Measuring and forecasting quality in English hospitals

Measuring and forecasting quality in English hospitals

Irene Papanicolas, Alistair McGuire

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

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

Abstract

Hospital performance metrics, often in the form of risk-adjusted hospital mortality rates, are increasingly being made available in the public domain to compare different hospitals. Despite the proliferation of these metrics, uncertainty remains regarding their validity and reliability given the noise surrounding their underlying measures. This paper considers a quality measure of hospital performance developed by McClellan and Staiger (1999) which smooths within hospitals and over time, while remaining computationally straightforward. The McClellan and Staiger method improves on others by incorporating different measures of outcome, eliminating systematic bias arising from the heterogeneous mix of hospital outputs and the noise inherent in other measures of quality. The technique also allows the forecasting of future quality. Using English Hospital Episode Statistics for the years 2000-2005 for Acute Myocardial Infarction (AMI) and Hip Replacement, we use this technique to return quality measures based on hospital fixed effects estimated from yearly cross-sectional patient level data, and Vector Autoregressions (VARs) estimated over time, which then combine information from different time periods and across conditions to produce robust hospital quality measures. Our results suggest that this method is well suited to measure and predict provider quality of care in the English setting.

Keywords: Measuring Quality; Vector Autoregressions; Health; Hospitals.

1 Introduction

Metrics of risk adjusted hospital mortality are increasingly being released into the public domain to assess health care provider performance, enabling patients to make informed choices, and help managers and clinicians to improve service delivery. The publication of this type of performance information aids policy, planning and research through providing forecasts and evaluations of health care services. Evaluations of performance information allow policy makers to better understand the consequences of policy actions or practices that have occurred in the past, while forecasting allows them to use data to look forward and make informed planning decisions, such as with the allocation of funding (Jones and Spiegelhalter, 2012; McClellan and Staiger, 1999).

To date the main measure used for these types of activities in the United Kingdom has been the Hospital Standardized Mortality Ratio (HSMR) and since 2011 the Summary Hospital-level Mortality Indicator (SHMI). However, these measures have limitations in their ability to inform policy and have been heavily criticized for their methodological and practical short comings (Campbell et al., 2012; Lilford and Pronovost, 2010). As 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 estimated using national data, after adjustment for factors affecting the risk of in-hospital death; essentially adjusting through age, gender, diagnosis and route of admission (Shojania and Forster, 2008). The SHMI builds upon recommendations for the improvement of HSMRs to increase robustness, such as using three years of past data to build risk adjustment models rather than one year, but follows a similar methodology (Information Centre, 2012). Despite their widespread use, many authors express concern over the degree of true quality information that these indicators hold and recommend caution when using them to draw conclusions (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). Mohammed et al. (2009), for example find direct, systematic associations between hospital mortality rates and the factors used to adjust for case-mix in England based on Hospital Episode Statistics (HES) data, suggesting that the employment of these measures as risk adjusters may actually increase the bias that they are intended to reduce (Lilford et al., 2007; Powell et al., 2003). While such concerns may represent skepticism over the ability of risk adjustment techniques to control for differences in case mix or chance variation, they also reflect concerns over the use of a particular, single dimension of mortality as a proxy for overall hospital quality (Shojania and Forster, 2008).

Notwithstanding these problems, policy makers increasingly focus on mortality rates associated with specific conditions or procedures, where the quality of care is known to have a large impact on patient outcomes (Lilford and Pronovost, 2010). Mortality rates for specific conditions or procedures have become popular measures as they are able to identify key areas where health system quality is more likely to influence specific outcomes, especially 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. The proven link between identified care processes and patient outcomes, for conditions such as AMI, allow researchers to be more confident in making judgments about quality and its association with treatment (Klazinga, 2011). A considerable body of work has used AMI mortality as a proxy for quality both in England (Bloom et al., 2010; Cooper et al., 2010; Propper et al., 2004, 2008) and internationally (Kessler and McClellan, 1996, 2011; McClellan and Staiger, 1999; Shen, 2003). Yet even where the relationship between treatments and outcomes is established, observed variation across hospitals and across time may continue to reflect considerable random variation and not true changes in quality (Dimick and Welch, 2008).

Other measures are often also presented alongside mortality indicators. At the hospital level another common proxy for quality is readmission rates, with higher emergency readmissions in particular, thought to be indicative of worse quality. This measure has become increasingly popular despite the fact that it cannot always be attributed to the overall quality of care delivered by the hospital. McClellan and Staiger (1999) note that high readmissions may not signal poor quality when hospital treatment is lowering mortality rates and more severely ill patients are surviving initial disease episodes. Under such circumstances higher readmission rates might be expected. Moreover, readmissions may reflect poor quality care in other parts of the health care system (e.g. the primary care sector), or individual behavioural factors beyond hospital control (e.g. poor adherence to medicines). 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 this indicator that can be attributed poor quality care. Their own study, using different readmission indicators for a range of conditions, estimated the variation for readmissions associated with improved quality of care to be between 9% and 50%. They did 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, these indicators may provide more appropriate measures of quality. Fischer et al. (2012) also suggest that there is little evidence to indicate that readmissions are related to quality of care. However, after initial use in the US,

there are now a growing number of European countries employing readmission rates as a health service outcome measure (Klazinga, 2011). At present, the use of hospital readmission to measure quality while increasing in practice, remains disputed. Nonetheless, in April 2011, the British government introduced controversial rules aimed at penalizing emergency readmissions by barring payment for them.

Thus, while there remains doubt over their acceptability, there is increasing employment in the UK of measures to monitor hospital performance based on routinely collected data. Given this environment, we wish to examine a method proposed by McClellan and Staiger (1999), which helps overcome a number of issues associated with measuring hospital quality. Hospital quality is difficult to observe directly. To measure true quality across a range of relevant dimensions and contain random noise can involve highly complex modeling, as discussed by Jones and Spiegelhalter (2012). The broad measures developed by McClellan and Staiger (1999) use data from specific conditions where latent quality is linked to particular outcomes, for example AMI 30-day mortality and emergency 28-day readmissions, in a relatively simple two-step modeling process. In the US setting they find that these measures appear to be good indicators of performance (Kessler and McClellan, 1996; McClellan and Staiger, 1999) and may be even as good as more detailed measures of performance which are more costly to obtain (Dranove et al., 2002). While no comparable research has yet been done in the UK to support either of these findings, as noted above crudely adjusted hospital mortality rates and AMI mortality rates have been used by numerous studies as indicators of performance in the UK and to draw conclusions on the effectiveness of policy, such as increasing hospital competition (Bloom et al., 2010; Cooper et al., 2010; Propper et al., 2004). Our contention is that the method proposed by McClellan and Staiger (1999) can improve on these crude signals of hospital quality at low cost in terms of data collection and computation. Moreover while the method can be used to assess past quality, it can also be used to predict future quality levels, an attractive attribute emphasized by Jones and Spiegelhalter (2012).

