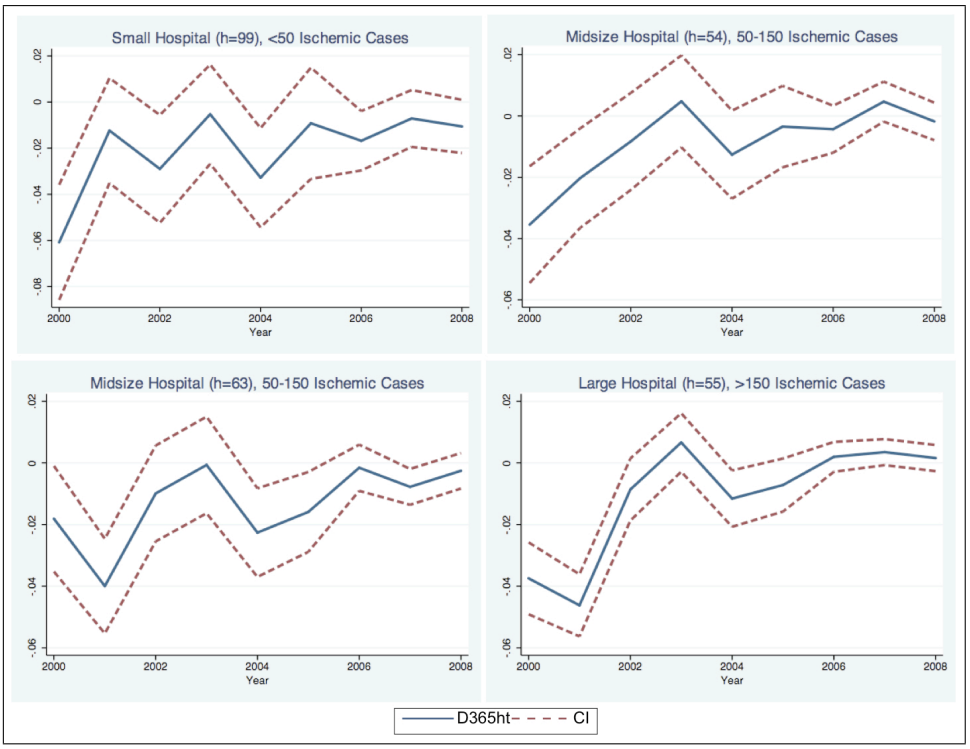


Figure A.11: Trends across years in Latent IHD 28-day readmissions for selected hospitals.

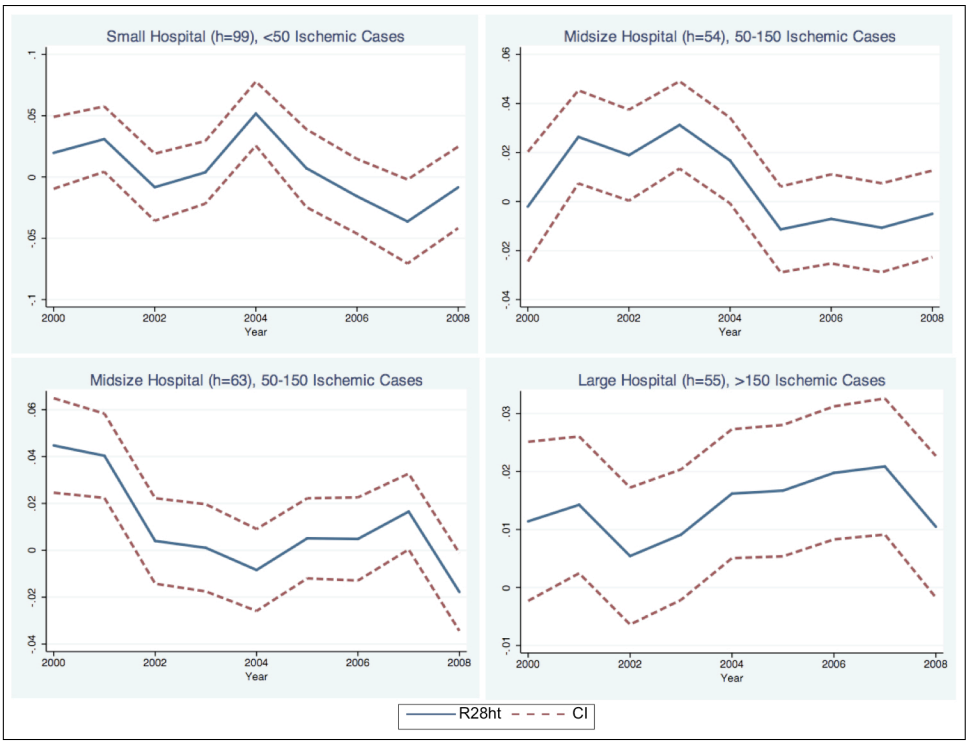
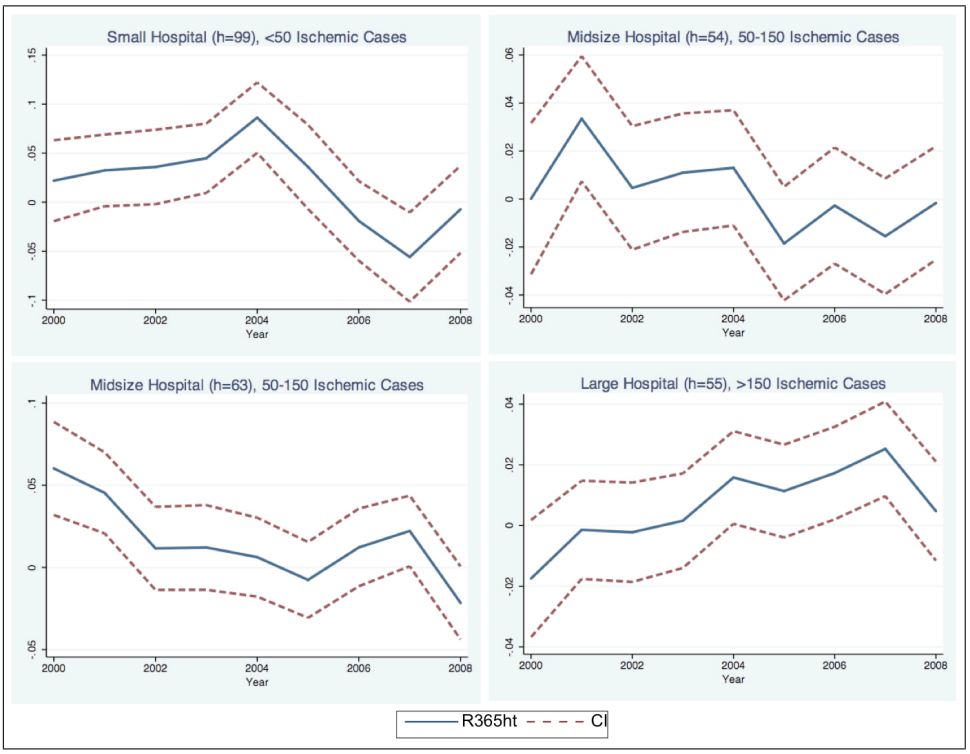


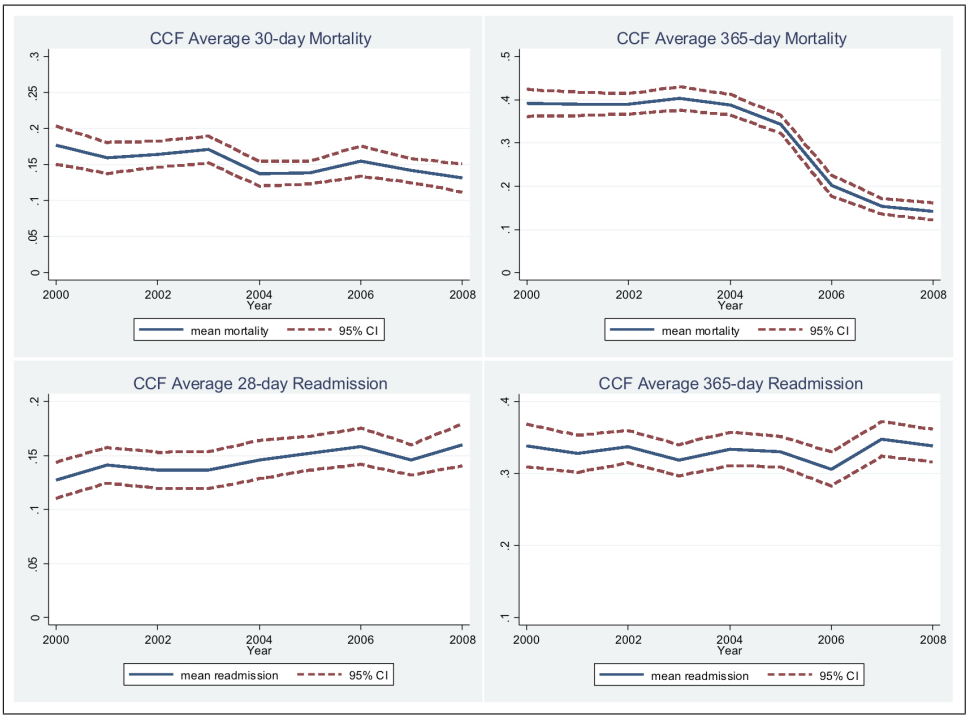
Figure A.12: Trends across years in Latent IHD 365-day readmissions for selected hospitals.



A.3 CCF

The descriptive statistics displayed in Figure A.13 show the pattern of average mortality in the period 2000-2008. As illustrated, 30-day mortality has declined over this time period. However, this mortality decline is very small, and has not been consistent throughout the years. Figure A.13, shows a much larger decline of 365-day mortality rates for CCF over the same time period. Similar to the trend in average long term mortality in the other conditions studied, this decline is particularly sharp between 2005-2006. The figure also indicates an increase in short term readmissions, which is mostly consistent aside from two small declines in 2002-2002 and 2006-2007. Average long term readmissions exhibit a relatively stable trend over the period studied. The confidence intervals for all four outcomes are wider than for the other conditions, suggesting more variation in CCF outcomes than in the other conditions.

Figure A.13: Trends across years in average CCF outcome measures across hospitals.



The results of the regressions run for the first model, presented in Table A.3, indicate what patient factors are significant predictors of mortality and readmissions. Age and co-morbidities are always highly significant predictors of mortality, and deprivation is sometimes significant. Where, older age, higher deprivation and more co-morbidities are associated with higher short and long term mortality. Age is not a significant determinant of both long and short term readmissions for most of the years studied, but co-morbidities remain highly significant in the same direction. Deprivation is significant in influencing

long-term readmissions, but not short-term readmissions, such that more deprived patients have higher rates of long-term readmission. While type of admission is not significant for the other outcomes, elective admissions were significantly associated with lower long term readmissions as compared to non-elective admissions. The trust dummies included for each hospital are highly significant for all four outcome measures.

Table A.3: Regression results for CCF Model 1.

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
30-Day Mortality							
2000	2475	0.004***	0.005	-0.003	0.033***	-0.038	yes
		(0.015)	(0.001)	(0.003)	(0.006)	(0.042)	
2001	3436	0.004***	0.009	-0.003	0.027***	0.0089	yes
		(0.012)	(0.001)	(0.003)	(0.005)	(0.041)	
2002	3541	0.004***	0.003	0.002	0.029***	-0.001	yes
		(0.012)	(0.001)	(0.002)	(0.005)	(0.037)	
2003	3590	0.004***	2.10e-04	-0.003	0.034***	-0.035	yes
		(0.013)	(0.001)	(0.003)	(0.005)	(0.034)	
2004	3694	0.004***	0.017	6.51e-04	0.022***	0.008	yes
		(0.012)	(0.001)	(0.001)	(0.004)	(0.035)	
2005	4102	0.004***	0.021*	-0.001	0.026***	0.039	yes
		(0.011)	(0.001)	(0.001)	(0.004)	(0.029)	
2006	4417	0.004***	-0.004	-0.004**	0.026***	0.032	yes
		(0.011)	(0.001)	(0.004)	(0.004)	(0.026)	
2007	4165	0.004***	0.008	-0.004*	0.031***	0.085***	yes
		(0.011)	(0.001)	(0.004)	(0.004)	(0.024)	
2008	3817	0.003***	-4.75e-04	-0.005***	0.028***	0.071***	yes
		(0.010)	(0.001)	(0.005)	(0.004)	(0.020)	
365-Day Mortality							
2000	2475	0.009***	0.037*	-0.003	0.066***	-0.102*	yes
		(0.001)	(0.020)	(0.004)	(0.008)	(0.055)	
2001	3436	0.008***	0.040**	-0.002	0.076***	0.027	yes
		(0.001)	(0.020)	(0.003)	(0.006)	(0.055)	
2002	3541	0.008***	0.011	0.004	0.062***	0.049	yes

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
		(0.001)	(0.020)	(0.003)	(0.007)	(0.050)	
2003	3590	0.009***	0.024	-0.005	0.073***	0.047	yes
		(0.000653)	(0.016)	(0.003)	(0.006)	(0.044)	
2004	3694	0.009***	0.024	0.003	0.057***	0.021	yes
		(0.001)	(0.016)	(0.003)	(0.006)	(0.047)	
2005	4102	0.008***	0.068***	-0.007***	0.048***	0.063	yes
		(0.001)	(0.015)	(0.003)	(0.005)	(0.040)	
2006	4417	0.005***	0.004	-0.004*	0.029***	0.058**	yes
		(0.000)	(0.012)	(0.002)	(0.004)	(0.029)	
2007	4165	0.004***	0.007	-0.003	0.030***	0.099***	yes
		(0.000)	(0.012)	(0.002)	(0.004)	(0.025)	
2008	3817	0.004***	0.002	-0.006***	0.031***	0.079***	yes
		(0.000)	(0.011)	(0.002)	(0.004)	(0.021)	
28-Day Readmission							
2000	2475	-0.002***	-0.005	1.69e-04	0.017***	0.098**	yes
		(0.001)	(0.015)	(0.003)	(0.006)	(0.042)	
2001	3436	-6.04e-04	0.021*	0.001	0.017***	0.075*	yes
		(0.001)	(0.013)	(0.003)	(0.005)	(0.043)	
2002	3541	-8.96e-04*	0.021*	0.005**	0.007	0.062*	yes
		(0.000)	(0.012)	(0.002)	(0.005)	(0.037)	
2003	3590	2.61e-04	0.003	2.64e-05	0.018***	0.064*	yes
		(0.001)	(0.012)	(0.003)	(0.005)	(0.034)	
2004	3694	-6.91e-04	-0.009	0.002	0.010**	0.045	yes
		(0.000)	(0.012)	(0.001)	(0.005)	(0.036)	
2005	4102	-5.07e-04	0.005	0.002	0.015***	0.066*	yes
		(0.000)	(0.012)	(0.001)	(0.004)	(0.032)	
2006	4417	7.95e-05	0.008	0.004**	0.001	0.046	yes
		(0.000)	(0.012)	(0.004)	(0.004)	(0.029)	
2007	4165	-8.60e-05	-0.005	0.002	0.011***	0.049*	yes
		(0.000)	(0.012)	(0.004)	(0.004)	(0.026)	
2008	3817	-1.53e-04	0.005	0.002	0.011**	0.039*	yes
		(0.000)	(0.012)	(0.005)	(0.004)	(0.023)	
365-Day Readmission							

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
2000	2475	-0.001	0.012	0.003	0.033***	0.143**	yes
		(0.001)	(0.020)	(0.004)	(0.009)	(0.057)	
2001	3436	-3.61e-04	0.010	0.010***	0.020***	0.117**	yes
		(0.001)	(0.017)	(0.003)	(0.007)	(0.056)	
2002	3541	2.88e-04	0.017	0.010***	0.015**	0.117**	yes
		(0.001)	(0.017)	(0.003)	(0.007)	(0.050)	
2003	3590	0.001	-0.009	0.007**	0.013**	0.162***	yes
		(0.001)	(0.016)	(0.003)	(0.006)	(0.044)	
2004	3694	1.55e-04	-0.013	0.007**	0.007	0.038	yes
		(0.001)	(0.016)	(0.003)	(0.006)	(0.047)	
2005	4102	-2.05e-04	-0.011	0.007**	0.017***	0.112***	yes
		(0.001)	(0.016)	(0.003)	(0.006)	(0.041)	
2006	4417	-1.15e-04	0.009	0.005**	0.003	0.113***	yes
		(0.001)	(0.015)	(0.003)	(0.005)	(0.036)	
2007	4165	1.29e-04**	-0.002	0.007***	0.003	0.002	yes
		(0.001)	(0.016)	(0.003)	(0.005)	(0.034)	
2008	3817	4.73e-04	-0.006	0.008***	0.017***	0.079***	yes
		(0.001)	(0.016)	(0.003)	(0.006)	(0.030)	

* Significant at $p \leq 0.1$

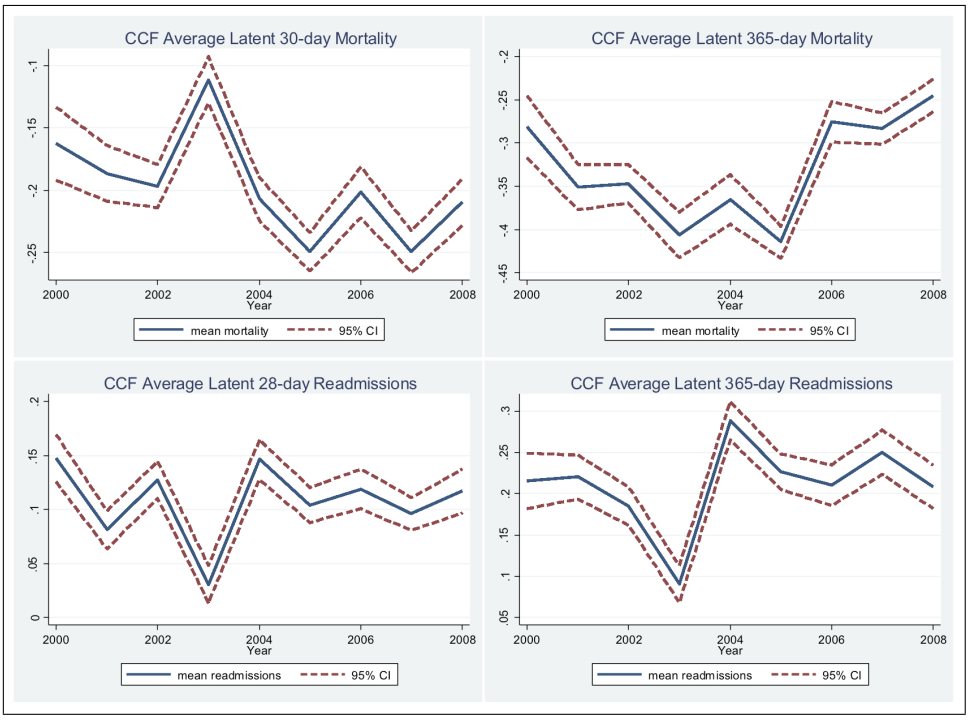
** Significant at $p \leq 0.05$

*** Significant at $p \leq 0.01$

The average hospital intercept for CCF is graphed for each year studied in Figures A.14. This indicates the effect hospitals have had on mortality and readmissions in each year, and how the rate of change of these outcomes over the time period studied. Both short and long term mortality latent averages are below zero. This indicates that the average effect hospitals have on mortality is decreasing over time. While the intercepts become more negative and less negative in different years throughout the sample, short term mortality seems to be decreasing at a decreasing rate for most years, while year-long mortality is decreasing at an increasing rate from 2005 onwards. The average hospital intercepts as calculated for short and long term readmissions are both positive over the entire period investigated. This suggests that the readmissions attributable to hospital quality are increasing over time. Again there is some variation in the magnitude of the average intercept over time, with what appears to be a dip in 2003 in both short and

long term readmissions. This suggests that readmissions were increasing at a decreasing rate over that period. For the short and long term mortality intercepts, and the long term readmission intercepts the confidence intervals become more narrow at the end of the period. This suggests that there is less variation amongst these outcomes in the later years of the sample. However, the opposite can be said about short term readmissions, where the confidence intervals seems slightly wider in the last years as compared to the first.

Figure A.14: Trends across years in average latent CCF outcome measures across hospitals.



Figures A.15 – A.18 show the latent mortality and readmission rates for four selected hospitals treating patients with CCF. For all four outcomes there is large within hospital and year-to-year variation. The confidence intervals suggest within hospital variations of over 20% above or below average for both short term and long term latent mortality estimates, and variations of about $10 - 15\%$ above or below average for both short term and long term latent readmission estimates. Long and short term mortality can vary substantially from year-to-year, in the case of the small hospital increasing by as much as 40%. However this fluctuation is most probably a result of the small number of patients treated for CCF annually. There is smaller year-to-year variation amongst the readmission measures, although this is commonly around 10%.

Figure A.15: Trends across years in latent CCF 30-day mortality for selected hospitals.

Figure A.16: Trends across years in latent CCF 365-day mortality for selected hospitals.

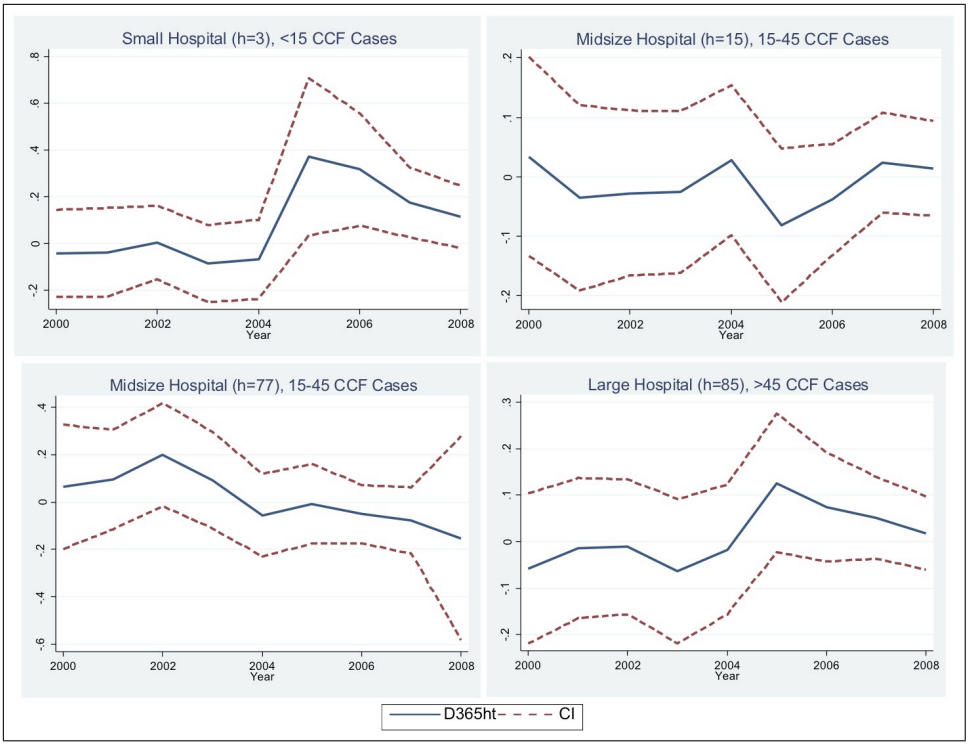


Figure A.17: Trends across years in Latent CCF 28-day readmissions for selected hospitals.

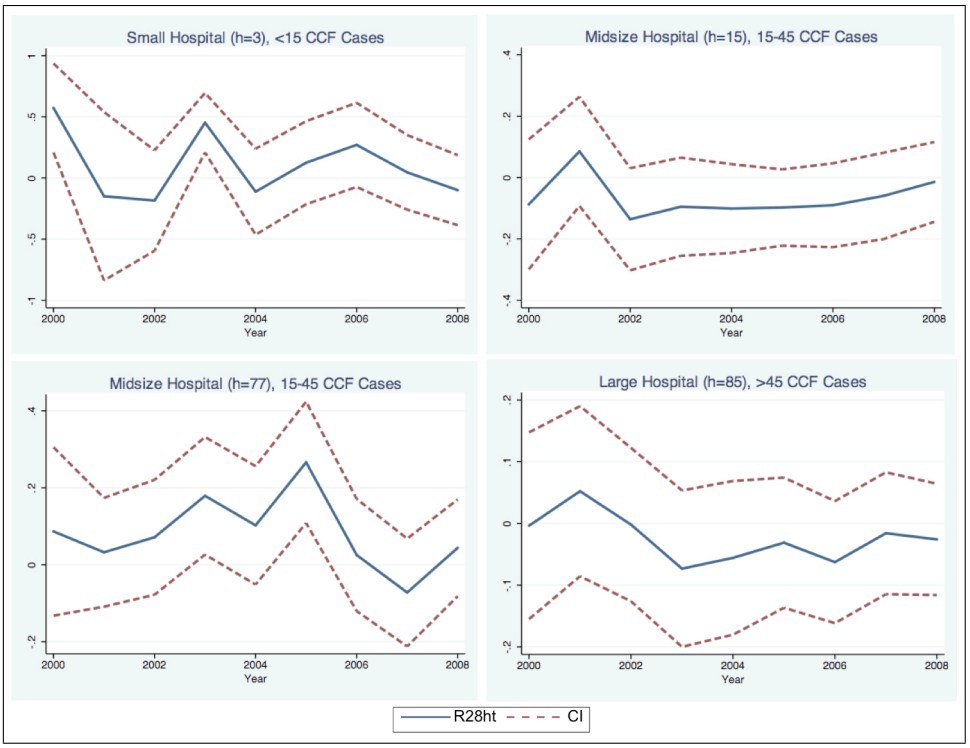
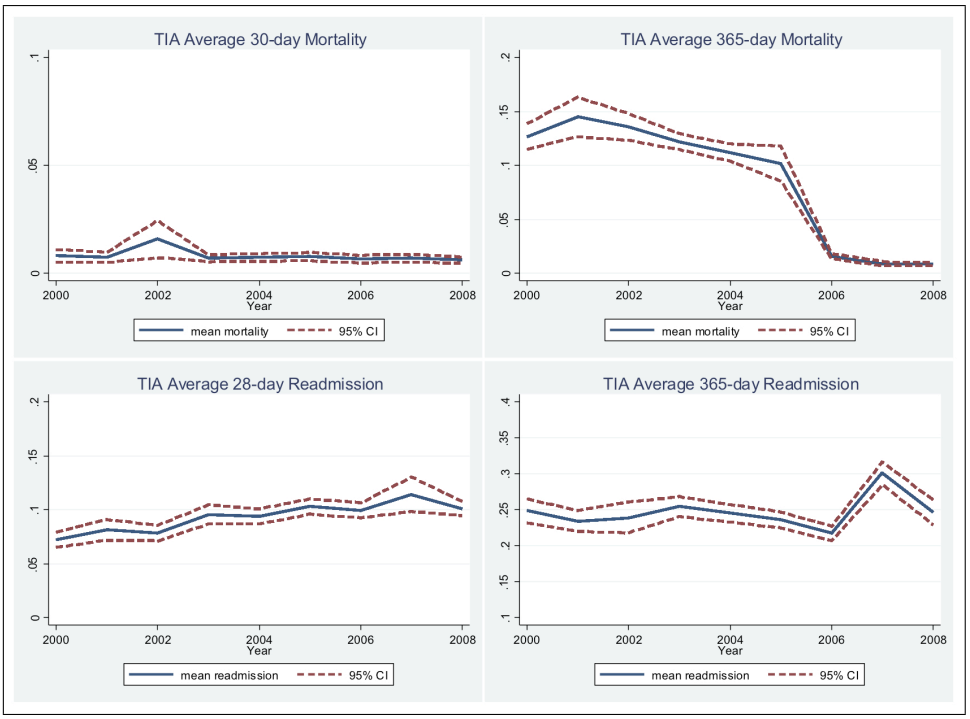


Figure A.18: Trends across years in Latent CCF 365-day readmissions for selected hospitals.

A.4 TIA

The average 30-day mortality for TIA amongst all hospitals in the sample is shown in Figure A.19. In the time period under investigation, the short-term mortality rates are relatively constant, apart from a slight jump in mortality in 2002. The average 365-day mortality rates for TIA, indicates a downwards trend in long term mortality. This trend is particularly pronounced the in year 2005-2006. Average TIA 28-day readmission exhibit a clear upwards trend over the 2000-2008 period is observed. The rate of increase is faster for some years than others, but for the most part is consistently rising throughout the time period. Finally the average 365-day readmission rates for TIA amongst all hospitals in the sample are relatively constant except for a jump in mortality rates in 2007.

Figure A.19: Trends across years in average TIA outcome measures across hospitals.



The results of the mortality regressions for Model 1, shown in Table A.4, indicate that age and co-morbidities are the only significant variables apart from the trust dummies. For both mortality regressions higher age and higher co-morbidity is associated with higher mortality. Table A.4 presents the results for the readmission regressions. Short and long term readmissions are influenced by age, deprivation, co-morbidities and type of admission, and long term readmissions are also influenced by gender. Older age, increased deprivation and the presence of co-morbidities are all associated with higher readmissions, while elective admissions are significantly associated with lower mortality as compared to non-elective admissions. Where significant gender indicates that men have slightly

higher 365-day readmissions then women. The trust dummies are also significant for both readmission regressions.

Table A.4: Regression results for TIA Model 1.

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
30-Day Mortality							
2000	9751	3.08e-04***	4.97e-04	-2.01e-05	0.002*	0.002	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.005)	
2001	12853	4.11e-04***	-0.001	1.26e-04	0.005***	0.005	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.004)	
2002	12375	4.90e-04***	-0.001	-1.60e-04	0.008***	0.007	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.005)	
2003	13618	3.81e-04***	-4.12e-04	2.24e-04	0.009***	0.001	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.005)	
2004	14263	4.18e-04***	3.81e-04	-5.71e-05	0.003***	0.002	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.004)	
2005	14857	3.22e-04***	-6.53e-04	-3.48e-04	0.005***	0.001	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.004)	
2006	15629	3.34e-04***	-0.001	9.85e-05	0.006***	0.003	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.004)	
2007	15620	2.98e-04***	-0.001	-6.60e-05	0.004***	0.001	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.004)	
2008	16577	2.82e-04***	-0.001	-2.13e-04	0.005***	0.004	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.004)	
365-Day Mortality							
2000	9751	0.005***	0.003	0.002	0.053***	-0.001	yes
		(0.000)	(0.007)	(0.001)	(0.004)	(0.017)	
2001	12853	0.006***	0.002	0.002*	0.054***	0.010	yes
		(0.000)	(0.006)	(0.001)	(0.004)	(0.016)	
2002	12375	0.006***	0.005	-2.55e-04	0.048***	0.028*	yes
		(0.000)	(0.006)	(0.001)	(0.004)	(0.017)	
2003	13618	0.00552***	-2.62e-04	0.002	0.055***	0.025	yes
		(0.000208)	(0.006)	(0.001)	(0.003)	(0.017)	

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
2004	14263	0.005***	0.012**	0.002*	0.052***	-0.005	yes
		(0.000)	(0.005)	(0.001)	(0.00318)	(0.016)	
2005	14857	0.004***	-0.007	-2.58e-04	0.043***	0.022	yes
		(0.000)	(0.005)	(0.001)	(0.003)	(0.015)	
2006	15629	0.001***	-0.003	1.40e-04	0.013***	0.0016	yes
		(0.000)	(0.002)	(0.000)	(0.001)	(0.006)	
2007	15620	4.67e-04***	-0.001	-8.17e-05	0.004***	0.003	yes
		(0.00)	(0.002)	(0.000)	(0.001)	(0.005)	
2008	16577	4.12e-04***	-0.001	-1.02e-04	0.007***	0.004	yes
		(0.00)	(0.001)	(0.000)	(0.001)	(0.004)	
28-Day Readmission							
2000	9751	6.54e-04***	0.008	0.003***	0.014***	0.026*	yes
		(0.000)	(0.006)	(0.002)	(0.000)	(0.005)	
2001	12853	3.35e-04*	-0.008*	0.001	0.021***	0.025*	yes
		(0.000)	(0.005)	(0.002)	(0.000)	(0.004)	
2002	12375	7.18e-04***	0.002	0.003***	0.021***	0.048***	yes
		(0.000)	(0.005)	(0.002)	(0.000)	(0.005)	
2003	13618	9.30e-04***	-0.006	0.004***	0.014***	0.046***	yes
		(0.000)	(0.005)	(0.002)	(0.000)	(0.005)	
2004	14263	4.47e-04***	-0.002	0.002**	0.012***	0.051***	yes
		(0.000)	(0.005)	(0.001)	(0.000)	(0.004)	
2005	14857	8.23e-04***	-0.001*	0.003***	0.009***	0.03***	yes
		(0.000)	(0.005)	(0.001)	(0.000)	(0.004)	
2006	15629	7.84e-04***	0.001	0.004***	0.014***	0.048***	yes
		(0.000)	(0.005)	(0.001)	(0.000)	(0.004)	
2007	15620	9.67e-04***	-0.002	0.002**	0.011***	0.041***	yes
		(0.000)	(0.005)	(0.001)	(0.000)	(0.004)	
2008	16577	1.03e-04***	-0.001	0.002**	0.016***	0.031**	yes
		(0.000)	(0.005)	(0.001)	(0.000)	(0.004)	
365-Day Readmission							
2000	9751	0.004***	0.006	0.008***	0.039***	0.067***	yes
		(0.000)	(0.009)	(0.002)	(0.006)	(0.023)	
2001	12853	0.003***	-0.026***	0.009***	0.035***	0.066***	yes

Year	N (total)	Age	Gender	Carstairs Score	Co-morbidity	Elective	Trust dummies
		(0.000)	(0.008)	(0.001)	(0.005)	(0.021)	
2002	12375	0.003***	-0.015*	0.008***	0.038***	0.100***	yes
		(0.000)	(0.007)	(0.001)	(0.005)	(0.023)	
2003	13618	0.004***	-0.012	0.010***	0.035***	0.106***	yes
		(0.000)	(0.008)	(0.001)	(0.005)	(0.023)	
2004	14263	0.003***	-0.008	0.007***	0.039***	0.085***	yes
		(0.000)	(0.007)	(0.001)	(0.004)	(0.023)	
2005	14857	0.003***	-0.024***	0.009***	0.037***	0.095***	yes
		(0.000)	(0.007)	(0.001)	(0.004)	(0.022)	
2006	15629	0.003***	-0.015**	0.009***	0.037***	0.078***	yes
		(0.000)	(0.007)	(0.001)	(0.00399)	(0.020)	
2007	15620	0.004***	-0.017**	0.009***	0.033***	0.088***	yes
		(0.000)	(0.007)	(0.001)	(0.00427)	(0.022)	
2008	16577	0.004***	-0.022***	0.007***	0.045***	0.059***	yes
		(0.000)	(0.007)	(0.001)	(0.00386)	(0.020)	

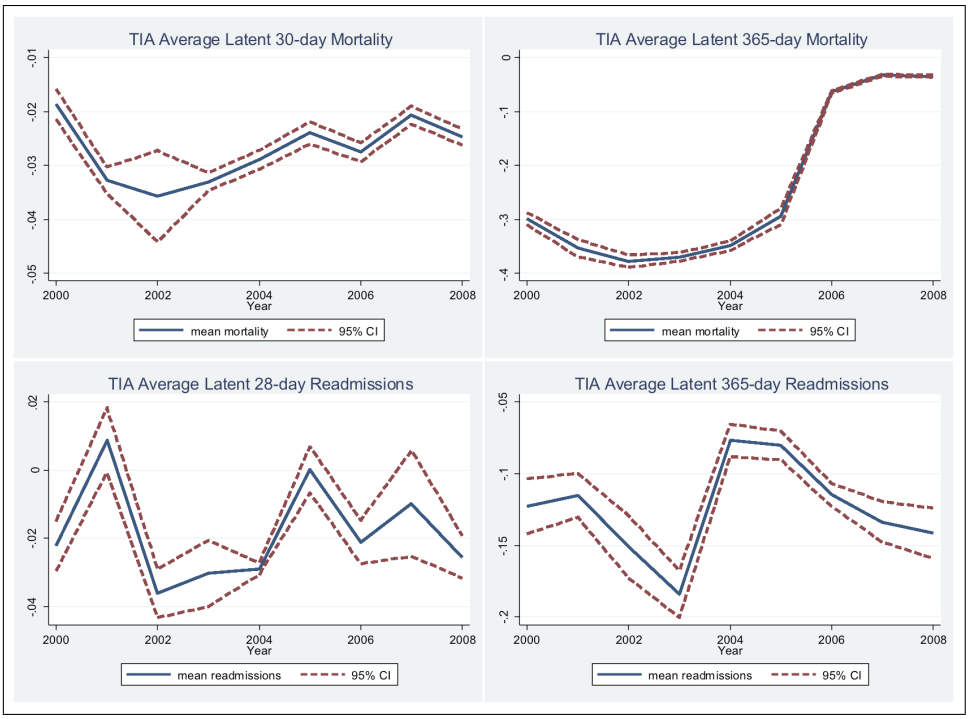
* Significant at $p \leq 0.1$

** Significant at $p \leq 0.05$

*** Significant at $p \leq 0.01$

Figure A.20 graphs the average 30-day latent TIA mortality for the 2000-2008 time period for mortality and readmissions. The short and long term latent mortality estimates are below 0 for the entire time period. This suggests that during these years mortality attributable to hospital quality is decreasing. For both short and long term mortality the change in the latent estimates over time suggests that initially, mortality is decreasing at a decreasing rate, but from 2004 onwards it begins to decrease at an increasing rate. The latent measures calculated for the readmission outcomes are also mostly negative. The latent estimates for short term readmissions are negative for all years apart from 2001, indicating decreasing TIA readmissions in most of the time period. The year-long intercepts are always negative, also indicating this decreasing trend, however the slope suggests that for some years both are decreasing at an increasing rate. Moreover the confidence intervals for the readmission estimates are becoming wider towards the end of the sample, indicating increasing variation amongst hospitals. For the mortality intercepts, a trend is only noticeable for year-long mortality where the confidence intervals become narrower.

