LSE Health

Working paper No: 22/2011 May 2011 LSE Health

Irene Papanicolas, Alistair McGuire

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

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

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

Irene Papanicolas, Alistair McGuire

First Published in May 2011

Working paper no. 22/2011

LSE Health
The London School of Economics and Political Science
Houghton Street
London WC2A 2AE

British Library Cataloguing in Publication Data A catalogue record for this publication is available from the British Library ISBN [978-0-85328-464-2]

Corresponding author
Irene Papanicolas and McGuire (2011)
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Email: I.N.Papanicolas@lse.ac.uk

Acknowledgments The authours would like to acknowledge Professor Elias Mossialos for his comments and support on this piece of work.

Abstract

In order to address the problem of poor quality information available to health care providers today, McClellan and Staiger (1999) developed a new method to measure quality, which addresses some key limitations of other approaches. Their method produces quality estimates that reflect different dimensions of quality and are able to eliminate systematic bias and noise inherent in these types of measures. While these measures are promising indicators, they have not been applied to other conditions or health systems since their publication. This paper attempts to replicate their 1999 method by calculating these quality measures for English Hospitals using Hospital Episode Statistics for the years 1996–2008 for Acute Myocardial Infarction (AMI) and Hip Replacement. Using the latent outcome measures calculated previously, Vector Autoregressions (VARs) are used to combine the information from different time periods and across measures within each condition. These measures are then used to compare current and past quality of care within and across NHS Acute Trusts. Our results support that this method is well suited to measure and predict provider quality of care in the English setting using the individual patient level data collected.

Keywords: Measuring Quality; Vector Autoregressions; Health.

Contents

1 Introduction

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 (**?**). In England, the routine use of Patient Reported Outcome 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 (Lisa I Iezzoni, Lisa I Iezzoni). 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 (Lisa I Iezzoni, Lisa I Iezzoni; 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 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,

[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.

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; as hospital technology improves patients may survive, but with worsened morbidity and subsequent episodes of hospital readmission. 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 surrounding 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. Previous work (Papanicolas and McGuire, 2011) has used this approach to measure the quality of different English NHS hospitals in providing services over the period 1996 –2008 for seven different conditions. 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.

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. These models are able to create smoothed out hospital rates of mortality and complications over time as well as to forecast future performance. This paper applies the McClellan and Staiger (1999) technique to our previously estimated latent estimates calculated in Papanicolas and McGuire (2011) and assesses the performance of the two measures in order to consider the advantages and disadvantages of the different methodologies.

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 outlined in Wooldridge (2002). 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 that it is atheorectical. 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[2]. 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:

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

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

Each equation in this system defines an outcome of interest and is estimated by Ordinary Least Squares (OLS). The outcomes are 30-day mortality, $D30_{ht}$, from the disease under consideration, one-year mortality, $D365_{ht}$, as well as 28-day readmissions $R28_{ht}$ and one-year readmissions, $R365_{ht}$. The subscript t denotes time in terms of years, and $t-1$ occruances in the past year. 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:

$$
\begin{aligned}
D30_{ht} &= \alpha + \gamma_1 D365_{ht} + \gamma_2 R28_{ht} + \gamma_3 R365_{ht} + \beta_2 D365_{h(t-1)} \\
&\quad + \beta_3 R28_{h(t-1)} + \beta_4 R365_{h(t-1)} + \epsilon_{D30ht} \\
D365_{ht} &= \alpha + \gamma_1 D30_{ht} + \gamma_2 R28_{ht} + \gamma_3 R365_{ht} + \beta_1 D30_{h(t-1)} + \beta_2 D365_{h(t-1)} \\
&\quad + \beta_3 R28_{h(t-1)} + \beta_4 R365_{h(t-1)} + \epsilon_{D365h} \\
R28_{ht} &= \alpha + \gamma_1 D30_{ht} + \gamma_2 D365_{ht} + \gamma_3 R365_{ht} + \gamma_3 R365_{ht} + \beta_1 D30_{h(t-1)} + \beta_2 D365_{h(t-1)} + \\
&\quad \beta_3 R28_{h(t-1)} + \beta_4 R365_{h(t-1)} + \epsilon_{R28ht} \\
R365_{ht} &= \alpha + \gamma_1 D30_{ht} + \gamma_2 D365_{ht} + \gamma_3 R28_{ht} + \beta_1 D30_{h(t-1)} + \beta_2 D365_{h(t-1)} \\
&\quad + \beta_3 R28_{h(t-1)} + \beta_4 R365_{h(t-1)} + \epsilon_{R365ht} \,. \qquad (2)
\end{aligned}
$$

Equations (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.

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. 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 replicate this process are reviewed in detail the following section before the results are presented and discussed.

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 Papanicolas and McGuire (2011), derives latent outcome measures at the hospital level (h) by estimating patient level (i) regressions 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.

$$Y^k_{iht} = \beta q^k_{1h} + \sum \phi X_{jht} + u_{iht}\,. \tag{3}$$

As explained in detail in Papanicolas and McGuire (2011), 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 (3)) 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:

$$q^k_{ht} = Q_{ht}\beta^k_{ht} + \omega_{iht} \tag{4}$$

They note that equation (4) 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 technique as outlined, for example, in Titterington et al. (1985).

While equation (4) 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

$$\beta = [E(Q'_hQ_h)]^{-1}E(Q'_hq_h) \,. \tag{5}$$

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

$$E(Q_h'Q_h) = E(q_h'q_h) + E(\epsilon_h'\epsilon_h) \tag{6}$$

$$E(Q_h'Q_h) = E(q_h'q_h)\,, \tag{7}$$

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 (5) allows the calculation of the desired least squares estimates, such that:

$$\hat{q}_{ht} = Q[E(Q_h'Q_h)]^{-1}E(Q_h'q_h) = Q_h[E(q_h'q_h) + E(\epsilon_h'\epsilon_h)]^{-1}E(q_h'q_h)\,. \tag{8}$$

Using estimates (6)and (7), the R-squared statistic can also be calculated, based on the least squared formula.