An important characteristic of the McClellan and Staiger (1999) technique it uses smoothing to improve the quality signal of noisy, naive estimators - where the smoothing is informed by the data structure both within individual providers and over time. As noted by Jones and Spiegelhalter (2012), smoothing of healthcare provider performance measures is known to lead to advantages in terms of predictive ability. Indeed, many outcomes measures reported by ONS, such as HSMRs or the SHMI, are smoothed over a three-year period for these reasons. Recently a number of such “bidirectional” smoothing or shrinkage estimators have been employed within this field (Jones and Spiegelhalter, 2012; Normand et al., 1997). Jones and Spiegelhalter (2012) assess the predictive performance

of a number of these types of smoothing estimators. They note that while some of these estimators can be applied with good predictive effect to UK data, the models are complex in terms of data manipulation, estimation and computation. In concluding their assessment they highlight the promise of hierarchical time series models, but they also note that such models are complex and tend to rely on specialized software, particularly if the within provider, time smoothing technique employs Markov chain Monte Carlo techniques to model dynamic processes. The McClellan and Staiger (1999) technique is closely related to these empirical Bayes modeling approaches, but is simpler to implement, easier to interpret and does not require anything other than standard software. As Jones and Spiegelhalter (2012) themselves note, these attributes alone warrant further investigation of this specific method. A primary objective of this paper is therefore to assess the robustness of this specific technique to determine whether it is useable within the English NHS hospital data setting given the ease of implementation. Indeed, one advantage gained in applying to the English hospital sector is that the data allow the possible extension of the method to incorporate co-morbidity data. The paper also assesses the robustness of the McClellan and Staiger (1999) approach by applying to a wider range of hospital outputs and to smaller sample sizes than was the case in their original study.

The paper continues as follows. Section 2 outlines the English data we use for the analysis. We use two treatment conditions in the application of this method within the English NHS setting; Acute Myocardial Infarction (AMI) and Hip Replacement drawing on Hospital Episode Statistics (HES). Section 2 also gives definition to our risk adjusters and the measures of hospital quality we analyze. For each of the two conditions, these are 30-day in-hospital mortality, year-long mortality, 28-day emergency readmission rates, and year-long readmission rates. Section 3 describes the methodology and the formal model used in the analysis. Section 4 reports our results. Section 5 gives concluding remarks highlighting the ease of application and the usefulness of the method in reporting a more credible measure of underlying hospital quality than cruder risk-adjusted measures.

2 Data

Hospital Episode Statistics (HES) are used to conduct this analysis. The HES database has been in existence since 1987 and is used by the Department of Health to provide performance information at the hospital level in England (Spiegelhalter et al., 2002). These data contain individual records for over 15 million NHS patients admitted to English hospitals each financial year (April 1 to March 31), with information on all medical and

surgical specialties, including private patients treated in NHS hospital trusts. Since the introduction of the internal market into the NHS, HES data have also been used for contracting between purchasers and providers. The data available in the HES database contains patient characteristic data (e.g. gender, age), clinical information (e.g. diagnoses, procedures undergone), mode of admission (emergency, elective), outcome data (mortality, readmission, discharge location) as well as detail 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. HES data can also be linked to other data sources such as the death registries, to provide additional information such as death rates at different intervals (30-days and yearly), readmission rates and further details on the patient, including information on co-morbidities and on socioeconomic characteristics. Diagnosis of patients are coded using ICD-10 (International Classification of Diseases, tenth revision) codes, while procedures use the former UK Office of Population Censuses and Surveys classification (OPCS4).

Data on gender and age are used as risk-adjusters in our analysis, as discussed below, as is information on whether the treatment undergone was an elective procedure or not. In addition 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 illness. This index is constructed by assigning a score to each of 22 conditions reflecting the risk of dying from that condition, and aggregating these scores through summation (Charlson et al., 1987). The conditions used are Myocardial Infarction, Congestive Cardiac Failure, Peripheral Vascular disease, Dementia, Cerebrovascular disease, Chronic Lung disease, Connective Tissue disease, Ulcer, Chronic Liver disease, Hemiplegia, moderate or severe Kidney disease, Diabetes, Diabetes with complications, Tumor, Leukemia, Lymphoma, moderate or severe Liver disease, Malignant Tumor, Metastasis and AIDS. Given the range of clinical conditions used, the index will inherently be a better indicator of co-morbidity for some conditions over others. For example, as it controls for many heart conditions specifically, it is more likely to be correlated with AMI than Hip Replacement. That said, even with this limitation it improves on the US data which were available to McClellan and Staiger (1999), allowing greater refinement in the risk-adjustment. Finally, socioeconomic status was measured using the Carstairs index of deprivation. This index is based on four census indicators: social class, car ownership or lack of, overcrowding in accommodation 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, calculated for the postcode area of the individual and then applied individually.

We restrict presentation to two conditions: AMI and Hip Replacement, although we have analyzed a wider set of conditions. These two conditions were chosen to illustrate the performance of this technique, as risk adjusted mortality and readmissions indicators for these conditions are commonly used in hospital report cards. Moreover, these two conditions have an established link between treatment and survival, a high number of admissions each year ensuring large sample sizes, and AMI cases represent emergency care, while Hip Replacement cases represent elective care. The data extraction was based on the ICD-10 and OPCS 4.3 classification codes as indicated in Table 1. Any hospital trust that had less than 10 admissions for these conditions throughout the entire period of analysis was dropped from the analysis. Given our focus on NHS hospital in-patient treatments NHS primary care trusts, private trusts and social care trusts, and all patients coded as day cases, were also excluded.

For the two conditions, AMI and Hip Replacement, we risk-adjust four crude outcome measures, namely 30-day in hospital mortality ($D30_{ht}$), year-long mortality ($D365_{ht}$), 28-day emergency readmission rates ($R28_{ht}$) and year-long readmission rates ($R365_{ht}$). These outcome measures are routinely available. The 30-day hospital mortality and the 28-day readmission rates are commonly used for the reporting of hospital quality, hence our focus on these. The longer term outcomes are important to capture on-going quality of treatment issues, but this does mean matching data from other sources, in this case to UK Office of National Statistics (ONS) mortality data which is linked across individual patients. Both matches are straightforward. Identification of short-term and long-term readmission rates also allows us to test whether any improvement in mortality does affect the signal of quality contained in readmission data. As noted, if more severely ill patients are surviving because of improved quality of treatment, this may subsequently increase rather than reduce the readmission rate.

The analysis is conducted using HES data as a pooled cross-section. As Table 1 reports, English hospital performance is considered over the period 2000 to 2005 for both AMI and Hip Replacement. This gave an average of 50,613 individual cases each year for hip replacement and 64,208 individual cases on average each year for AMI. A unique patient identifer links the data across years. There were 139 hospitals involved in treating the Hip Replacement patients and 177 hospitals associated with treating the AMI patients. The table also shows the crude mortality rates (CMR) and crude readmission rates (CR) for the sample of hospitals across all years, reported per 1,000 deaths or readmissions respectively. As expected, AMI mortality and readmissions are higher than Hip Replacement.

Tab. 1: Summary Statistics of the Sample.

Condition	ICD-10/ OPCS4.3 codes	Years Analyzed	Mean cases per year	Number of hospitals
AMI	ICD-10: 121	2000-2005	59,678	177
Hip	OPCS4.3: W37-W39 W46-W48 W58	2000-2005	47,472	139
Condition/ year	30-day mortality ($D30_{ht}$) CMR per 1,000 (σ)	365-day mortality ($D365_{ht}$) CMR per 1,000 (σ)	28 day readmissions ($R28_{ht}$) CR per 1,000 (σ)	365-day readmissions ($R365_{ht}$) CR per 1,000 (σ)
Hip 2000	3.36 (4.54)	15.66 (13.56)	35.67 (24.60)	71.42 (46.79)
Hip 2001	3.28 (4.05)	19.32 (15.39)	37.42 (26.77)	73.86 (49.20)
Hip 2002	3.13 (3.85)	18.48 (14.36)	36.29 (24.70)	74.26 (48.86)
Hip 2003	3.26 (4.11)	19.25 (14.67)	36.22 (25.31)	74.62 (48.69)
Hip 2004	3.04 (3.37)	18.45 (14.50)	36.65 (24.98)	75.68 (48.80)
Hip 2005	2.90 (3.48)	14.10 (12.19)	38.69 (26.70)	74.45 (49.22)
AMI 2000	73.08 (48.87)	119.57 (76.44)	56.21 (35.84)	125.17 (76.92)
AMI 2001	72.48 (46.65)	120.59 (75.10)	56.30 (34.85)	116.87 (70.82)
AMI 2002	72.73 (45.90)	124.24 (76.20)	58.55 (35.45)	121.69 (72.44)
AMI 2003	69.12 (44.21)	123.88 (76.05)	61.89 (37.15)	126.57 (74.43)
AMI 2004	63.94 (41.51)	116.67 (72.38)	59.31 (36.75)	122.79 (73.82)
AMI 2005	59.44 (40.80)	105.27 (67.79)	62.39 (37.87)	118.00 (70.00)