Figure A.20: Trends across years in average latent TIA outcome measures across hospitals.



Figures A.21 – A.24 show the latent mortality and readmission estimates for the four selected hospitals treating TIA patients. The variation in latent mortality within hospitals smaller than all other conditions aside from Hip Replacement, it ranges between 1-3% below and above both 30-day and 365-day aggregate mortality measures. Similar to the other conditions there is wider variation for the small hospital. While there is year-to-year variation observed within hospitals this does not exceed 2% in either direction of the estimate. The variation within hospitals is greater for the hospital latent readmission estimates, where the confidence intervals range from 5-10% below and above the 28-day aggregate readmission estimates, and around 15% below and above the 365-day aggregate readmission estimates. In Figure A.21 the small hospital has the widest confidence intervals. Interestingly, the confidence intervals of all hospitals in Figure A.24 experience a very pronounced convergence towards the mean from 2005 onwards, indicating much less variation in outcomes in the later years of the sample. There is also year-to-year variation amongst readmission rates, but rarely does this exceed 5% for either short term or long term readmissions.

Figure A.21: Trends across years in latent TIA 30-day mortality for selected hospitals.

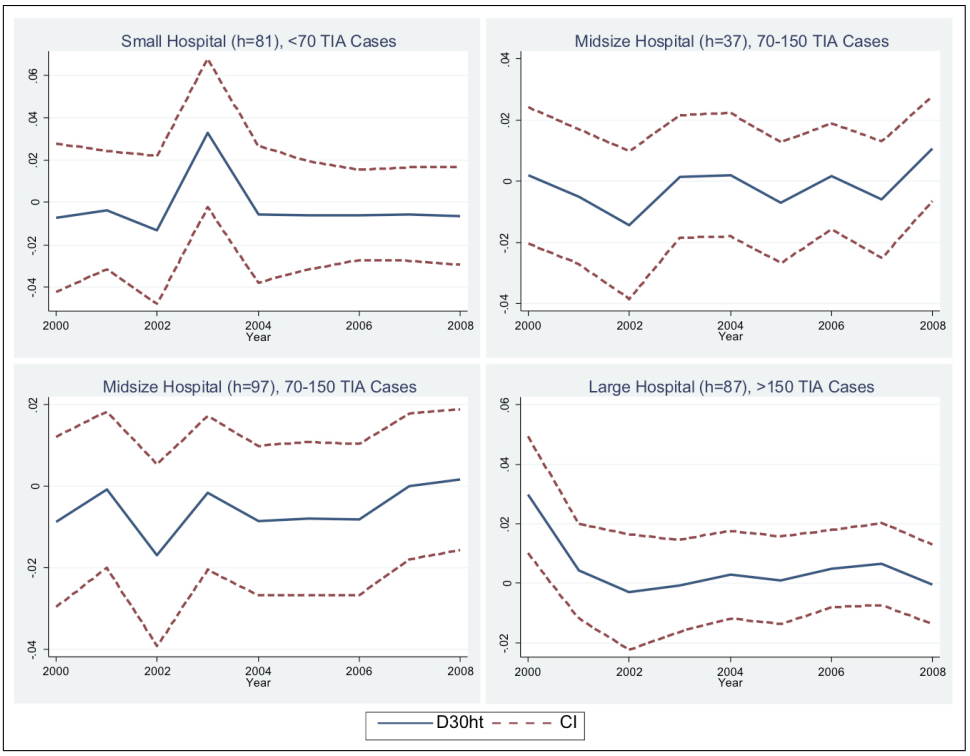


Figure A.22: Trends across years in latent TIA 365-day mortality for selected hospitals.

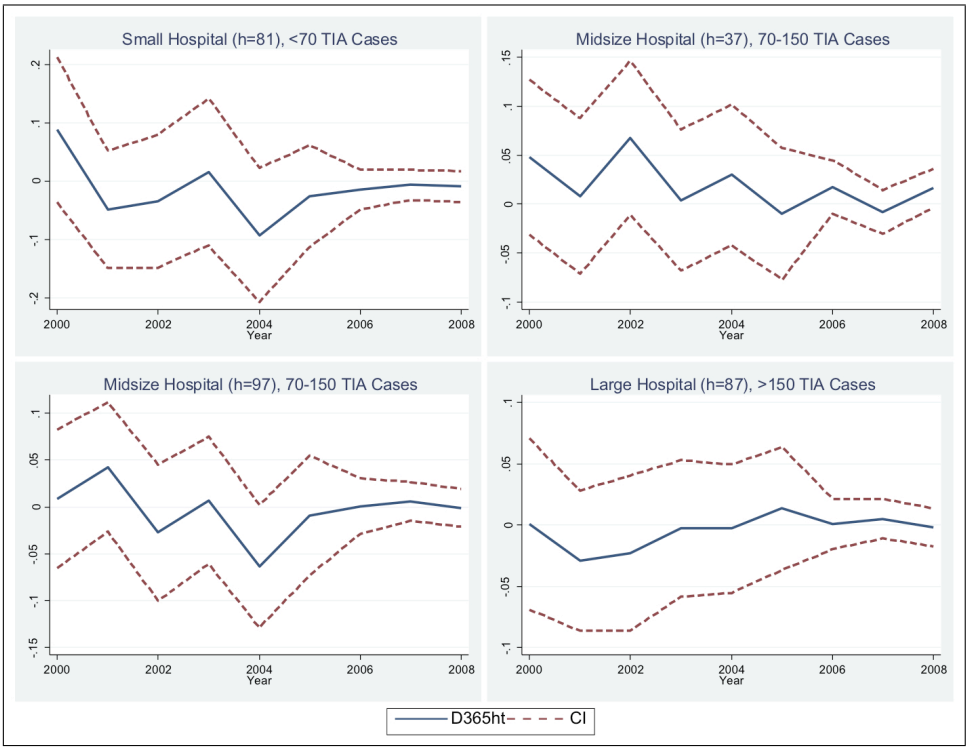


Figure A.23: Trends across years in Latent TIA 28-day readmissions for selected hospitals.

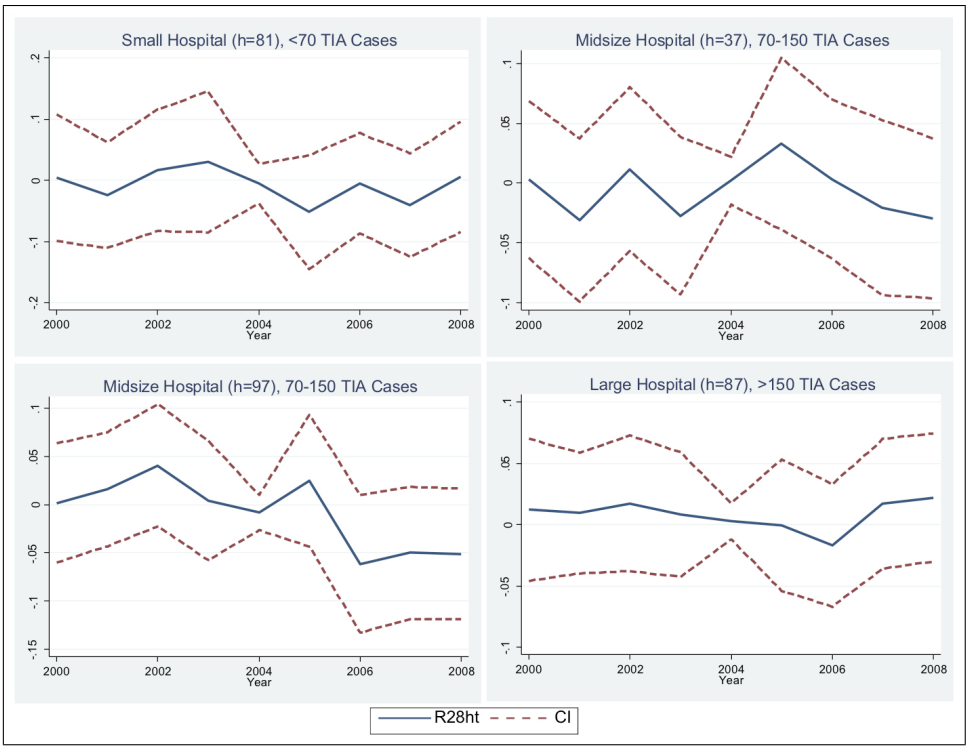
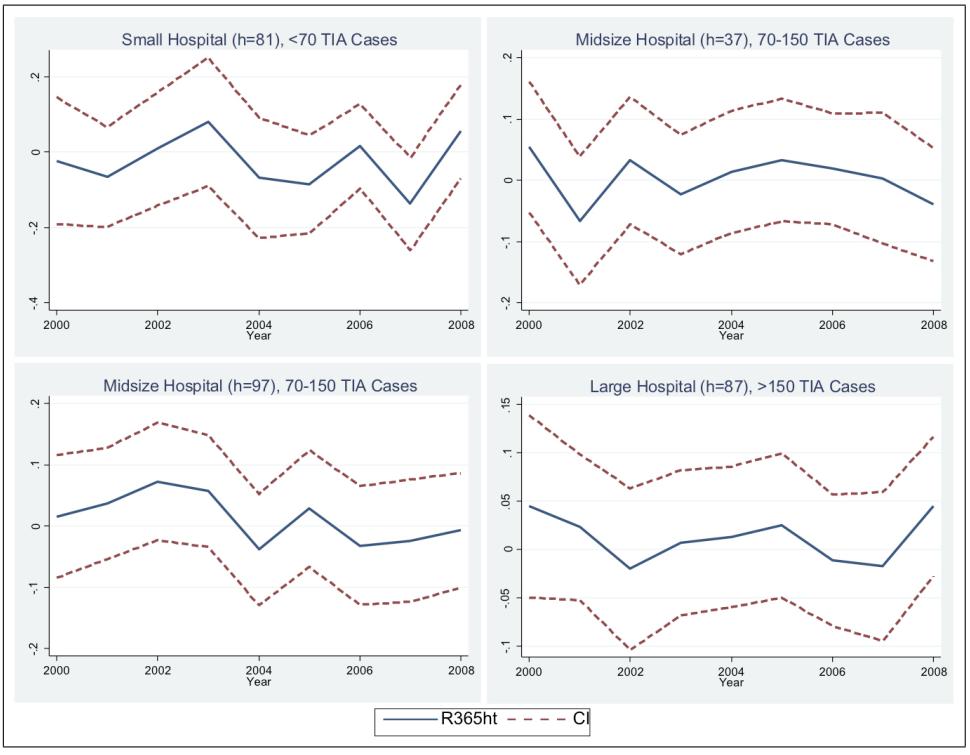


Figure A.24: Trends across years in Latent TIA 365-day readmissions for selected hospitals.



B | Results for Chapter 3

B.1 MI

Table B.1 shows the parameter estimates of the basic models run for patients classified as having MI. In these models, none of the quality indicators are very persistent. The lag coefficients of $D30_{ht}$ and $D365_{ht}$ range around $0.3 - 0.4$, and coefficients for $R28_{ht}$ and $R365_{ht}$ at around $0.04 - 0.06$. The sign on $R365_{ht}$ is always negative, and always positive for the other indicators. The variance of the indicators in the year 2000, indicates a standard deviation across hospitals of about 9% for $D30_{ht}$ and12% for $D365_{ht}$. Similarly, for readmissions the standard deviation across hospitals is about 7% for $R28_{ht}$ and 11% for $R365_{ht}$. The variance of the residuals indicate annual standard deviations in a similar range; corresponding to about 9%, 12.5%, 12.4% and8% for $D30_{ht}$, $D365_{ht}$, $R28_{ht}$ and $R365_{ht}$ respectively. The correlation coefficients on the variables in 2000, and the residuals, indicate a strong positive association between $D30_{ht}$ and $D365_{ht}$, as well as $R365_{ht}$ and $R28_{ht}$. All other associations are weak.

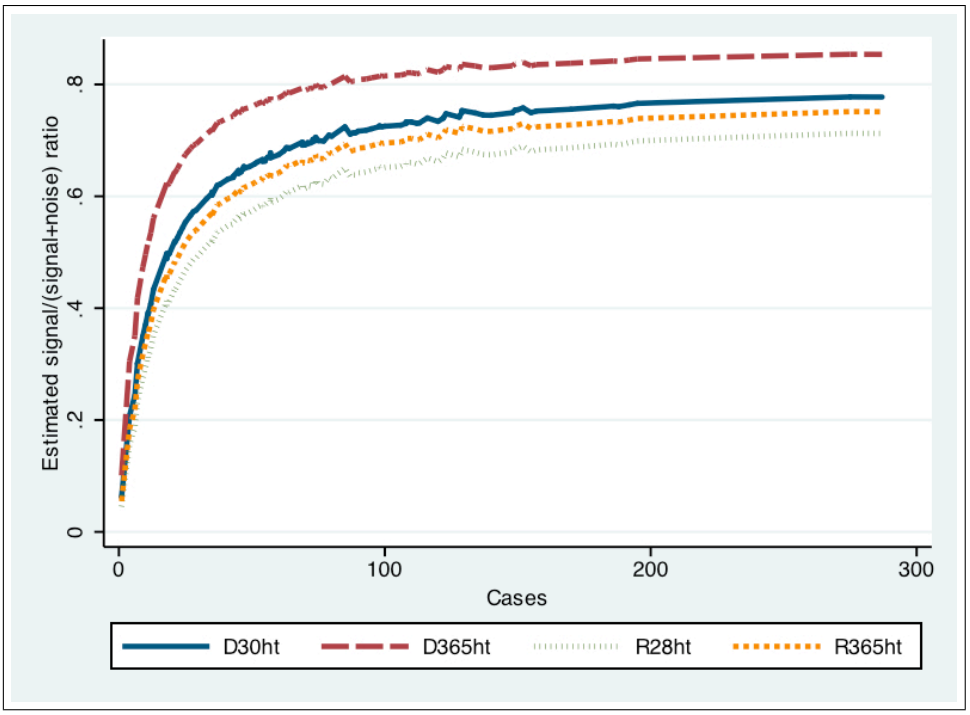
Table B.1: Estimates of MI multivariate VAR(1) parameters for hospital specific effects.

	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D30_{h(t-1)}$	0.255582	-0.068077	-0.380821	0.023963
	(0.03826)	(0.03517)	(0.06342)	(0.04909)
	[6.68086]	[-1.93562]	[-6.00438]	[0.48820]
$R28_{h(t-1)}$	-0.055568	0.025542	-0.150551	0.093971
	(0.04366)	(0.04014)	(0.07238)	(0.05602)
	[-1.27277]	[0.63636]	[-2.07997]	[1.67751]
$D365_{h(t-1)}$	0.050689	0.054887	0.832693	-0.102417
	(0.01978)	(0.01818)	(0.03279)	(0.02538)
	[2.56262]	[3.01825]	[25.3923]	[-4.03541]
$R365_{h(t-1)}$	-0.085410	0.022985	-0.363636	-0.061257
	(0.03453)	(0.03175)	(0.05725)	(0.04431)
	[-2.47338]	[0.72402]	[-6.35173]	[-1.38255]
Residuals				
S.D. dependent	0.094031	0.081331	0.197032	0.113909
Correlation of residuals ($D30_{ht}$)	1.000000	-0.141512	0.534637	-0.310228

	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
Correlation of residuals ($R28_{ht}$)	-0.141512	1.000000	0.155160	0.701684
Correlation of residuals ($D365_{ht}$)	0.534637	0.155160	1.000000	0.105374
Correlation of residuals ($R365ht$)	-0.310228	0.701684	0.105374	1.000000
Initial Conditions				
S.D. dependent in 2000	0.094472	0.073892	0.120822	0.113283
Correlation with $D30_{ht}$ in 2000	-	-0.0617	0.7028	-0.2464
Correlation with $R28_{ht}$ in 2000	-0.0617	-	0.1228	0.6479
Correlation with $D365_{ht}$ in 2000	0.7028	0.1228	-	0.1228
Correlation with $R365ht$ in 2000	-0.2464	0.6479	0.1228	-
Sample (adjusted): 2001 2008				
Included observations: 904 after adjustments				
Standard errors in ( ) & t-statistics in [ ]				

The signal to noise estimates for the four outcome measures of MI is plotted in Figure B.1 against the number of cases treated in each hospital. For this condition the signal to noise ratios are not as high as for AMI, Stroke and Hip Replacement. This suggests that the performance indicators for MI are noisier estimates of hospital performance. This may be the result of larger estimation error, combined with the smaller number of cases for this condition, resulting in a smaller sample from which to construct the estimates. While the signal to noise ratio for all measures improves as sample size increases, it still remains below the other conditions. Of the four measures both mortality estimates outperform the readmission estimates, and in both cases the year-long measure is better than its short term counterpart.

Figure B.1: Signal to noise ratio for the four MI outcome measures (year 2005).



Figures B.2– B.5 present the filtered MI measures, their 95% confidence intervals and the corresponding latent outcome measures derived in Appendix A.1 for selected hospitals. Similar observations can be made for the filtered measures were for AMI, Stroke and Hip Replacement above. The filtered measures are able to present much smoother estimates of performance over time, as compared to the latent measures which increase and decrease sharply from year to year. The confidence intervals for all conditions, including MI are nearly twice as large as they were for the latent measures, although this is most likely a result of the small sample of hospitals available in the English data versus the US data used by McClellan and Staiger (1999). Again in the small hospital, where the fewer cases result in much more erratic jumps in the latent measure, the smoothed indicator is much easier to use in order to make comparisons of performance over time.

Figure B.2: Filtered and latent estimates for MI $D30_{ht}$ for selected hospitals.

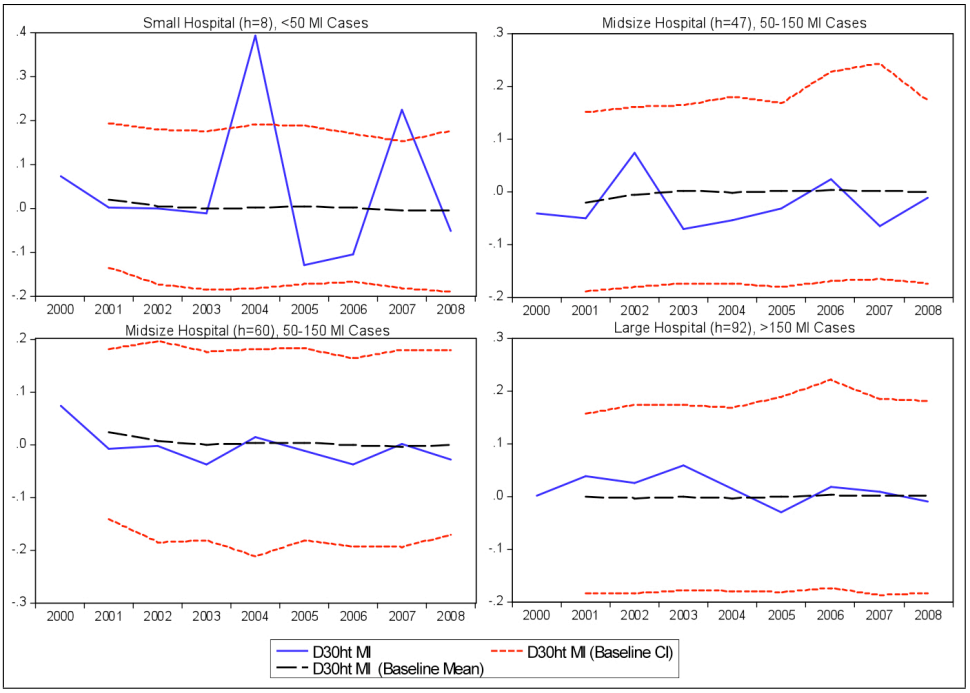


Figure B.3: Filtered and latent estimates for MI $D365_{ht}$ for selected hospitals.

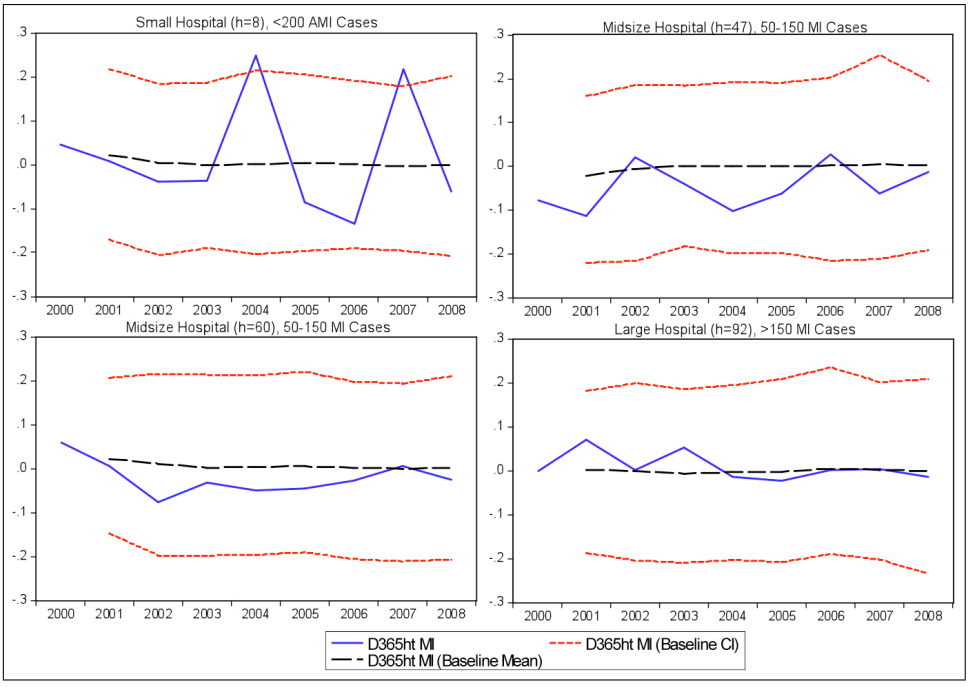


Figure B.4: Filtered and latent estimates for MI $R28_{ht}$ for selected hospitals.

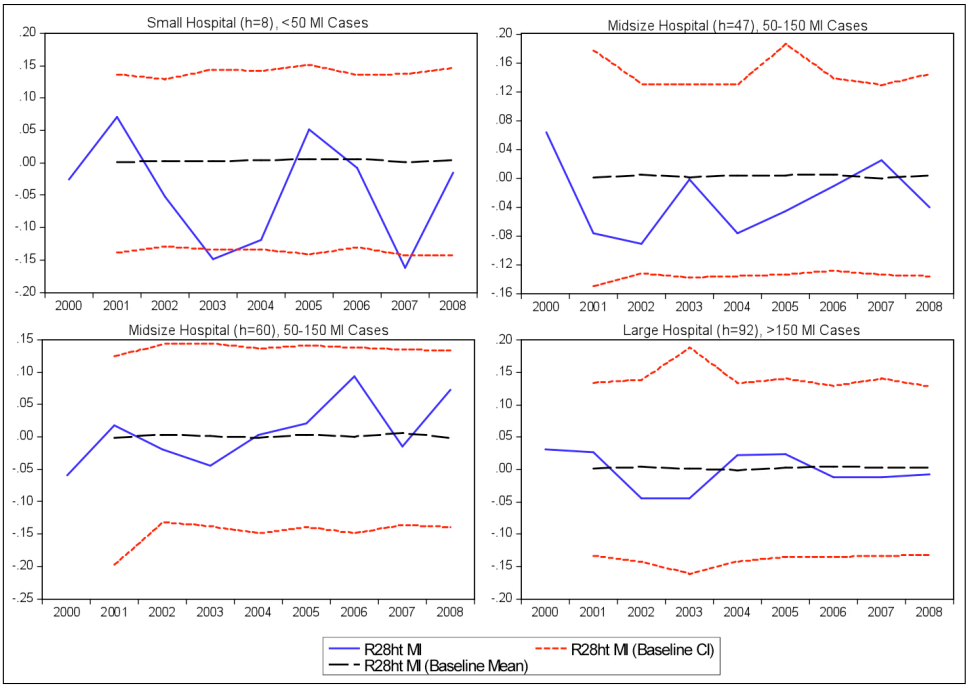


Figure B.5: Filtered and latent estimates for MI $R365_{ht}$ for selected hospitals.

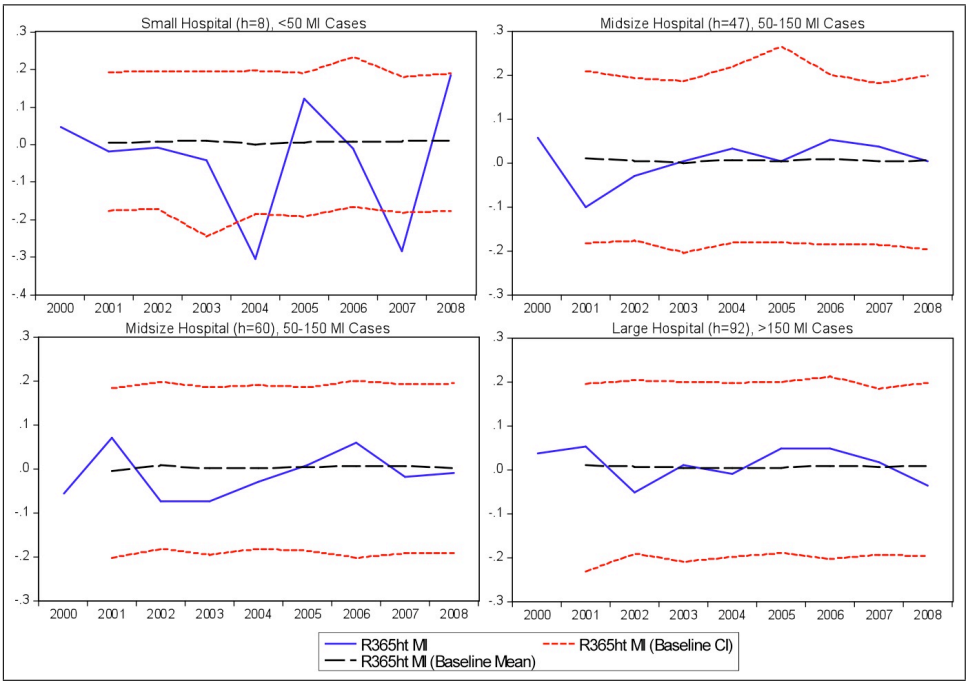


Table B.2 indicates the R-squared estimates for the predictions made for the MI filtered outcomes, using different amounts of past data. The R-squared values for MI are high, but not near perfect as they are for AMI and Hip Replacement. They suggest the estimates are

very good predictors of performance, especially for mortality. Moreover, they remain good predictors even when using only one year of data. The R-squared values for the outcome forecasts are presented in Table B.3. Both the actual and the expected R-squared values are very high, indicating that the forecasts are also able to predict the true values extremely well for up to two years after the end of the data set. The similar results for the VAR(1) and VAR(2) specifications suggest that the forecast performance is not sensitive to the lag choice specified in the VAR model.

Table B.2: Summary of estimated prediction accuracy using alternative methods of signal extraction. All estimates based on the VAR(1) model from Table B.1.

Expected R^2 prediction based on:						
	All 8 years		3 most recent years		Concurrent year	
	All outcomes	Same outcome	All outcomes	Same outcome	All outcomes	Same outcome
$D30_{ht}$						
2004	0.979257	0.979107	0.979253	0.979411	0.86238	0.977483
2006	0.981212	0.981341	0.982484	0.982588	0.98259	0.984905
$D365_{ht}$						
2004	0.882788	0.882535	0.882747	0.886962	0.984214	0.960961
2006	0.970821	0.970211	0.933748	0.934264	0.782221	0.777805
$R28_{ht}$						
2004	0.984656	0.984660	0.984659	0.984617	0.961512	0.984237
2006	0.992722	0.992707	0.992502	0.992492	0.992508	0.992510
$R365_{ht}$						
2004	0.864754	0.86511	0.864754	0.864274	0.977453	0.862513
2006	0.972337	0.972063	0.971442	0.971342	0.970737	0.970938

Table B.3: Summary of forecast accuracy using alternative forecasting models. Forecasting 2006-2008 values using data from 2000-2006.

	All outcomes	Same outcome	All outcomes	Same outcome
	VAR(1), forecasting with		VAR(2), forecasting with	
$D30_{ht}$				
2007(expected)	0.9986781	0.9987475	0.9988944	0.998772
2007 (actual)	0.9823117	0.9822435	0.9809794	0.9809223
2008(expected)	0.9974576	0.9975153	0.9974376	0.9975464
2008 (actual)	0.9882959	0.9881461	0.9880028	0.9879705

	All outcomes	Same outcome	All outcomes	Same outcome
$D365_{ht}$				
2007(expected)	0.9689792	0.9713811	0.9811399	0.9809792
2007 (actual)	0.960669	0.9594202	0.9705148	0.9705643
2008(expected)	0.9712172	0.9750321	0.98259	0.9842377
2008 (actual)	0.9507935	0.9486848	0.9723613	0.9717647
$R28_{ht}$				
2007(expected)	0.9954525	0.993432	0.9801697	0.979403
2007 (actual)	0.988189	0.9881358	0.9878182	0.987783
2008(expected)	0.955344	0.9565898	0.9661497	0.9630006
2008 (actual)	0.9538543	0.9537103	0.9532683	0.9533528
$R365_{ht}$				
2007(expected)	0.9917701	0.9915122	0.9746649	0.9715055
2007 (actual)	0.9582364	0.9581847	0.9587629	0.9587073
2008(expected)	0.9766199	0.9814568	0.9740965	0.9739562
2008 (actual)	0.9393019	0.9393542	0.9402394	0.9405488

B.2 IHD

The parameter estimates for IHD presented in Table B.4 show overall more persistent quality indicators that the previous conditions, especially for $D30_{ht}$. The coefficient on the $D30_{ht}$ lag varies between $0.67 - 0.68$, and it's initial variance, and variance of the residuals indicates very low standard deviations at around 1%; suggesting 30-day mortality is heavily influenced by its own past performance, with little differences across hospitals. The coefficients are lower for the other quality indicators at around 0.45 for $D365_{ht}$, 0.5 for $R28_{ht}$ and 0.3 for $R365_{ht}$. The initial variance and residual variance for $R28_{ht}$ both indicate standard deviations of around 2%. The initial variances and residual variances of $D365_{ht}$ and $R365_{ht}$ are higher, corresponding to a standard deviation of 3.6% and 4% and 2% and 4% respectively. The correlation coefficients between the indicators and residuals show strong positive associations between $D30_{ht}$ and $D365_{ht}$, as well as $R28_{ht}$ and $R365_{ht}$, but weak associations for all other quality sets.