Estimation of equation (8) 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 - given its past performance plus a contemporaneous shock - can be correlated across the different outcome measures. Thus a first order VAR model for $q_{ht}(1 \times K)$ is estimated, where:

$$q_{ht} = q_{h,t-1}\Phi + v_{ht}\,. \tag{9}$$

$Z = V(v_{ht})$ 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:

$$E(Q_h'Q_h) - S_h = E(q_h'q_h) = f(\Phi, Z, \Gamma)\,. \tag{10}$$

Using the parameters estimated from the VAR model we are able to estimate equation (10), 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 detail in Papanicolas and McGuire (2011), 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.

4 Data

Similar to Papanicolas and McGuire (2011), the data used in this paper is Hospital Episode Statistics (HES) accessed through Dr. Foster. 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. 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). 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.

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.

This paper analyses data provided for the financial years 1996 –2008, for the conditions of Acute myocardial infarction (AMI) 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. Due to problems with the sample sizes for some of the years before 2000 for AMI these years were not included in the analysis. Moreover, 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.

This paper builds on the methods used in Papanicolas and McGuire (2011) which used individual patient mortality rates and readmission rates at different intervals to contract latent outcome measures at the hospital level, as outlined in the methodology section above. 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. 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 1. Data were collected on seven conditions to assess the generalisability of the method. The seven conditions are Acute Myocardial Infarction (AMI), Myocardial Infarction (MI), Ischemic Heart Disease (IHD), Congestive Cardiac Failure (CCF), Stroke, Transient Ischemic Attack (TIA) and Hip Replacement. However, we reprot on two conditions in detail, AMI and Hip Replacement, for the sake of brevity and as the general conclusion hold for all the conditions. More detail can be obtained by contacting the authors.

Tab. 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
Hip	OPCS4.3: W37-W39 W46-W48 W58	1996 –2008	120	332

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 and Hip Replacement are presented in this section, by condition. The methodology was also used to study five other outcomes, namely Myocardial Infarction, Ischemic Heart Disease, Stroke, Congestive Cardiac Failure and Transient Ischemic Attack, in order to test feasibility across a wider range of conditions. As the methodology was applicable, and the results were similar for all conditions we chose not to present all the results in this paper due to the relatively large set of results which, if presented in totality, might obscure the main objective of this article 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 of the results 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 (9), 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 (10) above, to estimate the signal-to-noise ratio for each of the outcome measures, as specified in equation (11). 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).

$$\text{Signal}/(\text{Signal} + \text{Noise}) = V_{ht}/(V_{ht} + S_{ht)} \tag{11}$$

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 uses the VAR framework to further refine the latent outcome measures estimated, as done in Papanicolas and McGuire (2011) 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 Papanicolas and McGuire (2011). 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:

$$R^2 = 1 - \frac{\sum_{h=1}^{N} \hat{u}_h^2}{\sum_{h=1}^{N} q_h^2} \,. \tag{12}$$

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 6 and 7 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:

$$R^2 = 1 - \frac{\sum_{h=1}^{N} (\hat{u}_h^2 - S_h)}{\sum_{h=1}^{N} (Q_h^2 - S_h)} \,. \tag{13}$$

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 (12)), the actual R-squared measures, based on actual estimates (equation (13)) 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 (5.3), 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.

5.1 AMI

The parameter estimates of basic model coefficients in Table 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 mortal-

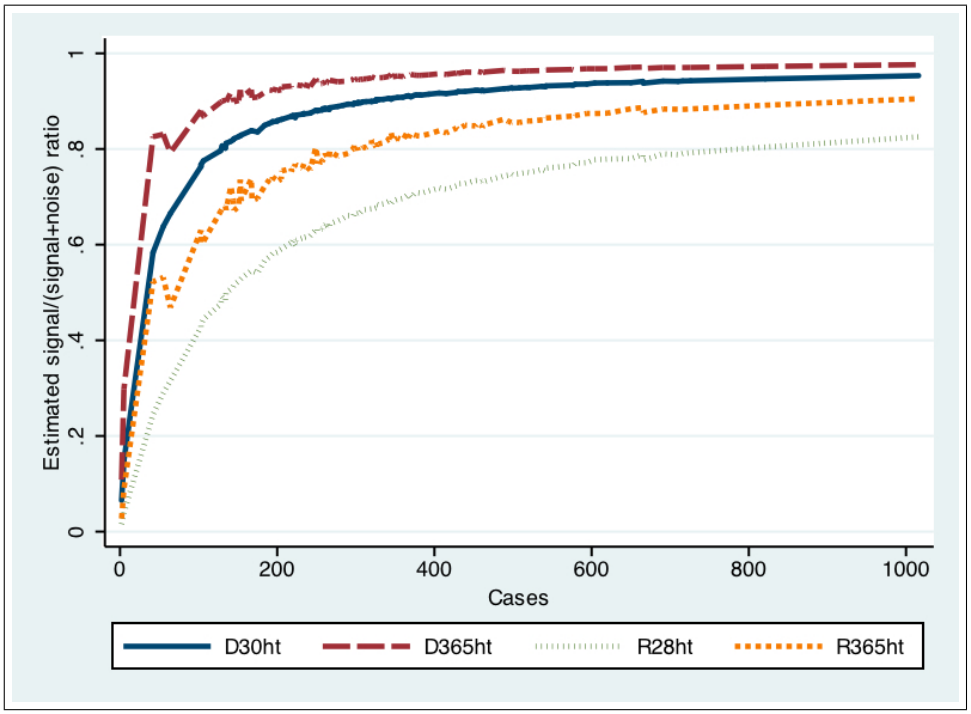
ity rates fluctuate more, varying around 10% annually. The correlation between variables in the year 2000, indicates a negative association between the outcome measures 30-day mortality, $D30_{ht}$ and short term re-admissions, $R28_{ht}$. The correlation of residuals indicates a similar negative association between $D365_{ht}$ and $R365_{ht}$, and a positive association between $R28_{ht}$ and $R365_{ht}$.