3 Empirical Model

McClellan and Staiger (1999) assume true hospital quality is a latent variable, but that valuable information on this latent measure of quality can be returned through a two step, smoothing procedure. Following their methodology, the first step of our analysis uses the four unadjusted individual outcome measures, [($D30_{ht}$), ($D365_{ht}$), ($R28_{ht}$) and ($R365_{ht}$)], for each of the two conditions, (AMI and Hip Replacement), and risk-adjusts them through linear regression against individual patient characteristics. This following first stage, risk-adjustment regression equation is run on each of the individual outcome measures for each year of the analysis separately:

$$Y_{ih} = \mu_h + \sum_{n=1}^{m} \phi_{im} X_{ih} + u_{ih}\,, \tag{1}$$

where Y represents the quality outcome measure, i indexes the individual patient, and h the hospital they were treated in. The X_i represents a set of m individual control

variables for patient characteristics (in our case age, gender, socioeconomic deprivation, co-morbidities and whether an elective or emergency hospital admission), across each hospital h they were treated in.

The equation is run with no constant term, and the term μ_h is of greatest interest as it returns a hospital fixed effect, which followingMcClellan and Staiger (1999), is taken to be a proxy measure of latent hospital quality. These μ_h are, therefore, estimates of true hospital quality for each of the four outcome measures gained by removing noise through risk-adjustment. As noted, the equation is run separately for each year, for each of the four unadjusted quality outcome measures, [$(D30_{ht})$, $(D365_{ht})$, $(R28_{ht})$ and $(R365_{ht})$], applied to each of the two conditions, (AMI and Hip Replacement). The fixed effects, returned from each yearly regression, are used to define a new vector, Q_h, of risk-adjusted hospital quality, for each of the four outcomes analysed and for each of the two conditions. Assuming T time-periods and K measures of quality, the hospital quality vector, Q_h, has dimensions $[1 \times TK]$. This vector, Q_h, is then assumed to represent the following relationship to the latent (true) hospital quality:[1]

$$Q_h = q_h + \epsilon_h \,, \tag{2}$$

where q_h represents the $[1 \times TK]$ vector of the true (latent) hospital quality for hospital h, and ϵ_h is the estimation error (assumed to have mean zero and be uncorrelated with q_h). Thus, equation (2) assumes that the estimated risk-adjusted hospital quality fixed effects, Q_h, are suitable predictors of hospital (latent) quality, and anything not captured by these estimates is incorporated in the error term, ϵ_h. It is the removal of the error term, ϵ_h, from the estimated hospital quality fixed effects, Q_h, which allows further improvement in the measures of hospital quality. The error term, ϵ_h, is related to the patient level regressions (equation (1)), in particular, to the variance-covariance of the regression estimates Q_h. That is:

$$E(\epsilon_h^{'}\epsilon_h) = S_h$$
$$E(\epsilon_h^{'}\epsilon_h) = 0$$

where S_h represents the variance-covariance matrix of the hospital effects estimates for

[1] In our case for each of the two conditions, AMI and Hip Replacement and with 4 measures of quality, [$(D30_{ht})$, $(D365_{ht})$, $(R28_{ht})$ and $(R365_{ht})$], and yearly observations for 2000-2005 the vector, Q_h, has dimension [1 x 24].

hospital h for each year. The true latent hospital quality measure, q_h, is not directly observable, but McClellan and Staiger (1999) outline a method to estimate q_h. They propose creating a linear combination of each hospital's observed risk-adjusted measure of hospital quality for each year, in such a way that it minimizes the mean squared error of the predictions. This could be conceptualized as running the following (hypothetical) regression for each year:

$$q_h = Q_h\beta_h + \omega_h \tag{3}$$

They note, however, that equation (3) cannot be estimated directly, precisely because q_h represents the true, unobserved (latent) quality for the defined outcomes, in each hospital, h, for each year. Assuming K measures of quality and T years, note that Q_h is a $[1 \times TK]$ vector and the optimal β for each quality measure, k, varies by hospital and year, given equation (2). The measurement challenge is to return the true hospital quality, q_h for each quality measure k in each year, from the noisy estimate Q_h. McClellan and Staiger (1999) use a shrinkage estimate of q_h to further reduce noise in the risk-adjusted measures of hospital quality, without distorting the true quality measure. This is analogous to the use of smoothing techniques as outlined, for example, in Jones and Spiegelhalter (2012); Titterington et al. (1985).

Their insight is that, while equation (3) can not be estimated directly as q_{ht} is not observed, the parameters of the hypothetical regression represented by (3) can be retrieved from the existing data. They proceed by noting that the minimum least squared estimate, for each of the k quality measures over each of the t time periods, can be given by:

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

where

$$\beta = [E(Q_h^{'}Q_h)]^{-1}E(Q_h^{'}q_h)\,. \tag{4}$$

This best linear estimate can be returned using the following definitions:

$$E(Q_h^{'}Q_h) = E(q_h^{'}q_h) + E(\epsilon_h^{'}\epsilon_h) \tag{5}$$

$$E(Q_h^{'}q_h) = E(q_h^{'}q_h)\,, \tag{6}$$

where $E(Q_h^{'}Q_h)$ is is the expected value of the products and cross-products of the hospi-

tal fixed effects, which is gained from the first-stage patient level regressions, and where $E(\epsilon_h^{'}\epsilon_h)$ is variance-covariance matrix of the distrubances associated with these fixed effects, which again is gained from the first-stage patient level regressions. Let us call this latter estimate S_h. Noting that the S_h vary among hospitals, then $E(q_h^{'}q_h)$ can be estimated by rearranging (5) such that $E(Q_h^{'}Q_h - S_h) = E(Q_h^{'}q_h)$. Subsequently (6) becomes $E(Q_h^{'}Q_h - S_h) = E(q_h^{'}q_h)$. Using these estimates and equalities and inserting the relevant estimates into equation (4) allows derivation of the desired least squares parameters in(3). The shrinkage estimates, q_h, can then be easily estimated by individual hospital for each year using observed values. McClellan and Staiger (1999) define this estmate as:

$$\hat{q}_h = Q_h[E(Q^{'}{}_hQ_h)]^{-1}E(Q_hq_h) = Q_h[E(q_h^{'}q_h) + E(\epsilon_h^{'}\epsilon_h)]^{-1}E(q_h^{'}q_h)\,. \tag{7}$$

where the $E(q_h^{'}q_h)$ have been calculated as above. McClellan and Staiger (1999) refer to these estimates as 'filtered estimates' as they optimally filter out the estimation error from the risk-adjusted quality measures.