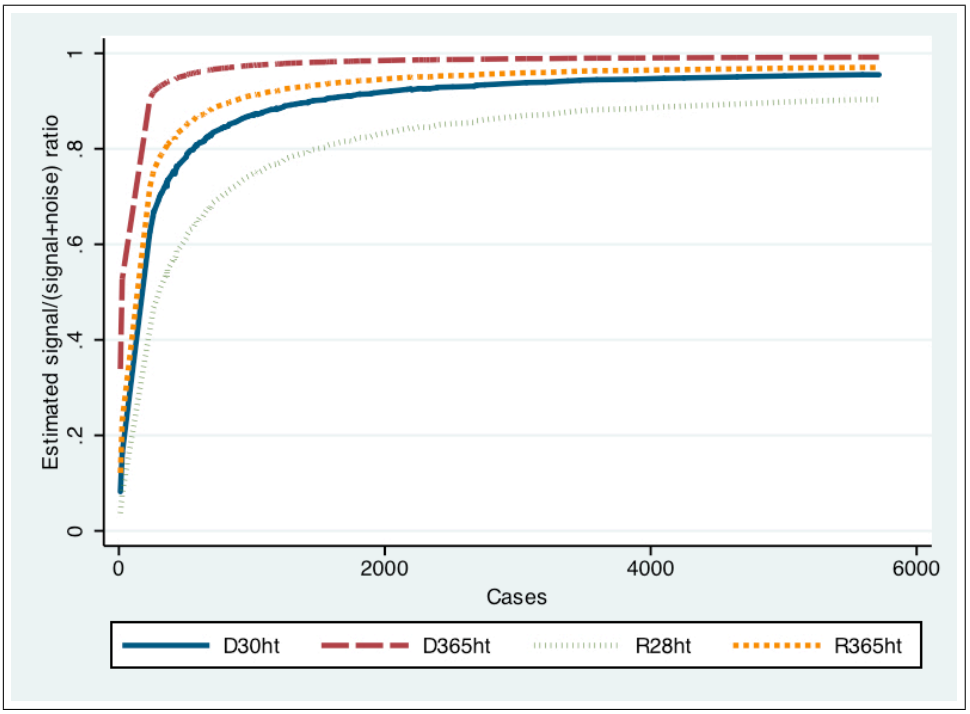
Table B.4: Estimates of IHD multivariate VAR(1) parameters for hospital specific effects.

	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D30_{h(t-1)}$	0.772371	-0.096571	1.166722	0.019439
	(0.02072)	(0.03869)	(0.17225)	(0.15913)
	[37.2845]	[-2.49620]	[6.77351]	[0.12216]
$R28_{h(t-1)}$	-0.002239	0.505702	-0.217399	0.048397
	(0.01521)	(0.02840)	(0.12646)	(0.11683)
	[-0.14723]	[17.8042]	[-1.71910]	[0.41426]
$D365_{h(t-1)}$	0.007407	0.004054	0.708119	0.066781
	(0.00328)	(0.00613)	(0.02731)	(0.02523)
	[2.25510]	[0.66098]	[25.9288]	[2.64686]
$R365_{h(t-1)}$	-0.014991	0.002601	-0.140492	0.527053
	(0.00601)	(0.01122)	(0.04998)	(0.04617)
	[-2.49410]	[0.23175]	[-2.81115]	[11.4155]
Residuals				
S.D. dependent	0.016641	0.021044	0.117278	0.078530
Correlation of residuals ($D30_{ht}$)	1.000000	0.023460	0.156514	-0.52385
Correlation of residuals ($R28_{ht}$)	0.023460	1.000000	-0.066481	-0.038846
Correlation of residuals ($D365_{ht}$)	0.156514	-0.066481	1.000000	0.029961
Correlation of residuals ($R365ht$)	-0.52385	-0.038846	0.029961	1.000000
Initial Conditions				
S.D. dependent in 2000	0.015556	0.021142	0.023958	0.042603
Correlation with $D30_{ht}$ in 2000	-	-0.1434	0.6986	-0.2566
Correlation with $R28_{ht}$ in 2000	-0.1434	-	0.0168	0.7812
Correlation with $D365_{ht}$ in 2000	0.6986	0.0168	-	-0.0335
Correlation with $R365ht$ in 2000	-0.2566	0.7812	-0.0335	-
Sample (adjusted): 2001 2008				
Included observations: 849 after adjustments				
Standard errors in ( ) & t-statistics in [ ]				

Figure B.6 plots the estimates of the signal to noise ratio against the number of cases for patients admitted for IHD. The ratios for the measures estimated for this condition are very high. The estimates have been constructed using a large sample of patients, which contributes to the strong ratios despite the weak signal variance observed in Table B.4. As cases increase, the signal for all four measures becomes stronger. Indeed, below 500 cases most of the measures are quite weak, although only a small sample of hospitals

have so few cases. Of the four measures, year long mortality and year long readmissions consistently out perform 30-day mortality and 28-day readmissions, of which the latter generally underperforms relative to the other three measures.

Figure B.6: Signal to noise ratio for the four IHD outcome measures (year 2005).



Figures B.7– B.10 present the filtered measures derived for IHD, their 95% confidence intervals and the corresponding latent outcome measures from Appendix A.2 for selected hospitals. Once again the filtered measures are able to provide smoother estimates over time than the latent measures. This makes it easier to observe an overall worsening in performance in all four hospitals, for all both mortality measures. Figures B.7 and B.8 suggest rising 30-day and 365-day mortality over time. In the same four hospitals 28-day emergency readmissions as well as 365-day readmissions are falling from above average or staying stable around the average rate. The confidence intervals surrounding these measures make it difficult to make any very conclusive results about whether performance is above or below average at any point in time, as they are very wide. This is most likely a result of the relatively few hospitals available in the sample.

Figure B.7: Filtered and latent estimates for IHD $D30_{ht}$ for selected hospitals.

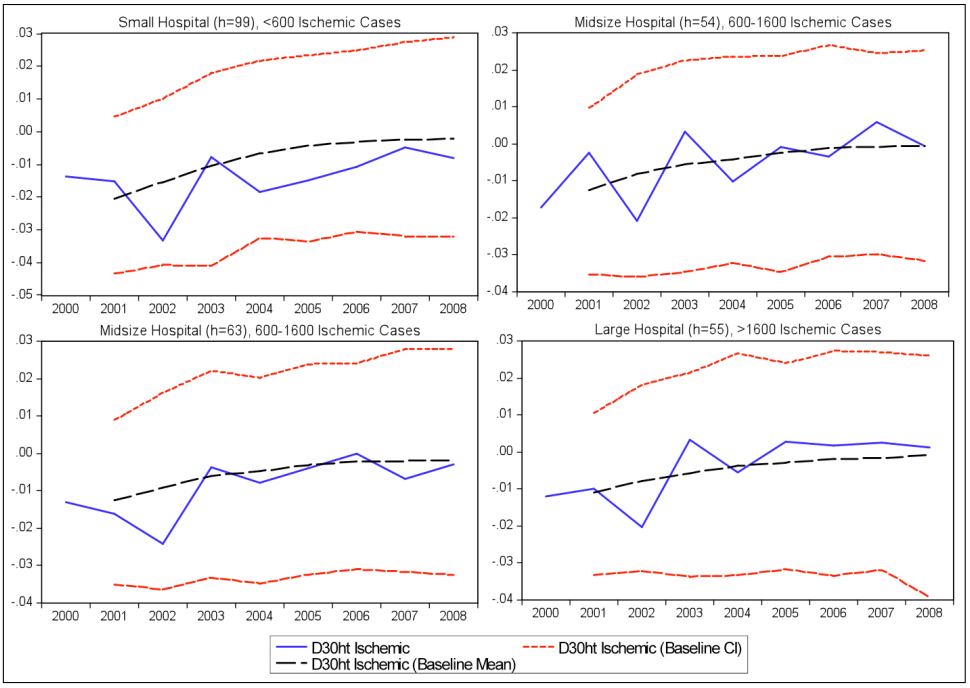


Figure B.8: Filtered and latent estimates for IHD $D365_{ht}$ for selected hospitals.

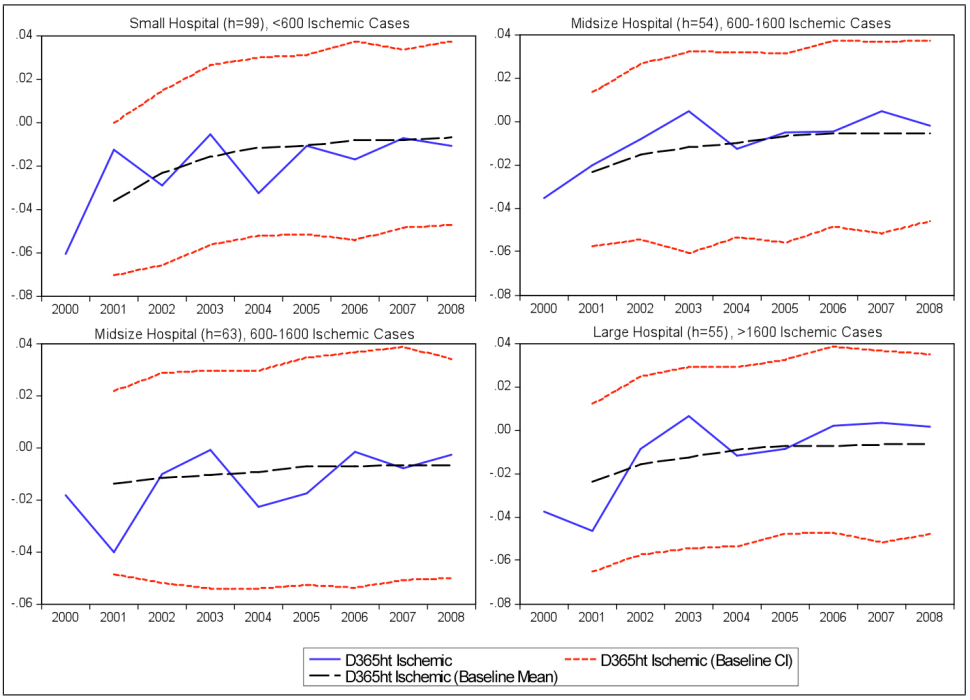


Figure B.9: Filtered and latent estimates for IHD $R28_{ht}$ for selected hospitals.

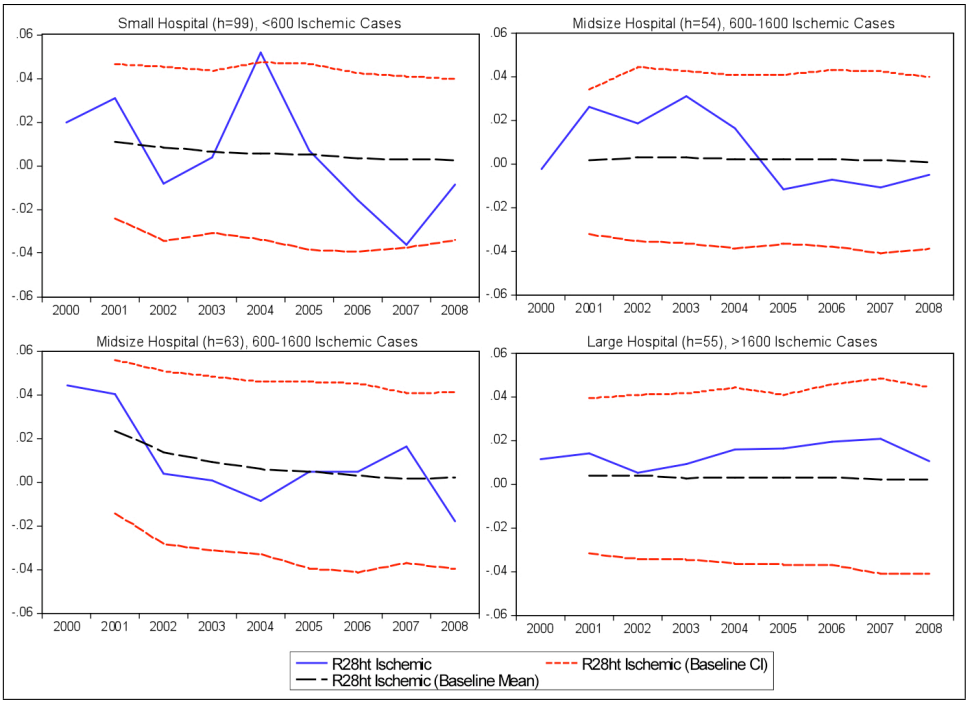


Figure B.10: Filtered and latent estimates for IHD $R365_{ht}$ for selected hospitals.

The R-squared measures for the predictions made for filtered outcomes of IHD, estimated using different amounts of past data are presented in Table B.5. The values are near perfect and suggest the estimates are very good predictors of performance, especially

for mortality. The values remain extremely high even when using only one year of data, but are weakest for the long-term measures. The R-squared estimates presented in Table B.6 for the forecasts are also very high, for both the VAR(1) and VAR(2) specifications of the model. This suggests that the filters are extremely good predictors of performance, as well as being able to make remarkably good forecasts. Moreover the model is not sensitive to the lag choice specified.

Table B.5: Summary of estimated prediction accuracy using alternative methods of signal extraction. All estimates based on the VAR(1) model from Table B.4.

Expected R^2 prediction based on:						
	All 8 years		3 most recent years		Concurrent year	
	All outcomes	Same outcome	All outcomes	Same outcome	All outcomes	Same outcome
$D30_{ht}$						
2004	0.999654	0.999652	0.999653	0.999652	0.999530	0.999531
2006	0.999782	0.999774	0.999702	0.999706	0.999792	0.999793
$D365_{ht}$						
2004	0.977067	0.976950	0.977257	0.976960	0.997215	0.997232
2006	0.977641	0.977988	0.907615	0.907003	0.859228	0.858978
$R28_{ht}$						
2004	0.998357	0.998349	0.998350	0.998352	0.998883	0.998909
2006	0.988082	0.988098	0.988040	0.988052	0.989795	0.989880
$R365_{ht}$						
2004	0.990350	0.990255	0.990264	0.990328	0.992136	0.992187
2006	0.957160	0.957129	0.957299	0.957102	0.959742	0.960013

Table B.6: Summary of forecast accuracy using alternative forecasting models. Forecasting 2006-2008 values using data from 2000-2006.

	All outcomes	Same outcome	All outcomes	Same outcome
	VAR(1), forecasting with		VAR(2), forecasting with	
$D30_{ht}$				
2007(expected)	0.9992606	0.999232	0.9990933	0.9990526
2007 (actual)	0.9997973	0.9997938	0.9997545	0.9997534
2008(expected)	0.9993752	0.9993606	0.9992415	0.9992353
2008 (actual)	0.9998603	0.9998556	0.9998349	0.9998349

	All outcomes	Same outcome	All outcomes	Same outcome
$D365_{ht}$				
2007(expected)	0.9988642	0.9987977	0.9959353	0.9950806
2007 (actual)	0.9827728	0.9826841	0.9591145	0.9593199
2008(expected)	0.8647367	0.8607932	0.8378298	0.8269623
2008 (actual)	0.992272	0.9919147	0.9709476	0.9715925
$R28_{ht}$				
2007(expected)	0.9991775	0.9992595	0.9974778	0.9974377
2007 (actual)	0.9895275	0.98956	0.9898381	0.9898477
2008(expected)	0.9972882	0.9976669	0.9899789	0.989258
2008 (actual)	0.9978728	0.9978783	0.9981748	0.9981768
$R365_{ht}$				
2007(expected)	0.9988965	0.9987113	0.9980229	0.9980314
2007 (actual)	0.9588806	0.9588513	0.959781	0.959805
2008(expected)	0.9993992	0.9993255	0.9993939	0.9993185
2008 (actual)	0.668666	0.6688468	0.6715124	0.671193

B.3 CCF

The VAR results for CCF, presented in Table B.7, show very low lag coefficients overall. Looking closely at each quality measure, the coefficients of each variable's own lags suggest that past performance has very little influence on current quality. The most dynamic quality measure of the lot is $D365_{ht}$, with a coefficient of around 0.17, with all others less that 0.1. The initial variance of indicators and the variance of their residuals, indicate higher variation across hospitals and annually. The standard deviation for both short term quality indicators ($D30_{ht}$ and $R28_{ht}$) in the year 2000 is 12%, and 10% amongst their residuals. The standard deviation is higher for the long term indicators, at 12% for $D365_{ht}$ 13.4% for $R365_{ht}$ in the year 2000, and at 16% and 17%for the residuals of $D365_{ht}$ and $R365_{ht}$ respectively. The correlation coefficients amongst indicators and residuals suggest a strong positive association between the pairs $D30_{ht}$ and $D365_{ht}$ as well as $R365_{ht}$ and $R28_{ht}$, and a mild negative association between $D30_{ht}$ and $R365_{ht}$. All other pairs have a weak negative correlation.

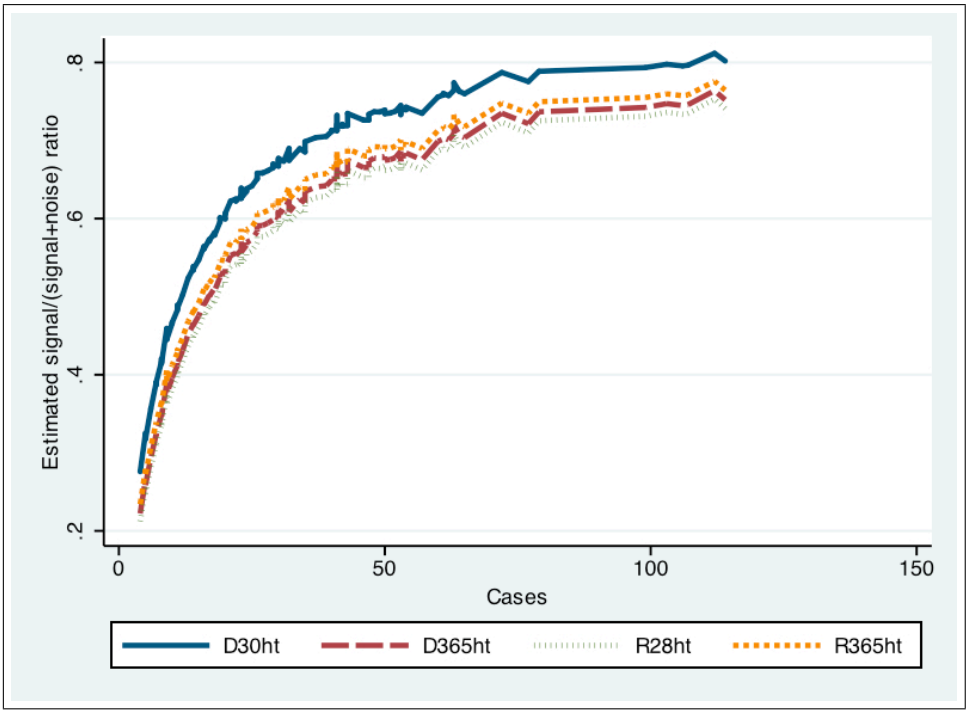
Table B.7: Estimates of CCF multivariate VAR(1) parameters for hospital specific effects.

	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D30_{h(t-1)}$	0.091842	0.021532	-0.101393	0.104027
	(0.04004)	(0.03691)	(0.04757)	(0.05055)
	[2.29368]	[0.58330]	[-2.13166]	[2.05785]
$R28_{h(t-1)}$	-0.033167	-0.075961	-0.129367	-0.125635
	(0.04308)	(0.03971)	(0.05117)	(0.05438)
	[-0.76999]	[-1.91283]	[-2.52822]	[-2.31024]
$D365_{h(t-1)}$	0.031140	-0.018943	0.234136	-0.063650
	(0.03081)	(0.02840)	(0.03660)	(0.03890)
	[1.01074]	[-0.66692]	[6.39752]	[-1.63643]
$R365_{h(t-1)}$	-0.026010	0.049197	0.042751	0.106006
	(0.03317)	(0.03058)	(0.03940)	(0.04188)
	[-0.78411]	[1.60873]	[1.08492]	[2.53124]
Residuals				
S.D. dependent	0.107777	0.098509	0.129701	0.135376
Correlation of residuals ($D30_{ht}$)	1.000000	-0.269111	0.521298	-0.383715
Correlation of residuals ($R28_{ht}$)	-0.269111	1.000000	-0.014197	0.652999
Correlation of residuals ($D365_{ht}$)	0.521298	-0.014197	1.000000	-0.060731
Correlation of residuals ($R365ht$)	-0.383715	0.652999	-0.060731	1.000000
Initial Conditions				
S.D. dependent in 2000	0.123641	0.121285	0.162693	0.174356
Correlation with $D30_{ht}$ in 2000	-	-0.2937	0.5604	-0.3892
Correlation with $R28_{ht}$ in 2000	-0.2937	-	0.1169	0.5774
Correlation with $D365_{ht}$ in 2000	0.5604	0.1169	-	-0.0888
Correlation with $R365ht$ in 2000	-0.3892	0.5774	-0.0888	-
Sample (adjusted): 2001 2008				
Included observations: 960 after adjustments				
Standard errors in ( ) & t-statistics in [ ]				

The signal to noise estimates for CCF is presented in FigureB.11, plotted against the number of cases treated in each hospital. The ratios are relatively low when compared to those of the other conditions, however they perform better than the signal to noise ratios of other conditions with such few cases. This is probably because the signal variance of all measures is relatively high (TableB.7). Unlike the other conditions, 30-day mortality has the highest signal to noise ratio, followed by year-long readmissions and the other

variables close behind.

Figure B.11: Signal to noise ratio for the four CCF outcome measures (year 2005).



Figures B.12– B.15 present the CCF filtered outcome measures, their 95% confidence intervals and the latent outcome measures presented in Appendix A.3. As is observed by the filtered measures for the other conditions, they are able to smooth out the values of the latent variables often allowing an easier interpretation of performance at a single point in time. This is particularly useful in cases where a smaller sample size may lead to more erratic latent measures, such as with the small hospitals. This is the case in the figures below, where the filtered estimates for hospital 3, in the upper left hand panel, have used the time series information from the latent measures to provide a smooth estimate across time. However, the confidence intervals for the filtered estimates are much larger than of the latent estimates due to the small sample of hospitals from which they are derived.

Figure B.12: Filtered and latent estimates for CCF $D30_{ht}$ for selected hospitals.

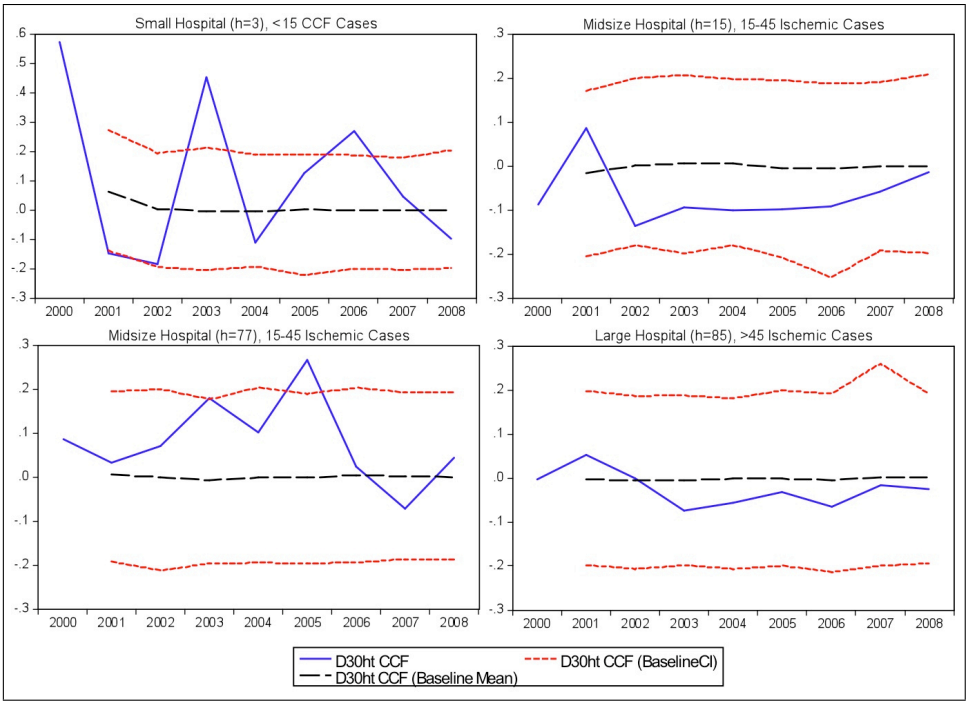


Figure B.13: Filtered and latent estimates for CCF $D365_{ht}$ for selected hospitals.

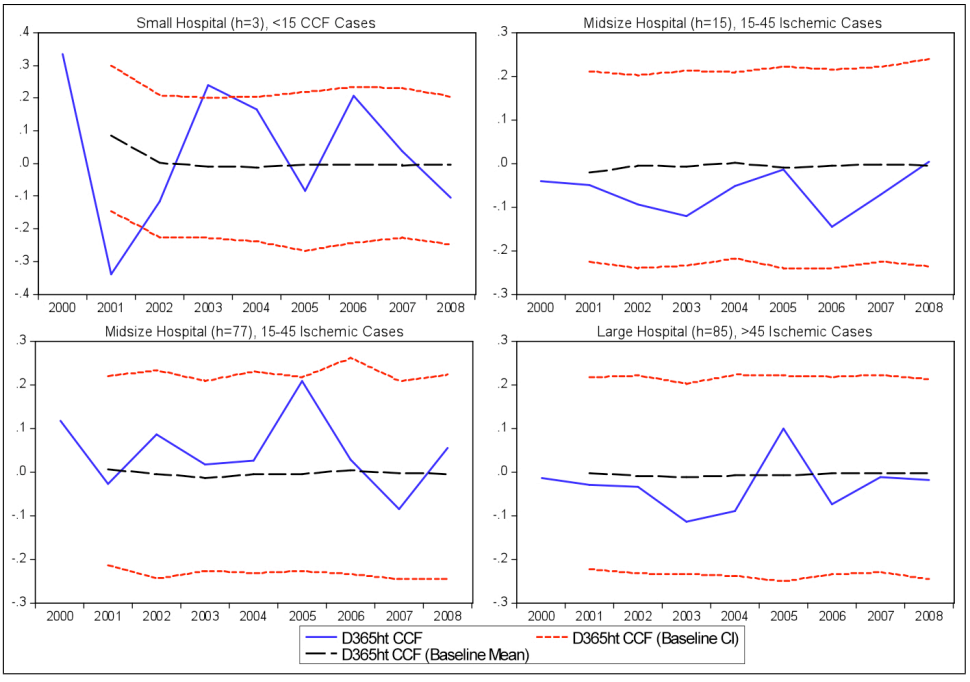


Figure B.14: Filtered and latent estimates for CCF $R28_{ht}$ for selected hospitals.

Figure B.15: Filtered and latent estimates for CCF $R365_{ht}$ for selected hospitals.

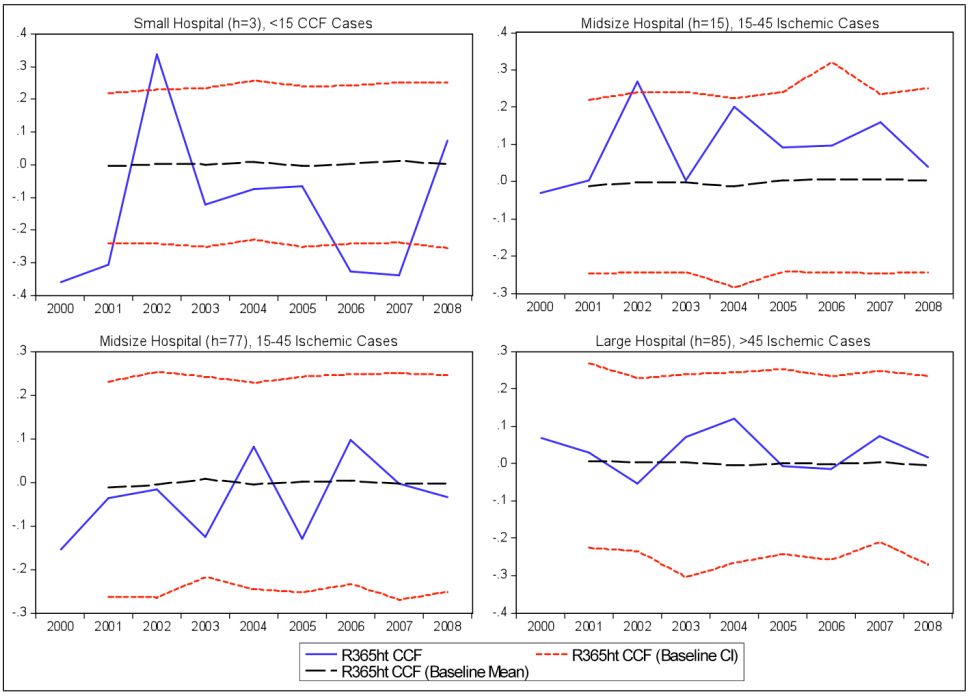


Table B.8 indicates the R-squared measures for the predictions CCF filtered outcomes of, estimated using different amounts of past data. The R-squared high, but not near perfect as they are for some of the other conditions. They suggest that the filtered outcomes

are good predictors of performance for both mortality measures, but slightly less so for the 28-day readmission measure. The values remain almost identical even when using only one year of data, but are weakest for the long-term measures. The R-squared estimates presented in Table B.9 for the forecasts are very high, suggesting that the filters are able to make remarkably good forecasts in addition to providing good predictions. Moreover the model is not sensitive to the lag choice specified as the results are very close both the VAR(1) and VAR(2) specifications of the model.

Table B.8: Summary of estimated prediction accuracy using alternative methods of signal extraction. All estimates based on the VAR(1) model from Table B.7.

Expected R^2 prediction based on:						
	All 8 years		3 most recent years		Concurrent year	
	All outcomes	Same outcome	All outcomes	Same outcome	All outcomes	Same outcome
$D30_{ht}$						
2004	0.975314	0.975353	0.975196	0.975670	0.972799	0.972910
2006	0.975988	0.976293	0.932016	0.976108	0.978864	0.978588
$D365_{ht}$						
2004	0.945162	0.945132	0.944651	0.945136	0.953034	0.952415
2006	0.966899	0.967043	0.885099	0.966757	0.959355	0.958919
$R28_{ht}$						
2004	0.924133	0.923822	0.923969	0.924196	0.924090	0.924214
2006	0.885111	0.885208	0.966626	0.885587	0.885823	0.885651
$R365_{ht}$						
2004	0.914019	0.913348	0.912940	0.914041	0.9177160	0.918538
2006	0.93241	0.932143	0.975844	0.932243	0.932146	0.931479

Table B.9: Summary of forecast accuracy using alternative forecasting models. Forecasting 2006-2008 values using data from 2000-2006.

	All outcomes	Same outcome	All outcomes	Same outcome
	VAR(1), forecasting with		VAR(2), forecasting with	
$D30_{ht}$				
2007(expected)	0.9737893	0.9740673	0.9760579	0.9763862
2007 (actual)	0.9648712	0.9647909	0.9652303	0.9652056
2008(expected)	0.975627	0.9750905	0.9743466	0.9739609
2008 (actual)	0.9767619	0.9767731	0.9769464	0.9766793

	All outcomes	Same outcome	All outcomes	Same outcome
$D365_{ht}$				
2007(expected)	0.9898384	0.9892356	0.9885849	0.9887405
2007 (actual)	0.9688362	0.9685165	0.9710999	0.9709966
2008(expected)	0.9909006	0.9908916	0.9904666	0.9910005
2008 (actual)	0.9499599	0.95004	0.9568327	0.9567122
$R28_{ht}$				
2007(expected)	0.9873314	0.9872443	0.9799542	0.9780822
2007 (actual)	0.973367	0.9734743	0.9737522	0.9738179
2008(expected)	0.9615405	0.9644274	0.9605659	0.9638652
2008 (actual)	0.961602	0.9620075	0.9619093	0.9618942
$R365_{ht}$				
2007(expected)	0.9982671	0.9985245	0.9977947	0.9975429
2007 (actual)	0.9384079	0.9380842	0.9652056	0.9368232
2008(expected)	0.9412284	0.9414987	0.9340023	0.9359227
2008 (actual)	0.9266524	0.9271426	0.9766793	0.9260207

B.4 TIA

The results for TIA shown in Table B.10, indicate that none of the quality measures are particularly influenced by their past values. Of these the most persistent dimension of hospital quality is 365-day mortality, which has lag coefficients ranging from $0.3 - 0.4$. All other quality measures have lag coefficients of 0.1 or less. The variance of quality measures in the year 2000, shows very little variation across hospitals for $D30_{ht}$ with a standard deviation of around 1%. There is higher variation for $D365_{ht}$, $R28_{ht}$ and $R365_{ht}$, with standard deviations of 4%, 4% and 6% respectively. There is a similar pattern in the variance of residuals, indicating little annual variation in $D30_{ht}$, and higher variation in the other quality indicators. The annual standard deviation of $D30_{ht}$ corresponds about 1%, and is around 6%, 4% and 7% for $D365_{ht}$, $R28_{ht}$ and $R365_{ht}$. The correlation coefficients on the quality indicators and their residuals suggest a positive correlation between $R28_{ht}$ and $R365_{ht}$, but no other strong association between amongst the other pairs.

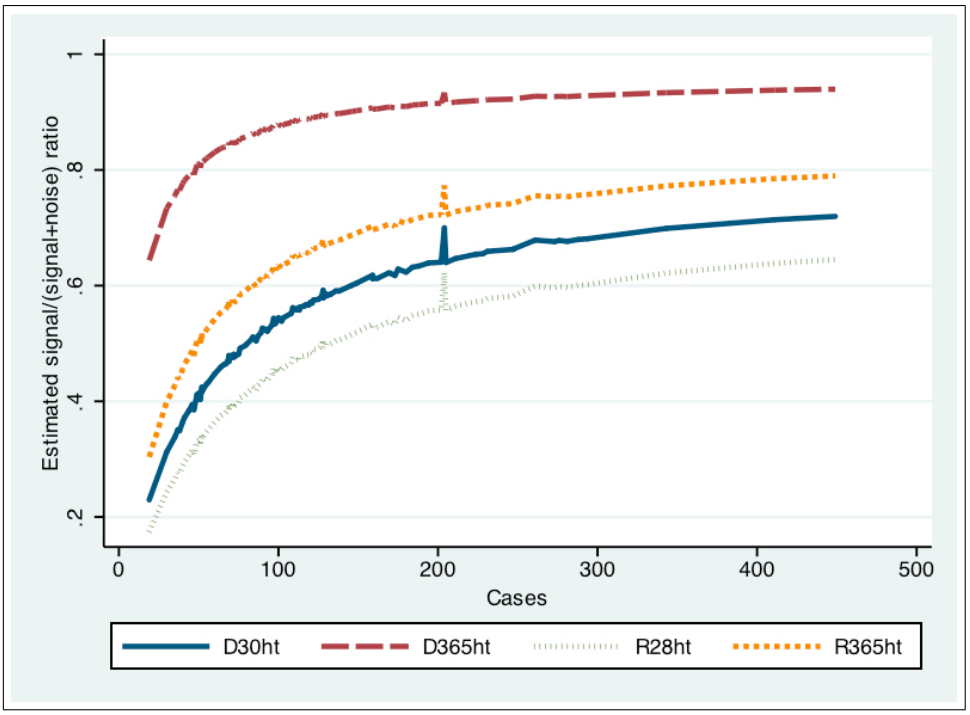
Table B.10: Estimates of TIA multivariate VAR(1) parameters for hospital specific effects.