Tab. 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]
$D365_{h(t-1)}$	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 ($R365_{ht}$)	-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 $R365_{ht}$ in 2000	0.0641	0.0334	-0.0431	-
Sample (adjusted): 2001 2008				
Included observations: 952 after adjustments				
Standard errors in ( ) & t-statistics in [ ]				

Figure 1 presents the signal to noise ratio of the four AMI outcome measures. This is calculated as specified by equation (11) 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. What is immediately apparent from Figure 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 to the higher variance across hospitals in the true effects observed in Table 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.

Fig. 1: Signal to noise ratio for the four AMI outcome measures (year 2005).



Figures 2 – 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 as derived in Papanicolas and McGuire (2011) (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 the filtered and latent outcome indicators are the confidence intervals which are much wider for the filtered measures than they were for the latent variables as estimated in Papanicolas and McGuire (2011). 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.

Fig. 2: Filtered and latent estimates for AMI $D30_{ht}$ for selected hospitals.

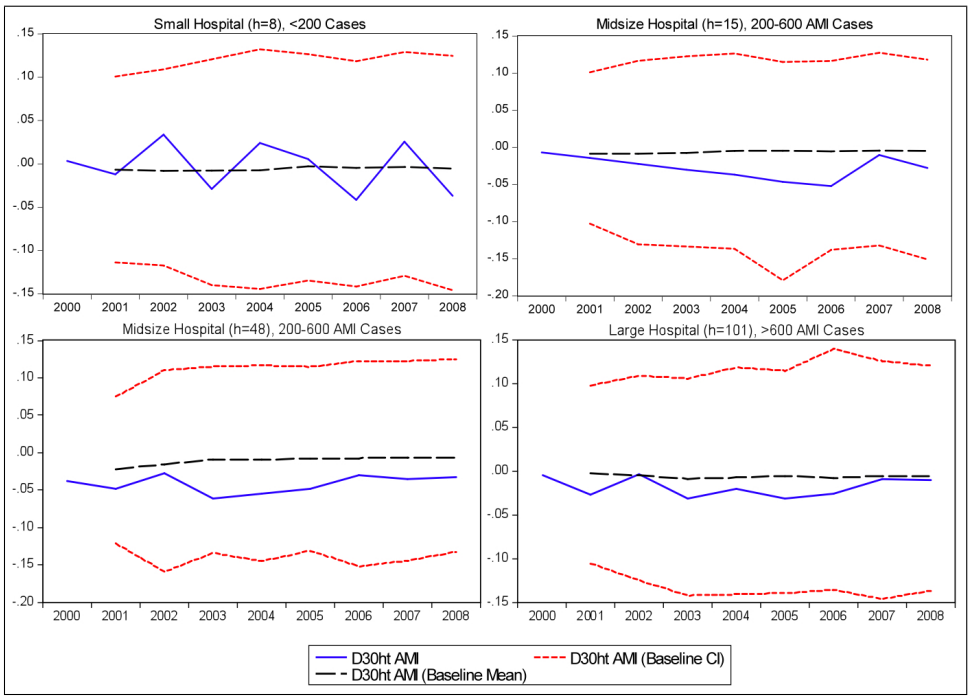


Fig. 3: Filtered and latent estimates for AMI $D365_{ht}$ for selected hospitals.

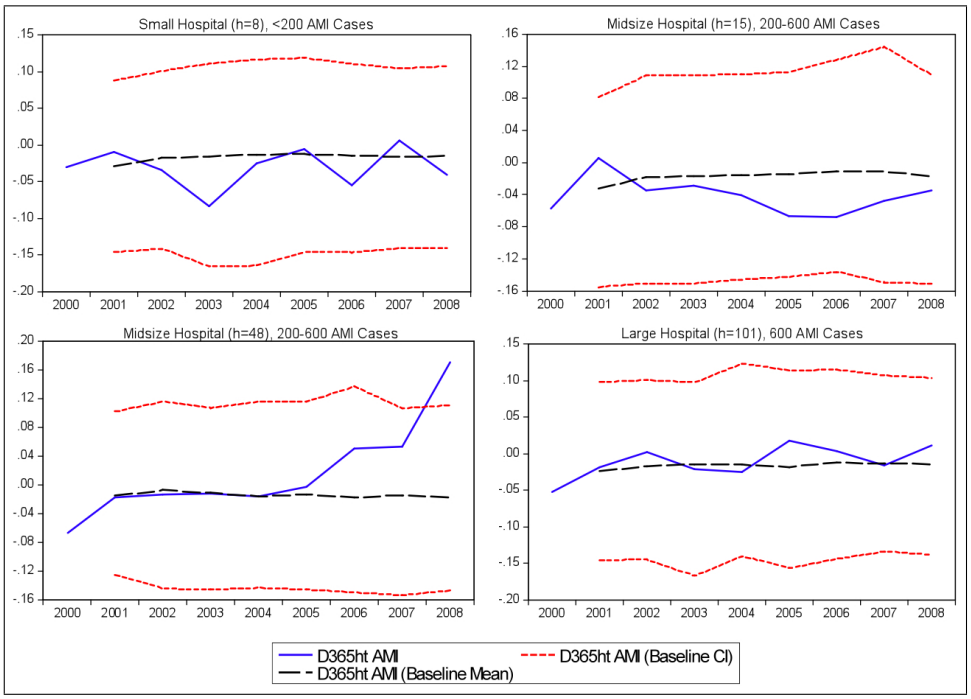


Fig. 4: Filtered and latent estimates for AMI $R28_{ht}$ for selected hospitals.