These filtered estimates have a number of attractive properties. First, they allow information for many different outcome indicators to be combined in a systematic manner; as noted we use four outcome measures as defined above [$(D30_{ht})$, $(D365_{ht})$, $(R28_{ht})$ and $(R365_{ht})$], which we apply to each of the two conditions, (AMI and Hip Replacement). 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 (we use STATA). Note further that using estimates from (5) and (6), the R-squared statistic can be calculated, based on the least squared formula, returning a simple measure of explained variation. However, the filtered estimates will be sensitive to first stage estimation. In particular, any unobserved characteristics which systematically affect the variance-covariance matrix may influence the size of the filetered estimate.

As equation (1) is run with no constant, in it important to ensure that the assumptions made above are still plausible. In particular, that the estimation error in (2) has mean error zero or that the covariance matrix for the parameter estimates, Q_h, are unbiased. We test this assumption first by re-running equation (1) with a constant and use a t-test to examine whether the constant is statistically significant, thus suggesting it is significantly different from zero. Our results show that significance of the constant is not maintained (at well below conventioan standards) in any of the AMI models, and is only significant in the 2001 models for Hip Replacement. We further test this assumption by extracting the residuals from the models, run as specified in (1) and check their descriptives to ensure that that the these assumptions hold. We find that in all cases the mean is sufficiently

near zero.

Estimation of equation (7) is then used to further smooth these filtered hospital quality measures through a further estimation step. This further step utilizes the information across the different time periods to additionally improve these risk-adjusted, filtered quality outcome measures. Hence the method can be considered a form of bi-directional smoothing estimation, where the measures are able to reduce noise within individual hospitals, and across time periods. This is undertaken using a Vector Autoregression (VAR) model, with further structure imposed on the filtered quality estimates by assuming that each quality measure is reflective of its past performance, plus a contemporaneous shock that may be correlated across the different outcome measures.

Noting that we have K measures of quality, which are inter-related and contain signals from past performance, and T years, a first order VAR model is specified to return the estimate $\hat{q}_k^{(ht)}$, which is a $(1 \times K)$ vector incorporating values from time periods t and $t-1$. That is:

$$\hat{q}_k^{(ht)} = q_k^{(h,t-1)}\Phi + v_k^{(ht)}\,. \tag{8}$$

We define Φ as a $(K \times K)$ matrix containing the estimates of the lag coefficients. We can further estimate $Z = V(v_{ht})$, the $(K \times K)$ variance matrix of the residuals, and $\Gamma = V(q_{ht})$, the $(K \times K)$ initial variance matrix from the first year of the data sample. The VAR structure and (7) implies:

$$E(Q_k^{'(h)}Q_k^{(h)}) - S^{(h)} = E(q_k^{'(h)}q_k^{(h)}) = f(\Phi, Z, \Gamma)\,. \tag{9}$$

where the only undefined term is S_h which is the estimation error. Using the parameters estimated from the VAR model, we then estimate equation (9) to return non-stochastic smoothed estimates of quality incorporating the times series data, we refer to these as 'smoothed outcome measures' which are the time-smoothed, filtered estimates. These bidirectional smoothed estimates McClellan and Staiger (1999) refer to as 'predicted' estimates, but we prefer the term (bidirectional) smoothed estimators, which we adopt from here onwards.

Incorporating the VAR approach allows the further production of forecast measures of hospital quality. This, as noted by Jones and Spiegelhalter (2012), is particularly important as it is the prediction of future hospital quality that is of main interest to current observers. The first stage of the analysis, to return the $\hat{q}_h^{(k)}$ (equation 7) was performed

using the statistical package STATA, and the second stage, to return $\hat{q}_k^{(ht)}$ (equation 8) was undertaken in eViews merely because this includes more options to perform time-series analyses, being especially straightforward in the estimation of VAR models and forecasts.

To assess the statistical performance of the estimated quality outcome measures we then use a series of metrics. The filtered estimates are assessed by a signal to noise ratio estimate, while two constructed R-squared measures allow us to assess the fit of the smoothed and forecasted outcomes values. We also use the Root Mean Squared Error (RMSE) to assess the forecasting ability of the smoothed estimators.

The first two performance measures were proposed by McClellan and Staiger (1999). As the VAR model is able to extract the underlying quality signal of each outcome measure, and hence the signal variance, V_{ht}, from the original hospital data, this allows us to define a signal-to-noise ratio for each of the quality outcome measures. Using the signal variance together with the estimation error contained in each measure, defined as S_h in equation (9) above, defines the signal to noise ratio as:

$$\text{Signal}/(\text{Signal} + \text{Noise}) = V_{(ht)}/(V_{(ht)} + S_{(ht)}) \tag{10}$$

In order to assess the ability of the smoothed quality outcome measures to estimate variation in true quality, McClellan and Staiger (1999) construct an R-squared measure drawing on the standard R-squared formula:

$$R^2 = 1 - \frac{\sum_{h=1}^{N}(\hat{u}_k^{(ht)})^2}{\sum_{h=1}^{N}(q_k^{(ht)})^2}\,. \tag{11}$$

As the purpose of this goodness of fit measure is to assess the degree to which the estimated outcomes minimize the mean square error of the prediction, the numerator should measure prediction error, such that:

$$\hat{u}_k^{(ht)} = q_k^{(ht)} - \hat{q}_k^{(ht)}.$$

Since $q_k^{(ht)}$ is not observed, estimates must be used for both the numerator and the denominator. McClellan and Staiger (1999) propose the use of $E(q_k^{'(h)} q_k^{(h)})$ for the denominator and $E(q_k^{(h)} - \hat{q}_k^{(h)})'(q_k^{(h)} - \hat{q}_k^{(h)})$ for the numerator. Both of these can be estimated for each year using 5 and 6 above. The R-squared measures are calculated for the smoothed quality outcome estimates, and presented separately for each treatment condition.

To evaluate the usefulness of the bivariate smoothed estimator in forecasting future events we follow Jones and Spiegelhalter (2012) and use the empirical root-mean-squared error (RMSE):

$$RMSE = \sqrt{\{\frac{1}{n}\sum_{h=1}^{n}(q_k^{(h)} - \hat{q}_k^{(h)})^2\}} \tag{12}$$

where $\hat{q}_k^{(h)}$ indicates the forecasted values of the quality estimates, $q_k^{(h)}$, for the n hospitals in the sample.

4 Results

The methodology outlined above was applied to data on our four hospital outcome measures, (30-day in-hospital mortality ($D30_{ht}$), year-long mortality ($D365_{ht}$), 28-day emergency readmission rates ($R28_{ht}$) and year-long readmission rates ($R365_{ht}$)), for each of the two treatment conditions, AMI and Hip Replacement. The results illustrate how well the bivariate smoothed estimates perform in measuring within sample hospital quality through the use of the filtered and bivariate smoothed estimates, and forecasting out of sample hospital quality through the use of the forecast estimates. Diagrammatically comparing the filtered and smoothed measures of hospital quality, measuring the signal to noise ratio of the filtered estimates and estimating the goodness of fit measures of both smoothed and forecasted outcome measures suggests this method is a simple and robust means of evaluating underlying true hospital quality. These results to support this conclusion are now presented.

For each of the two conditions, AMI and Hip Replacement, Tables 2 and 3 report the basic parameters associated with the bivariate smoothed estimator, as based on the VAR estimates of interest which are constructed using the full, available data: that is, the lag coefficients to assess persistence; the variance and correlation between the residuals for each effect; and the initial variance and correlation of the time varying effects with the first sample year. The VAR parameters for each of the two conditions are estimated using the information on all of the four outcome measures, (i.e. 30-day in hospital mortality ($D30_{ht}$), year-long mortality ($D365_{ht}$), 28-day emergency readmission rates ($R28_{ht}$) and year-long readmission rates $R365_{ht}$). The VAR specification is as given in equation (8), although other specifications, with different lag lengths were tested. The inclusion of additional lags

yielded similar scores, sometimes marginally better, as judged by the Akaike information criterion and the Schwartz criterion. Given the small improvement, we chose to use the VAR(1) specification for all models as it promotes ease of interpretation and is relatively parsimonious with the data. All VAR models were tested for stability and passed unit root tests for stationarity.