	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D30_{h(t-1)}$	0.119608	0.336481	0.741973	0.361246
	(0.03143)	(0.09688)	(0.29679)	(0.24303)
	[3.80529]	[3.47316]	[2.49997]	[1.48645]
$R28_{h(t-1)}$	-0.044904	0.014876	0.213408	-0.158981
	(0.01072)	(0.03303)	(0.10119)	(0.08286)
	[-4.19027]	[0.45038]	[2.10904]	[-1.91875]
$D365_{h(t-1)}$	0.023533	-0.009371	0.814464	-0.039244
	(0.00288)	(0.00888)	(0.02720)	(0.02227)
	[8.17034]	[-1.05562]	[29.9470]	[-1.76219]
$R365_{h(t-1)}$	-7.01E-05	0.038492	0.163180	0.146632
	(0.00474)	(0.01462)	(0.04480)	(0.03669)
	[-0.01478]	[2.63193]	[3.64212]	[3.99685]
Residuals				
S.D. dependent	0.011979	0.034944	0.159175	0.087696
Correlation of residuals ($D30_{ht}$)	1.000000	0.019324	0.217958	0.036735
Correlation of residuals ($R28_{ht}$)	0.019324	1.000000	-0.026832	0.146502
Correlation of residuals ($D365_{ht}$)	0.217958	-0.026832	1.000000	0.072153
Correlation of residuals ($R365ht$)	0.036735	0.146502	0.072153	1.000000
Initial Conditions				
S.D. dependent in 2000	0.011136	0.038484	0.043116	0.038484
Correlation with $D30_{ht}$ in 2000	-	-0.0960	0.1432	-0.1025
Correlation with $R28_{ht}$ in 2000	-0.0960	-	0.2381	0.4184
Correlation with $D365_{ht}$ in 2000	0.1432	0.2381	-	0.3878
Correlation with $R365ht$ in 2000	-0.1025	0.4184	0.3878	-
Sample (adjusted): 2001 2008				
Included observations: 881 after adjustments				
Standard errors in ( ) & t-statistics in [ ]				

Figure B.16 illustrates the signal to noise estimates in the observed hospital outcome measures for TIA plotted against the number of cases treated in each hospital. While the signal to noise ratios improve when more cases are analysed, they perform relatively worse than the signal to noise ratios of other conditions estimated for the same amount of cases. Short term readmissions and short term mortality perform the worst, reflecting the little signal variation in the VAR parameters. Like most of the other conditions year-

long mortality has the highest signal to noise ratio the four measures. The poor overall performance is most probably related to the weak signals and small sample size used to construct these estimates.

Figure B.16: Signal to noise ratio for the four TIA outcome measures (year 2005).



Figures B.17–B.20 present the filtered outcome measures for TIA, their 95% confidence intervals and the latent outcome measures presented in Appendix A.4. The filtered estimates for TIA present the same properties observed in the previous conditions. The filtered estimates are able to smooth out the latent measures using information from previous time periods and the other outcome measures. Using these measures make it easier to interpret a single hospital's relative performance than using the latent measures which jump sharply from one period to the next. However, the filtered estimates of all conditions, and TIA, have much larger confidence intervals surrounding them. This suggests a larger degree of efficiency, and is most likely attributable to the small sample size.

Figure B.17: Filtered and latent estimates for TIA $D30_{ht}$ for selected hospitals.

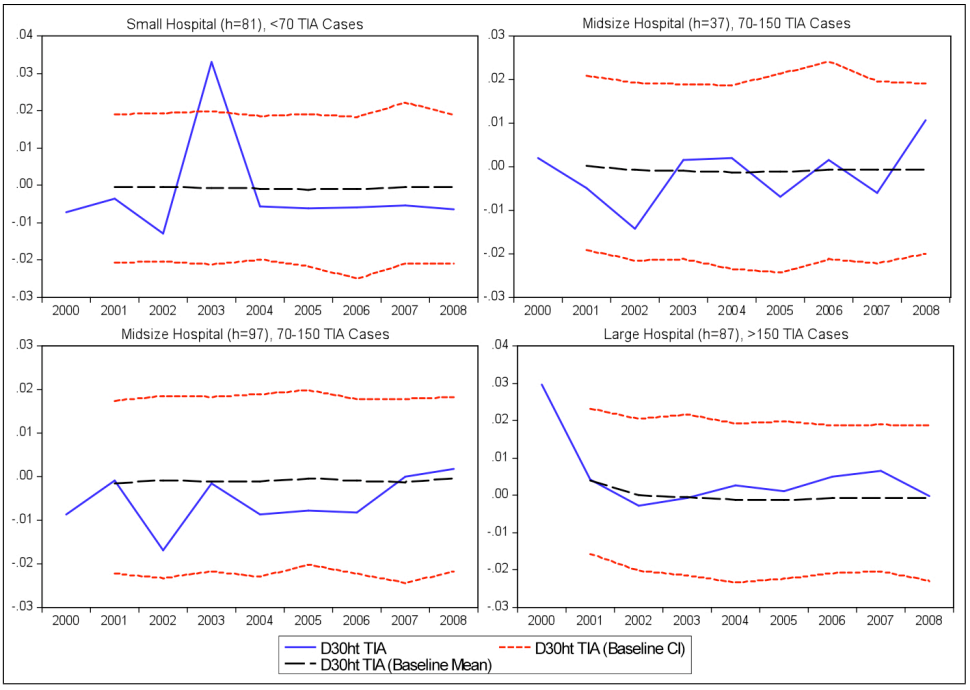


Figure B.18: Filtered and latent estimates for TIA $D365_{ht}$ for selected hospitals.

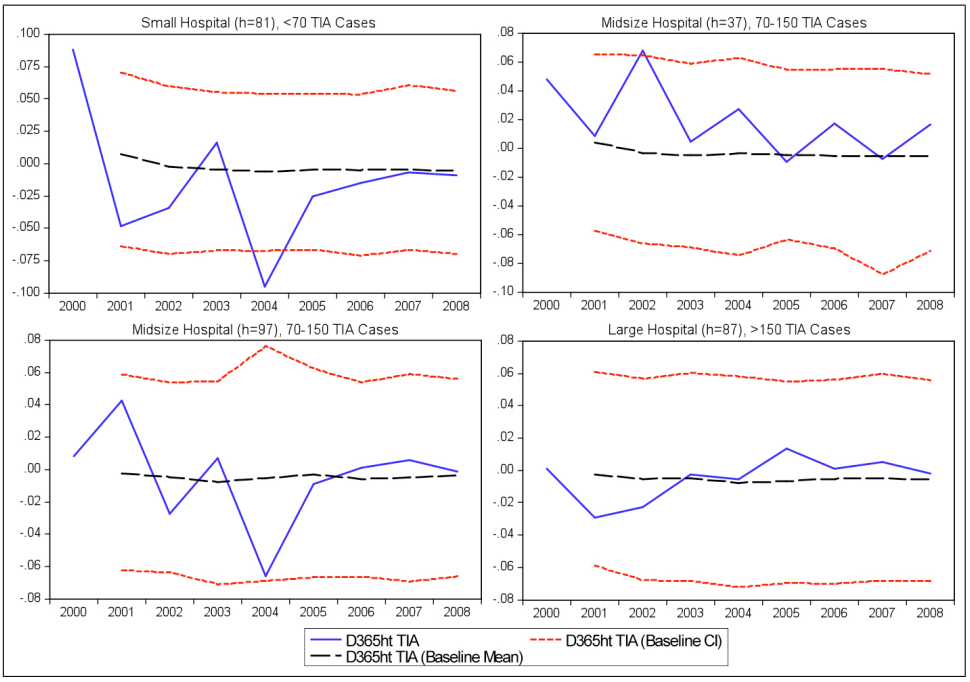


Figure B.19: Filtered and latent estimates for TIA $R28_{ht}$ for selected hospitals.

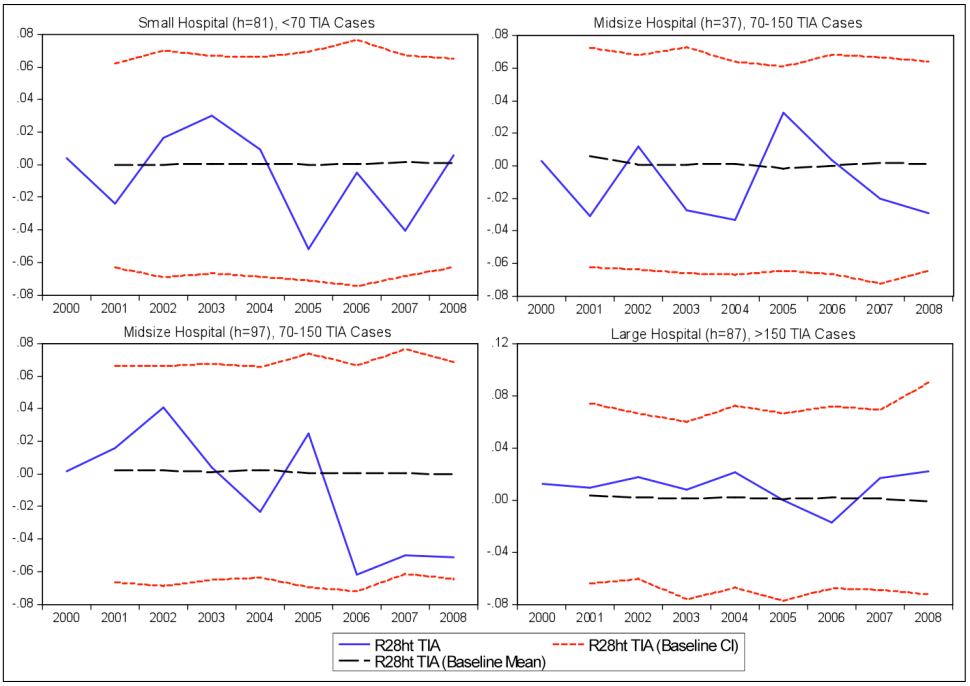


Figure B.20: Filtered and latent estimates for TIA $R365_{ht}$ for selected hospitals.

Table B.11 indicates the R-squared measures for the TIA filtered outcome predictions, estimated using different amounts of past data. The results suggest that the predictions made are extremely good, and almost perfect for some years. Only the one-year mortality measures in 2006 exhibits a particularly low prediction, when predicted with 3-years and

1-year of data. However, overall the data is able to predict the true quality remarkably well. Similarly, the forecasts, presented in Table B.12, are also very high, for both the VAR(1) and VAR(2) specifications of the model. The year-long mortality forecasts for 2007 and 2008 are the lowest in the VAR(1) specification, however this is improved with the VAR(2) specification. While the model is not sensitive to the lag choice specified for all other conditions and years, it appears to improve both the forecasts for this variable. However, the predictions under this specification fall by nearly 10%.

Table B.11: Summary of estimated prediction accuracy using alternative methods of signal extraction. All estimates based on the VAR(1) model from Table B.10.

Expected R^2 prediction based on:						
	All 8 years		3 most recent years		Concurrent year	
	All outcomes	Same outcome	All outcomes	Same outcome	All outcomes	Same outcome
$D30_{ht}$						
2004	0.999752	0.985212	0.999751	0.999755	0.999755	0.999757
2006	0.999780	0.999781	.9998308	0.999832	0.999812	0.999814
$D365_{ht}$						
2004	0.926829	0.999450	0.926862	0.92724	0.989374	0.989280
2006	0.938555	0.938117	.3441075	0.34271	0.015640	0.017167
$R28_{ht}$						
2004	0.999460	0.927045	0.999448	0.999453	0.999646	0.999648
2006	0.994492	0.994541	.9944965	0.994513	0.994335	0.994315
$R365_{ht}$						
2004	0.985387	0.999755	0.985705	0.98547	0.984629	0.984583
2006	0.991698	0.991751	.9919978	0.991985	0.992142	0.992124

Table B.12: Summary of forecast accuracy using alternative forecasting models. Forecasting 2006-2008 values using data from 2000-2006.

	All outcomes	Same outcome	All outcomes	Same outcome
	VAR(1), forecasting with		VAR(2), forecasting with	
$D30_{ht}$				
2007(expected)	0.9979086	0.9975985	0.9995478	0.9995313
2007 (actual)	0.9998127	0.9998112	0.9997479	0..9997458
2008(expected)	0.9984913	0.998384	0.9993958	0.9993626
2008 (actual)	0.9998712	0.9998671	0.9998718	0.9998712

	All outcomes	Same outcome	All outcomes	Same outcome
$D365_{ht}$				
2007(expected)	0.6524382	0.6156154	0.987202	0.9866609
2007 (actual)	0.8305044	0.8277816	0.7114277	0.7084467
2008(expected)	0.68696	0.6435999	0.9817656	0.9847195
2008 (actual)	0.922959	0.9217521	0.7958283	0.7915719
$R28_{ht}$				
2007(expected)	0.9847434	0.9836687	0.9933212	0.9930748
2007 (actual)	0.9960541	0.9960372	0.9961804	0.9961928
2008(expected)	0.9930163	0.9944037	0.9968603	0.9970341
2008 (actual)	0.9953419	0.9953497	0.9955111	0.995513
$R365_{ht}$				
2007(expected)	0.99931	0.9992769	0.9994361	.9994153
2007 (actual)	0.9807389	0.9805277	0.9809509	0.9809542
2008(expected)	0.9981527	0.9981514	0.9975603	0.9971483
2008 (actual)	0.9365411	0.9367729	0.9370904	0.9371955

B.5 Comparison of Indicators

Table B.13: Rankings of 2005 AMI $D365_{ht}$ measures.

Ranking	Mean $D365_{ht}$	Hospital	Latent $D365_{ht}$	Hospital	Filtered $D365_{ht}$	Hospital
Top 10						
1	0.087248	89	-10.9951	83	-3.58184	17
2	0.108949	55	-7.33457	119	-3.03305	54
3	0.113143	119	-7.01737	47	-2.52628	22
4	0.116531	52	-6.30999	42	-2.45368	3
5	0.123737	97	-5.81326	45	-2.3618	103
6	0.129032	62	-5.72524	15	-2.03254	18
7	0.141892	45	-5.43161	80	-2.01344	7
8	0.149923	112	-5.26463	91	-1.99536	107
9	0.150538	88	-4.79057	22	-1.79006	44
10	0.155303	19	-4.63651	62	-1.75694	40
Bottom 10						

Ranking	Mean $D365_{ht}$	Hospital	Latent $D365_{ht}$	Hospital	Filtered $D365_{ht}$	Hospital
110	0.27566	21	4.33436	21	0.998402	114
111	0.276094	76	4.432666	90	1.038727	41
112	0.276423	61	4.860979	96	1.257631	38
113	0.285	53	4.952693	53	1.294286	33
114	0.291228	41	4.981547	36	1.342004	35
115	0.29912	96	5.641765	107	1.400486	99
116	0.306338	71	7.110268	10	1.55153	66
117	0.312139	3	7.266694	3	2.121651	27
118	0.4	66	7.715625	71	2.203174	9
119	0.426573	43	18.70868	43	2.538532	56

Table B.14: Rankings of 2005 AMI $R28_{ht}$ measures.

Ranking	Mean $R28_{ht}$	Hospital	Latent $R28_{ht}$	Hospital	Filtered $R28_{ht}$	Hospital
Top 10						
1	0	66	-17.15334	66	-0.5097684	56
2	0.0410959	80	-13.77112	83	-0.4249609	27
3	0.0537634	62	-9.429364	62	-0.4037885	9
4	0.0758123	57	-7.974833	80	-0.3244067	38
5	0.0769231	43	-6.535955	43	-0.2796607	99
6	0.0824373	88	-4.354861	57	-0.2724753	41
7	0.0873786	113	-3.504739	113	-0.2207526	116
8	0.09375	36	-3.356516	36	-0.2168493	33
9	0.0989209	63	-2.987282	88	-0.1949843	35
10	0.0990415	51	-2.770071	45	-0.1943502	53
Bottom 10						
110	0.1564246	85	3.137839	23	0.5282587	83
111	0.1594203	27	3.158059	6	0.5671023	107
112	0.1598916	14	3.614229	27	0.5836549	106
113	0.1601423	23	3.619557	72	0.589515	40
114	0.1606061	16	3.697118	9	0.5981762	3
115	0.1615721	59	4.066902	14	0.6017366	18
116	0.164486	6	4.984622	46	0.7226745	22
117	0.1715976	9	5.005144	16	0.7290823	103
118	0.1856061	19	5.010148	19	0.8328696	17

Ranking	Mean $R28_{ht}$	Hospital	Latent $R28_{ht}$	Hospital	Filtered $R28_{ht}$	Hospital
119	0.1933962	46	5.822222	59	0.8452681	54

Table B.15: Rankings of 2005 AMI $R365_{ht}$ measures.

Ranking	Mean $R365_{ht}$	Hospital	Latent $R365_{ht}$	Hospital	Filtered $R365_{ht}$	Hospital
Top 10						
1	0.118881	43	-12.698	43	-0.6166016	56
2	0.167785	89	-12.05263	83	-0.5771485	33
3	0.169675	57	-7.481782	62	-0.4943517	99
4	0.172043	62	-7.190022	89	-0.483924	38
5	0.172524	51	-5.687592	113	-0.4257711	116
6	0.181004	88	-5.204206	33	-0.3740641	9
7	0.182222	99	-4.925087	99	-0.3105961	62
8	0.18932	113	-4.808173	51	-0.2949201	66
9	0.197425	33	-4.342741	88	-0.2844733	41
10	0.200557	102	-4.314243	58	-0.2767854	118
Bottom 10						
110	0.278986	27	4.798292	95	0.4114325	5
111	0.280397	78	4.828352	28	0.4298307	67
112	0.283951	72	4.928086	80	0.4988003	3
113	0.284734	95	5.156519	72	0.5044565	50
114	0.285266	86	5.201916	23	0.5073113	77
115	0.288256	23	5.52123	71	0.5493379	18
116	0.292254	71	5.75459	11	0.5793964	17
117	0.292553	11	5.975807	4	0.6079986	40
118	0.301887	46	7.396799	46	0.6340984	106
119	0.4	66	19.4526	66	0.6462436	54

C | Comments for Chapter 4

C.1 MI

The forecast error variance decompositions for MI indicate that its performance is highly dynamic. For all mortality and readmissions outcomes MI is able to predict over 90% of its own variance in a 10 year horizon. The results from the MI seemingly unrelated regressions confirm that performance is very highly dynamic, with all three lags being significant predictors of current outcomes, such that high outcomes in a previous period are highly associated with high outcomes in the current period. The Granger causality results for MI suggest that the lagged values of other conditions do significantly help to predict some of the MI outcomes, such as AMI, IHD, CCF and Hip Replacement. While the variance decomposition percentages show that MI explains most of its own forecast error, these conditions help to explain between 2-6% of the variance for some outcomes.

MI & IHD

The relationship between MI and AMI is discussed in the results section of the chapter. Moreover the relationship between MI and IHD is very similar to that of AMI and IHD, such that they have a dynamic reinforcing relationship for 30-day mortality. However, there is no significant Granger causality conditions as in the case of AMI and IHD. In the case of short term readmissions there is bidirectional Granger causality between the two conditions, and IHD explains over 5% of the variation in the forecast error of MI. The SUR results suggest that they are contemporaneously correlated, such that lower readmissions in one condition are associated with lower conditions in the other. The coefficient of the IHD variable in the MI model is very high, indicating that this effect explains a large amount of the variation in MI readmissions. However, the MI coefficient in the IHD model shows that MI readmissions explain only a small amount of the variance in IHD readmissions. Moreover lagged MI is also significant in the IHD model, although with a negative sign, such that lower lagged MI readmissions lead to higher IHD readmissions. Neither long term mortality nor long term readmissions are significantly associated between MI and IHD.

MI & CCF

What is most interesting about the MI results is that most relationships between MI and the other conditions is very different from the relationship of AMI with those conditions. The relationship between CCF and MI is such that the Granger causality estimates indicate no significant correlation between short or long term mortality. The SUR model also shows no effect between these two conditions for 30-day mortality. The long term mortality model indicates that lagged MI mortality is significantly positively associated with CCF mortality. The readmission models both indicate a dynamic competing relationship such the two conditions are negatively contemporaneously correlated, but the lag of one is positively associated with the other. This is different from the relationship we saw between AMI and CCF, where there was a dynamic reinforcing relationship.

MI & Stroke

Stroke and MI mortality have a dynamic reinforcing relationship for short term mortality, and a dynamic competing relationship for long term mortality. The SUR readmission models so no association between MI and Stroke. The Granger causality estimates are not significant and do not help is in understanding the relationship between the two conditions. However, exactly the same relationship is observed between TIA and MI. Given the similarities between TIA and Stroke it is likely that the same factors are driving the relationship between MI and these two conditions.

MI & TIA & Hip Replacement

Finally, the SUR model for MI indicates that contemporaneous Hip Replacement mortality is a very strong positive predictor of 30-day mortality. Moreover, in the Hip Replacement equation contemporaneous MI is insignificant, but lagged MI is positively associated with hip, such that increasing lagged MI mortality is associated with decreasing hip mortality, although only weakly. Interestingly, the long-term mortality SUR model indices no association between the two conditions. The 28-day readmissions model indicates a dynamic competing relationship between the two conditions, which is much stronger for MI than Hip Replacement. The long term readmissions model also shows a negative contemporaneous effect of Hip Replacement on MI but only at the 10% level. Moreover, lagged effects are not significant, nor is the contemporaneous effect of MI on Hip Replacement.

C.2 IHD

The forecast error variance decompositions for IHD indicate that its performance is highly dynamic. For all mortality and readmissions outcomes IHD is able to predict over 90% of its own variance in a 10 year horizon. The results from the IHD seemingly unrelated regressions confirm that performance is very highly dynamic, with all three lags being significant predictors of current outcomes, such that high outcomes in a previous period are highly associated with high outcomes in the current period. The Granger causality results for IHD suggest that the lagged values of many of the other conditions do significantly help to predict some of the IHD outcomes. While the variance decomposition percentages show that AMI, Stroke and Hip Replacement explains most of the IHD forecast error, at about 3-4%.

IHD & CCF

The relationship between IHD and AMI and IHD and MI is explained above. IHD and CCF are also related, such that IHD short term mortality has a very strong dynamic reinforcing effect on CCF mortality. That is IHD outcomes are positively associated with contemporaneous CCF outcomes, and their lag is negatively associated with CCF outcomes. This is the same effect IHD outcomes have on CCF, only the size of the coefficients is much smaller, indicating the the effect explains a much smaller amount of the IHD variance. The SUR model for long term mortality and short term readmissions only shows a contemporaneous reinforcing effect of one condition on the other. The coefficients on the long-term mortality variables are lower than the short term mortality model's, although in the 28-day readmission model the contemporaneous effect of IHD on CCF is quite strong. Finally the long term readmissions model indicates a dynamic competing effect of CCF on IHD, while only IHD lags are significant in positively influencing CCF. However, the coefficient of the lagged IHD variable is very strong indicating that it explains a good amount of the variation in CCF.

IHD & Stroke

IHD and Stroke 30-day mortality outcomes, and 28-day readmission outcomes are not significantly associated. However, year-long IHD mortality and Stroke mortality have a competing contemporaneous relationship. The coefficient of one condition on another is very low, however, indicating that this effect is weak. Yet the Granger causality estimates suggest that lagged outcomes of Stroke are significant in influencing IHD. Moreover, the year-long readmissions SUR indicates the two conditions have a weak reinforcing effect.

The effect indicates that lagged Stroke readmissions also influence IHD but not the other way around. Thus, the relationship between IHD and Stroke is such that some Stroke patients are readmitted with IHD conditions, hence if Stroke mortality is low in one year it is likely that IHD mortality is slightly higher. Yet, this will only apply to a small amount of cases.

IHD & TIA & Hip Replacement

TIA and IHD have a dynamic competing relationship for short and long term mortality, and short term readmissions. Moreover, the Granger causality estimates suggest that TIA Granger causes short term IHD mortality, while IHD Granger causes short term TIA readmissions. The long-term readmissions model indicates that IHD is negatively associated with contemporaneous TIA readmissions, but only at 10% significance. Finally, the only significant relationship between IHD and Hip Replacement is when looking at long-term readmissions. Hip Replacement is positively associated with IHD contemporaneous readmissions, while contemporaneous and lagged IHD is positively associated with hip readmissions. The value of the coefficients is very low in all cases though, indicating this only applies to a small number of cases.

C.3 CCF

CCF performance is highly dynamic as shown by the forecast error variance decompositions. For all mortality and readmissions outcomes CCF is able to predict over 90% of its own variance in a 10 year horizon. The results from the CCF seemingly unrelated regressions confirm that performance is very highly dynamic, with all three lags being significant predictors of current outcomes, such that high outcomes in a previous periods are highly associated with high outcomes in the current period. The Granger causality results for CCF suggest that the lagged values of many of the other conditions do significantly help to predict some of the variance in the CCF outcomes. While the variance decomposition percentages show that most of the other conditions do not predict more than 2% of the forecast error variance. The relationship between CCF, AMI, MI and IHD is discussed in the previous sections.

The Granger causality estimates indicate that CCF and Stroke mortality exhibit unidirectional causality such that lagged Stroke short and long term mortality influences CCF outcomes. The SUR models for these outcomes shows a dynamic competing relationship between the two variables, with higher coefficients in the CCF model. The results for the readmission indicators are slightly different. The Granger causality estimates indicate

no significant causality amongst the short term readmissions, but do show that lagged year-long CCF readmissions influence Stroke readmissions. The short term readmissions SUR model indicates a dynamic reinforcing effect of Stroke on CCF, but no effect of CCF on Stroke. While, the long term readmissions SUR model indicates no positive effects.

CCF & Stroke

Heart failure is a known risk factor for Stroke. Thus the competing dynamic relationship indicates that higher mortality in one of these conditions will result in lower mortality in the other, however the improvements in the lagged effect of one will translate to improvements in the other. As for short-term readmissions, the dynamic reinforcing effect of Stroke on heart failure, simply indicates that if emergency readmissions increase for Stroke, a certain amount of that will be from heart failure patients, and thus they are positively associated with CCF readmissions. While increased lagged CCF readmissions will result in fewer Stroke readmissions, presumably because there are fewer patients with this risk factor in the next year.

CCF & TIA & Hip Replacement

TIA and heart failure have a similar relationship, where the Granger causality estimates suggest that lagged values of CCF Granger cause TIA year-long mortality, and short term readmissions. While lagged values of TIA Granger cause CCF year-long readmissions. Moreover, the SUR results for the short term mortality model show a competing dynamic effect, which is much greater for TIA on CCF than the reverse. In the long run mortality model, however, only lagged CCF outcomes are significant in positively influencing TIA. The long term readmissions model shows no significant association between the conditions. Yet, the short-term model indicates the dynamic competing effect of TIA on CCF, and only a negative contemporaneous association of CCF on TIA.

The results of the SUR model indicate no significant relationship between CCF and Hip Replacement apart from a very small effect lagged CCF 30-day mortality has on short term hip mortality. However this effect is only significant at the 10% level, and has an extremely small coefficient.

C.4 Stroke

Similar to the other conditions Stroke is highly dynamic. The variance decomposition estimates for the four conditions indicate that in most cases it is able to predict over 95% of its own forecast error in a ten year forecast period. This is slightly less for year-long

mortality, but still high at 88%. The Granger causality estimates indicate that its lags are significantly associated with some of the other conditions being investigated such as TIA, and also that its can be explained by lags of the other conditions. The results of the SUR models, the Granger causality estimates, and the variance decomposition percentages allow us to understand some of the relationships between Stroke and the other conditions in more detail. The relationship with AMI, MI, IHD and CCF have already been discussed in the previous sections.

Stroke & TIA & Hip Replacement

The Granger causality estimates indicate that lagged values of TIA 30-day mortality are causally linked to Stroke, whereas lagged values of Stroke 30-day mortality are causally linked to TIA. The 30-day mortality SUR model indicates no significant effect between the two variables, while the year-long mortality SUR suggests a dynamic competing relationships between them. Moreover, the coefficients indicate that the effect of Stroke on TIA is much stronger than the reverse. While the Granger causality estimates are not significant for any of the readmission indicators, the SUR models indicate a dynamic competing relationship for short term readmissions and a dynamic reinforcing relationship for long term readmissions. As mentioned above, a TIA is exactly the same as a Stroke except the patient will recover in 24 hours. If the patient does not receive appropriate treatment shortly after they are at a great risk of having a Stroke in the next few weeks. Thus, the dynamic competing relationships we see for long term mortality and short term readmissions may be capturing the contemporaneous differences between providers in treating TIA, while also capturing the lagged effect which indicates that improvement in one area should filter through to the other. The reverse relationship is indicated for long term readmissions, which could be capturing the same effect the lagged short term readmissions are.

Stroke and Hip Replacement exhibit no significant relationship in either of the SUR mortality models. However, they do show a dynamic reinforcing relationship for short-term readmissions and a dynamic competing relationship in the long-term readmissions SUR. The value on the coefficients of the short term model are very low, however the Stroke coefficients on the hip model are quite high. This may indicate that some hip patients are readmitted with Stroke.

C.5 TIA & Hip Replacement

TIA and Hip Replacement are a very dynamic condition, explaining over 87% and 91% of the variance is their forecast error over a 10-period respectively. The Granger causality

and variance decomposition percentages indicate which conditions are significantly associated with TIA and Hip Replacement outcomes, and how much of the variance they influence respectively. The relationship between both conditions and the other five conditions included in the models have been explained in detail above. Hip and TIA are not very strongly related. However, contemporaneous short-term TIA mortality is negatively associated with 30-day hip mortality. Similarly year-long hip readmissions are negatively associated with TIA readmissions.

D | Results for Chapter 5

D.1 MI

The average latent and filtered hospital performance for MI is presented in Figure D.1. The curves illustrated are calculated from the latent and filtered estimates calculated in Chapters 2 and 3, which represent the marginal effect each hospital is having on different outcomes. Figure 4.1.1 indicates a decline throughout the period being investigated for 30-day in hospital mortality of about 0.5%, but an increase of about 0.5% for year-long mortality. In both mortality panels the filtered curves are much smoother than the latent curves which show a larger decline in mortality, of about 1.5% in 2003, that later improved. The filtered curves in the bottom two panels, representing short and long term readmissions, are also much smoother representations. The curves indicate an increase in both short and long term readmissions of about 2% and 4% respectively.

Figure D.1: Average hospital quality over time for MI.

The latent and filtered indicators are used in models 1 and 2 to determine how what effect PbR had on outcomes. The results in Table D.1 suggest that in most cases they are unable to predict over 8% of the variance in the dependent variable. Moreover, the PbR

dummy is only significant for filtered short and long term readmissions. The coefficient indicates that since PbR filtered short term readmissions have increased by about 2% and filtered long term readmissions by about 4%.

The tariff variable is only significant for the filtered 28-day readmission model, and only at 10%. It indicates that hospitals receiving a higher average tariff have lower short term readmissions. The average age variable is negatively associated with latent 30-day mortality, such that hospitals with a higher average age have lower latent mortality outcomes. Average length of stay is significant in all filtered mortality models, and the filtered 28-day readmission model. In all mortality models an increase in the average LOS of patients is associated with lower filtered mortality outcomes, but higher 28-day readmissions. Caseload is significant in all latent mortality models, as well as the latent year-long readmissions model. The coefficients suggest that an increase in caseload is associated with a decline in mortality, and an increase in readmissions. co-morbidity is not significant in any of the models, while average deprivation is significant for all the filtered models. The signs on the average deprivation coefficients for all the filtered mortality models suggest that hospitals with higher levels of deprived patients have higher mortalities, and lower readmissions. Finally, the dummy variable for foundation trust is significant for latent 28-day and year-long readmissions, as well as filtered year-long readmissions. In all cases it suggests that foundation trusts have higher readmissions.

Table D.1: MI Models 1 & 2.