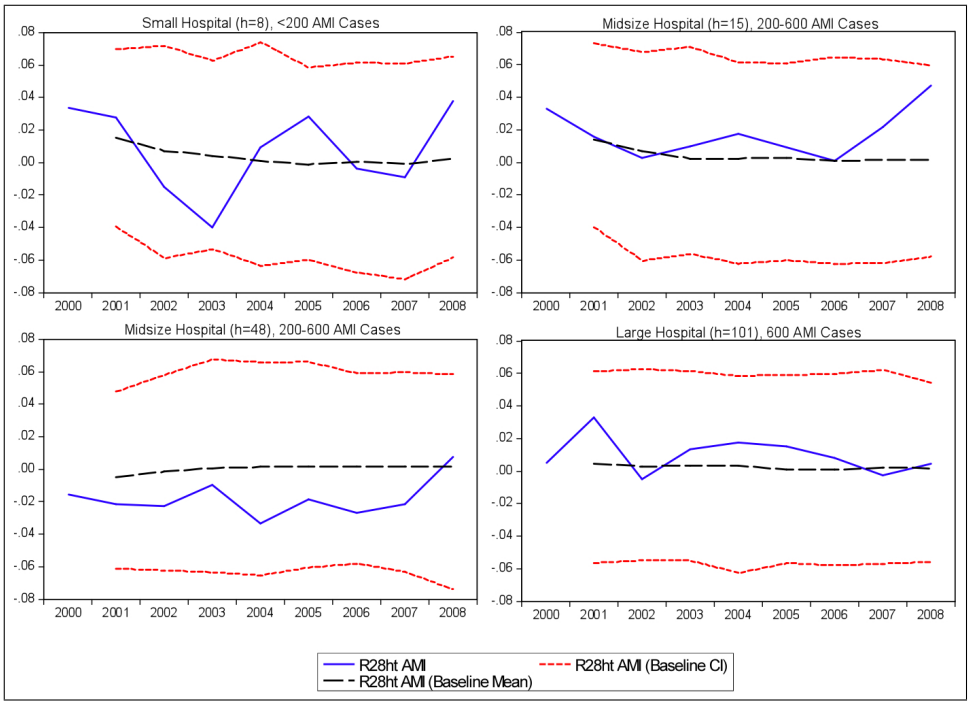


Fig. 5: Filtered and latent estimates for AMI $R365_{ht}$ for selected hospitals.

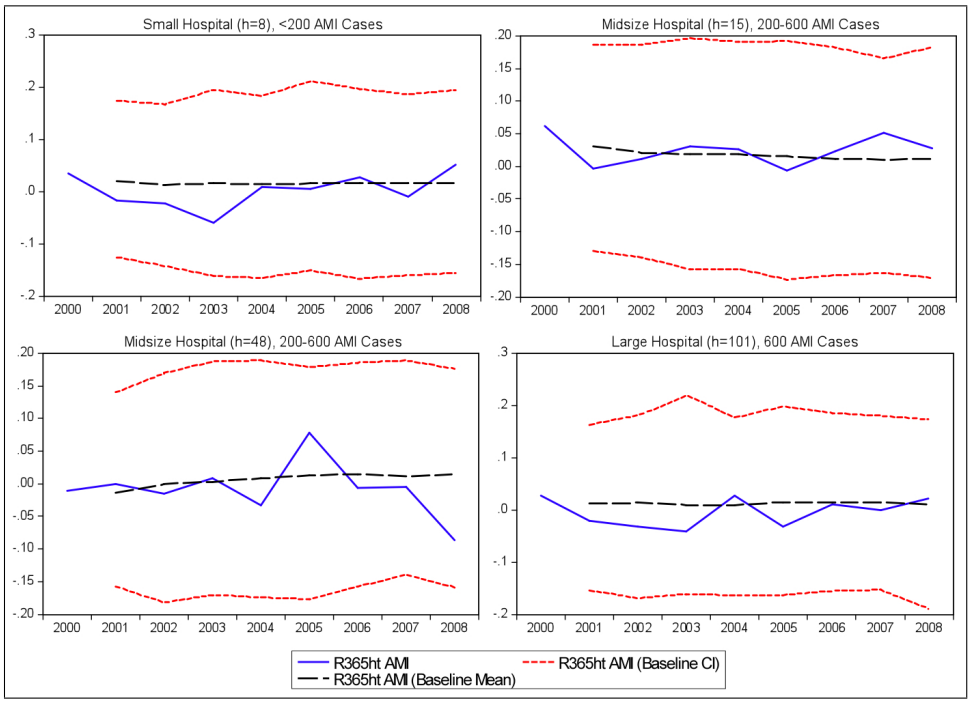


Table 3 indicates the R-squared estimates as calculated from equation (12) discussed above. These are presented for the predictions made of the different outcome measures, using different amounts of past data. The table indicates very high R-squared values for all measures, suggesting that the filtered 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 also indicates that the filtered estimates are able to predict just as well using fewer years of data.