The (bidirectional) smoothed parameter estimates reported in Tables 2 and 3 indicate the effect that past values of each of the four outcome measures have on their own performance for the two conditions, AMI and Hip Replacement respectively. The results suggest that none of the outcome measures are strongly persistent. For AMI, long term hospital mortality, $D365_{ht}$, is slightly more persistent than the other four outcome indicators with the value of the coefficient on its own lag taking a value of approximately0.18. In the case of Hip Replacement, short term mortality, $D30_{ht}$ is slightly more persistent than the other outcomes and take a value of 0.10.

Two robustness checks were performed to determine to what extent the persistence estimates reported in Tables 2 and 3 are influenced by the removal of more of the noise from the risk adjusted estimates to create the filtered estimates, and by incorporating other outcome measures into the VAR models. The first test was undertaken by running the same VAR model specification using simple risk adjusted metrics - the fixed effects from equation 1, for each year of the sample - instead of the filtered estimates. The results on this check show the AMI results to be very similar, although it suggests that all outcomes apart from short term mortality, $D30_{ht}$ are more persistent if the filtered measures are used. However, the Hip Replacement results suggest that the persistence for all measures is lower when the filtered measures are used, but particularly for the long term outcomes, $D365_{ht}$ and $R365_{ht}$. The second test re-estimates the VAR using only each individual outcome measure and not including the other three outcome measures. This tests how much estimated persistence is influenced by other past outcomes. The results when running the VAR on the individual outcomes alone do change the coefficients, in some cases making them more persistent while in others less, however in all instances the effect is small. For example in AMI, running the VAR on each of the outcomes alone results in more persistent readmission outcomes, and less persistent mortality outcomes. For Hip Replacement, the effects are much smaller on all outcome measures aside from short term readmissions, $R28_{ht}$, which come out to be more persistent.

As illustrated in Table 2, the standard deviation of the residuals indicates a variation of about 0.05 in short-term AMI mortality rates for the individual hospitals over time, and a variation of nearly 0.06 in long-term mortality rates. As reported in Table 3,

there is much less variation in short- and long-term mortality following Hip Replacement, with a variation of almost 0.1 for short-term mortality, and approximately 0.04 in long-term mortality, possibly reflecting the low absolute mortality rates for this condition. For both conditions, readmissions are subject to less variation than mortality across the sampled hospitals. The variation in emergency 28-day readmissions is notably 0.03 for both conditions, while the variation in long-term readmissions is slightly larger for both conditions; at just under 0.05.

The time-varying standard deviations from the initial year of the sample, provide information on the annual variation across hospitals associated with each outcome measure. Short-term mortality varies the most for AMI at around 0.85 and is only 0.08 for Hip Replacement. Long-term mortality has a higher standard deviation for Hip Replacement of approximately 0.2, but remains at approximately 0.8 across hospitals for AMI. The results for AMI no doubt reflect the effect of improved treatments over time. Readmissions have a relatively low standard deviation for both conditions, with short-term readmissions indicating a variation of about 0.02 for both conditions, and long-term variation of approximately 0.05 for AMI and 0.1 for Hip Replacement. Taken at face value, the explained variation in quality across time and hospitals therefore appears relatively high in the majority of outcomes.

The correlation across the four different outcome indicators is also of interest, and can be assessed both in terms of the VAR residuals, as well as over time by comparing later measures to the initial year of the sample. The correlation of residuals for AMI indicate a negative association between $D30_{ht}$ and $R28_{ht}$, and a negative association between $D30_{ht}$ and $R365_{ht}$. The Hip Replacement results show a negative association between short term mortality $D30$ and long term readmissions $R365_{ht}$, although not between short term mortality and short term readmissions, $R28_{ht}$. For both conditions, there is a positive association between $R28_{ht}$ and $R365_{ht}$ and $D30_{ht}$ and $D365_{ht}$. These are important results, especially given the UK government's linking of emergency re-admission rates to financial penalties. The negative correlations between mortality and re-admission rates, possibly reflects that treatment improvements over time are leading to "less healthy" individuals surviving, with subsequent increases in re-admissions. The correlation between outcome measures with the outcomes in the initial year of the sample, also indicates a negative association between the outcome measures 30-day mortality and both short and long term re-admissions for AMI. While the positive association between $D30_{ht}$ and $D365_{ht}$, and between $R28_{ht}$ and $R365_{ht}$ is still observed in the comparisons with the initial samples for both conditions, indicating that the short term outcomes are associated with the long term outcomes.

Tab. 2: Estimates of multivariate VAR(1) parameters for hospital specific effects (AMI).

	AMI			
	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D30_{h(t-1)}$	0.0919	0.0645	-0.1199	-0.1421
	(0.0470)	(0.0311)	(0.0578)	(0.0685)
$R28_{h(t-1)}$	0.0490	-0.0389	-0.2065	-0.0024
	(0.0610)	(0.0404)	(0.0750)	(0.0890)
$D365_{h(t-1)}$	-0.0838	-0.1125	0.1807	0.2086
	(0.0368)	(0.0244)	(0.0452)	(0.0536)
$R365_{h(t-1)}$	-0.0690	0.0078	0.0016	0.0676
	(0.0283)	(0.0187)	(0.0348)	(0.0413)
Constant	-0.0690	0.0078	0.0016	0.0676
	(0.0017)	(0.0012)	(0.0021)	(0.0025)
Residuals				
S.D. dependent	0.0487	0.0324	0.0603	0.0712
Correlation of residuals ($D30_{ht}$)	-	-0.1895	0.5717	-0.3184
Correlation of residuals ($R28_{ht}$)	-0.1895	-	-0.1428	0.4208
Correlation of residuals ($D365_{ht}$)	0.5717	-0.1428	-	-0.1068
Correlation of residuals ($R365_{ht}$)	-0.3184	0.4208	-0.1068	-
Initial Conditions				
S.D. dependent in 2000	0.0447	0.0309	0.0564	0.0472
Correlation with $D30_{ht}$ 2000	-	-0.3030	0.8486	-0.3571
Correlation with $R28_{ht}$ 2000	-0.3030	-	-0.2328	0.6043
Correlation with $D365_{ht}$ 2000	0.8486	-0.2328	-	-0.1831
Correlation with $R365_{ht}$ 2000	-0.3571	0.6043	-0.1831	-
Included observations (Hospitals):	165			
Included observations (Individuals):	770			

Sample (adjusted): 2001 2005

Standard errors in ( )

Tab. 3: Estimates of multivariate VAR(1) parameters for hospital specific effects (Hip).