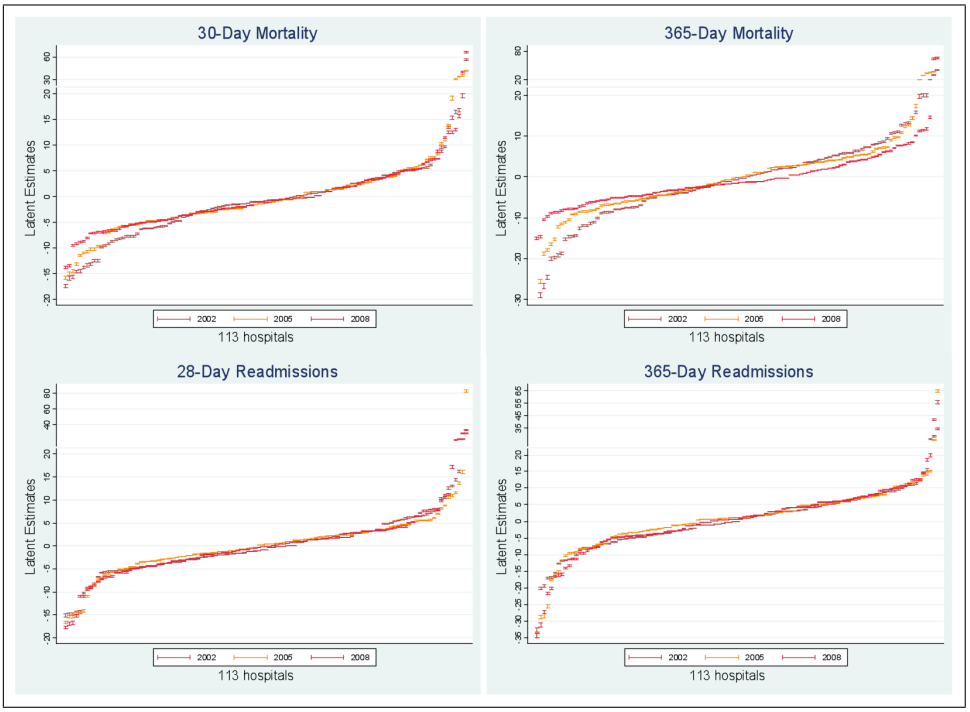
	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	-0.00268	0.000240	-0.00395	0.000285	0.000715	-9.94e-05*	0.000793	-1.32e-05
	(0.00249)	(0.000227)	(0.00395)	(0.000238)	(0.00121)	(5.07e-05)	(0.00168)	(6.21e-05)
Age	-0.512**	-0.00839	-0.0788	-0.00776	0.365	0.000623	0.424	-0.000909
	(0.248)	(0.0388)	(0.325)	(0.0415)	(0.236)	(0.0158)	(0.260)	(0.0138)
LOS	0.111	-0.0503**	0.317	-0.0542**	0.125	0.0118**	0.0748	0.00889
	(0.213)	(0.0218)	(0.351)	(0.0249)	(0.146)	(0.00555)	(0.155)	(0.00624)
Cases	-0.0692***	-0.000714	-0.0653**	-0.00101	0.0238	-3.95e-05	0.0672**	0.000170
	(0.0262)	(0.00245)	(0.0269)	(0.00266)	(0.0168)	(0.000723)	(0.0273)	(0.00114)
Co-morbidity	0.594	0.276	1.321	0.365	-0.0609	-0.0117	-0.0902	-0.0650
	(3.724)	(0.229)	(3.542)	(0.268)	(2.055)	(0.0731)	(2.945)	(0.0803)
Deprivation	0.901	0.590***	0.981	0.578***	-0.136	-0.157**	-1.472	-0.196***
	(1.495)	(0.172)	(1.417)	(0.186)	(0.893)	(0.0661)	(1.348)	(0.0745)
FT	-5.012	-0.207	-1.458	-0.258	5.060***	0.112	9.173***	0.256**

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
	(3.436)	(0.225)	(3.718)	(0.238)	(1.804)	(0.0768)	(2.893)	(0.105)
PbR(06)	7.140	-0.447	4.541	0.474	-2.990	1.753***	-4.866	3.591***
	(6.036)	(0.601)	(5.680)	(0.648)	(2.384)	(0.173)	(4.303)	(0.192)
Constant	46.19**	-0.293	17.21	-0.579	-30.01**	0.240	-35.85*	0.158
	(23.24)	(2.749)	(23.65)	(2.968)	(15.02)	(1.143)	(20.04)	(1.014)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,017	1,017	1,017	1,017	1,017	1,017	1,017	1,017
$\boldsymbol{R}^2$	0.034	0.059	0.022	0.059	0.050	0.529	0.070	0.798
Hospitals	113	113	113	113	113	113	113	113

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Figures D.2 and D.3 illustrate relative hospital performance for the years 2002, 2005, 2008 using the latent and filtered outcomes. The plots for the latent measures in Figure D.2 have many similar features to the corresponding plots in the other conditions. The outcomes fall in a large range, and have many outliers at either end. Moreover, all hospitals have very small confidence intervals for the point estimates of each hospital. The top two panels show the performance of short and long term mortality – and both indicate a slight convergence to the mean. For short term mortality this is driven mostly by an improvement in mortality from the below average hospitals, while for long-term mortality there is also a decline in the mortality of above average hospitals. The bottom two panels show hospitals plotted by readmission outcomes, in these panels the curves overlap indicating that performance in these areas has not changed much in the years investigated.

Figure D.2: Relative hospital performance over time for MI (normalized latent outcome indicators).



Hospital performance is plotted according to the filtered outcome measures in Figure D.3 for the years 2002, 2005 and 2008. As compared to the latent estimates, there less extreme outliers in either direction, and the range of outcomes is much smaller. The confidence intervals are also larger for the individual hospital estimates. The top two panels show the performance of hospitals with regards to short and long term mortalities. In both panels, relative hospital performance with regards to mortality has converged towards the mean after 2002, and not changed much from 2005 to 2008. The convergence towards the mean has been such that there are fewer hospitals with less than average mortality but also fewer hospitals with more than average mortality in the latter years. There has been a much smaller convergence to the mean in short term readmissions, and almost no convergence for long term readmissions.

Figure D.3: Relative hospital performance over time for MI (normalized filtered outcome indicators).

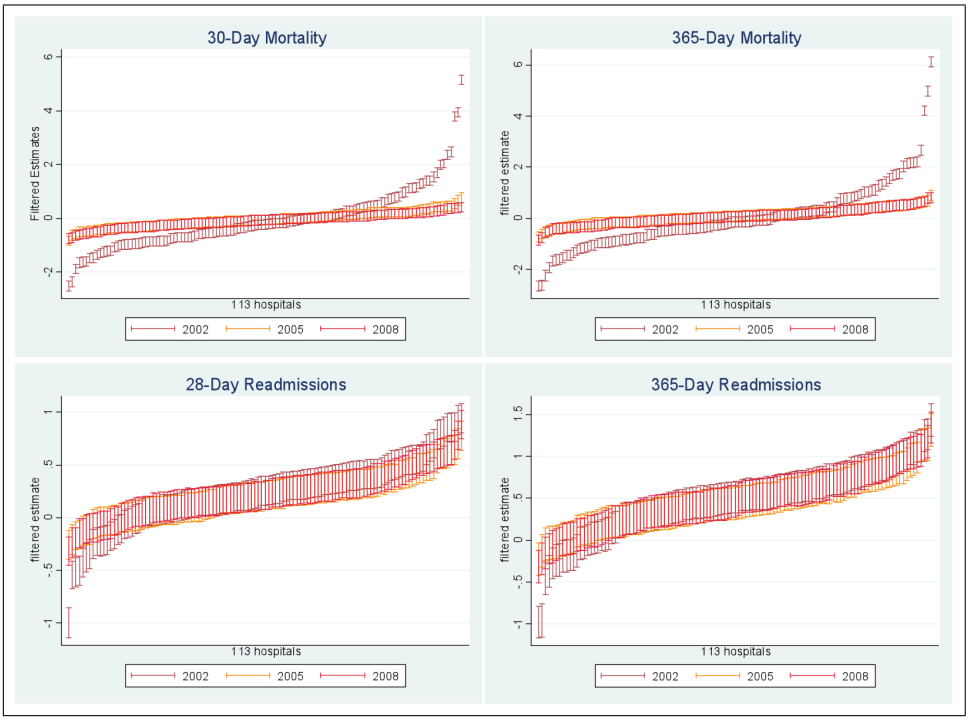


Table D.2 presents the results for models 3 and 4 which examine the factors which influence the normalized latent and filtered outcome measures plotted above. The R-squared estimates indicated that most models explain 6% or less of the variance in the dependent variable, apart from the filtered 30-day mortality model which is able to account for nearly 21%. The PbR dummy is not significant for any of the models. This suggests that PbR has not had a significant effect on the change relative hospital performance over this time period. Average age of the treated population and hospital status are also insignificant. Tariff is only significant at 10% for 30-day in hospital and overall latent mortality, but also for 30-day filtered mortality. In all three cases it has a negative coefficient indicating that a higher tariff is associated with lower mortality. These three models are also all negatively associated with caseload, such that more cases reduces mortality. Length of stay is positively associated with year-long mortality and negatively associated with latent and filtered year-long readmissions. This indicate that patients with a higher length of stay have a higher mortality and lower levels of year-long readmissions. Finally, average deprivation is positively associated with long-term readmissions, such that hospitals with a more deprived patient population have higher year-long readmission rates.

Table D.2: MI Models 3 & 4.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	-0.00345**	-0.00329*	-0.000104	1.47e-05	2.61e-06	1.89e-05	-2.76e-05	-4.35e-05
	(0.00174)	(0.00176)	(0.00172)	(0.000221)	(0.000218)	(0.000233)	(7.70e-05)	(7.48e-05)
Age	0.184	0.197	0.238	0.00419	0.00473	0.0147	-0.00272	0.00599
	(0.166)	(0.181)	(0.240)	(0.0258)	(0.0253)	(0.0288)	(0.0101)	(0.0106)
LOS	-0.106	-0.136	0.215*	0.0207	0.0211	0.0236	-0.0123**	-0.0181***
	(0.124)	(0.128)	(0.119)	(0.0189)	(0.0190)	(0.0225)	(0.00527)	(0.00668)
Cases	-0.0308***	-0.0286***	0.00734	-0.000237	-0.000299	-0.000463	-6.56e-06	-0.000199
	(0.00812)	(0.00850)	(0.00964)	(0.00148)	(0.00145)	(0.00159)	(0.000585)	(0.000658)
Co-morbidity	1.341	0.700	0.430	-0.189	-0.158	-0.281	0.194*	0.109
	(1.619)	(1.771)	(3.170)	(0.223)	(0.221)	(0.243)	(0.107)	(0.101)
Deprivation	0.718	0.578	-0.638	-0.175	-0.164	-0.189	0.118***	0.0694
	(1.139)	(1.147)	(1.418)	(0.108)	(0.108)	(0.121)	(0.0423)	(0.0573)
FT	1.731	1.854	-0.692	0.202	0.199	0.238	-0.0196	-0.0613
	(1.098)	(1.182)	(1.451)	(0.154)	(0.151)	(0.166)	(0.0585)	(0.0640)
PbR(06)	1.196	2.361	-0.759	-0.000598	0.00378	0.0386	0.00228	-0.00125
	(1.562)	(1.656)	(1.968)	(0.0546)	(0.0553)	(0.0607)	(0.0335)	(0.0398)
Constant	-1.365	-27.31**	-19.21	-0.238	-0.232	-0.741	0.256	0.172
	(12.46)	(13.42)	(16.39)	(1.652)	(1.611)	(1.835)	(0.650)	(0.696)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,017	1,017	1,017	904	904	904	904	904
R^2	0.068	0.208	0.016	0.015	0.014	0.021	0.037	0.030
Hospitals	113	113	113	113	113	113	113	113

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

D.2 IHD

The average filtered and latent outcome measures for IHD over the period 2000-2008 are presented in Figure D.4. The top two panels of the figure show the trend in 30-day in hospital and year-long mortality, both of which have fallen over the period studied. Short term mortality has fallen by about 4%, while year-long mortality by about 8%. The filtered estimates smooth out the different rates of change in the years, which in both cases are largest from 2000-2004. The bottom panel indicates the readmission indicators, and

suggests that 28-day readmissions have risen by about 2.5%, while year-long readmissions have risen by about 3%. In all four panels the filtered estimates smooth out some of the variations in outcomes that appear in the latent measures.

Figure D.4: Average hospital quality over time for IHD.

Models 1 and 2 use the latent and filtered outcome measures presented above as dependent variables, the results of these models are presented in Table D.3. The R-squared estimates vary considerably by model, estimating over 50% of the variance in most of the filtered models, and less than 35% for most of the latent models. The coefficients of the PbR dummy suggest it is significant for most models, apart latent year-long readmissions. In all latent mortality models, the PbR dummy is negative indicating that since the implementation of the policy, mortality rates have fallen. The magnitude of the coefficient indicates that the fall is about 3% and 6% for 30-day and year-long mortality respectively. The coefficient on the PbR dummy in the filtered 30-day and year-long mortality models are also negative, and also suggest the magnitude of declining mortality is 3% and 6% respectively. The PbR dummy is positively associated with latent 28-day readmissions, indicating a near 3% increase since the implementation of the policy. Finally the PbR dummy is also positively associated with filtered short and long term readmissions, such that they increase by 2% and 2.5% respectively.

The average tariff of hospitals is significant for both latent and filtered year-long mortality models, as well as in the latent 28-day readmission model. The sign on the coefficient

indicates that an increase in average tariff is associated with lower mortality rates and higher readmission rates. Average age is significant in all the filtered mortality models, and the latent year-long mortality model. In all cases the sign on the coefficient suggests that an increase in the average age is associated with a decline in mortality. Average LOS is not significant in any of the models. Average deprivation is negatively associated with all of the readmission models, such that hospitals with higher average deprivation have lower readmissions. Caseload is not significant in any of the models, and foundation trust status in only significant in the filtered readmission models at 10%. The sign on the foundation trust dummy variable indicates that hospitals with foundation trust status have higher readmissions.

Table D.3: IHD Models 1 & 2.

	Latent	Filtered	Latent	Filtered	Latent	Filtered	Latent	Filtered
	$D30_{ht}$	$D30_{ht}$	$D365_{ht}$	$D365_{ht}$	$R28_{ht}$	$R28_{ht}$	$R365_{ht}$	$R365_{ht}$
Tariff	-5.05e-05	-1.09e-05	-0.0018***	-0.000194**	0.00169***	0.000126	-0.000257	0.000107
	(0.000298)	(7.90e-05)	(0.000500)	(8.81e-05)	(0.000286)	(0.000109)	(0.000404)	(8.79e-05)
Age	-0.213	-0.202***	-0.557**	-0.275***	0.189	-0.0240	0.353	0.0514
	(0.223)	(0.0740)	(0.270)	(0.0757)	(0.376)	(0.0455)	(0.606)	(0.0692)
LOS	0.0248	-0.0229	0.663***	0.00626	0.218	-0.0109	0.512	-0.00474
	(0.105)	(0.0377)	(0.168)	(0.0420)	(0.194)	(0.0462)	(0.543)	(0.0448)
Cases	-0.000759	-0.000200	-0.000678	-0.000259	-0.000310	-8.74e-05	-0.000832	-1.92e-05
	(0.000736)	(0.000231)	(0.000848)	(0.000244)	(0.000683)	(0.000135)	(0.00121)	(0.000216)
Co-morbidity	-2.212	1.154	-0.955	1.691**	-1.702	-0.175	1.151	-0.528
	(2.439)	(0.777)	(2.590)	(0.792)	(4.459)	(0.391)	(4.824)	(0.690)
Deprivation	0.598	0.183	0.915	0.112	-1.852**	-0.458**	-11.16***	-0.484*
	(0.527)	(0.188)	(0.720)	(0.189)	(0.727)	(0.204)	(2.515)	(0.249)
FT	-0.605	-0.253	-0.750	-0.226	0.754	0.238*	2.109	0.346*
	(0.473)	(0.188)	(0.628)	(0.201)	(0.925)	(0.124)	(1.459)	(0.180)
PbR(06)	-2.987**	-2.995***	-6.066***	-6.104***	2.705***	1.846***	1.913	2.468***
	(1.315)	(0.423)	(1.480)	(0.452)	(0.907)	(0.243)	(1.806)	(0.390)
Constant	15.87	13.32***	38.31**	18.20***	-16.02	1.608	-24.46	-3.328
	(15.69)	(5.025)	(18.81)	(5.116)	(23.41)	(3.040)	(37.44)	(4.694)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,089	1,089	1,089	1,089	1,089	1,089	1,089	1,089
$\boldsymbol{R}^2$	0.222	0.555	0.353	0.796	0.095	0.529	0.161	0.516
Hospitals	121	121	121	121	121	121	121	121

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

The latent outcome indicators for IHD, ranked by hospital are presented in Figure D.5. While the diagrams do indicate outliers for all outcome measures, there are not as many or as extreme values as in the conditions presented above. However, similar to the other conditions the confidence intervals for each hospital's IHD latent outcome estimate are very small. In the top left hand panel we can see the performance of hospitals with regards to latent 30-day in hospital mortality. In this figure it is clear that performance in this area has improved greatly in 2005 and 2008 as compared to 2002, where it was largely below average. In the right hand panel illustrating year-long mortality, we also see an improvement in mortality from 2002 to 2008, only it is more gradual, and mostly felt in the worst performers. The bottom two panels indicating the performance of the readmission measures, show some increase in short term readmissions, and a slight decline in 365-day readmissions, as both indicators converge to their mean value.

Figure D.5: Relative hospital performance over time for IHD (normalized latent outcome indicators).

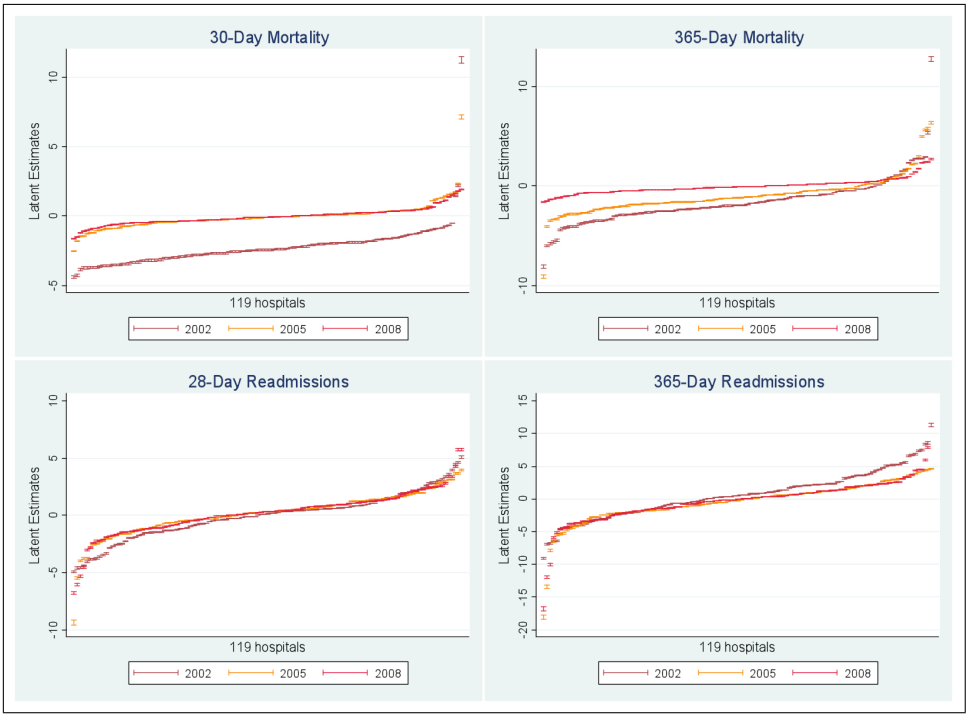
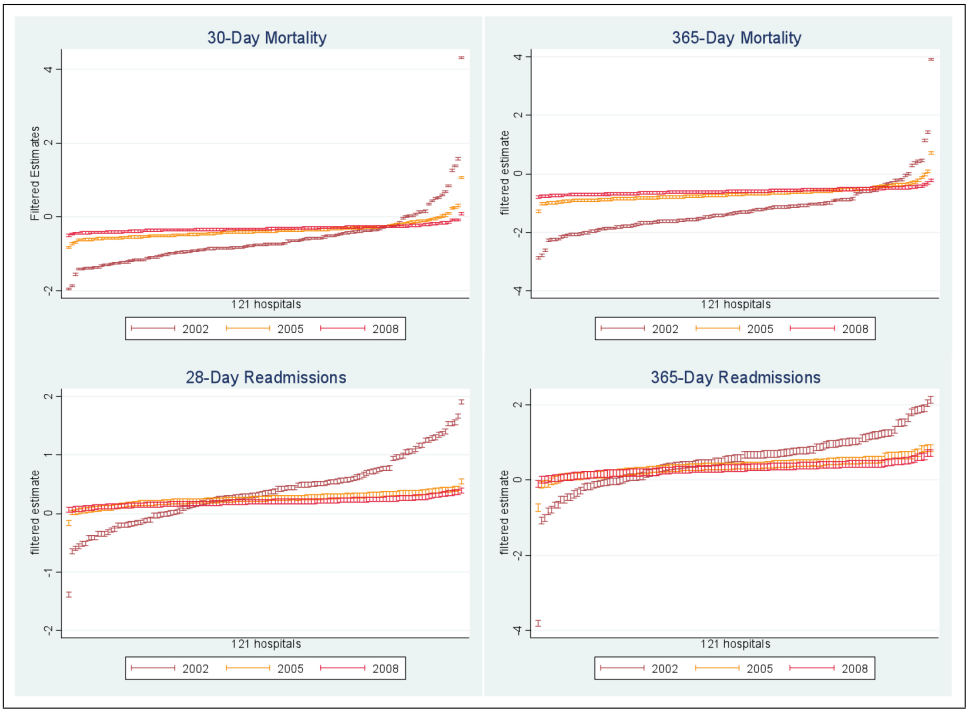


Figure D.6 indicates the relative hospital performance throughout time as measured by the normalized filtered outcome indicators. There are much fewer outliers in these measures, and the confidence intervals are also very small, indeed they are noticeably smaller than in the corresponding diagrams for the other conditions. In all four panels of Figure D.6, the indicators appear to converge to the mean, such that there is less variation in relative performance in the latter years of the sample. The change in relative

performance between 2002 to 2005 is much greater, than the change from 2005 to 2008 in all panels. For short term and long term mortality, this is exhibited mostly by an improvement in the below average performers, but there is also a decline in the above average outliers. For both short term and long term readmissions the convergence to the mean is also apparent. Many of the above average readmissions are falling to mean levels, but there is also an increase in readmissions that were below average. Also in all cases, the convergence does not seem to be at 0, indeed for mortality it is at some level below zero, while for readmissions it is for some level slightly above zero.

Figure D.6: Relative hospital performance over time for IHD (normalized filtered outcome indicators).



The normalized indicators plotted in Figures D.5 and D.6 are used as the dependent variables for models 3 and 4 are presented in Table D.4. The R-squared values vary considerably by model, ranging from a low of 5% to nearly 55%. The readmissions models, are generally weaker than the mortality models. The PbR dummy is significant and positive for all mortality indicators. This indicates that in since PbR relative mortality has been increasing, such that there are fewer hospitals with below average mortality. Of the readmission models, PbR is only significant for latent 28-day readmissions, where it indicates that relative readmissions have fallen since the introduction of PbR. The tariff variable is only significant for latent short term and long term readmissions, such that a higher average tariff is positively associated with an increase in short term readmissions,

and a decrease in long-term readmissions. The average age of patients is significant and positive for all mortality indicators. This indicates that an increase in average age of patients is associated with an increase in mortality. Average age is also positively associated with latent short and long term readmissions, such that higher average age is correlated with higher readmissions. Average LOS is only associated with latent year-long mortality, indicating that higher average LOS leads to higher year-long mortality.

Average co-morbidity is negatively associated with all latent mortalities, and also year-long filtered mortality. This indicates that higher co-morbidity is correlated with lower mortality. Average deprivation is also negatively correlated with latent short term mortality, and positively correlated with both filtered readmission models. This indicates that hospitals with more deprived patients have higher readmissions and lower 30-day mortality. Caseload is not significant in any of the models. Foundation trust status in only significant in the filtered readmission models, and indicates that foundation trusts have lower readmissions than other trusts.

Table D.4: IHD Models 3 & 4.

	Latent	Filtered	Latent	Filtered	Latent	Filtered	Latent	Filtered
	$D30_{ht}$	$D30_{ht}$	$D365_{ht}$	$D365_{ht}$	$R28_{ht}$	$R28_{ht}$	$R365_{ht}$	$R365_{ht}$
Tariff	-8.96e-05	4.85e-05	-4.67e-05	8.76e-05	0.000285**	4.33e-05	-0.00184*	1.32e-05
	(6.37e-05)	(9.52e-05)	(8.62e-05)	(0.000122)	(0.000139)	(9.53e-05)	(0.00103)	(0.000119)
Age	0.193**	0.0606**	0.429***	0.105***	0.289***	0.0426	0.556***	0.0150
	(0.0861)	(0.0273)	(0.0958)	(0.0292)	(0.0791)	(0.0282)	(0.148)	(0.0350)
LOS	0.0103	0.00546	0.203***	-0.000669	-0.0392	-0.0278	0.00169	-0.0201
	(0.0556)	(0.0211)	(0.0601)	(0.0266)	(0.0659)	(0.0314)	(0.340)	(0.0354)
Cases	2.65e-05	6.42e-05	0.000102	8.03e-05	0.000222	1.29e-05	0.000529	-5.01e-06
	(0.000173)	(8.86e-05)	(0.000213)	(0.000103)	(0.000202)	(6.45e-05)	(0.000558)	(9.58e-05)
Co-morbidity	-1.227*	-0.344	-4.367***	-0.656**	-1.118	0.0676	1.745	0.0918
	(0.632)	(0.279)	(0.939)	(0.297)	(0.838)	(0.215)	(3.940)	(0.318)
Deprivation	-0.456*	-0.0205	-0.793	0.00800	-0.865	0.246**	-7.728	0.288**
	(0.241)	(0.0523)	(0.574)	(0.0891)	(1.229)	(0.112)	(5.551)	(0.140)
FT	0.0491	0.106	-0.00991	0.111	-0.0351	-0.124*	-0.403	-0.154*
	(0.132)	(0.0660)	(0.212)	(0.0807)	(0.241)	(0.0691)	(0.635)	(0.0866)
PbR(06)	1.247***	0.529***	3.814***	1.326***	-0.966***	-0.225	-1.114	-0.110
	(0.260)	(0.167)	(0.331)	(0.201)	(0.321)	(0.142)	(1.875)	(0.193)
Constant	-13.41**	-4.957**	-30.97***	-8.899***	-18.55***	-2.415	-34.56***	-0.505
	(5.661)	(1.956)	(6.447)	(2.076)	(5.051)	(1.859)	(8.777)	(2.401)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

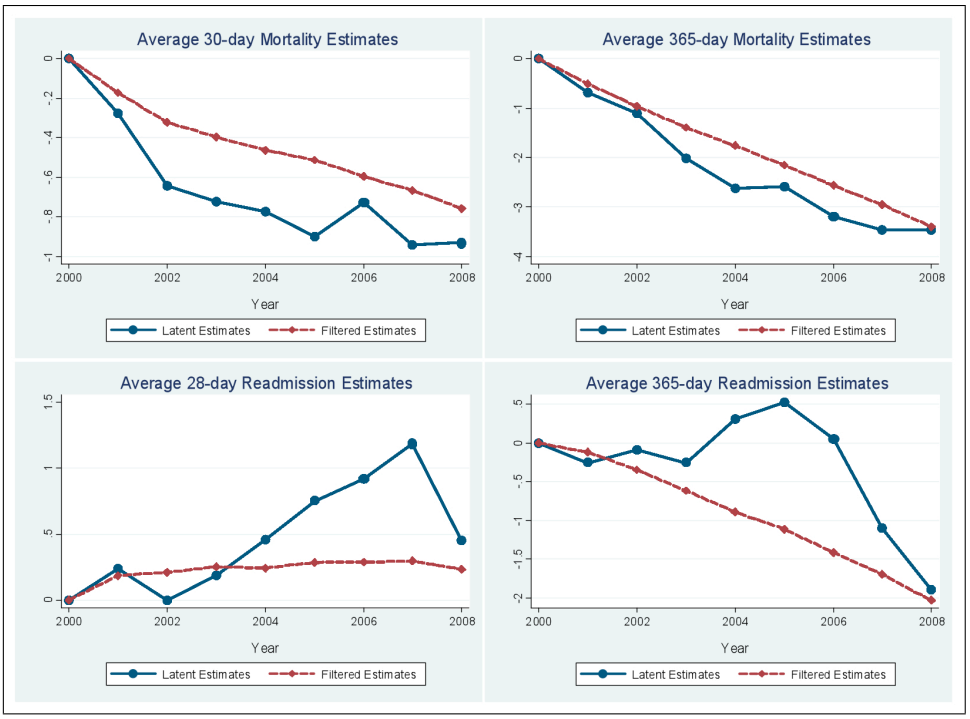
	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
N	1,089	968	1,089	968	1,089	968	1,089	968
$\boldsymbol{R}^2$	0.444	0.216	0.540	0.496	0.143	0.085	0.184	0.054
Hospitals	121	121	121	121	121	121	121	121

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

D.3 CCF

Figure D.7 indicates the latent and filtered CCF outcomes over time. Both the filtered and the latent mortality indicators show a decline in 30-day mortality during the period 2000-2008. The former has gradually declined by about a 1%, while the latter has gradually declined by about 3.5%. While both the filtered and the latent measures are declining steadily, there is much more year-to-year variation in the latent measures than in the filtered measures. The bottom two panels show the performance in the readmission estimates over the same time period. Both the latent and the filtered measures show an increase in 28-day readmissions of around 0.5% However the latent measures show an almost double increase up till 2007, before they fall to 0.5%, while the filtered estimates show a smoother curve over the years. A similar difference between the latent and filtered curves is indicates in the year-long readmission estimates, where both the filtered and latent measures show a decline of 2%, but following a different trajectory through time.

Figure D.7: Average hospital quality over time for CCF..



The latent and filtered indicators plotted in Figure D.7 are used as dependent variables in models 1 and 2, and the results are presented in Table D.5. These R-squared value of the models is below 7% in most cases, with the exception of year-long filtered mortality and year-long filtered readmissions. In all cases the filtered outcome measures are able to explain more of the variance in the dependent variable than the latent measure. The PbR dummy is significant for filtered 30-day in hospital mortality and filtered year-long mortality. In all cases it is negative, indicating a fall in mortality since the introduction of PbR, by 0.6% and 2.6% respectively. Of the readmission models, the PbR dummy is only significant for year-long readmissions, where it is also negative and indicates a 1.5% decline since the implementation of the policy.

Most of the other explanatory variables are not significant. Average tariff, average co-morbidity and average deprivation are not significant for any of the models. Average age is significant at 10% for the filtered 30-day mortality and year-long readmission models. The coefficients suggest that hospitals treating an older age group have higher short term mortality and lower long term readmissions. Average LOS is only significant in the latent 30-day mortality model, where it is positively associated with the dependent variable. Caseload is significant at the 10% level for latent 30-day mortality and both filtered readmission measures, such that higher caseload is associated with lower mortality and increased readmissions. Finally hospitals with foundation trust status are positively

associated with filtered 28-day readmissions, but only at the 10% significance level.

Table D.5: CCF Models 1 & 2.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.000477	7.17e-05	-0.000543	0.000137	-3.78e-05	4.18e-05	-0.000236	8.23e-05
	(0.00116)	(9.43e-05)	(0.00130)	(9.54e-05)	(0.000901)	(3.11e-05)	(0.00125)	(5.76e-05)
Age	-0.0709	0.0196*	-0.175	0.0232	0.0492	-0.00617	-0.0823	-0.0221*
	(0.216)	(0.0113)	(0.233)	(0.0144)	(0.186)	(0.00751)	(0.207)	(0.0118)
LOS	0.298	0.00577	0.544**	0.00543	-0.0380	-0.00102	-0.0375	-0.00678
	(0.219)	(0.0157)	(0.255)	(0.0179)	(0.176)	(0.00636)	(0.254)	(0.0114)
Cases	-0.113*	-0.000144	-0.0986	0.00111	0.0537	0.00425**	0.0770	0.00549*
	(0.0601)	(0.00227)	(0.0695)	(0.00294)	(0.0402)	(0.00209)	(0.0641)	(0.00317)
Co-morbidity	0.979	-0.0949	0.113	0.0390	0.555	0.0826	1.110	-0.0263
	(2.149)	(0.106)	(2.182)	(0.127)	(1.853)	(0.0673)	(2.079)	(0.0844)
Deprivation	0.933	-0.0602	-0.493	-0.0752	0.0212	0.00195	-0.320	-0.0121
	(0.853)	(0.0576)	(1.125)	(0.0711)	(0.750)	(0.0288)	(0.867)	(0.0564)
FT	1.518	-0.0642	-0.857	-0.0569	1.127	0.149*	0.817	0.170
	(3.248)	(0.134)	(3.676)	(0.187)	(2.420)	(0.0841)	(4.000)	(0.134)
PbR(06)	0.578	-0.601**	-1.200	-2.642***	-0.483	0.175	-1.528	-1.478***
	(3.337)	(0.240)	(4.030)	(0.298)	(2.671)	(0.154)	(3.434)	(0.237)
Constant	1.589	-1.518**	10.31	-2.120**	-4.838	0.163	4.129	1.379*
	(15.03)	(0.763)	(16.66)	(0.956)	(12.20)	(0.509)	(14.65)	(0.830)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,080	1,080	1,080	1,080	1,080	1,080	1,080	1,080
$\boldsymbol{R^2}$	0.017	0.077	0.020	0.507	0.005	0.037	0.006	0.387
Hospitals	120	120	120	120	120	120	120	120

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

In order to better understand the changes in relative performance in our outcomes of interest for CCF, we plot the normalized latent and filtered outcome indicators for different years. Figure D.8 shows the relative performance of all hospitals on four latent outcome measures, plotted for the years 2002, 2005 and 2008. Similar to the latent outcome indicator plots in many of the other conditions, there are many outliers at either end, and the confidence intervals for each hospital measure are very small. The top left hand panel shows the short term mortality measures, the curves for the three years are very similar

and there does not appear to be a distinctive trend in mortality. However, for year-long mortality, shown in the top right hand panel, we can see a gradual convergence towards the mean from 2002 to 2008, such that in 2008 there are fewer outliers and more hospitals are performing closer to average than in the other years. The two bottom panels show the latent outcome measures for short and long term readmissions. In both of these cases it is difficult to observe any trend as the curves are largely overlapping.

Figure D.8: Relative hospital performance over time for CCF (normalized latent outcome indicators).