Tab. 3: Summary of estimated prediction accuracy using alternative methods of signal extraction. All estimates based on the VAR(1) model from Table 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 4. The expected R-squared values are derived using equation (13) 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.

Tab. 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

	All outcomes	Same outcome	All outcomes	Same outcome
$D365_{ht}$				
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

5.2 Hip Replacement

The parameter estimates of the basic model run for Hip Replacement are presented in Table 5. 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 $R28_{ht}$.

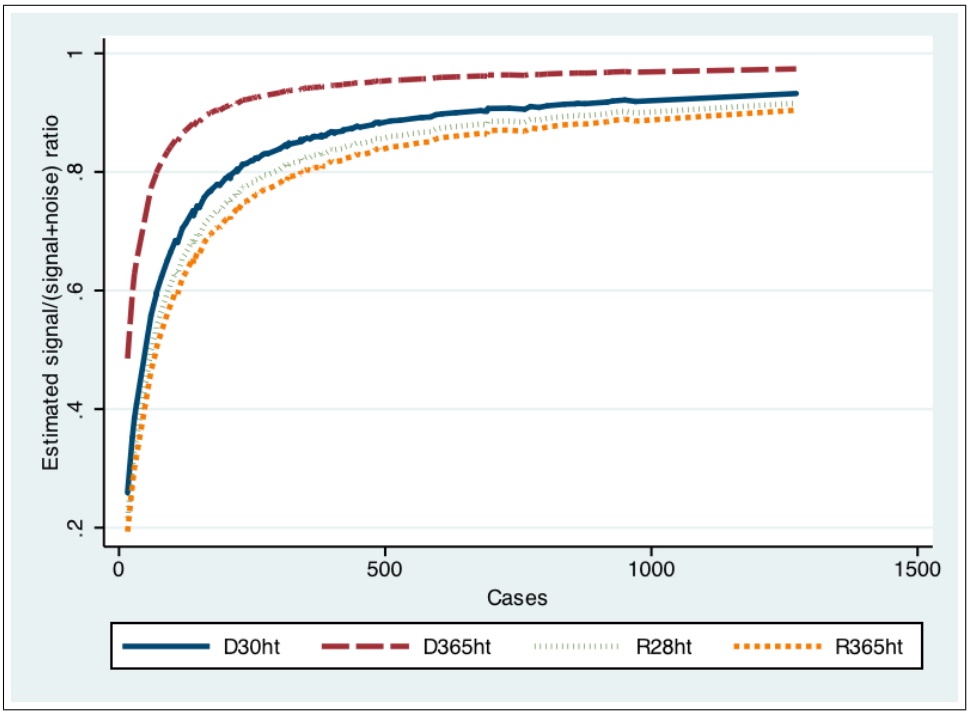
Tab. 5: Estimates of Hip Replacement multivariate VAR(1) parameters for hospital specific effects.

	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D30_{h(t-1)}$	-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 6 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. While year-long readmissions have a higher signal variance than year-long mortality (Table 5), they most probably have higher amounts in the variance of the estimation error, causing them to perform the worst of the four measures.

Fig. 6: Signal to noise ratio for the four Hip Replacement outcome measures (year 2005) .



Figures 7 – 10 present the filtered Hip outcome measures, their 95% confidence intervals and the corresponding latent outcome measures derived in Papanicolas and McGuire (2011) for selected hospitals. The sample for Hip Replacement is longer than for AMI, 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.

Fig. 7: Filtered and latent estimates for Hip Replacement $D30_{ht}$.

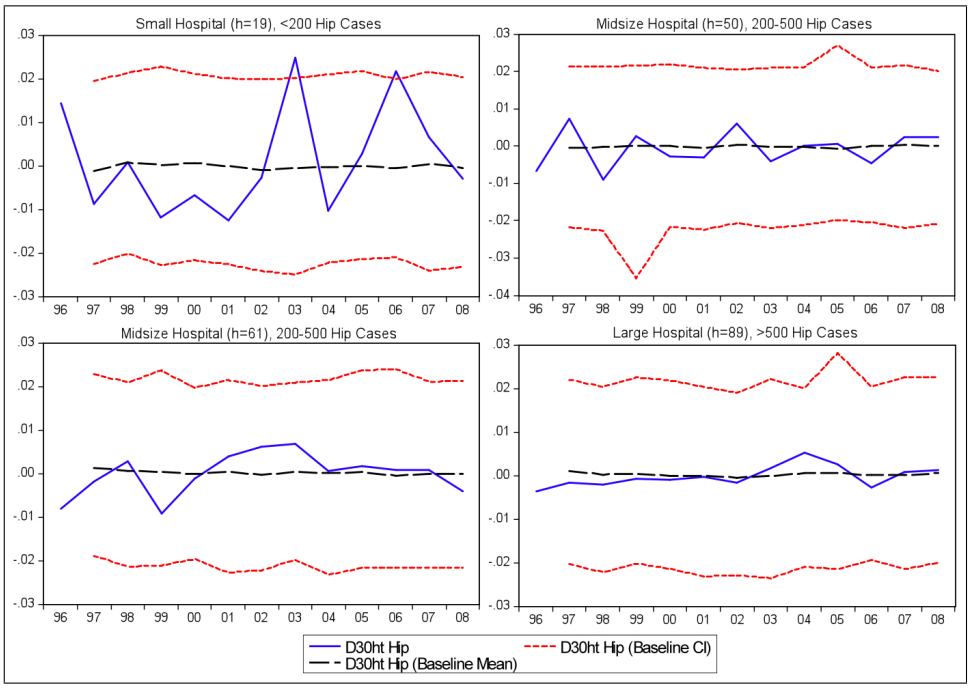


Fig. 8: Filtered and latent estimates for Hip Replacement $D365_{ht}$.