	Hip Replacement			
	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D30_{h(t-1)}$	0.1002	0.4626	0.3970	-0.0427
	(0.0385)	(0.1219)	(0.1794)	(0.2410)
$R28_{h(t-1)}$	-0.0410	0.0713	0.0986	0.2294
	(0.0125)	(0.0396)	(0.0582)	(0.0782)
$D365_{h(t-1)}$	0.0067	-0.0721	0.0133	-0.1432
	(0.0092)	(0.0292)	(0.0429)	(0.0577)
$R365_{h(t-1)}$	-0.0153	0.1234	-0.2062	0.04462
	(0.0075)	(0.0237)	(0.0348)	(0.0468)
Constant	-4.81e-05	-0.0051	7.18e-05	0.0003
	(0.0003)	(0.0009)	(0.0013)	(0.0018)
Residuals				
S.D. dependent	0.0077	0.0247	0.0363	0.0477
Correlation of residuals ($D30_{ht}$)	1	0.0143	0.2036	-0.0499
Correlation of residuals ($R28_{ht}$)	0.0143	1	0.1007	0.4116
Correlation of residuals ($D365_{ht}$)	0.2036	0.1007	1	0.4370
Correlation of residuals ($R365_{ht}$)	-0.0499	0.4116	0.4370	1
Initial Conditions				
S.D. dependent in 2000	0.0076	0.0217	0.0204	0.0402
Correlation with $D30_{ht}$ 2000	1	0.0830	0.6241	0.0810
Correlation with $R28_{ht}$ 2000	0.0830	1	0.0766	0.5269
Correlation with $D365_{ht}$ 2000	0.6241	0.0766	1	0.3084
Correlation with $R365_{ht}$ 2000	0.0810	0.5269	0.3084	1
Included observations (Hospitals):	138			
Included observations (Individuals):	685			

Sample (adjusted): 2001 2005

Standard errors in ( )

Figures 1 and 2 present the signal to noise ratio for the four outcome measures across the two conditions, AMI and Hip Replacement for the year 2005. By plotting 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), this plot provides statistical information on the level of "true" signal contained in each of the quality measures relative to the underlying noise in the estimates. Both Figures indicate that the signal to noise ratios observed for all four outcomes rise as the number of cases increase. In some cases the smoothed quality outcome measures represent relatively robust estimates of true underlying hospital quality, once cases go above 200, such as the readmission metrics for AMI, and most of the Hip Replacement outcomes. Of the four outcome measures, the two mortality outcomes have the weakest signal for AMI, while for Hip Replacement the signal to noise ratios performed better for long term mortality. These results suggest that as the sample exceeds 300 patient, the indicators can be used to reliably detect a large amount of quality outcome variation across hospitals. However, short term outcomes, such as 30 day in hospital mortality are still subject to a largest degree of noise for both conditions, even when looking across a high number of cases.

Fig. 1: Signal to noise ratio for smoothed AMI measures (year 2005).

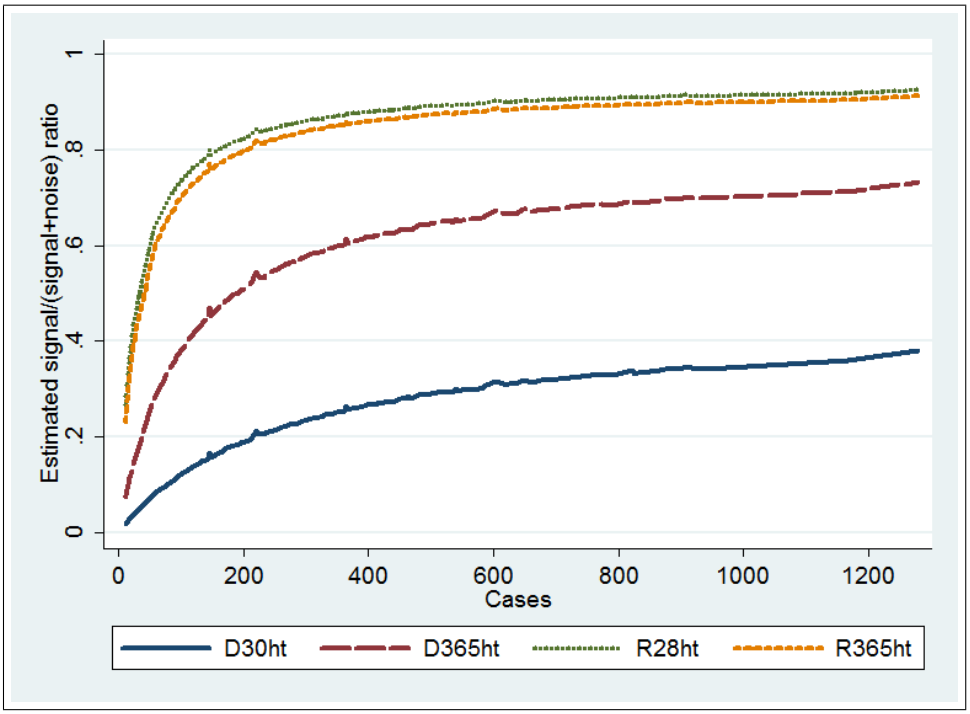


Fig. 2: Signal to noise ratio for smoothed Hip measures (year 2005).

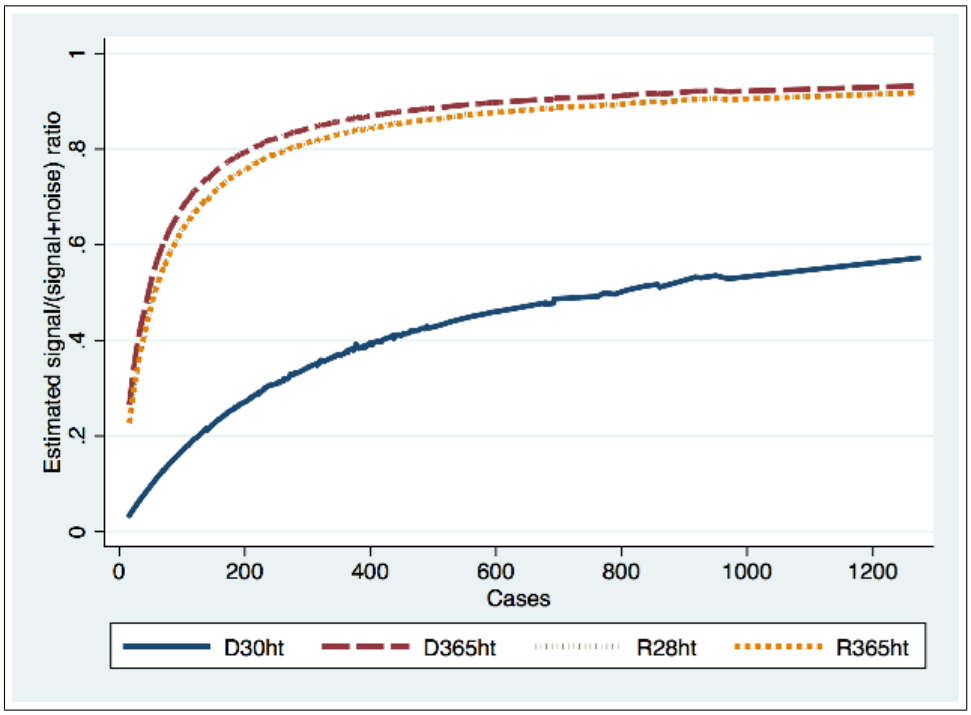
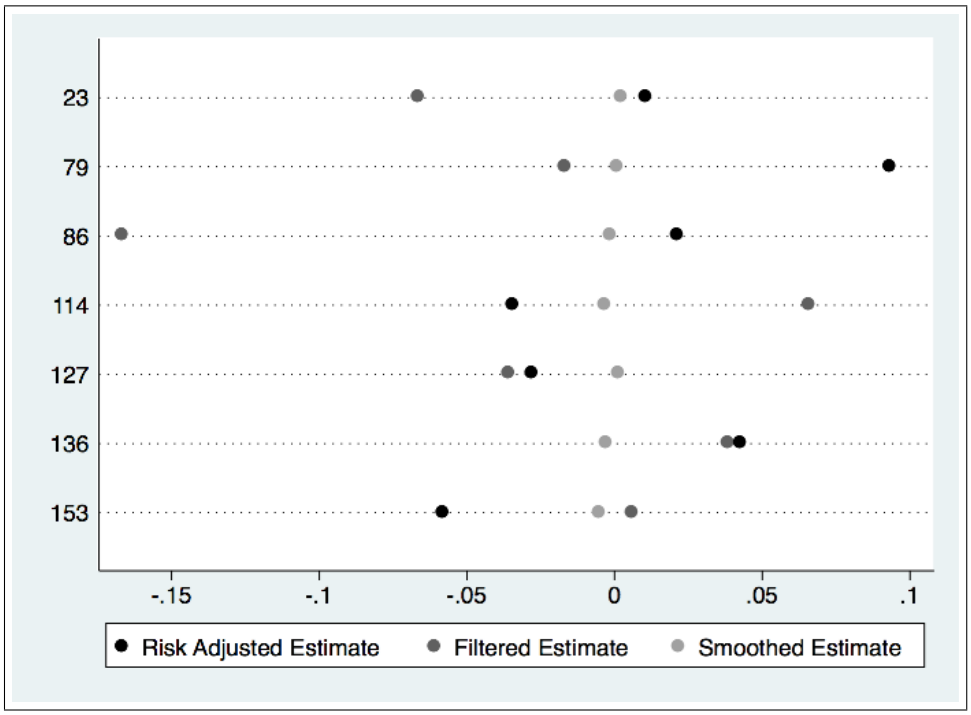