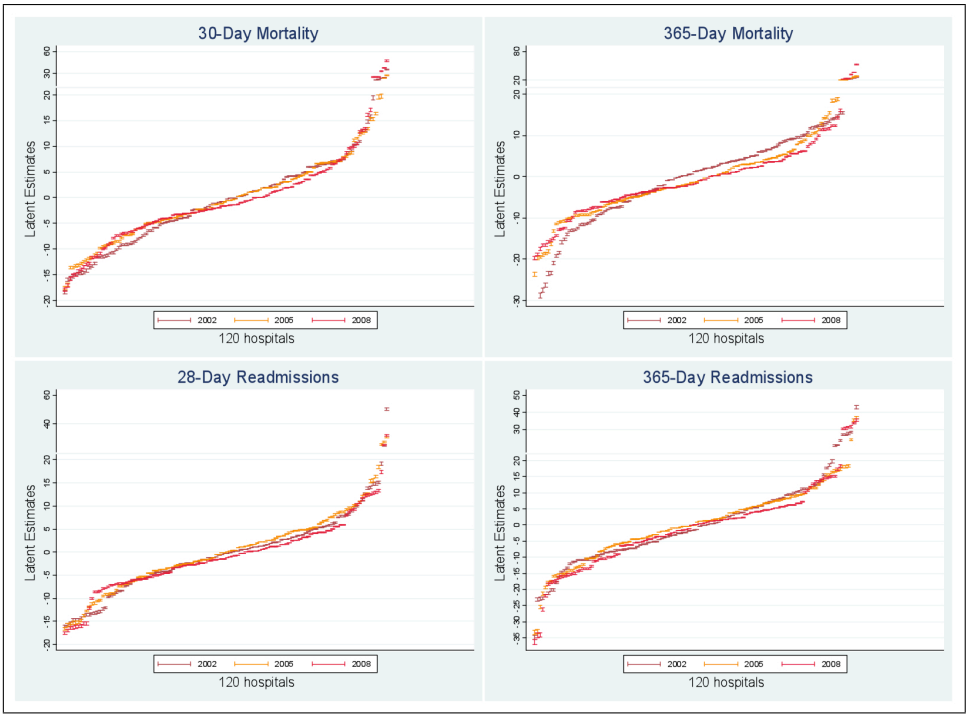
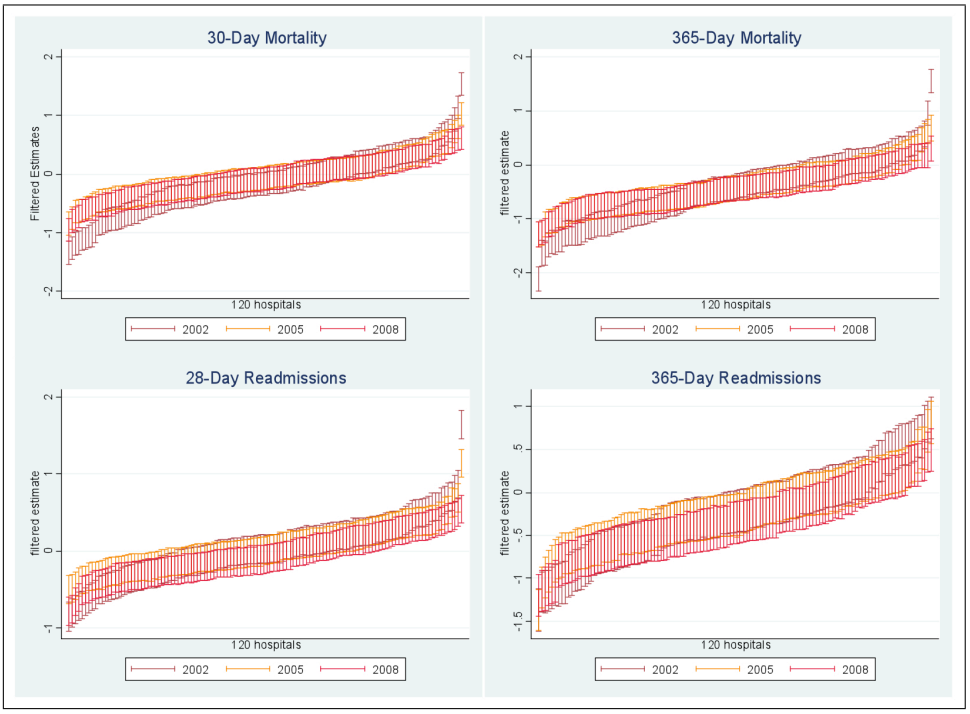


Figure D.9 shows the performance of the hospitals with regards to the filtered outcome indicators calculated for CCF. Again the range of estimates is much smaller for the filtered measures than the latent indicators, and there are fewer outliers. Moreover, the confidence intervals for the hospital estimates are larger than they are for latent outcome measures. It is difficult to identify a trend in relative performance for any of the indicators. Although in the mortality panels, there appears to be a very slight flattening of the curve, indicating less variation amongst hospitals. This is most distinct for the hospitals with below average mortality, where the line appears to have moved closer to the mean for 2005 and 2008. There is some evidence of this behaviour for long term readmissions, although mostly amongst hospitals with higher readmission rates in 2002. While there appears to be a similar trend for short term mortality, it is much smaller.

Figure D.9: Relative hospital performance over time for CCF (normalized filtered outcome indicators).



Models 3 and 4 explore the factors influencing the normalized latent and filtered outcome indicators, including the effects of the PbR policy. The results in Table D.6 indicate that the PbR dummy is not significant in any of the models. The tariff variable is only significant for year-long mortality, both latent and filtered, where it is associated with a decline in mortality. Average length of stay is associated with higher latent year-long mortality, and lower filtered short-term readmissions. Caseload is significant for both filtered readmission measures. In both instances, more cases are associated with lower readmissions. Average deprivation of the patients being treated in each hospital is insignificant, while average co-morbidity is associated with higher latent mortality in all intervals. Finally foundation trust status is significant for short term readmissions, such that they have lower readmissions. The R-squared estimates indicate that most models explain only around $1-3\%$ of the variance in the dependent variables.

Table D.6: CCF Models 3 & 4.

	Latent	Filtered	Latent	Filtered	Latent	Filtered	Latent	Filtered
	$D30_{ht}$	$D30_{ht}$	$D365_{ht}$	$D365_{ht}$	$R28_{ht}$	$R28_{ht}$	$R365_{ht}$	$R365_{ht}$
Tariff	-0.000100	-9.96e-05	-0.00164**	-0.000143*	0.000117	4.21e-05	-0.000847	1.87e-05
	(0.000751)	(6.25e-05)	(0.000761)	(8.05e-05)	(0.000754)	(4.94e-05)	(0.000794)	(8.12e-05)

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Age	-0.228	0.0126	-0.164	0.0232	-0.0924	-0.00701	0.0406	-0.0136
	(0.142)	(0.0125)	(0.136)	(0.0156)	(0.118)	(0.00870)	(0.140)	(0.0126)
LOS	-0.133	0.00887	0.324**	0.0101	-0.0383	-0.0130*	-0.173	-0.0158
	(0.162)	(0.0115)	(0.145)	(0.0147)	(0.156)	(0.00756)	(0.205)	(0.0121)
Cases	-0.0300	0.00250	-0.0171	0.000903	0.0289	-0.00342**	0.0234	-0.00455*
	(0.0248)	(0.00215)	(0.0242)	(0.00272)	(0.0190)	(0.00171)	(0.0281)	(0.00255)
Co-morbidity	3.713***	-0.0475	2.880**	-0.135	-1.755	0.0564	-0.555	0.0172
	(1.359)	(0.120)	(1.343)	(0.146)	(1.109)	(0.0603)	(1.459)	(0.0911)
Deprivation	0.832	-0.0267	-0.781	-0.0175	0.254	-0.0407	-0.537	-0.00969
	(0.646)	(0.0590)	(0.767)	(0.0697)	(0.767)	(0.0422)	(0.904)	(0.0550)
FT	1.627	0.108	1.124	0.0991	-2.816**	-0.0937	-2.127	-0.0405
	(1.090)	(0.0974)	(1.216)	(0.112)	(1.131)	(0.0765)	(1.685)	(0.111)
PbR(06)	0.551	0.00824	-0.602	0.00536	0.898	-0.205	-0.0104	-0.169
	(1.534)	(0.185)	(1.771)	(0.235)	(1.522)	(0.137)	(2.036)	(0.195)
Constant	13.27	-0.931	9.138	-1.785*	8.432	0.769	1.173	1.104
	(10.47)	(0.823)	(9.844)	(1.038)	(7.847)	(0.631)	(10.07)	(0.924)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,080	960	1,080	960	1,080	960	1,080	960
$\boldsymbol{R}^2$	0.033	0.013	0.019	0.017	0.015	0.029	0.012	0.019
Hospitals	120	120	120	120	120	120	120	120

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

D.4 TIA

The average TIA latent and filtered indicators for the period 2000-2008 are plotted in Figure D.10. From this Figure we can observe that TIA mortality has been falling. The filtered and latent 30-day mortality panel indicates that it has fallen by about 0.6%, where as year-long mortality has fallen by about 3%. Similar to the other conditions, the trajectory of the filtered and latent estimates are different, as the filtered estimates smooth over the large jumps in mortality for some of the years. The bottom two plots illustrate the change in readmissions over the same period, and indicate that they have been relatively stable. 28-day readmissions only fall by about 0.2%, while year long readmissions have almost the same value in 2008 as they did in 2000. However in both cases, the latent curve indicates some fluctuation from year to year, yet this is never greater than a change of

0.7%.

Figure D.10: Average hospital quality over time for TIA..

The latent and filtered indicators plotted in FigureD.10 are used as dependent variables in models 1 and 2. The results for models are presented in Table D.7. The R-squared values for the models are very mixed, ranging from 1.5%-91%. All readmission models are relatively poor at explaining the variance in the dependent variables, and in most cases the filtered mortality models are better than the latent ones. The PbR dummy is significant for filtered 30-day mortality and both filtered and latent year-long mortality. In all these cases it is associated with a decline in mortality, of 0.6 for the the short term model, and around 3% for the year-long models. Of the four readmission models, the PbR dummy is only significant for the 28-day readmission model, where it is associated with an 0.4% decline in readmissions. Many of the other explanatory variables are insignificant for the models, including average tariff, average age, average LOS, caseload and average co-morbidity. Average deprivation is only significant in the latent year-long readmissions model where, higher levels of average deprivation are associated with a decline in readmissions. Foundation trust status is associated filtered year-long readmissions, such that foundation trusts have lower long-term readmissions.

Table D.7: TIA Models 1 & 2.

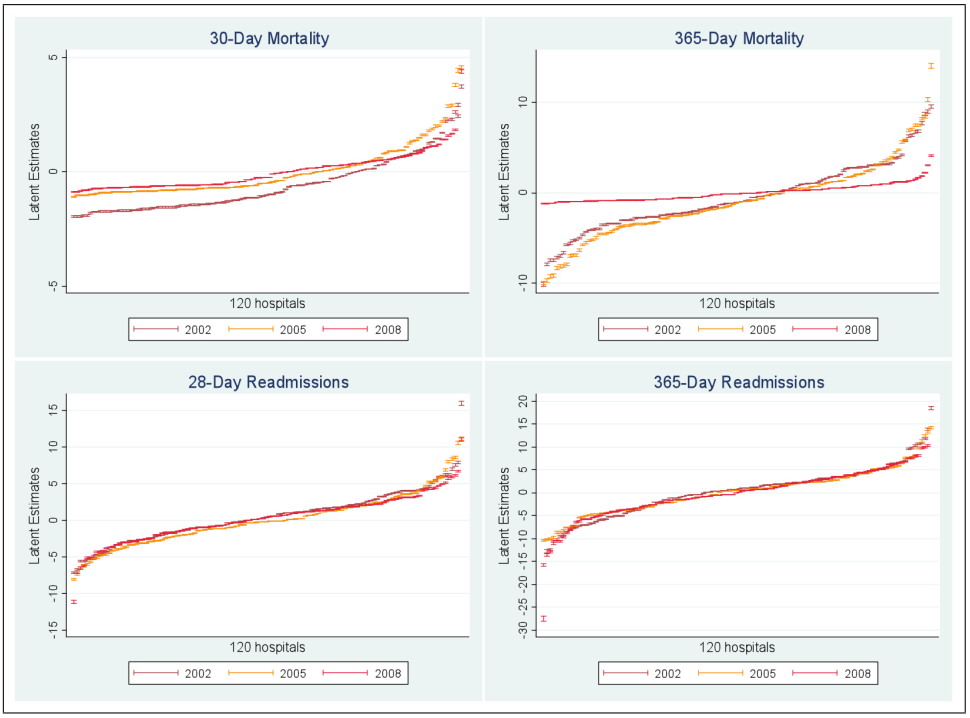
	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.000664	-2.11e-06	-0.000834	-4.38e-05	-0.000286	-3.99e-05	0.000334	-0.000168
	(0.000565)	(2.44e-05)	(0.00177)	(0.000104)	(0.00169)	(9.05e-05)	(0.00344)	(0.000137)
Age	-0.0269	-8.28e-05	-0.0381	-0.00380	-0.0163	-0.00561	0.255	0.000619
	(0.0413)	(0.00172)	(0.120)	(0.00841)	(0.0966)	(0.00590)	(0.155)	(0.00991)
LOS	-0.0735	-0.00186	0.0855	0.000260	-0.126	0.00480	-0.317	0.0235
	(0.0874)	(0.00435)	(0.298)	(0.0178)	(0.260)	(0.0148)	(0.568)	(0.0227)
Cases	-0.00319	4.31e-05	-0.00367	3.81e-05	0.00828	-0.000408	0.0130	-8.36e-05
	(0.00322)	(0.000105)	(0.0110)	(0.000450)	(0.00840)	(0.000424)	(0.0160)	(0.000756)
Co-morbidity	0.368	-0.0174	4.469**	-0.0134	2.359	0.0976	4.694	0.0518
	(0.603)	(0.0220)	(2.068)	(0.0914)	(1.692)	(0.0937)	(3.260)	(0.143)
Deprivation	0.0897	-0.00342	-0.663	0.0190	-0.577	-0.0279	-2.909***	-0.0440
	(0.159)	(0.00710)	(0.554)	(0.0361)	(0.411)	(0.0311)	(1.069)	(0.0506)
FT	0.254	0.00260	0.215	-0.0422	0.423	-0.0483	-1.613	-0.141**
	(0.314)	(0.0112)	(0.931)	(0.0384)	(0.913)	(0.0343)	(1.698)	(0.0579)
PbR(06)	-0.563	-0.611***	-3.637***	-2.817***	-2.037	-0.370***	-2.123	-0.00429
	(0.461)	(0.0207)	(1.345)	(0.0914)	(1.416)	(0.0746)	(2.474)	(0.125)
Constant	1.124	0.0432	-2.209	0.361	-1.254	0.337	-23.92**	0.00230
	(2.789)	(0.124)	(8.793)	(0.611)	(6.627)	(0.430)	(11.23)	(0.769)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,080	1,080	1,080	1,080	1,080	1,080	1,080	1,080
R^2	0.047	0.883	0.051	0.912	0.018	0.263	0.028	0.014
Hospitals	120	120	120	120	120	120	120	120

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

The second part of the methodology plots the relative performance of latent and filtered outcome indicators for TIA in order to investigate trends in performance over time. Figure D.11 shows the relative performance of all hospitals on four latent outcome measures, plotted for the years 2002, 2005 and 2008. All of the latent measure plots have very narrow confidence intervals for each of the hospital estimates. In the top left hand panel the plots for 30-day mortality show a gradual move towards the mean from 2002 to 2008. This change from 2002 to 2005 shows most hospitals with below average mortality moving towards the mean while hospitals with above average mortality also exhibit an increase in mortality, however from 2005-2008, all hospitals have moved closer to the mean. In the top

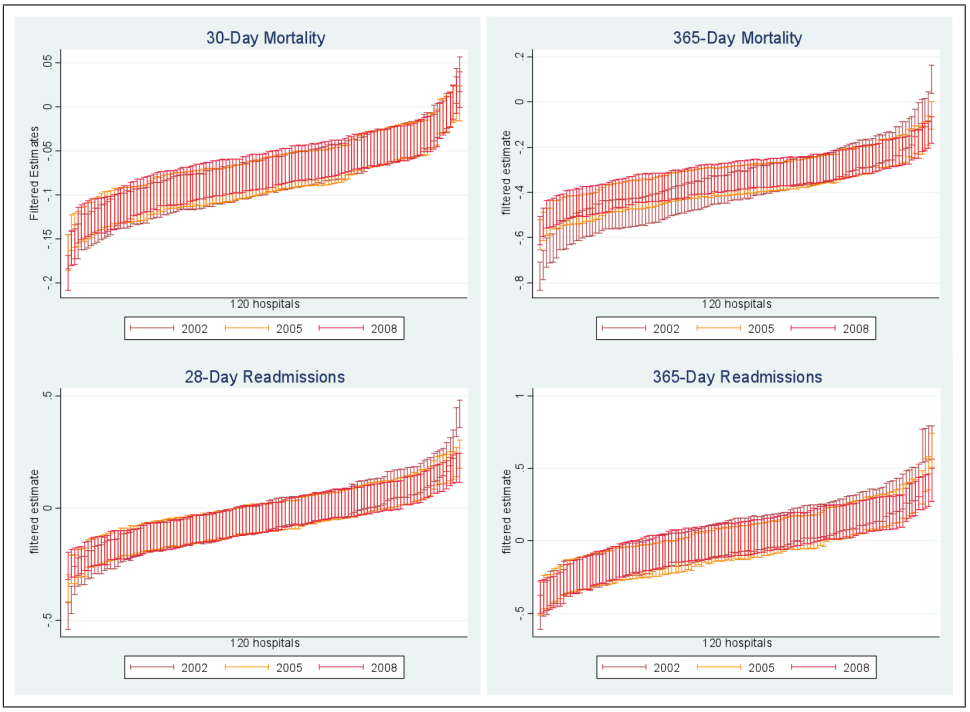
right hand panel, showing the plots for year-long mortality we see a fall in all mortalities from 2002-2005, and then a large shit towards the mean values from 2005-2008. The readmission indicators, in the bottom two panels, do not show any large change between the different years plotted.

Figure D.11: Relative hospital performance over time for TIA (normalized latent outcome indicators).



The filtered indicators are plotted in Figure D.12. The confidence intervals for the filtered estimates are larger than those of the latent measure Moreover the range of the filtered indicators is a lot smaller, and there are fewer outliers than for the latent measures. For most of the indicators it is difficult to observe any obvious change in the indicators from one year to the next. However for the year-long mortality panel, there does seem to be a very gradual change in the relative performance of hospitals from 2002 to 2008, such that they are all moving closer to the mean.

Figure D.12: Relative hospital performance over time for TIA (normalized filtered outcome indicators).



Models 3 and 4 use the normalized latent and filtered outcome indicators plotted above as dependent variables. The results of these models presented in Table 4.4.2, and investigate the effect different explanatory variables have had on the latent and filtered indicators. The R-squared estimates of all models are quite low, around 2%. The PbR dummy is only significant for the latent mortality models, where it is positively associated with the dependent variables. This indicates that since PbR there has been an increase in mortality. We see this effect graphically in Figure D.11. The tariff variable is not significant for any of the models however. Of the hospital characteristics average age is significant in the latent and filtered year-long readmission model, where and increase in the average age of patients is associated with an increase in readmissions. Average LOS is also significant for latent year-long readmissions, although the sign on the coefficient is negative, such that an increase in LOS is associated with a decline in readmissions.

Caseload is significant for latent 30-day mortality, latent year-long mortality and both short and long term latent readmissions. The signs on the coefficients indicate that an increase in activity is associated with a decline in mortality, but with an increase in readmissions. Average co-morbidity is negatively associated with latent year-long mortality, such that an increase in the average co-morbidity of the patients is associated with a decline in mortality. Average deprivation is positively associated with year-long filtered

readmissions, such that hospitals with a higher number of deprived patients have higher readmission rates. Foundation trust status is only significant for short term latent readmissions, at 1%, such that foundation trusts have higher readmission rates than other hospitals.

Table D.8: TIA Models 3 & 4.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.000461	3.45e-05	0.00100	0.000158	-0.000218	4.64e-05	0.00292	9.55e-05
	(0.000319)	(2.45e-05)	(0.00111)	(0.000104)	(0.00101)	(7.76e-05)	(0.00197)	(0.000129)
Age	0.00496	0.00247	-0.0413	0.0122*	0.0127	0.00145	0.327***	0.0225**
	(0.0222)	(0.00151)	(0.0821)	(0.00688)	(0.0725)	(0.00572)	(0.109)	(0.00999)
LOS	-0.0134	-0.00668	0.0218	-0.0309	-0.000263	-0.00818	-0.544*	-0.0187
	(0.0524)	(0.00438)	(0.185)	(0.0201)	(0.153)	(0.0130)	(0.278)	(0.0231)
Cases	-0.00166*	4.90e-05	-0.00531*	0.000117	0.00803*	0.000217	0.0139*	0.000293
	(0.000900)	(9.84e-05)	(0.00299)	(0.000382)	(0.00408)	(0.000301)	(0.00740)	(0.000539)
Co-morbidity	-0.465	0.00943	-2.659**	-0.0762	-0.495	-0.0449	-1.571	-0.146
	(0.366)	(0.0225)	(1.049)	(0.0967)	(1.074)	(0.0678)	(1.694)	(0.113)
Deprivation	0.0696	0.00735	-0.475	0.0340	-0.126	0.0395	-0.964	0.139**
	(0.120)	(0.00912)	(0.381)	(0.0397)	(0.394)	(0.0285)	(1.092)	(0.0676)
FT	-0.0437	0.00727	-0.191	0.0101	0.849*	-0.00725	0.651	0.0216
	(0.140)	(0.00771)	(0.356)	(0.0302)	(0.439)	(0.0221)	(0.755)	(0.0430)
PbR(06)	0.469***	-0.000855	1.877***	0.0158	-1.164*	-0.0112	-1.544	-0.0558
	(0.166)	(0.00802)	(0.548)	(0.0319)	(0.668)	(0.0257)	(1.681)	(0.0472)
Constant	-0.403	-0.283**	4.480	-1.198**	-0.0604	-0.118	-23.08***	-1.446*
	(1.507)	(0.113)	(5.681)	(0.529)	(5.610)	(0.404)	(8.422)	(0.744)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,080	960	1,080	960	1,080	960	1,080	960
$\boldsymbol{R}^2$	0.058	0.023	0.048	0.025	0.024	0.010	0.025	0.026
Hospitals	120	120	120	120	120	120	120	120

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

D.5 Sensitivity Analysis

Model 1&2 Random Effects

Table D.9: AMI Models 1 & 2 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.000327	0.000333	0.00147	0.000158	-0.00258**	-0.000155	-0.00118	-2.14e-06
	(0.000890)	(0.000599)	(0.00197)	(0.000744)	(0.00113)	(0.000188)	(0.00205)	(0.000137)
Age	-0.0555	-0.00456	0.392	-0.0131	-0.509***	0.00597	0.0422	0.0105
	(0.337)	(0.131)	(0.380)	(0.193)	(0.176)	(0.0505)	(0.348)	(0.0372)
LOS	0.259**	0.0537	0.946***	0.138	-0.144	-0.0212	-0.238	-0.0477***
	(0.129)	(0.0782)	(0.159)	(0.0951)	(0.100)	(0.0251)	(0.339)	(0.0172)
Cases	-0.00493	0.00254	-0.00676	0.00466*	0.0103***	-0.00105	0.0110***	-0.000583
	(0.00373)	(0.00180)	(0.00495)	(0.00265)	(0.00278)	(0.000690)	(0.00422)	(0.000619)
Co-morbidity	-8.107***	2.067*	0.771	3.383**	-7.050	-0.787*	0.999	-0.976**
	(2.668)	(1.086)	(6.268)	(1.611)	(4.387)	(0.442)	(2.209)	(0.467)
Deprivation	3.285	0.689**	6.208	1.072**	-2.125	-0.267**	-2.139	-0.223**
	(2.416)	(0.306)	(4.309)	(0.429)	(1.389)	(0.114)	(1.544)	(0.0953)
FT	-2.068	-0.336	-2.155	-0.372	0.849	0.125	1.465	0.162
	(1.444)	(0.690)	(1.941)	(1.012)	(1.271)	(0.256)	(1.855)	(0.194)
Teach	-3.660	-1.101	-13.06	-1.751	4.665	0.523	6.816**	0.377
	(4.682)	(1.110)	(8.786)	(1.649)	(3.268)	(0.523)	(3.363)	(0.477)
PbR(05)	-2.549	-5.673***	-5.400***	-5.307***	4.647***	1.839***	0.710	1.289***
	(2.408)	(0.868)	(1.709)	(1.300)	(1.147)	(0.335)	(2.344)	(0.263)
Constant	13.81	-4.968	-38.66	-7.064	52.05***	1.708	-2.020	1.310
	(25.23)	(9.507)	(35.25)	(13.71)	(18.64)	(3.670)	(25.87)	(2.698)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,071	1,071	1,071	1,071	1,071	1,071	1,071	1,071
Hospitals	119	119	119	119	119	119	119	119

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.10: MI Models 1 & 2 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	-0.00303	0.000232	-0.00454	0.000279	0.000237	-0.000102*	0.000497	-1.84e-05
	(0.00233)	(0.000228)	(0.00363)	(0.000238)	(0.00113)	(5.24e-05)	(0.00167)	(6.49e-05)
Age	-0.525**	-0.00466	-0.0341	-0.00365	0.359	-0.000470	0.487**	-0.00142
	(0.235)	(0.0390)	(0.289)	(0.0416)	(0.231)	(0.0158)	(0.244)	(0.0139)
LOS	0.102	-0.0484**	0.316	-0.0527**	0.152	0.0115**	0.0945	0.00882
	(0.202)	(0.0214)	(0.313)	(0.0246)	(0.147)	(0.00570)	(0.155)	(0.00643)
Cases	-0.0643**	-0.00189	-0.0520**	-0.00214	0.0317**	0.000364	0.0755***	0.000977
	(0.0252)	(0.00263)	(0.0254)	(0.00287)	(0.0150)	(0.000781)	(0.0232)	(0.00118)
Co-morbidity	1.248	0.334	1.881	0.424	-0.865	-0.0272	-0.623	-0.0865
	(3.543)	(0.227)	(3.494)	(0.265)	(2.000)	(0.0726)	(2.777)	(0.0782)
Deprivation	1.709	0.571***	1.800	0.562***	0.0731	-0.145**	-1.413	-0.183***
	(1.307)	(0.166)	(1.158)	(0.179)	(0.647)	(0.0614)	(1.127)	(0.0692)
FT	-5.199	-0.229	-1.350	-0.282	5.191***	0.116	9.685***	0.269**
	(3.523)	(0.238)	(3.849)	(0.252)	(1.934)	(0.0809)	(3.021)	(0.109)
Teach	-6.323	-2.671***	-10.23	-2.689***	-0.802	0.507	4.357	0.538
	(5.003)	(0.972)	(7.857)	(1.019)	(3.515)	(0.399)	(3.853)	(0.371)
PbR(06)	7.048	-0.440	3.925	0.478	-2.688	1.752***	-5.260	3.573***
	(6.147)	(0.611)	(5.784)	(0.660)	(2.336)	(0.176)	(4.332)	(0.195)
Constant	47.92**	-0.192	15.90	-0.510	-27.27*	0.259	-39.61**	0.126
	(21.86)	(2.752)	(21.48)	(2.962)	(14.78)	(1.111)	(18.92)	(1.004)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,017	1,017	1,017	1,017	1,017	1,017	1,017	1,017
Hospitals	113	113	113	113	113	113	113	113

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.11: IHD Models 1 & 2 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	5.26e-06	-2.50e-05	-0.00151***	-0.00021**	0.00142***	9.68e-05	-1.26e-05	7.68e-05
	(0.000277)	(8.49e-05)	(0.000474)	(9.21e-05)	(0.000296)	(0.000113)	(0.000465)	(9.20e-05)
Age	-0.163	-0.179***	-0.358	-0.236***	0.0333	-0.0257	0.00157	0.0459
	(0.182)	(0.0659)	(0.229)	(0.0695)	(0.193)	(0.0404)	(0.362)	(0.0600)
LOS	0.0467	-0.0144	0.680***	0.0178	0.123	-0.0146	0.523	-0.0126
	(0.103)	(0.0371)	(0.181)	(0.0415)	(0.171)	(0.0443)	(0.361)	(0.0433)

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Cases	-0.000632	-0.000252	-0.000429	-0.000308	0.000167	-7.06e-05	9.73e-05	2.10e-05
	(0.000629)	(0.000204)	(0.000712)	(0.000217)	(0.000502)	(0.000132)	(0.000921)	(0.000196)
Co-morbidity	-2.412	0.855	-2.087	1.281	-1.170	-0.153	0.477	-0.406
	(2.492)	(0.792)	(2.638)	(0.814)	(3.691)	(0.379)	(4.392)	(0.688)
Deprivation	0.217	0.0965	0.305	0.0637	-0.905**	-0.266***	-4.107**	-0.265**
	(0.155)	(0.0983)	(0.505)	(0.128)	(0.364)	(0.0957)	(1.774)	(0.105)
FT	-0.663	-0.273	-0.850	-0.249	0.856	0.257*	2.339	0.373*
	(0.511)	(0.202)	(0.684)	(0.217)	(0.976)	(0.133)	(1.514)	(0.194)
Teach	-0.115	0.0205	-0.289	0.110	-0.340	0.214	4.175	0.0798
	(0.805)	(0.394)	(1.464)	(0.482)	(1.543)	(0.536)	(3.525)	(0.552)
PbR(06)	-2.799*	-3.513***	-5.242***	-7.169***	2.312**	2.158***	2.690	2.938***
	(1.506)	(0.478)	(1.730)	(0.515)	(1.103)	(0.274)	(1.990)	(0.434)
Constant	12.30	11.96***	24.78	15.76***	-5.260	1.723	-3.412	-2.983
	(13.28)	(4.629)	(16.33)	(4.837)	(12.55)	(2.790)	(23.86)	(4.243)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,089	1,089	1,089	1,089	1,089	1,089	1,089	1,089
Hospitals	121	121	121	121	121	121	121	121

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.12: CCF Models 1 & 2 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.000688	5.64e-05	-0.000635	0.000117	-0.000133	4.14e-05	-0.000704	8.06e-05
	(0.000939)	(9.51e-05)	(0.00118)	(9.54e-05)	(0.000767)	(3.25e-05)	(0.000983)	(5.60e-05)
Age	-0.0226	0.0182	-0.123	0.0217	0.0442	-0.00529	-0.0765	-0.0213*
	(0.209)	(0.0111)	(0.226)	(0.0142)	(0.179)	(0.00733)	(0.200)	(0.0118)
LOS	0.273	0.00737	0.535**	0.00724	-0.0523	-0.00169	-0.0456	-0.00798
	(0.205)	(0.0158)	(0.242)	(0.0178)	(0.164)	(0.00652)	(0.236)	(0.0113)
Cases	-0.0835	-9.32e-05	-0.0814	0.00130	0.0638*	0.00414**	0.0826	0.00539*
	(0.0536)	(0.00238)	(0.0629)	(0.00313)	(0.0382)	(0.00207)	(0.0574)	(0.00319)
Co-morbidity	0.937	-0.0901	-0.398	0.0409	-0.00934	0.0741	0.303	-0.0396
	(2.100)	(0.108)	(2.152)	(0.130)	(1.824)	(0.0670)	(1.995)	(0.0854)
Deprivation	0.308	-0.0612	-0.655	-0.0715	0.402	0.00554	0.173	-0.00173
	(0.748)	(0.0543)	(1.130)	(0.0683)	(0.679)	(0.0270)	(0.736)	(0.0598)

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
FT	1.602	-0.0682	-0.698	-0.0556	1.749	0.159*	1.245	0.175
	(3.440)	(0.141)	(3.860)	(0.196)	(2.572)	(0.0890)	(4.250)	(0.142)
Teach	-11.97***	-0.570	-2.930	-0.797	-0.956	0.225	3.564	0.609
	(4.027)	(0.474)	(5.420)	(0.536)	(3.227)	(0.364)	(4.086)	(0.482)
PbR(06)	-0.934	-0.749***	-1.982	-3.458***	-1.056	0.105	-3.246	-2.121***
	(3.990)	(0.249)	(4.798)	(0.309)	(2.964)	(0.169)	(4.021)	(0.265)
Constant	-0.515	-1.307*	7.657	-1.858*	-3.602	0.0813	5.052	1.255
	(14.31)	(0.733)	(15.84)	(0.962)	(12.25)	(0.494)	(13.88)	(0.859)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,080	1,080	1,080	1,080	1,080	1,080	1,080	1,080
Hospitals	120	120	120	120	120	120	120	120

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.13: Stroke Models 1 & 2 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.00277	-1.33e-05	0.00117	-7.92e-05	0.000347	3.43e-05	0.000133	5.96e-05
	(0.00260)	(0.000437)	(0.00230)	(0.000648)	(0.000654)	(9.57e-05)	(0.00114)	(0.000227)
Age	0.181	-0.0250	0.00351	-0.0444	-0.148	0.00659	-0.0385	0.00901
	(0.216)	(0.0905)	(0.228)	(0.140)	(0.124)	(0.0218)	(0.192)	(0.0417)
LOS	-0.529**	0.0206	-0.326	0.0202	-0.118*	-0.00163	-0.213	0.0112
	(0.267)	(0.0678)	(0.238)	(0.0979)	(0.0659)	(0.0132)	(0.160)	(0.0319)
Cases	0.00791	0.000692	0.00629	0.000373	0.000384	-0.000154	0.00170	0.000563
	(0.00522)	(0.00140)	(0.00455)	(0.00206)	(0.00138)	(0.000317)	(0.00258)	(0.000740)
Co-morbidity	-12.76***	0.551	-12.17***	1.598	1.738	0.126	3.810*	-0.325
	(3.356)	(0.901)	(3.621)	(1.324)	(1.405)	(0.223)	(2.208)	(0.480)
Deprivation	-0.0642	0.118	-0.326	0.233	0.291	0.0556	1.313**	0.0688
	(0.915)	(0.315)	(0.997)	(0.486)	(0.275)	(0.0711)	(0.545)	(0.149)
FT	0.888	0.740	-0.0523	1.367	-0.469	0.135	1.632	0.246
	(2.178)	(0.590)	(2.452)	(0.902)	(0.746)	(0.133)	(1.444)	(0.258)
Teach	-8.915***	-1.156	-9.269***	-1.153	0.892	-0.0237	0.166	-0.710
	(2.375)	(0.985)	(2.746)	(1.460)	(1.195)	(0.211)	(2.376)	(0.529)
PbR(06)	-1.885	1.580*	-10.10**	-5.096***	-1.944**	0.503**	-2.313	2.208***
	(4.460)	(0.947)	(4.209)	(1.407)	(0.768)	(0.200)	(1.893)	(0.434)