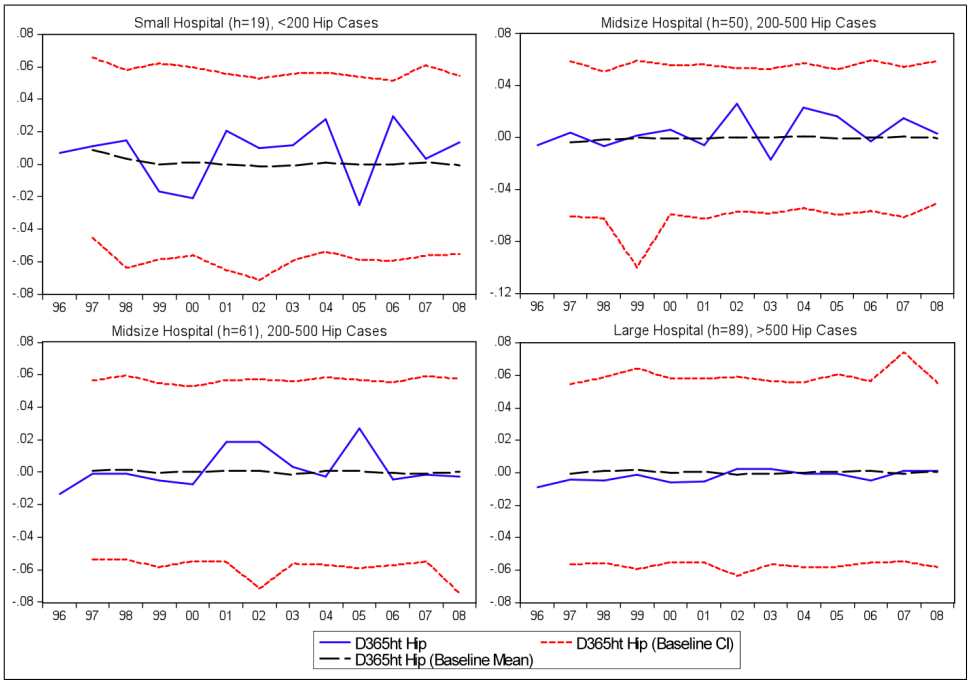


Fig. 9: Filtered and latent estimates of Hip Replacement $R28_{ht}$.

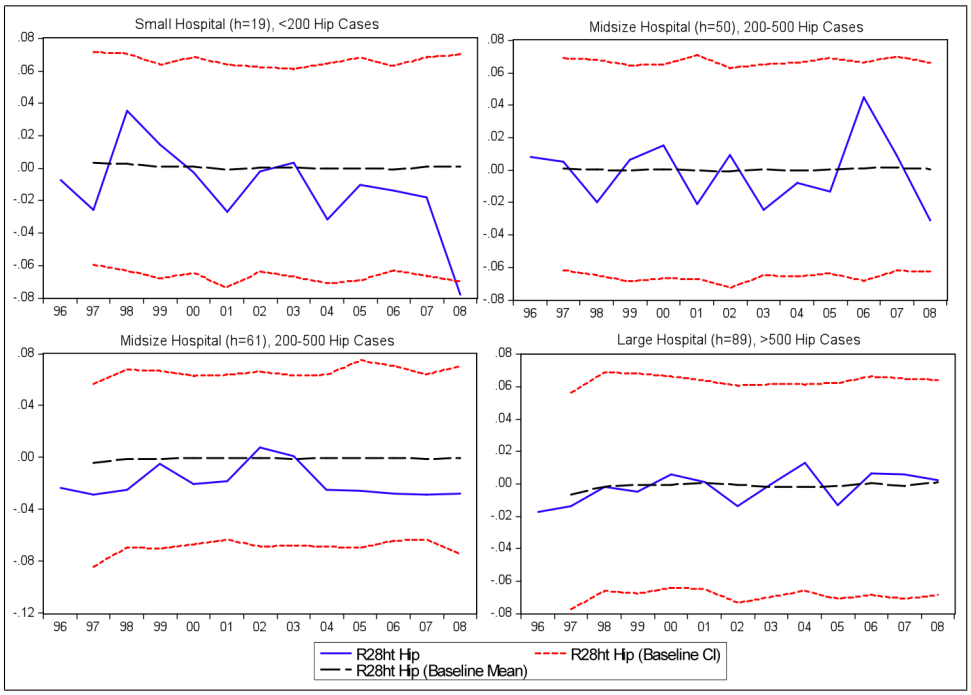


Fig. 10: Filtered and latent estimates of Hip Replacement $R365_{ht}$.