Figure 3 presents the filtered outcome measures alongside the smoothed measures of 30 day mortality for a random subset of hospitals for the year 2005, as well as comparing them to the simple risk adjusted rate (or the fixed effects from equation 1). Each horizontal line of the figure reprsents the three estimates for a particular hospital. The outcome measures are normalized such that the mean value is equal to zero, where a value below zero indicates the hospital has below average mortality and vice versa. The hospitals are listed in rank order based on the risk adjusted estimates. Moreover, the risk adjusted estimates can be interpreted as absolute outcome differences; 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, after controlling for patient characteristics. The filtered estimates are estimated using equation 7, and represent cleaned estimates of the risk adjusted quality measures. The smoothed estimates, are derived from the multivariate VAR model, applied to AMI and as presented in equation 2, thus incorporating all of the individual hospital's data from 2000-2005 for all four outcome measures applied to the one condition, AMI. Both the risk adjusted and the filtered estimates exhibit considerably more variation than the smoothed estimates. The figure also indicates that despite being derived from the same underlying mortality rates, the three quality measures do not present a consistent ranking of hospitals; in other words noise does affect ranking.

Fig. 3: Risk Adjusted, Filtered and Smoothed estimates for AMI $D30_{ht}$ (year 2005).



As a further test of the model, we run the VAR separately small and large hospitals. This relates to one of the assumptions made regarding the calculation of the filtered estimates, namely that they will be influenced by any bias present through unaccounted unobservables. Not only does this related to omitted variables but also to any heteroskedasticity related to hopsital size. Seperating the analysis this way will limit the smoothing to hospitals of a similar size, which allows us to explore the effect of heteroskedasticity. We classify hospitals into small and large according to the overall mean cases over the years studied: hospitals with less than an average of 300 cases throughout the period are termed small hospitals, hospitals with averaging over 300 cases over the period are consider large hospitals. We use the number 250 based on our signal to noise ratios, which suggest more volatility in the signal to noise ratios for cases less than 200-300. Our findings suggest that there is more variation across hospitals, particularly for AMI outcomes, across VAR estimates derived from the sample of small hospitals. The technique is weakest when applied to the smaller hospitals as the background noise, possibly as reflecting changes in mean patient characteristics over time, will be greatest in these hospitals.

Tab. 4: Estimates of multivariate VAR(1) parameters for small and large hospitals (AMI).

	Small Hospitals (n<250)				Large Hospitals (n>250)			
	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D30_{h(t-1)}$	0.2415	0.0472	-0.0219	-0.1144	0.0750	-0.0265	0.0796	-0.1088
	(0.0751)	(0.0495)	(0.0935)	(0.0718)	(0.1040)	(0.0720)	(0.1242)	(0.1095)
$R28_{h(t-1)}$	0.0089	0.0049	-0.1242	0.1188	-0.0205	-0.0074	-0.0579	-0.0422
	(0.1248)	(0.0822)	(0.1555)	(0.1194)	(0.1161)	(0.0803)	(0.1387)	(0.1222)
$D365_{h(t-1)}$	-0.1227	-0.0963	0.1626	-0.0050	-0.0642	-0.0260	-0.0733	0.0297
	(0.0577)	(0.0380)	(0.0719)	(0.0552)	(0.0839)	(0.0581)	(0.1002)	(0.0885)
$R365_{h(t-1)}$	0.1111	-0.0780	0.0877	-0.1501	0.0254	0.0512	0.1025	0.1374
	(0.0891)	(0.0587)	(0.1110)	(0.0852)	(0.0811)	(0.0561)	(0.0969)	(0.0854)
Constant	-6.85E-05	-7.15E-05	3.63E-06	4.25E-05	1.88E-09	4.11E-11	-3.75E-09	7.82E-10
	(0.0042)	(0.00274)	(0.00517)	(0.00397)	(0.0014)	(0.0010)	(0.0017)	(0.0015)
Residuals								
S.D. dependent	0.0715	0.0469	0.0885	0.0677	0.0266	0.0185	0.0319	0.0284
Correlation of residuals ($D30_{ht}$)	-	-0.1993	0.5211	-0.3746	-	-0.1110	0.8294	-0.2573
Correlation of residuals ($R28_{ht}$)	-0.1993	-	-0.1754	0.5914	-0.1110	-	0.0292	0.7448
Correlation of residuals ($D365_{ht}$)	0.5211	-0.1754	-	-0.1506	0.8294	0.0292	-	-0.0411
Correlation of residuals ($R365_{ht}$)	-0.3746	0.5914	-0.1506	-	-0.2573	0.7448	-0.0411	-
Initial Conditions								
S.D. dependent in 2000	0.0680	0.0450	0.0851	0.0654	0.0263	0.0191	0.0312	0.2900
Correlation with $D30_{ht}$ 2000	-	-0.2070	0.5368	-0.3771	-	-0.1628	0.8182	-0.2836
Correlation with $R28_{ht}$ 2000	-0.2070	-	-0.2159	0.6080	-0.1628	-	0.0055	0.7542
Correlation with $D365_{ht}$ 2000	0.5368	-0.2159	-	-0.1694	0.8182	0.0055	-	-0.0374
Correlation with $R365_{ht}$ 2000	-0.3771	0.6080	-0.1694	-	-0.2836	0.7542	-0.0374	-
Included observations (Hospitals):								
Included observations (Individuals):								

Sample (adjusted): 2001 2005

Standard errors in ( )

Tab. 5: Estimates of multivariate VAR(1) parameters for small and large hospitals (Hip).