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Constant	5.326	0.408	20.52	0.663	9.235	-0.734	1.091	-0.853
	(19.55)	(6.805)	(20.48)	(10.46)	(9.733)	(1.664)	(14.87)	(3.317)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,035	1,035	1,035	1,035	1,035	1,035	1,035	1,035
Hospitals	115	115	115	115	115	115	115	115

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.14: TIA Models 1 & 2 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.000693	-1.09e-06	-0.00151	-4.44e-05	-0.000440	-5.50e-05	-0.000796	-0.000172
	(0.000555)	(2.44e-05)	(0.00171)	(0.000104)	(0.00169)	(9.22e-05)	(0.00368)	(0.000139)
Age	-0.0238	-5.85e-05	-0.0311	-0.00254	-0.0449	-0.00474	0.199	0.00196
	(0.0385)	(0.00165)	(0.111)	(0.00805)	(0.0938)	(0.00574)	(0.150)	(0.00961)
LOS	-0.0678	-0.00179	0.259	0.000371	-0.0731	0.00734	-0.0442	0.0235
	(0.0829)	(0.00426)	(0.277)	(0.0174)	(0.251)	(0.0148)	(0.583)	(0.0226)
Cases	-0.00167	8.92e-05	-0.00376	0.000160	0.00710	-0.000218	0.0168	-4.77e-06
	(0.00284)	(0.000111)	(0.00959)	(0.000443)	(0.00661)	(0.000408)	(0.0120)	(0.000733)
Co-morbidity	0.114	-0.0180	2.442	-0.0301	1.976	0.0724	4.126	0.0179
	(0.552)	(0.0212)	(1.814)	(0.0877)	(1.453)	(0.0861)	(2.985)	(0.138)
Deprivation	0.148	-0.00450	0.0142	-0.00600	-0.0458	-0.0390	-1.091*	-0.0745*
	(0.108)	(0.00632)	(0.416)	(0.0285)	(0.303)	(0.0255)	(0.599)	(0.0439)
FT	0.249	0.00247	0.202	-0.0434	0.284	-0.0459	-1.873	-0.140**
	(0.331)	(0.0118)	(1.022)	(0.0410)	(0.944)	(0.0361)	(1.767)	(0.0616)
Teach	-0.312	-0.00324	-1.268	-0.0750	0.349	-0.0101	0.475	-0.0256
	(0.476)	(0.0364)	(1.856)	(0.167)	(1.041)	(0.131)	(2.665)	(0.236)
PbR(06)	-0.574	-0.614***	-2.950**	-2.825***	-1.661	-0.378***	-1.474	-0.0106
	(0.457)	(0.0212)	(1.412)	(0.0941)	(1.370)	(0.0760)	(2.372)	(0.127)
Constant	1.080	0.0369	0.0900	0.295	1.226	0.302	-19.66*	-0.0445
	(2.653)	(0.120)	(8.150)	(0.589)	(6.310)	(0.414)	(10.84)	(0.761)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,080	1,080	1,080	1,080	1,080	1,080	1,080	1,080
Hospitals	120	120	120	120	120	120	120	120

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.15: Hip Replacement Models 1 & 2 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.000239	1.78e-07	-0.000822	-2.69e-05	-2.89e-05	2.47e-05	-0.000501	6.75e-05
	(0.000278)	(9.23e-06)	(0.000599)	(2.18e-05)	(0.00104)	(3.41e-05)	(0.00149)	(4.73e-05)
Age	0.0375	0.00280**	0.0693	0.000793	0.199	-0.00557	0.0333	-0.0116
	(0.0551)	(0.00141)	(0.165)	(0.00349)	(0.280)	(0.00600)	(0.298)	(0.00843)
LOS	-0.104*	0.00253	-0.0597	0.00510	0.0690	-0.00909	0.281	-0.0140
	(0.0579)	(0.00185)	(0.143)	(0.00412)	(0.256)	(0.00737)	(0.375)	(0.0104)
Cases	-0.000831	6.65e-05**	-0.00396**	7.42e-05	-0.000469	-0.000157	-0.00156	-3.72e-05
	(0.000650)	(3.22e-05)	(0.00164)	(9.59e-05)	(0.00250)	(0.000154)	(0.00392)	(0.000204)
Co-morbidity	1.184	0.0560*	3.277**	0.151*	0.964	-0.0986	0.890	-0.203
	(0.735)	(0.0326)	(1.442)	(0.0795)	(3.226)	(0.129)	(4.700)	(0.158)
Deprivation	-0.517	0.0190	-0.884	0.0432*	-0.502	-0.0106	-1.132	-0.0531
	(0.321)	(0.0121)	(1.005)	(0.0244)	(0.719)	(0.0463)	(1.016)	(0.0554)
FT	-0.113	-0.00209	-0.188	0.0352*	1.678*	0.0668**	2.245	0.0807**
	(0.167)	(0.00867)	(0.345)	(0.0206)	(0.994)	(0.0319)	(1.450)	(0.0411)
Teach	3.000	0.160	11.52	0.126	3.766	-0.0840	3.123	-0.121
	(1.909)	(0.334)	(7.217)	(0.865)	(5.319)	(0.617)	(5.288)	(1.002)
PbR(06)	-0.226*	0.0153*	-0.0941	0.0374**	-0.560	-0.159***	-0.486	-0.144***
	(0.136)	(0.00822)	(0.267)	(0.0177)	(0.748)	(0.0303)	(1.033)	(0.0401)
Constant	-3.605	-0.279**	-0.770	0.0269	-16.02	0.0853	-3.444	0.399
	(4.741)	(0.123)	(8.896)	(0.312)	(18.93)	(0.507)	(20.57)	(0.692)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	826	826	826	826	826	826	826	826
Hospitals	118	118	118	118	118	118	118	118

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Models 3&4 Random Effects

Table D.16: AMI Models 3 & 4 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	-0.000782*	2.63e-05	-0.00141*	-0.000251	-0.000381	5.49e-05	0.000778	0.000191
	(0.000465)	(8.26e-05)	(0.000849)	(0.000263)	(0.000282)	(7.30e-05)	(0.00149)	(0.000200)

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Age	0.160	0.0315	0.185	0.0378	-0.104	-0.0177	0.386*	-0.0157
	(0.115)	(0.0207)	(0.113)	(0.0367)	(0.146)	(0.0109)	(0.200)	(0.0113)
LOS	0.0354	-0.0142	0.704***	0.0214	-0.0396	-0.00615	-0.331	-0.0346
	(0.0358)	(0.0105)	(0.185)	(0.0349)	(0.0420)	(0.00838)	(0.232)	(0.0250)
Cases	-0.0035***	-9.58e-05	-0.0034***	-0.000632	0.00311***	6.36e-05	0.00408***	0.000207
	(0.000818)	(0.000387)	(0.00124)	(0.000610)	(0.000911)	(0.000153)	(0.00100)	(0.000141)
Co-morbidity	2.498	-0.398**	-2.844	-0.732**	-2.319**	0.157*	-2.652	0.162
	(3.137)	(0.159)	(2.796)	(0.297)	(0.908)	(0.0821)	(1.651)	(0.113)
Deprivation	0.868	-0.00961	0.670*	-0.00401	-0.117	-0.0156	0.235	-0.0476
	(0.568)	(0.0625)	(0.370)	(0.101)	(0.138)	(0.0320)	(0.145)	(0.0328)
FT	-1.438*	0.0141	-1.255**	0.0422	0.234	-0.0358	0.388	-0.0345
	(0.871)	(0.106)	(0.570)	(0.178)	(0.323)	(0.0506)	(0.483)	(0.0456)
Teach	-0.285	0.0519	-1.270	0.148	0.638*	-0.0134	1.423**	-0.0260
	(0.663)	(0.230)	(0.833)	(0.402)	(0.368)	(0.122)	(0.694)	(0.146)
PbR(05)	1.711***	0.415***	3.930***	0.741***	0.0584	-0.175***	-0.934	-0.220***
	(0.578)	(0.0890)	(0.786)	(0.140)	(0.511)	(0.0381)	(0.779)	(0.0596)
Constant	-12.79	-2.122	-11.95**	-1.118	11.98	1.052	-22.70	0.586
	(10.40)	(1.454)	(6.039)	(2.623)	(10.41)	(0.748)	(14.31)	(0.885)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,071	952	1,071	952	1,071	952	1,071	952
Hospitals	119	119	119	119	119	119	119	119

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.17: MI Models 3 & 4 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	-0.00312**	-1.57e-05	-0.00313**	3.43e-06	-0.000414	-1.32e-05	2.59e-05	-1.53e-05
	(0.00150)	(0.000163)	(0.00140)	(0.000172)	(0.000521)	(5.17e-05)	(0.00132)	(4.78e-05)
Age	0.162	0.00673	0.297	0.0165	0.164	-0.00312	0.239	0.00589
	(0.142)	(0.0175)	(0.223)	(0.0200)	(0.153)	(0.00692)	(0.193)	(0.00646)
LOS	-0.114	0.0227	-0.0597	0.0233	0.232	-0.0122***	0.162*	-0.0163***
	(0.107)	(0.0170)	(0.137)	(0.0198)	(0.141)	(0.00444)	(0.0942)	(0.00531)
Cases	-0.0299***	-0.00181**	-0.0236***	-0.00188**	0.0146***	0.000578*	0.0261***	0.000851**
	(0.00728)	(0.000758)	(0.00850)	(0.000800)	(0.00479)	(0.000332)	(0.00614)	(0.000332)

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Co-morbidity	1.834	0.0268	1.384	-0.0494	-2.408	0.0830	-0.582	0.00679
	(1.386)	(0.179)	(2.584)	(0.196)	(1.525)	(0.0832)	(2.395)	(0.0712)
Deprivation	0.975*	0.00495	0.725*	0.00548	0.190	0.0119	-0.244	-0.00215
	(0.535)	(0.0527)	(0.416)	(0.0568)	(0.254)	(0.0161)	(0.460)	(0.0190)
FT	0.973	0.136*	-0.0984	0.162*	1.037	-0.0182	1.217	-0.00390
	(0.997)	(0.0817)	(1.096)	(0.0885)	(0.746)	(0.0282)	(1.167)	(0.0373)
Teach	0.174	-0.312	-0.0920	-0.293	-0.116	0.0720	1.187	0.140**
	(1.669)	(0.201)	(1.531)	(0.210)	(0.722)	(0.0628)	(1.385)	(0.0702)
PbR(06)	1.250	-0.0494	0.654	-0.0272	-0.275	-0.00660	-1.756	-0.00182
	(1.496)	(0.0708)	(1.958)	(0.0780)	(1.041)	(0.0311)	(1.608)	(0.0415)
Constant	-1.792	-0.603	-10.78	-1.130	-8.848	0.406	-18.86	0.156
	(10.75)	(1.302)	(13.38)	(1.442)	(9.643)	(0.499)	(13.83)	(0.491)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,017	904	1,017	904	1,017	904	1,017	904
Hospitals	113	113	113	113	113	113	113	113

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.18: IHD Models 3 & 4 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	-2.58e-05	8.56e-06	-1.82e-05	1.41e-06	0.000143	-1.78e-05	- 0.000819**	-1.92e-05
	(6.07e-05)	(4.04e-05)	(6.45e-05)	(4.77e-05)	(0.000122)	(3.90e-05)	(0.000391)	(4.42e-05)
Age	0.130**	0.0399*	0.281***	0.0739***	0.0980	0.0387	0.192	0.0301
	(0.0651)	(0.0234)	(0.0748)	(0.0243)	(0.125)	(0.0245)	(0.152)	(0.0275)
LOS	0.0282	0.0152	0.227***	0.0173	-0.0782	-0.0391*	-0.000417	-0.0400
	(0.0442)	(0.0150)	(0.0553)	(0.0179)	(0.0639)	(0.0211)	(0.135)	(0.0243)
Cases	6.74e-05	1.19e-05	0.000154*	1.97e-05	0.000282***	2.34e-05	0.000830***	2.72e-05
	(7.54e-05)	(3.28e-05)	(8.53e-05)	(3.62e-05)	(0.000104)	(2.90e-05)	(0.000206)	(3.59e-05)
Co-morbidity	-1.223***	-0.445**	-4.251***	-0.766***	-0.531	0.173	0.547	0.264
	(0.465)	(0.210)	(0.714)	(0.215)	(0.659)	(0.180)	(2.134)	(0.235)
Deprivation	-0.0314	0.0403*	0.111	0.0783**	-0.133	0.0443	-0.484	0.0383
	(0.0913)	(0.0229)	(0.149)	(0.0313)	(0.300)	(0.0301)	(0.678)	(0.0320)
FT	0.0302	0.0893	-0.0208	0.0876	0.0502	-0.0838*	-0.0614	-0.106

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
	(0.122)	(0.0556)	(0.194)	(0.0658)	(0.245)	(0.0491)	(0.410)	(0.0644)
Teach	0.265	0.0547	0.145	0.0876	0.123	0.0722	1.140	0.0213
	(0.240)	(0.0904)	(0.250)	(0.104)	(0.393)	(0.0842)	(0.950)	(0.0942)
PbR(06)	1.341***	0.0692	4.175***	0.124**	-0.856***	-0.0776**	0.334	-0.0842
	(0.271)	(0.0451)	(0.318)	(0.0508)	(0.329)	(0.0359)	(0.827)	(0.0519)
Constant	-9.545**	-2.926*	-21.47***	-5.367***	-5.725	-2.272	-12.93	-1.613
	(4.297)	(1.641)	(5.024)	(1.694)	(8.418)	(1.608)	(9.764)	(1.848)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,089	968	1,089	968	1,089	968	1,089	968
Hospitals	121	121	121	121	121	121	121	121

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.19: CCF Models 3 & 4 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	-5.69e-05	-5.28e-05**	-0.00092**	-6.58e-05**	-4.14e-05	1.37e-05	-0.000512	1.15e-05
	(0.000412)	(2.18e-05)	(0.000414)	(3.33e-05)	(0.000301)	(2.06e-05)	(0.000414)	(3.21e-05)
Age	-0.113	-0.000390	-0.0969	0.00646	-0.0730	-0.000165	0.00251	-0.00814
	(0.124)	(0.00961)	(0.118)	(0.0120)	(0.0903)	(0.00688)	(0.118)	(0.00998)
LOS	-0.115	0.00572	0.248*	0.00493	-0.0685	-0.0112**	-0.257	-0.0183**
	(0.153)	(0.00937)	(0.131)	(0.0119)	(0.124)	(0.00554)	(0.188)	(0.00840)
Cases	-0.00846	0.000862	-0.00566	0.000740	0.0290***	-0.000494	0.0339**	-0.000812
	(0.0149)	(0.00104)	(0.0161)	(0.00119)	(0.0109)	(0.000734)	(0.0167)	(0.00109)
Co-morbidity	2.856**	-0.0218	1.584	-0.0772	-1.722*	0.00643	-1.418	-0.0470
	(1.195)	(0.0814)	(1.191)	(0.105)	(0.886)	(0.0457)	(1.194)	(0.0761)
Deprivation	0.117	-0.0226	-0.567	-0.0152	0.214	0.00467	0.159	0.0185
	(0.323)	(0.0212)	(0.403)	(0.0260)	(0.227)	(0.0164)	(0.358)	(0.0246)
FT	1.387	0.0861*	1.381	0.0885*	-1.032	-0.0287	-0.842	-0.0371
	(0.968)	(0.0488)	(0.960)	(0.0528)	(0.831)	(0.0372)	(1.402)	(0.0550)
Teach	-3.364***	-0.0948	-0.863	-0.0951	-0.601	0.0797	1.229	0.136
	(1.026)	(0.0842)	(1.097)	(0.102)	(0.735)	(0.0636)	(1.027)	(0.0856)
PbR(06)	-0.290	-0.0596	-0.255	-0.0697	-0.162	-0.101***	-0.0965	-0.109**
	(1.619)	(0.0453)	(1.918)	(0.0505)	(1.497)	(0.0371)	(2.196)	(0.0484)

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Constant	6.045	0.0577	4.616	-0.669	7.827	0.139	4.322	0.629
	(9.192)	(0.701)	(8.665)	(0.885)	(6.449)	(0.522)	(9.027)	(0.782)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,080	960	1,080	960	1,080	960	1,080	960
Hospitals	120	120	120	120	120	120	120	120

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.20: Stroke Models 3 & 4 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.000827**	-0.000174	0.00110**	-7.36e-05	-0.000251	2.96e-05	-0.000690**	-6.01e-05
	(0.000352)	(0.000140)	(0.000504)	(0.000161)	(0.000279)	(2.77e-05)	(0.000308)	(8.69e-05)
Age	-0.0122	0.0161	0.116	0.0461	-0.0708	0.00533	0.0694	-0.00886
	(0.0602)	(0.0289)	(0.105)	(0.0391)	(0.0708)	(0.00591)	(0.0770)	(0.0136)
LOS	-0.188***	0.00573	-0.169**	0.00333	0.000942	-0.00121	0.0423	-0.00256
	(0.0512)	(0.0153)	(0.0776)	(0.0207)	(0.0259)	(0.00328)	(0.0475)	(0.00668)
Cases	-5.90e-05	0.000244	-0.000110	0.000423	0.000128	1.90e-05	0.00131**	6.18e-05
	(0.000927)	(0.000269)	(0.000960)	(0.000379)	(0.000361)	(5.46e-05)	(0.000655)	(0.000128)
Co-morbidity	-1.056	0.0418	-3.500***	-0.364	-0.00933	-0.156*	-0.801	0.143
	(0.687)	(0.287)	(0.680)	(0.378)	(0.342)	(0.0803)	(0.745)	(0.190)
Deprivation	-0.131	0.0826	0.0980	0.141	0.159*	0.0228	0.568***	0.0168
	(0.174)	(0.0865)	(0.294)	(0.127)	(0.0813)	(0.0181)	(0.174)	(0.0352)
FT	0.827**	-0.0746	0.702*	-0.246	-0.416*	-0.0537**	-0.186	-0.00652
	(0.360)	(0.124)	(0.413)	(0.182)	(0.223)	(0.0259)	(0.433)	(0.0648)
Teach	-2.150***	-0.157	-1.949***	-0.0759	-0.0672	-0.0213	0.135	-0.189*
	(0.646)	(0.212)	(0.691)	(0.321)	(0.416)	(0.0488)	(0.598)	(0.106)
PbR(06)	-4.217***	-0.393***	-2.279***	-0.436***	0.506*	-0.0453*	0.465	-0.190***
	(0.545)	(0.0818)	(0.646)	(0.111)	(0.285)	(0.0258)	(0.609)	(0.0524)
Constant	6.105	-0.721	-3.086	-3.685	6.162	-0.0615	-2.530	1.115
	(4.965)	(2.103)	(8.000)	(2.805)	(5.641)	(0.465)	(6.762)	(1.080)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,035	920	1,035	920	1,035	920	1,035	920
Hospitals	115	115	115	115	115	115	115	115

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.21: TIA Models 3 & 4 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	0.000390	2.83e-05	-0.000192	0.000136	-0.000469	-3.74e-05	-0.000557	3.78e-05
	(0.000316)	(1.94e-05)	(0.000950)	(8.92e-05)	(0.000901)	(5.47e-05)	(0.00153)	(0.000104)
Age	0.00798	0.00221*	-0.000455	0.0126**	-0.00565	0.00237	0.222**	0.0204***
	(0.0175)	(0.00122)	(0.0618)	(0.00501)	(0.0507)	(0.00414)	(0.0892)	(0.00752)
LOS	-0.0102	-0.00511	0.187	-0.0273*	0.0418	0.00486	0.175	-0.0120
	(0.0512)	(0.00327)	(0.152)	(0.0163)	(0.135)	(0.00942)	(0.218)	(0.0187)
Cases	9.18e-06	3.91e-05	3.46e-05	0.000174*	0.00237*	0.000249***	0.00729***	0.000277*
	(0.000363)	(2.56e-05)	(0.00129)	(9.72e-05)	(0.00144)	(7.52e-05)	(0.00234)	(0.000148)
Co-morbidity	-0.598**	-0.00206	-3.777***	-0.139*	-0.319	-0.124**	-2.126	-0.242***
	(0.254)	(0.0187)	(0.768)	(0.0791)	(0.788)	(0.0568)	(1.443)	(0.0908)
Deprivation	0.0514	0.000221	0.0190	0.00116	0.0302	-0.000919	0.226	0.0104
	(0.0346)	(0.00206)	(0.109)	(0.0106)	(0.104)	(0.00606)	(0.169)	(0.0128)
FT	-0.0105	0.00236	-0.0621	0.00328	0.313	-0.0106	-0.246	0.0173
	(0.107)	(0.00417)	(0.215)	(0.0163)	(0.386)	(0.0124)	(0.593)	(0.0235)
Teach	-0.0934	0.00484	-0.367	0.0156	0.238	0.0147	0.280	-0.00245
	(0.0940)	(0.00700)	(0.338)	(0.0350)	(0.296)	(0.0198)	(0.675)	(0.0404)
PbR(06)	0.355***	-0.000498	1.771***	0.00693	-0.555	0.0124	0.366	0.0192
	(0.129)	(0.00542)	(0.435)	(0.0174)	(0.586)	(0.0159)	(1.470)	(0.0269)
Constant	-0.464	-0.251***	3.466	-1.130***	1.544	-0.0447	-13.74**	-1.170**
	(1.201)	(0.0864)	(4.275)	(0.380)	(3.819)	(0.308)	(6.732)	(0.565)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,080	960	1,080	960	1,080	960	1,080	960
Hospitals	120	120	120	120	120	120	120	120

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.22: Hip Replacement Models 3 & 4 Random Effects.

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
Tariff	-0.000156	5.82e-06*	-0.000347	1.16e-05	0.000709*	-1.16e-05	0.00212***	-2.11e-05
	(0.000118)	(3.36e-06)	(0.000350)	(7.68e-06)	(0.000428)	(1.11e-05)	(0.000819)	(1.61e-05)
Age	0.0585*	0.000290	0.157**	0.00230	0.227*	0.00277	0.256	-0.00293
	(0.0310)	(0.000548)	(0.0714)	(0.00150)	(0.117)	(0.00222)	(0.160)	(0.00253)

	Latent $D30_{ht}$	Filtered $D30_{ht}$	Latent $D365_{ht}$	Filtered $D365_{ht}$	Latent $R28_{ht}$	Filtered $R28_{ht}$	Latent $R365_{ht}$	Filtered $R365_{ht}$
LOS	0.0468*	-0.000273	0.125**	-0.00137	-0.0225	-0.00437**	-0.132	-0.00201
	(0.0246)	(0.000686)	(0.0584)	(0.00181)	(0.141)	(0.00215)	(0.141)	(0.00304)
Cases	0.000121	-2.44e-06	-8.13e-08	-9.13e-06	-0.000419	-1.89e-05	-0.00114	-1.34e-05
	(0.000127)	(5.64e-06)	(0.000291)	(1.49e-05)	(0.000655)	(2.13e-05)	(0.000901)	(3.01e-05)
Co-morbidity	0.141	0.0179	1.229	0.0152	-0.521	0.0276	-2.091	0.0316
	(0.593)	(0.0119)	(2.225)	(0.0320)	(2.515)	(0.0433)	(4.288)	(0.0578)
Deprivation	0.0807***	-0.00107	0.0509	-0.000721	-0.133	0.00896**	-0.408*	0.00478
	(0.0289)	(0.00140)	(0.117)	(0.00259)	(0.158)	(0.00429)	(0.246)	(0.00598)
FT	-0.0871	-0.00173	-0.178	0.00299	0.0384	0.0164	0.428	0.0385**
	(0.0937)	(0.00352)	(0.194)	(0.00875)	(0.364)	(0.0122)	(0.424)	(0.0162)
Teach	0.155	-0.00441	0.790*	-0.00893	1.105*	-0.00270	1.631**	-0.0111
	(0.200)	(0.00415)	(0.421)	(0.00828)	(0.589)	(0.0149)	(0.826)	(0.0189)
PbR(06)	0.122	0.00709	0.251	-0.00368	-0.0382	-0.0294*	-0.394	-0.0295
	(0.0881)	(0.00466)	(0.178)	(0.0126)	(0.700)	(0.0156)	(0.667)	(0.0198)
Constant	-3.714	-0.0553	-10.68**	-0.196*	-19.96**	-0.0956	-28.93**	0.334*
	(2.356)	(0.0398)	(4.454)	(0.110)	(9.678)	(0.176)	(13.83)	(0.197)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	826	826	826	826	826	826	826	826
Hospitals	118	118	118	118	118	118	118	118

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Model 2 with Interaction Dummy Variables

Table D.23: AMI Model 2 with interactions.

	Filtered $D30_{ht}$	Filtered $D365_{ht}$	Filtered $R28_{ht}$	Filtered $R365_{ht}$
Tariff	0.000212	-4.34e-05	-9.16e-05	6.58e-05
	(0.000631)	(0.000753)	(0.000196)	(0.000123)
Age	-0.0298	-0.0650	0.0210	0.0312
	(0.130)	(0.190)	(0.0479)	(0.0348)
LOS	0.0829	0.184*	-0.0323	-0.0564***
	(0.0854)	(0.100)	(0.0272)	(0.0167)
Cases	0.00384*	0.00631*	-0.00155*	-0.00103
	(0.00219)	(0.00323)	(0.000816)	(0.000677)
Co-morbidity	3.710**	6.157**	-1.206*	-1.278**

	Filtered $D30_{ht}$	Filtered $D365_{ht}$	Filtered $R28_{ht}$	Filtered $R365_{ht}$
	(1.796)	(2.580)	(0.629)	(0.549)
Deprivation	0.418	0.660	-0.132	-0.124
	(0.304)	(0.463)	(0.117)	(0.111)
FT	-0.651	-0.857	0.246	0.245
	(0.589)	(0.866)	(0.219)	(0.168)
Deprivation* PbR	0.626*	0.951*	-0.242*	-0.176
	(0.372)	(0.532)	(0.139)	(0.110)
PbR(05)	-2.481	-4.336	0.755	0.641
	(2.552)	(3.489)	(0.858)	(0.529)
co-morbidity*PbR	-1.642	1.750	0.567	0.188
	(4.129)	(5.618)	(1.385)	(0.830)
Constant	-6.048	-8.055	1.403	0.361
	(9.050)	(13.12)	(3.400)	(2.483)
Year Dummies	Yes	Yes	Yes	Yes
N	1,071	1,071	1,071	1,071
R^2	0.326	0.194	0.258	0.209
Hospitals	119	119	119	119

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.24: MI Model 2 with interactions.

	Filtered $D30_{ht}$	Filtered $D365_{ht}$	Filtered $R28_{ht}$	Filtered $R365_{ht}$
Tariff	0.000240	0.000284	-0.000100*	-1.30e-05
	(0.000227)	(0.000236)	(5.11e-05)	(6.21e-05)
Age	-0.00874	-0.00769	0.000860	-0.00119
	(0.0377)	(0.0401)	(0.0156)	(0.0136)
LOS	-0.0507**	-0.0543**	0.0120**	0.00865
	(0.0220)	(0.0252)	(0.00556)	(0.00624)
Cases	-0.000684	-0.000986	-4.41e-05	0.000188
	(0.00248)	(0.00270)	(0.000722)	(0.00114)
Co-morbidity	0.276	0.341	-0.0233	-0.0602
	(0.303)	(0.356)	(0.0953)	(0.104)
Deprivation	0.595***	0.581***	-0.159**	-0.193**
	(0.177)	(0.192)	(0.0684)	(0.0777)
FT	-0.194	-0.251	0.107	0.264**
	(0.247)	(0.260)	(0.0815)	(0.106)

	Filtered $D30_{ht}$	Filtered $D365_{ht}$	Filtered $R28_{ht}$	Filtered $R365_{ht}$
Deprivation* PbR	-0.0288	-0.0198	0.00727	-0.0185
	(0.113)	(0.123)	(0.0356)	(0.0460)
PbR(06)	0.0173	0.103	0.0394	-0.00638
	(0.447)	(0.507)	(0.129)	(0.158)
co-morbidity*PbR	-0.487	0.271	1.679***	3.599***
	(1.015)	(1.134)	(0.282)	(0.345)
Constant	-0.268	-0.543	0.243	0.170
	(2.765)	(2.984)	(1.142)	(1.009)
Year Dummies	Yes	Yes	Yes	Yes
N	1,017	1,017	1,017	1,017
R^2	0.059	0.059	0.529	0.798
Hospitals	113	113	113	113

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.25: IHD Model 2 with interactions.

	Filtered $D30_{ht}$	Filtered $D365_{ht}$	Filtered $R28_{ht}$	Filtered $R365_{ht}$
Tariff	-6.28e-05	-0.000255***	0.000143	0.000141*
	(8.63e-05)	(8.79e-05)	(0.000104)	(8.51e-05)
Age	-0.197***	-0.269***	-0.0240	0.0491
	(0.0717)	(0.0724)	(0.0450)	(0.0683)
LOS	-0.0305	-0.00275	-0.00805	0.000539
	(0.0352)	(0.0393)	(0.0456)	(0.0427)
Cases	-0.000211	-0.000273	-8.36e-05	-1.15e-05
	(0.000228)	(0.000239)	(0.000135)	(0.000216)
Co-morbidity	1.970***	2.663***	-0.440	-1.069*
	(0.641)	(0.654)	(0.419)	(0.634)
Deprivation	0.113	0.0301	-0.460**	-0.453
	(0.215)	(0.203)	(0.221)	(0.276)
FT	-0.306*	-0.289	0.243**	0.373**
	(0.184)	(0.195)	(0.117)	(0.179)
Deprivation* PbR	0.120*	0.141*	-0.00934	-0.0607
	(0.0650)	(0.0733)	(0.0553)	(0.0690)
PbR(06)	-2.595***	-3.087***	0.767	1.673**
	(0.883)	(0.956)	(0.482)	(0.814)
co-morbidity*PbR	-1.656**	-4.512***	1.455***	1.608**

	Filtered $D30_{ht}$	Filtered $D365_{ht}$	Filtered $R28_{ht}$	Filtered $R365_{ht}$
	(0.757)	(0.793)	(0.361)	(0.672)
Constant	12.78***	17.57***	1.680	-3.036
	(4.825)	(4.844)	(3.001)	(4.607)
Year Dummies	Yes	Yes	Yes	Yes
N	1,089	1,089	1,089	1,089
R^2	0.565	0.802	0.532	0.521
Hospitals	121	121	121	121

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.26: CCF Model 2 with interactions.

	Filtered $D30_{ht}$	Filtered $D365_{ht}$	Filtered $R28_{ht}$	Filtered $R365_{ht}$
Tariff	7.41e-05	0.000140	4.01e-05	8.17e-05
	(9.33e-05)	(9.39e-05)	(3.08e-05)	(5.84e-05)
Age	0.0200*	0.0236	-0.00641	-0.0209*
	(0.0113)	(0.0143)	(0.00760)	(0.0120)
LOS	0.00539	0.00494	-0.000768	-0.00707
	(0.0157)	(0.0177)	(0.00635)	(0.0114)
Cases	-0.000433	0.000760	0.00444**	0.00501
	(0.00233)	(0.00298)	(0.00204)	(0.00314)
Co-morbidity	-0.162	-0.0569	0.131	0.0296
	(0.132)	(0.149)	(0.0905)	(0.110)
Deprivation	-0.0642	-0.0791	0.00408	-0.0304
	(0.0623)	(0.0757)	(0.0310)	(0.0580)
FT	-0.0684	-0.0597	0.151*	0.133
	(0.133)	(0.190)	(0.0865)	(0.135)
Deprivation* PbR	0.0108	0.00882	-0.00506	0.0728*
	(0.0510)	(0.0631)	(0.0313)	(0.0434)
PbR(06)	0.180	0.258	-0.131	-0.149
	(0.138)	(0.172)	(0.104)	(0.140)
co-morbidity*PbR	-0.997***	-3.810***	0.286	-1.917***
	(0.289)	(0.350)	(0.227)	(0.354)
Constant	-1.462*	-2.033**	0.119	1.249
	(0.767)	(0.969)	(0.521)	(0.836)
Year Dummies	Yes	Yes	Yes	Yes
N	1,080	1,080	1,080	1,080

	Filtered $D30_{ht}$	Filtered $D365_{ht}$	Filtered $R28_{ht}$	Filtered $R365_{ht}$
R^2	0.079	0.508	0.038	0.390
Hospitals	120	120	120	120

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.27: Stroke Model 2 with interactions.

	Filtered $D30_{ht}$	Filtered $D365_{ht}$	Filtered $R28_{ht}$	Filtered $R365_{ht}$
Tariff	-1.76e-05	-6.65e-05	3.01e-05	5.61e-05
	(0.000454)	(0.000670)	(9.85e-05)	(0.000230)
Age	-0.0200	-0.0401	-9.97e-05	0.0105
	(0.0948)	(0.147)	(0.0240)	(0.0441)
LOS	0.0396	0.0390	0.00329	0.0260
	(0.0724)	(0.105)	(0.0142)	(0.0334)
Cases	0.000142	-0.000289	-0.000387	0.000210
	(0.00167)	(0.00258)	(0.000407)	(0.000831)
Co-morbidity	1.155	2.534*	0.310	0.00294
	(0.952)	(1.424)	(0.239)	(0.496)
Deprivation	-0.160	-0.220	0.0410	0.0171
	(0.560)	(0.874)	(0.128)	(0.236)
FT	0.536	1.052	0.0964	0.145
	(0.541)	(0.830)	(0.124)	(0.233)
Deprivation* PbR	0.162	0.238	0.0116	0.0914
	(0.242)	(0.370)	(0.0562)	(0.106)
PbR(06)	-2.953*	-4.241*	-0.742*	-1.914**
	(1.559)	(2.345)	(0.391)	(0.738)
co-morbidity*PbR	6.677**	2.162	1.797**	5.532***
	(2.718)	(4.250)	(0.690)	(1.204)
Constant	-1.329	-1.547	-0.547	-1.820
	(7.409)	(11.43)	(1.869)	(3.619)
Year Dummies	Yes	Yes	Yes	Yes
N	1,035	1,035	1,035	1,035
R^2	0.136	0.230	0.418	0.442
Hospitals	115	115	115	115

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.28: TIA Model 2 with interactions.