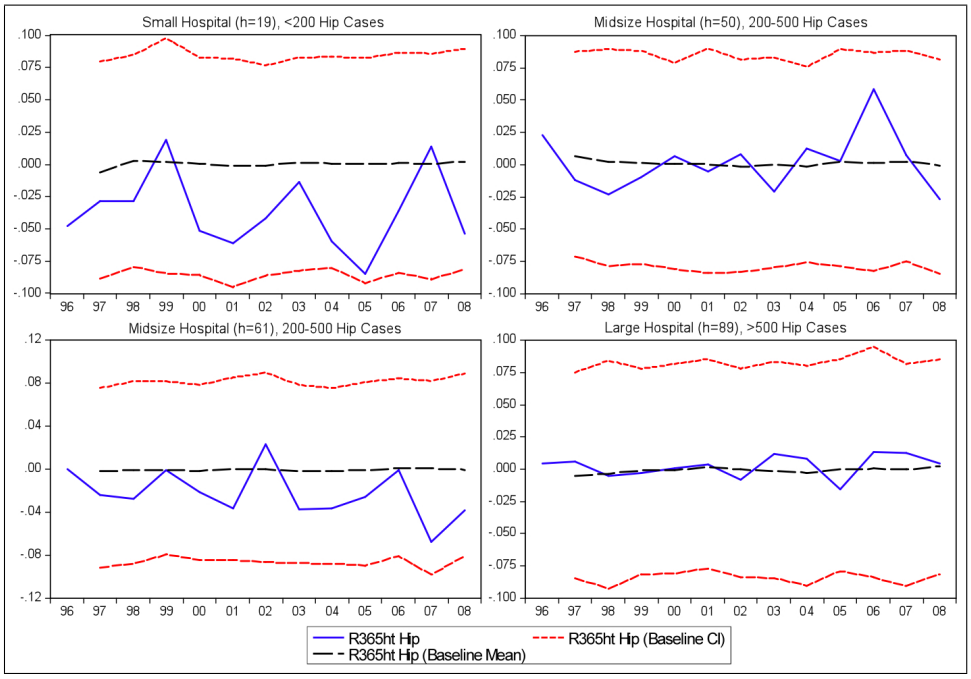


Table 6 indicates the R-squared estimates for the predictions made for the Hip filtered outcomes, using different amounts of past data. The R-squared values for Hip are extremely

high, indicating a near perfect prediction for all measures, even when using only one year of data. Table 7 indicates the R-squared values for the outcome forecasts, estimated using equation (13), and predictions estimated using equation (12). These are also near perfect for both the forecasts and predictions, and both the VAR(1) and VAR(2) specifications. This indicates 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.

Tab. 6: Summary of estimated prediction accuracy using alternative methods of signal extraction. All estimates based on the VAR(1) model from Table 5.

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

Tab. 7: 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
$D365_{ht}$				
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

	All outcomes	Same outcome	All outcomes	Same outcome
$R28_{ht}$				
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

5.3 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, while the third measure is the filtered 30-day mortality rate estimated using the McClellan and Staiger (1999) methodology. The hospitals are also ranked by the other outcomes and these are reported in Appendix A 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 and not the other conditions, as the results are very similar and do not provide further insight.

Tab. 8: 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
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 8 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 11 illustrates the different rankings for the first 15 hospitals in the sample. What is immediately apparent from both Table 8 and Figure 11 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.

Fig. 11: Rankings of 2005 AMI quality measures for $D30_{ht}$.

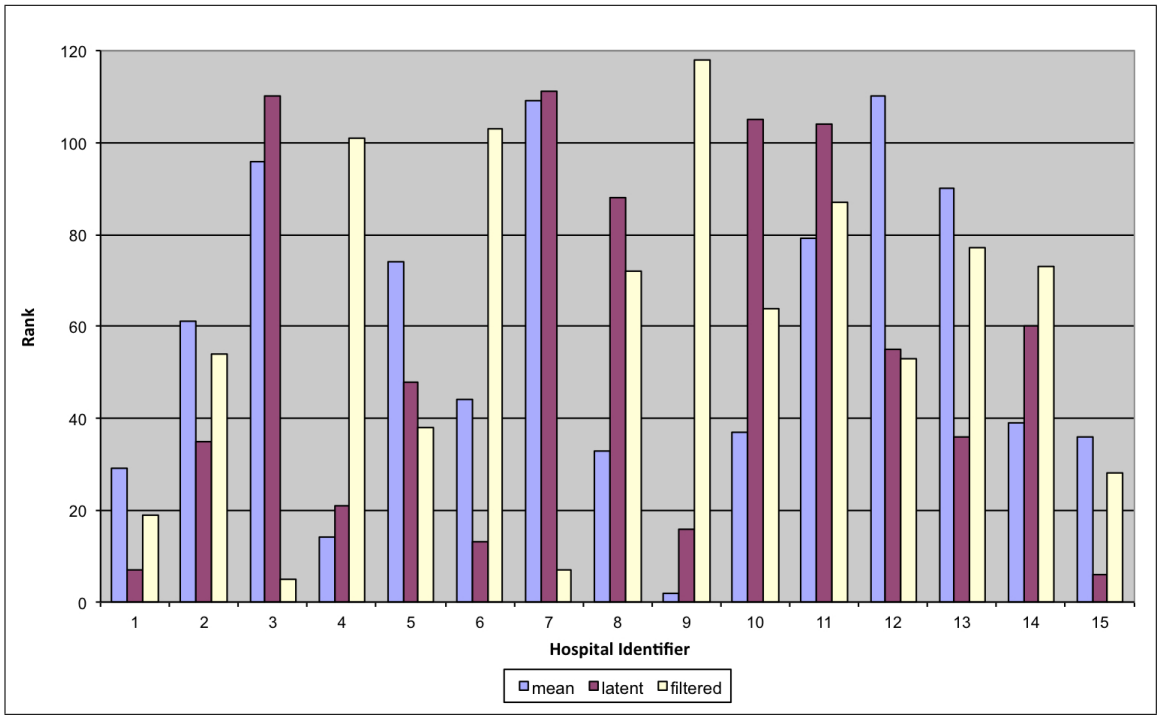
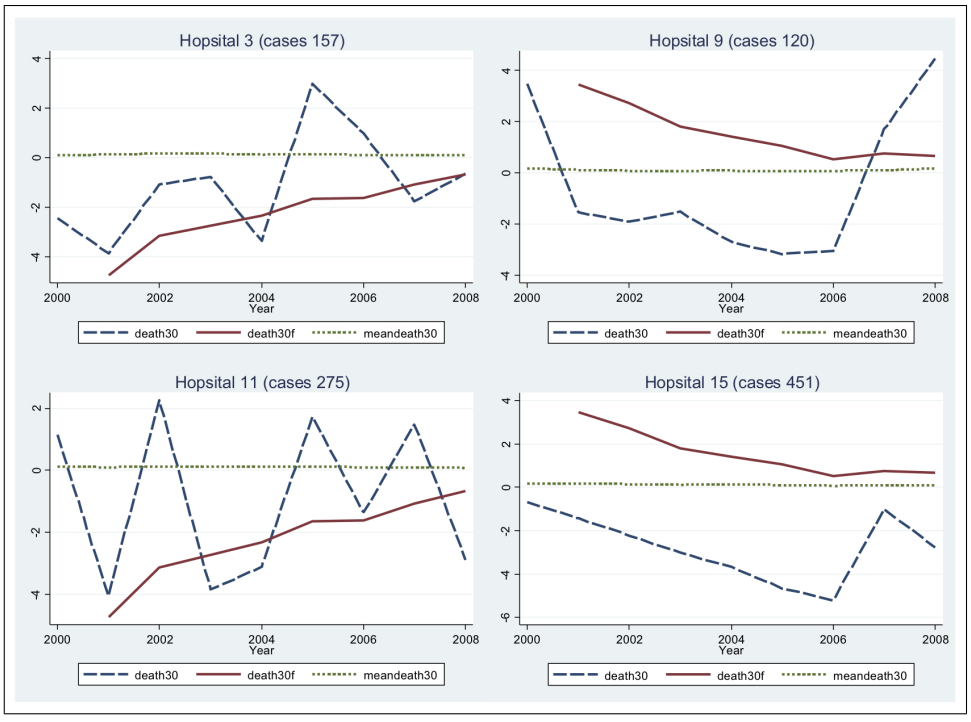


Figure 12 presents the full time series of the three different performance indicators for hospitals 3, 9, 11 and 15. This alternative presentation of the data can help to better understand why the rankings are different from one another. In the upper left hand corner the trajectories of hospital 3's 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 time 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.

Fig. 12: 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 finding was also reported in Papanicolas and McGuire (2011), where latent estimates calculated for 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.

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[3], or for other conditions. This paper attempts to replicate their analysis using English patient level data for a wider range of conditions. The paper 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 Papanicolas and McGuire (2011). 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 paper 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 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. In all conditions for which this methodology was used, it suggested a strong dynamic presence for the year-long mortality

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

indicator. For most conditions, except Hip Replacement, this indicator also exhibits a high standard deviation across hospitals, ranging from 20% to 5%. 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 varied considerably more by condition, while it was quite low for AMI and Hip Replacement which are the results reported in this paper. The variation of 30-day mortality across hospitals also varies considerably by condition, and again was low for AMI, at around 6% and Hip Replacement, at around 1%. 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, previous analyses using the latent indicators in Papanicolas and McGuire (2011) 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 and Hip Replacement, while year-long readmissions are not persistent for either condition. The variation in short and long term readmissions varies more considerably by condition. The standard deviation on both indicators is around 5% for AMI and Hip Replacement but ranges from between 10-2% for some of the other conditions estimated.

The AMI model indicates 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 model suggests 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, 30-day mortality is negatively associated with both short and long term readmissions but year-long mortality has a positive association. In most of the conditions, all associations are weak. However, for no condition were all associations positive, 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. Short and long term readmissions have strong positive correlations with each other in the AMI and Hip Replacement models, but very weak associations for most of 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. 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 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, although this was not the case for 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 to 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, 2010). 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. 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[4]. 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. 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,

[4] 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 paper 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 (13)) 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 paper 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 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 paper 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.

A Appendix: Comparison of Indicators

Tab. 9: 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						
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

Tab. 10: 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

Ranking	Mean $R28_{ht}$	Hospital	Latent $R28_{ht}$	Hospital	Filtered $R28_{ht}$	Hospital
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
119	0.1933962	46	5.822222	59	0.8452681	54

Tab. 11: 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

Ranking	Mean $R365_{ht}$	Hospital	Latent $R365_{ht}$	Hospital	Filtered $R365_{ht}$	Hospital
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

References

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*.

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.

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.

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.

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.

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., 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.

Enders, W. (2004). *Applied econometric time series*. J. Wiley.

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.

Hawkes, N. (2010, April). Patient coding and the ratings game. *BMJ 340*(apr23 2), c2153–c2153.

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. (2003, June). *Risk adjustment for measuring health care outcomes*. Health Administration 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., 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.

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.

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 *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.

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.

Lisa I Iezzoni, L. Risk adjustment for performance measurement. In *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects*. Cambridge: Cambridge University Press.

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.

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.

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.

Papanicolas, I. and A. McGuire (2011, May). Using a latent variable approach to measure the quality of English NHS Hospitals. Technical Report 21, London School of Economics, London.

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.

Reerink, E. (1990). Defining quality of care: Mission Impossible? *International Journal for Quality in Health Care 2*(3-4), 197 –202.

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.

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.

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.

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.

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.

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.

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.

Titterington, D., A. Smith, and U. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press.

Wright, J. and K. G. Shojania (2009, March). Measuring the quality of hospital care. *BMJ 338*(mar18 2), b569–b569.

For further information on this or any of the Health publications contact:

Naho Ollason
Managing Editor
LSE Health
The London School of Economics and Political Science
Houghton Street
London WC2A 2AE

Tel: + 44 (0)20 7955 3733
Fax: + 44 (0)20 7955 6090
Email: n.ollason@lse.ac.uk
Website: www.lse.ac.uk/collections/LSEHealth/