	Filtered $D30_{ht}$	Filtered $D365_{ht}$	Filtered $R28_{ht}$	Filtered $R365_{ht}$
Tariff	-2.20e-06	-4.29e-05	-3.85e-05	-0.000166
	(2.43e-05)	(0.000104)	(9.04e-05)	(0.000137)
Age	-7.54e-05	-0.00384	-0.00567	0.000574
	(0.00174)	(0.00826)	(0.00576)	(0.00979)
LOS	-0.00175	-0.000480	0.00371	0.0224
	(0.00432)	(0.0178)	(0.0148)	(0.0226)
Cases	5.23e-05	-2.36e-05	-0.000499	-0.000181
	(0.000109)	(0.000441)	(0.000403)	(0.000740)
Co-morbidity	-0.0186	0.00430	0.125	0.0953
	(0.0280)	(0.109)	(0.112)	(0.163)
Deprivation	-0.00456	0.0257	-0.0181	-0.0352
	(0.00735)	(0.0373)	(0.0306)	(0.0499)
FT	0.000714	-0.0318	-0.0331	-0.128**
	(0.0119)	(0.0375)	(0.0315)	(0.0554)
Deprivation* PbR	0.00402	-0.0227	-0.0330**	-0.0287
	(0.00553)	(0.0189)	(0.0131)	(0.0205)
PbR(06)	-0.000402	-0.0390	-0.0633	-0.130
	(0.0475)	(0.140)	(0.129)	(0.214)
co-morbidity*PbR	-0.610***	-2.767***	-0.287	0.179
	(0.0766)	(0.235)	(0.178)	(0.333)
Constant	0.0430	0.348	0.317	-0.0401
	(0.127)	(0.613)	(0.418)	(0.777)
Year Dummies	Yes	Yes	Yes	Yes
N	1,080	1,080	1,080	1,080
R^2	0.883	0.913	0.270	0.017
Hospitals	120	120	120	120

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table D.29: Hip Replacement Model 2 with interactions.

	Filtered $D30_{ht}$	Filtered $D365_{ht}$	Filtered $R28_{ht}$	Filtered $R365_{ht}$
Tariff	-1.65e-06	-3.37e-05	3.13e-05	7.51e-05
	(8.59e-06)	(2.12e-05)	(3.29e-05)	(4.63e-05)
Age	0.00272**	7.16e-05	-0.00507	-0.0113

	Filtered $D30_{ht}$	Filtered $D365_{ht}$	Filtered $R28_{ht}$	Filtered $R365_{ht}$
	(0.00124)	(0.00310)	(0.00565)	(0.00794)
LOS	0.00249	0.00535	-0.00926	-0.0141
	(0.00177)	(0.00402)	(0.00706)	(0.0101)
Cases	6.97e-05**	7.30e-05	-0.000156	-4.35e-05
	(3.17e-05)	(9.55e-05)	(0.000151)	(0.000200)
Co-morbidity	0.0248	0.107	-0.0422	-0.0944
	(0.0319)	(0.0832)	(0.149)	(0.177)
Deprivation	0.0202	0.0471*	-0.0172	-0.0616
	(0.0133)	(0.0255)	(0.0486)	(0.0564)
FT	-0.00204	0.0380*	0.0639*	0.0788*
	(0.00831)	(0.0199)	(0.0330)	(0.0414)
Deprivation* PbR	-0.00347	-0.0133	0.0132	0.0147
	(0.00457)	(0.00862)	(0.0139)	(0.0183)
PbR(06)	0.0778*	0.122	-0.150	-0.270
	(0.0435)	(0.106)	(0.181)	(0.220)
co-morbidity*PbR	dropped	dropped	dropped	dropped
Constant	-0.237**	0.144	-0.127	0.227
	(0.101)	(0.262)	(0.459)	(0.624)
Year Dummies	Yes	Yes	Yes	Yes
N	826	826	826	826
R^2	0.066	0.136	0.230	0.115
Hospitals	118	118	118	118

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Bibliography

Arellano, M. and S. Bond (1991, April). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies 58*(2), 277 –297.

Arellano, M. and O. Bover (1995, July). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics 68*(1), 29–51.

Audit Commission (2004, July). Introducing payment by results -.

Audit Commission (2005, October). Early lessons from payment by results.

Audit Commission (2008, August). PbR data assurance framework 2007/08.

Aylin, P., B. Alves, A. Cook, J. Bennett, A. Bottle, N. Best, and B. Catena (1999). Analysis of hospial episode statistics for the bristol royal infirmary inquiry.

Aylin, P., A. Bottle, and A. Majeed (2007, May). Use of administrative data or clinical databases as predictors of risk of death in hospital: comparison of models. *BMJ 334*(7602), 1044.

Benbassat, J. and M. Taragin (2000, April). Hospital readmissions as a measure of quality of health care: Advantages and limitations. *Arch Intern Med 160*(8), 1074–1081.

Bentler, P. M. (1980, January). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology 31*(1), 419–456.

Bevan, G. and R. Hamblin (2009, January). 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*(1), 161–190.

Birkmeyer, J. D., J. B. Dimick, and D. O. Staiger (2006, March). Operative mortality and procedure volume as predictors of subsequent hospital performance. *Annals of Surgery 243*(3), 411–417. PMID: 16495708 PMCID: 1448928.

Birkmeyer, J. D., A. E. Siewers, E. V. A. Finlayson, T. A. Stukel, F. L. Lucas, I. Batista, H. G. Welch, and D. E. Wennberg (2002). Hospital volume and surgical mortality in the united states. *New England Journal of Medicine 346*(15), 1128–1137.

Bloom, N., C. Propper, S. Seiler, and J. V. Reenen (2010, May). The impact of competition on management quality: Evidence from public hospitals. *National Bureau of Economic Research Working Paper Series No. 16032*.

Blundell, R. and S. Bond (1998, August). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics 87*(1), 115–143.

Böcking, W., U. Ahrens, W. Kirch, and M. Milakovic (2005, April). First results of the introduction of DRGs in germany and overview of experience from other DRG countries. *Journal of Public Health 13*(3), 128–137.

Boyle, S. (2009). Payment by results in england. *Euro Observer 13*(1).

Boyle, S. (2011). *United Kingdom of Great Britain and Northern Ireland HiT / England*. Health Systems in Transition (HiT). European Observatory on Health Systems and Policies.

Brook, R. H., M. E. A., and P. D. Cleary (1996, September). Measuring quality of care — NEJM. *New England Journal of Medicine 335*(13), 966–970.

Busse, R., J. Schreyögg, and P. C. Smith (2006, August). Hospital case payment systems in europe. *Health Care Management Science 9*(3), 211–213. PMID: 17016926.

Capewell, S., C. E. Morrison, and J. J. McMurray (1999, April). Contribution of modern cardiovascular treatment and risk factor changes to the decline in coronary heart disease mortality in scotland between 1975 and 1994. *Heart 81*(4), 380 –386.

Carter, G. M., J. P. Newhouse, and D. A. Relles (1990). How much change in the case mix index is DRG creep? *Journal of Health Economics 9*(4), 411–428.

Charlson, M. E., P. Pompei, K. L. Ales, and C. R. MacKenzie (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases 40*(5), 373–383. PMID: 3558716.

Christiansen, C. L. and C. N. Morris (1997, October). Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine 127*(2), 764 –768.

Chulis, G. S. (1991). Assessing medicare's prospective payment system for hospitals. *Medical Care Research and Review 48*(2), 167 –206.

Clarke, A. (1996, September). Why are we trying to reduce length of stay? evaluation of the costs and benefits of reducing time in hospital must start from the objectives

that govern change. *Quality in Health Care 5*(3), 172–179. PMID: 10161532 PMCID: 1055402.

Cohen, P., J. Cohen, J. Teresi, M. Marchi, and C. N. Velez (1990, June). Problems in the measurement of latent variables in structural equations causal models. *Applied Psychological Measurement 14*(2), 183 –196.

Cox, J. and N. Koutroumanos (2010, May). Comparing coding between interventional radiologists and hospital coding departments. *Clinical Audit*, 33.

Cutler, D. M. (1995, May). The cost and financing of health care. *The American Economic Review 85*(2), 32–37. ArticleType: research-article / Issue Title: Papers and Proceedings of the Hundredth and Seventh Annual Meeting of the American Economic Association Washington, DC, January 6-8, 1995 / Full publication date: May, 1995 / Copyright © 1995 American Economic Association.

Cutler, D. M. and S. J. Reber (1998, April). Paying for health insurance: The Trade-Off between competition and adverse selection*. *Quarterly Journal of Economics 113*(2), 433–466.

Davies, H. (2005, November). Measuring and reporting the quality of health care: issues and evidence from the international research literature.

Davis, C. K. and D. J. Rhodes (1988, April). The impact of DRGs on the cost and quality of health care in the united states. *Health Policy 9*(2), 117–131.

Department of Health (2008). NHS costing manual 2007/08.

Diderichsen, F. (1995, June). Market reforms in health care and sustainability of the welfare state: lessons from sweden. *Health Policy (Amsterdam, Netherlands) 32*(1-3), 141–153. PMID: 10156635.

Dimick, J. and H. Welch (2008, January). The zero mortality paradox in surgery. *Journal of the American College of Surgeons 206*(1), 13–16.

Dimick, J. B., D. O. Staiger, and J. D. Birkmeyer (2006, August). Are mortality rates for different operations related? implications for measuring the quality of non-cardiac surgery. *Medical care 44*(8), 774–778. PMID: 16862040 PMCID: 2121187.

Dimick, J. B., H. G. Welch, and J. D. Birkmeyer (2004). Surgical mortality as an indicator of hospital quality. *JAMA: The Journal of the American Medical Association 292*(7), 847 –851.

Donabedian, A. (1966, July). Evaluating the quality of medical care. *The Milbank Memorial Fund Quarterly 44*(3), 166–206. ArticleType: research-article / Issue Title: Part 2: Health Services Research I. A Series of Papers Commissioned by the Health Services Research Study Section of the United States Public Health Service. Discussed at a Conference Held in Chicago, October 15-16, 1965 / Full publication date: Jul., 1966 / Copyright © 1966 Milbank Memorial Fund.

Donabedian, A. (1988, September). The quality of care. how can it be assessed? *JAMA: The Journal of the American Medical Association 260*(12), 1743–1748. PMID: 3045356.

Dranove, D. and W. D. White (1987). Agency and the organization of health care delivery. *Inquiry: A Journal of Medical Care Organization, Provision and Financing 24*(4), 405–415. PMID: 2961701.

Duckett, S. J. (1995, November). Hospital payment arrangements to encourage efficiency: the case of victoria, australia. *Health Policy (Amsterdam, Netherlands) 34*(2), 113–134. PMID: 10153481.

Duckett, S. J. and T. J. Jackson (2000, May). The new health insurance rebate: an inefficient way of assisting public hospitals. *The Medical Journal of Australia 172*(9), 439–442. PMID: 10870538.

Dufour, J. and E. Renault (1998). Short run and long run causality in time series: Theory. *Econometrica 66*(5), 1099–1125. ArticleType: research-article / Full publication date: Sep., 1998 / Copyright © 1998 The Econometric Society.

Ellis, R. P. and T. G. McGuire (1996, June). Hospital response to prospective payment: Moral hazard, selection, and practice-style effects. *Journal of Health Economics 15*(3), 257–277.

Ellis, R. P. and M. Vidal-Fernandez (2007). Activity-Based payments and reforms of the english hospital payment system. *Health Economics, Policy and Law 2*(04), 435–444.

Enders, W. (2004). *Applied econometric time series*. J. Wiley.

Farrar, S., D. Yi, M. Sutton, M. Chalkley, J. Sussex, and A. Scott (2009). Has payment by results affected the way that english hospitals provide care? difference-in-differences analysis. *339*. PMID: 19713233 PMCID: 2733950.

Feder, J., J. Hadley, and S. Zuckerman (1987, October). How did medicare's prospective payment system affect hospitals? *The New England Journal of Medicine 317*(14), 867–873. PMID: 3306387.

Feigin, V. L., C. M. M. Lawes, D. A. Bennett, S. L. Barker-Collo, and V. Parag (2009, April). Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. *Lancet Neurology 8*(4), 355–369. PMID: 19233729.

Fitzpatric, R. (2009). Patient-reported outcome measures and performnce measurement. In *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects*. Cambridge: Cambridge University Press.

Forsberg, E., R. Axelsson, and B. Arnetz (2001, December). Financial incentives in health care. the impact of performance-based reimbursement. *Health Policy (Amsterdam, Netherlands) 58*(3), 243–262. PMID: 11641002.

Freeman, T. (2002, May). Using performance indicators to improve health care quality in the public sector: a review of the literature. *Health Serv Manage Res 15*(2), 126–137.

Gerbode, F. and A. Selzer (1948, September). Experimental cardiac hypertrophy; the acute effect of pulmonic and aortic stenosis. *Surgery 24*(3), 505–511. PMID: 18884126.

Gil, M., J. Marrugat, J. Sala, R. Masia, R. Elosua, X. Albert, A. Pena, J. Vila, M. Pavesi, and G. Perez (1999, April). Relationship of therapeutic improvements and 28-Day case fatality in patients hospitalized with acute myocardial infarction between 1978 and 1993 in the REGICOR study, gerona, spain. *Circulation 99*(13), 1767–1773.

Ginsburg, P. B. and G. M. Carter (1986). Medicare case-mix index increase. *Health Care Financing Review 7*(4), 51–65. PMID: 10311672.

Ginsburg, P. B. and J. M. Grossman (2005). When the price isn't right: How inadvertent payment incentives drive medical care. *Health Affairs*.

Gujarati, D. N. (2003). *Basic econometrics*. McGraw Hill.

Guterman, S., P. W. Eggers, G. Riley, T. F. Greene, and S. A. Terrell (1988). The first 3 years of medicare prospective payment: an overview. *Health Care Financing Review 9*(3), 67–77. PMID: 10312519.

Ham, C. (2004, October). *Health Policy in Britain : The Politics and Organisation of the National Health Service; Fifth Edition (Public Policy and Politics)*. Palgrave Macmillan.

Ham, C. (2009, June). *Health Policy in Britain*. Palgrave Macmillan.

Hambleton, R. K. and L. L. Cook (1977, July). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement 14*(2), 75–96. ArticleType: research-article / Issue Title: Applications of Latent Trait Models / Full

publication date: Summer, 1977 / Copyright © 1977 National Council on Measurement in Education.

Hawkes, N. (2010a, April). The coding maze: mortality ratios and real life | straight statistics. http://www.straightstatistics.org/article/coding-maze-mortality-ratios-and-real-life.

Hawkes, N. (2010b, April). Patient coding and the ratings game. *BMJ 340*(apr23 2), c2153–c2153.

Healy, J., E. Sharman, and B. Lokuge (2006). *Australia HiT (2006)*. Health Systems in Transition (HiT). European Observatory on Health Systems and Policies.

Hensen, P., S. Beissert, L. Bruckner-Tuderman, T. A. Luger, N. Roeder, and M. L. Müller (2008, February). Introduction of diagnosis-related groups in germany: evaluation of impact on in-patient care in a dermatological setting. *The European Journal of Public Health 18*(1), 85 –91.

Holland, W. W. (1988). *European Community atlas of "avoidable death"*. Oxford University Press.

HOPE (2006, December). HOPE DRG report. Technical report, European Hospital and Healthcare Federation.

Hsia, D. C., W. M. Krushat, A. B. Fagan, J. A. Tebbutt, and R. P. Kusserow (1988, February). Accuracy of diagnostic coding for medicare patients under the prospective-payment system. *The New England Journal of Medicine 318*(6), 352–355. PMID: 3123929.

Iezzoni, L. I. (1994, December). Using risk-adjusted outcomes to assess clinical practice: An overview of issues pertaining to risk adjustment. *The Annals of Thoracic Surgery 58*(6), 1822–1826.

Iezzoni, L. I. (1997, October). Assessing quality using administrative data. *Annals of Internal Medicine 127*(2), 666 –674.

Iezzoni, L. I. (2003, June). *Risk adjustment for measuring health care outcomes*. Health Administration Press.

Iezzoni, L. I. (2009). Risk adjustment for performance measurement. In *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects*. Cambridge: Cambridge University Press.

Iezzoni, L. I., A. S. Ash, M. Shwartz, J. Daley, J. S. Hughes, and Y. D. Mackiernan (1996, October). Judging hospitals by severity-adjusted mortality rates: the influence of the severity-adjustment method. *Am J Public Health 86*(10), 1379–1387.

Jarman, B., A. Bottle, P. Aylin, and M. Browne (2005, February). Monitoring changes in hospital standardised mortality ratios. *BMJ 330*(7487), 329.

Jarman, B., S. Gault, B. Alves, A. Hider, S. Dolan, A. Cook, B. Hurwitz, and L. I. Iezzoni (1999, June). Explaining differences in english hospital death rates using routinely collected data. *BMJ 318*(7197), 1515 –1520.

Jegers, M., K. Kesteloot, D. D. Graeve, and W. Gilles (2002, June). A typology for provider payment systems in health care. *Health Policy 60*, 255–273.

Kahn, K. L., D. Draver, E. B. Keeler, W. H. Rogers, L. V. Rubenstein, J. Kosecoff, M. J. Sherwood, E. J. Reinisch, M. F. Carney, C. Kamberg, S. M. Bentow, K. Wells, M. A. Allen, D. Reboussin, C. P. Roth, C. Chew, and R. H. Brook (1992). *The Effects of the DRG-Based Prospective Payment System on Quality of Care for Hospitalized Medicare Patients*. RAND Corporation.

Kahn, K. L., W. H. Rogers, L. V. Rubenstein, M. J. Sherwood, E. J. Reinisch, E. B. Keeler, D. Draper, J. Kosecoff, and R. H. Brook (1990, October). Measuring quality of care with explicit process criteria before and after implementation of the DRG-Based prospective payment system. *JAMA: The Journal of the American Medical Association 264*(15), 1969 –1973.

Kahn, K. L., L. V. Rubenstein, D. Draper, J. Kosecoff, W. H. Rogers, E. B. Keeler, and R. H. Brook (1990, October). The effects of the DRG-Based prospective payment system on quality of care for hospitalized medicare patients. *JAMA: The Journal of the American Medical Association 264*(15), 1953 –1955.

Kane, T. J., D. O. Staiger, D. Grissmer, and H. F. Ladd (2002, January). Volatility in school test scores: Implications for Test-Based accountability systems. *Brookings Papers on Education Policy* (5), 235–283. ArticleType: research-article / Full publication date: 2002 / Copyright © 2002 The Brookings Institution.

Kastberg, G. and S. Siverbo (2007, January). Activity-based financing of health care-experiences from sweden financing of health care. experiences from sweden. *The International Journal of Health Planning and Management 22*(1), 25–44.

Keeler, E. B., K. L. Kahn, D. Draper, M. J. Sherwood, L. V. Rubenstein, E. J. Reinisch, J. Kosecoff, and R. H. Brook (1990, October). Changes in sickness at admission following the introduction of the prospective payment system. *JAMA: The Journal of the American Medical Association 264*(15), 1962 –1968.

Kessler, D. and M. McClellan (1996, May). Do doctors practice defensive medicine? *The Quarterly Journal of Economics 111*(2), 353 –390.

Kessler, D. P. and M. B. McClellan (2011, April). Is hospital competition socially wasteful?*. *Quarterly Journal of Economics 115*(2), 577–615.

Khush, K., A. Kopelnik, P. Tung, N. Banki, M. Dae, M. Lawton, W. Smith, B. Drew, E. Foster, and J. Zaroff (2005, February). Age and aneurysm position predict patterns of left ventricular dysfunction after subarachnoid hemorrhage. *Journal of the American Society of Echocardiography 18*(2), 168–174.

Klazinga, N. (2011). Health service outcomes. In P. C. Smith, I. Papanicolas, and J. Figueras (Eds.), *Health system perfomrance comparison: an agenda for policy, information and research*. European Observatory on Health Systems and Policies.

Landrum, M. B., S. E. Bronskill, and S. T. Normand (2000). Analytic methods for constructing Cross-Sectional profiles of health care providers. *Health Services and Outcomes Research Methodology 1*(1), 23–47.

Langhorne, P., D. J. Stott, L. Robertson, J. MacDonald, L. Jones, C. McAlpine, F. Dick, G. S. Taylor, and G. Murray (2000, June). Medical complications after stroke : A multicenter study. *Stroke 31*(6), 1223–1229.

Lilford, R., M. A. Mohammed, D. Spiegelhalter, and R. Thomson (2004, April). Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *The Lancet 363*(9415), 1147–1154.

Lilford, R. and P. Pronovost (2010, April). Using hospital mortality rates to judge hospital performance: a bad idea that just won't go away. *BMJ 340*(apr19 2), c2016–c2016.

Lilford, R. J., C. A. Brown, and J. Nicholl (2007, September). Use of process measures to monitor the quality of clinical practice. *BMJ : British Medical Journal 335*(7621), 648–650. PMID: 17901516 PMCID: 1995522.

Lingsma, H., E. Steyerberg, M. Eijkemans, D. Dippel, W. S. O. Reimer, H. V. Houwelingen, and T. N. S. S. Investigators (2010, February). Comparing and ranking hospitals based on outcome: results from the netherlands stroke survey. *QJM 103*(2), 99 –108.

Lütkepohl, H. (2006, June). *New introduction to multiple time series analysis*. Birkhäuser.

Mannion, R., G. Marini, and A. Street (2008). Implementing payment by results in the english NHS: changing incentives and the role of information. *Journal of Health Organisation and Management 22*(1), 79–88.

Mannion, R. and A. Street (2006). Payment by results and demand management : learning from the south yorkshire laboratory. Technical Report 14, Centre for Health Economics, York.

Manton, K. G., M. A. Woodbury, J. C. Vertrees, and E. Stallard (1993, August). Use of medicare services before and after introduction of the prospective payment system. *Health Services Research 28*(3), 269–292. PMID: 8344820 PMCID: 1069936.

Martin, S. and P. C. Smith (2005, October). Multiple public service performance indicators: Toward an integrated statistical approach. *Journal of Public Administration Research and Theory 15*(4), 599 –613.

Mason, A., A. Street, and R. Verzulli (2010, August). Private sector treatment centres are treating less complex patients than the NHS. *103*(8), 322–331. PMID: 20610618 PMCID: 2913062.

Maynard, A. (2008). Payment for performance (P4P): international experience and a cautionary proposal for Estonia.

McClellan, M. and D. Staiger (1999, August). The quality of health care providers. *National Bureau of Economic Research Working Paper Series No. 7327*. published as McClellan, Mark and Douglas Staiger. "Comparing The Quality Of Health Care Providers," Forum for Health Economics and Policy, 2000, v3, Article 6. Mark McClellan & Douglas Staiger, 2000. "Comparing the Quality of Health Care Providers," NBER Chapters, in: Frontiers in Health Policy Research, Volume 3, pages 113-136 National Bureau of Economic Research, Inc.

McGovern, P. G., D. R. Jacobs, E. Shahar, D. K. Arnett, A. R. Folsom, H. Blackburn, and R. V. Luepker (2001, July). Trends in acute coronary heart disease mortality, morbidity, and medical care from 1985 through 1997 : The minnesota heart survey. *Circulation 104*(1), 19–24.

McKee, M., J. Coles, and P. James (1999, December). 'Failure to rescue' as a measure of quality of hospital care: the limitations of secondary diagnosis coding in english hospital data. *Journal of Public Health 21*(4), 453 –458.

McKee, M. and P. James (1997, October). Using routine data to evaluate quality of care in british hospitals. *Medical Care 35*(10), OS102–OS111. ArticleType: research-article / Issue Title: Supplement: Hospital Restructuring in North America and Europe: Patient Outcomes and Workforce Implications / Full publication date: Oct., 1997 / Copyright © 1997 Lippincott Williams & Wilkins.

Meltzer, D., J. Chung, and A. Basu (2002, October). Does competition under medicare prospective payment selectively reduce expenditures on High-Cost patients? *The RAND Journal of Economics 33*(3), 447–468. ArticleType: research-article / Full publication date: Autumn, 2002 / Copyright © 2002 The RAND Corporation.

Mikkola, H., I. Keskimäki, and U. Häkkinen (2002, January). DRG-related prices applied in a public health care system–can finland learn from norway and sweden? *Health Policy (Amsterdam, Netherlands) 59*(1), 37–51. PMID: 11786173.

Mohammed, M., J. Raftery, M. LeatherBarrow, M. Harley, and T. Marshall (2004, April). Counting in-hospital deaths in england: a comparison of hospital computer systems and mortuary registers. *J Health Serv Res Policy 9*(2), 100–103.

Mohammed, M. A., J. J. Deeks, A. Girling, G. Rudge, M. Carmalt, A. J. Stevens, and R. J. Lilford (2009). Evidence of methodological bias in hospital standardised mortality ratios: retrospective database study of english hospitals. *BMJ : British Medical Journal 338*. PMID: 19297447 PMCID: 2659855.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro unit. *The Review of Economics and Statistics 72*(2), 334–38.

Muller, A. (1993, April). Medicare prospective payment reforms and hospital utilization. temporary or lasting effects? *Medical Care 31*(4), 296–308. PMID: 8464247.

Muthén, B. O. (1989, September). Latent variable modeling in heterogeneous populations. *Psychometrika 54*(4), 557–585.

National Health Service (2005). A guide to the PbR algorithm.

Newhouse, J. P. (1989). Do unprofitable patients face access problems? *Health Care Financing Review 11*(2), 33–42. PMID: 10313456.

Newhouse, J. P. (2002). *Pricing the priceless: a health care conundrum*. MIT Press.

Newhouse, J. P. and D. J. Byrne (1988, December). Did medicare's prospective payment system cause length of stay to fall? *Journal of Health Economics 7*(4), 413–416.

Newhouse, J. P., W. G. Manning, E. B. Keeler, and E. M. Sloss (1989). Adjusting capitation rates using objective health measures and prior utilization. *Health Care Financing Review 10*(3), 41–54. PMID: 10313096.

NHS executive (1994). Comparative cost data: the use of costed HRGs to inform the contracting process.

Nickell, S. (1981, November). Biases in dynamic models with fixed effects. *Econometrica 49*(6), 1417–1426. ArticleType: research-article / Full publication date: Nov., 1981 / Copyright © 1981 The Econometric Society.

Nolte, E., M. McKee, N. T. for Research, and P. S. in Health Services (2004). *Does health care save lives?: avoidable mortality revisited*. Nuffield Trust.

Normand, S. T., M. E. Glickman, and C. A. Gatsonis (1997). Statistical methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association 92*(439), 803–814. ArticleType: research-article / Full publication date: Sep., 1997 / Copyright © 1997 American Statistical Association.

Normand, S. T., R. E. Wolf, J. Z. Ayanian, and B. J. McNeil. Assessing the accuracy of hospital clinical performance measures. *Medical Decision Making 27*(1), 9 –20.

O'Brien, S. (2002, August). An outcome study on average lenght of stay following total hip and knee replacement. *Journal of Orthopaedic Nursing 6*(3), 161–169.

Powell, A. E., H. T. O. Davies, and R. G. Thomson (2003, April). Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls. *Quality and Safety in Health Care 12*(2), 122 –128.

Propper, C., S. Burgess, and D. Gossage (2008, January). Competition and quality: Evidence from the NHS internal market 1991–9*. *The Economic Journal 118*(525), 138–170.

Propper, C., S. Burgess, and K. Green (2004, July). Does competition between hospitals improve the quality of care?: Hospital death rates and the NHS internal market. *Journal of Public Economics 88*(7-8), 1247–1272.

Rauner, M. S., A. Zeiles, M. Schaffhauser-Linzatti, and K. Hornik (2003, May). Modelling the effects of the austrian inpatient reimbursement system on length-of-stay distributions. *OR Spectrum 25*(2), 183–206.

Reerink, E. (1990). Defining quality of care: Mission impossible? *International Journal for Quality in Health Care 2*(3-4), 197 –202.

Rogers, R., S. Williams, B. Jarman, and P. Aylin (2005, March). "HRG drift" and payment by results. *BMJ 330*(7491), 563.

Roodman, D. (2006, December). How to do xtabond2: An introduction to "Difference" and "System" GMM in stata. Working Paper 103, Center for Global Development.

Rosenberg, Marjorie A., B. M. J. (2001). The impact of the inpatient prospective payment system and Diagnosis-Related groups: A survey of the literature. *North American Actuarial Journal 5*(4), 84–94.

Schreyögg, J., T. Stargardt, O. Tiemann, and R. Busse (2006, July). Methods to determine reimbursement rates for diagnosis related groups (DRG): a comparison of nine european countries. *Health Care Management Science 9*(3), 215–223.

Schreyögg, J., O. Tiemann, and B. Reinhard (2005). The DRG reimbursement system in germany. *Euro Observer 7*(4), 4–6.

Scottish Executive (2004). National health service reform (Scotland) act 2004. An Act of the Scottish Parliament to make provision in relation to the organisation and operation of the National Health Service and the promotion of health improvement; and for connected purposes.

Secretary of State for Health (1997, December). The new NHS: modern, dependable.

Secretary of State for Health (2002, April). Delivering the NHS plan.

Shahian, D. M., S. Normand, D. F. Torchiana, S. M. Lewis, J. O. Pastore, R. E. Kuntz, and P. I. Dreyer (2001, December). Cardiac surgery report cards: comprehensive review and statistical critique. *The Annals of Thoracic Surgery 72*(6), 2155–2168.

Shahian, D. M., R. E. Wolf, L. I. Iezzoni, L. Kirle, and S. T. Normand (2010, December). Variability in the measurement of hospital-wide mortality rates. *New England Journal of Medicine 363*(26), 2530–2539.

Shen, Y. (2003, March). The effect of financial pressure on the quality of care in hospitals. *Journal of Health Economics 22*(2), 243–269.

Shleifer, A. (1985, October). A theory of yardstick competition. *The RAND Journal of Economics 16*(3), 319–327. ArticleType: research-article / Full publication date: Autumn, 1985 / Copyright © 1985 The RAND Corporation.

Shojania, K. G. and A. J. Forster (2008, November). Hospital standardized mortality ratios. *CMAJ 179*(10), 1037.

Silber, J. H., P. R. Rosenbaum, and R. N. Ross (1995, March). Comparing the contributions of groups of predictors: Which outcomes vary with hospital rather than patient characteristics. *Journal of the American Statistical Association 90*(429), 7–18. ArticleType: research-article / Full publication date: Mar., 1995 / Copyright © 1995 American Statistical Association.

Silverman, E. and J. Skinner (2004, March). Medicare upcoding and hospital ownership. *Journal of Health Economics 23*(2), 369–389. PMID: 15019762.

Sims, C. A. (1980, January). Macroeconomics and reality. *Econometrica 48*(1), 1–48. ArticleType: research-article / Full publication date: Jan., 1980 / Copyright © 1980 The Econometric Society.

Sloan, F. A., M. A. Morrisey, and J. Valvona (1988, May). Case shifting and the medicare prospective payment system. *Am J Public Health 78*(5), 553–556.

Sommersguter-Reichmann, M. (2000). The impact of the austrian hospital financing reform on hospital productivity: empirical evidence on efficiency and technology changes using a non-parametric input-based malmquist approach. *Health Care Management Science 3*(4), 309–321.

Spiegelhalter, D. J., P. Aylin, N. G. Best, S. J. W. Evans, and G. D. Murray (2002, June). Commissioned analysis of surgical performance using routine data: lessons from the bristol inquiry. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 165*(2), 191–221.

Stark, J., S. Gallivan, J. Lovegrove, J. Hamilton, J. Monro, J. Pollock, and K. Watterson (2000, March). Mortality rates after surgery for congenital heart defects in children and surgeons' performance. *The Lancet 355*(9208), 1004–1007.

Stock, J. H. and M. W. Watson (2001, October). Vector autoregressions. *The Journal of Economic Perspectives 15*(4), 101–115. ArticleType: research-article / Full publication date: Autumn, 2001 / Copyright © 2001 American Economic Association.

Street, A. and D. Dawson (2002). Costing hospital activity: the experience with healthcare resource groups in england. *The European Journal of Health Economics: HEPAC: Health Economics in Prevention and Care 3*(1), 3–9. PMID: 15609112.

Street, A. and A. Maynard (2007). Activity based financing in england: The need for continual refinement of payment by results. *Health Economics, Policy and Law 2*(04), 419–427.

Sussex, J. and S. Farrar (2008, August). Activity-based funding for national health service hospitals in england: managers' experience and expectations. *The European Journal of Health Economics 10*(2), 197–206.

Terris, Darcey, D. and C. Aron, David (2009). Attribution and causality in health-care performance measurement. In *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects*. Cambridge: Cambridge University Press.

The British Medical Association (2008, March). Payment by results - breifing note.

Theunissen, N. C. M., T. G. C. Vogels, H. M. Koopman, G. H. W. Verrips, K. A. H. Zwinderman, S. P. Verloove-Vanhorick, and J. M. Wit (1998, July). The proxy problem: child report versus parent report in health-related quality of life research. *Quality of Life Research 7*(5), 387–397.

Theurl, E. and H. Winner (2007, August). The impact of hospital financing on the length of stay: evidence from austria. *Health Policy (Amsterdam, Netherlands) 82*(3), 375–389. PMID: 17166618.

Titterington, D., A. Smith, and U. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley.

Tovey, D. (2007). *Clinical evidence handbook: the international source of the best available evidence for effective health care*. BMJ Publishing Group.

Treier, S. and S. Jackman (2008, January). Democracy as a latent variable. *American Journal of Political Science 52*(1), 201–217.

Wells, K., W. Rogers, L. Davis, K. Kahn, G. Norquist, E. Keeler, J. Kosecoff, and R. Brook (1993, December). Quality of care for hospitalized depressed elderly patients before and after implementation of the medicare prospective payment system. *Am J Psychiatry 150*(12), 1799–1805.

Wennberg, J. E., K. McPherson, and P. Caper (1984, August). Will payment based on diagnosis-related groups control hospital costs? *The New England Journal of Medicine 311*(5), 295–300. PMID: 6429534.

Westaby, S., N. Archer, N. Manning, S. Adwani, C. Grebenik, O. Ormerod, R. Pillai, and N. Wilson (2007, October). Comparison of hospital episode statistics and central cardiac audit database in public reporting of congenital heart surgery mortality. *BMJ 335*(7623), 759.

Williams, J. and R. Mann (2002, January). Hospital episode statistics: time for clinicians to get involved? *Clinical Medicine, Journal of the Royal College of Physicians 2*, 34–37.

Windmeijer, F. (2005, May). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics 126*(1), 25–51.

Wright, J. and K. G. Shojania (2009, March). Measuring the quality of hospital care. *BMJ 338*(mar18 2), b569–b569.