	Small Hospitals (N<250)				Large Hospitals (N>250)			
	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$D30_{h(t-1)}$	0.0229	-0.3335	-0.3531	-0.2028	0.0680	0.0374	0.3420	0.2659
	(0.0641)	(0.3118)	(0.1658)	(0.4360)	(0.0727)	(0.2017)	(0.1787)	(0.3226)

	Small Hospitals (N<250)				Large Hospitals (N>250)			
	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$	$D30_{ht}$	$R28_{ht}$	$D365_{ht}$	$R365_{ht}$
$R28_{h(t-1)}$	0.0499	-0.0477	0.1281	-0.0422	0.0350	0.0497	0.1182	0.0546
	(0.0210)	(0.1020)	(0.0543)	(0.1427)	(0.0363)	(0.1008)	(0.0892)	(0.1613)
$D365_{h(t-1)}$	0.0051	0.1478	0.1134	0.1978	0.0138	-0.0106	0.0264	-0.0782
	(0.0237)	(0.1151)	(0.0612)	(0.1610)	(0.0331)	(0.0919)	(0.0813)	(0.1470)
$R365_{h(t-1)}$	-0.0256	0.0745	-0.0794	0.1479	-0.0594	-0.0541	-0.1174	-0.0730
	(0.0149)	(0.0724)	(0.0385)	(0.1013)	(0.0234)	(0.0649)	(0.0574)	(0.1038)
Constant	0.0001	-0.0001	0.0001	-0.0005	-1.52E-05	0.0003	-0.0001	0.0009
	(0.0002)	(0.0011)	(0.0006)	(0.0015)	(0.0006)	(0.0017)	(0.0015)	(0.0027)
Residuals								
S.D. dependent	0.0040	0.0194	0.0104	0.0272	0.0105	0.0285	0.0257	0.0456
Correlation of residuals ($D30_{ht}$)	-	-0.1345	0.4664	-0.1403	-	0.0197	0.5367	-0.0981
Correlation of residuals ($R28_{ht}$)	-0.1345	-	-0.0005	0.8452	0.0197	-	0.1403	0.6923
Correlation of residuals ($D365_{ht}$)	0.4664	-0.0005	-	0.0168	0.5367	0.1403	-	0.0075
Correlation of residuals ($R365_{ht}$)	-0.1403	0.8452	0.0168	-	-0.0981	0.6923	0.0075	-
Initial Conditions								
S.D. dependent in 2000	0.0068	0.0262	0.0158	0.0393	0.0039	0.0188	0.0105	0.0269
Correlation with $D30_{ht}$ 2000	-	0.0233	0.4345	0.1789	-	-0.1450	0.4943	-0.0112
Correlation with $R28_{ht}$ 2000	0.0233	-	0.0842	0.7542	-0.1450	-	-0.0258	0.8405
Correlation with $D365_{ht}$ 2000	0.4345	0.0842	-	0.1789	0.4943	-0.0258	-	-0.0112
Correlation with $R365_{ht}$ 2000	0.1789	0.7542	0.1789	-	-0.0112	0.8405	-0.0112	-
Included observations (Hospitals):								
Included observations (Individuals):								

Sample (adjusted): 2001 2005

Standard errors in ( )

To assess the performance of the smoothed estimates, Tables 6 and 7 indicate the R-squared estimates as calculated from equation (11) and applied to both conditions. These are presented for the four different quality outcome smoothed measures, using different amounts of past data, and running the models with all outcomes and for models with a single outcome. The R-squared estimates are also calculated for a year in the middle of the sample (2003) and and the last year of the sample (2005) to test how well the smoothing performs for years with no future data. In all cases the R-squared values for the smoothed estimates appear to capture much of the true variation across hospitals for each of the four outcome measures, even when only using one-year of data. For both AMI and Hip,

the R-squared estimates are particularly high for the $D30_{ht}$, $D365_{ht}$ and $R28_{ht}$ at over 90% for both the 2003 and 2005 values. There is little difference between the R-squared estimates for the 2003 values and the 2005 values, apart from the AMI $R365_{ht}$ estimates, which are notably weaker for 2003 as compared to 2005.

Tab. 6: Summary of estimated R-squared values for smoothed estimates using alternative methods of signal extraction. All estimates based on the AMI VAR(1) model from Table 2.

Expected R^2 prediction based on:						
	All 5 years		3 most recent years		Concurrent year	
	All outcomes	Same outcome	All outcomes	Same outcome	All outcomes	Same outcome
$D30_{ht}$						
2003	0.9815	0.9815	0.9815	0.9815	0.9827	0.9821
2005	0.9692	0.9691	0.9692	0.9692	0.9696	0.9692
$D365_{ht}$						
2003	0.9631	0.9631	0.9631	0.9631	0.9699	0.9651
2005	0.9246	0.9246	0.9246	0.9246	0.9248	0.9247
$R28_{ht}$						
2003	0.9936	0.9936	0.9935	0.9935	0.9936	0.9936
2005	0.9915	0.9915	0.9915	0.9915	0.9920	0.9915
$R365_{ht}$						
2003	0.8958	0.8958	0.8948	0.8948	0.7334	0.8948
2005	0.9711	0.9711	0.9711	0.9711	0.9714	0.9711

Tab. 7: Summary of estimated R-squared values for smoothed estimates using alternative methods of signal extraction. All estimates based on the Hip VAR(1) model from Table 3.

Expected R^2 prediction based on:						
	All 5 years		3 most recent years		Concurrent year	
	All outcomes	Same outcome	All outcomes	Same outcome	All outcomes	Same outcome
$D30_{ht}$						
2003	0.9919	0.9919	0.9919	0.9919	0.9943	0.9923
2005	0.9994	0.9994	0.9994	0.9994	0.9994	0.9994
$D365_{ht}$						
2003	0.9830	0.9830	0.9830	0.9830	0.9886	0.9830
2005	0.9960	0.9960	0.9960	0.9960	0.9964	0.9961
$R28_{ht}$						
2003	0.9980	0.9981	0.9981	0.9981	0.9984	0.9982
2005	0.9975	0.9975	0.9975	0.9975	0.9979	0.9978
$R365_{ht}$						
2003	0.9952	0.9952	0.9952	0.9952	0.9958	0.9953
2005	0.9890	0.9890	0.9890	0.9890	0.9999	0.99001

Turning to the ability to use the smoothed estimates for forecasting, the expected RSME values are derived using equation (12) and represent how well the forecast estimates are able to forecast out-of-sample values. The results indicate that the models forecast well for both AMI and Hip Replacement, yet in the outcomes across both conditions there are differences in the model's predictive ability. Overall, Hip Replacement forecasts are found to be closer to the true values, however this likely reflects the lack of variation in outcomes in this treatment area from one year to the next. For both conditions the values suggest that the forecasts for $D30_{ht}$ followed by $D365_{ht}$, are close to the true values for both years, and also that the expected and actual values are close in predictive power. Across all outcomes the Hip model is better able to predict future values than the AMI model.

Tab. 8: Summary of forecast accuracy using Root Square Mean Error (RSME) estimates. Forecasting 2000-2005 AMI and Hip Replacement values using data from 2000-2003.

	AMI		Hip Replacement	
	RSME estimate based on:		RSME estimate based on:	
	All outcomes	Same outcome	All outcomes	Same outcome
$D30_{ht}$				
2004 (expected)	0.0563	0.0565	0.0060	0.0061
2004 (actual)	0.0564	0.0564	0.0063	0.0063
2005 (expected)	0.0488	0.0492	0.0066	0.0066
2005 (actual)	0.0496	0.0492	0.0065	0.0065
$D365_{ht}$				
2004 (expected)	0.0725	0.0683	0.0202	0.0209
2004 (actual)	0.0682	0.0684	0.0193	0.0200
2005 (expected)	0.0660	0.0652	0.0180	0.0176
2005 (actual)	0.0651	0.0652	0.0174	0.0176
$R28_{ht}$				
2004 (expected)	0.0322	0.0322	0.0230	0.0236
2004 (actual)	0.0361	0.0324	0.0244	0.0244
2005 (expected)	0.0344	0.0348	0.0247	0.0249
2005 (actual)	0.0346	0.0347	0.0248	0.0250
$R365_{ht}$				
2004 (expected)	0.0579	0.0530	0.0369	0.0361
2004 (actual)	0.0525	0.0526	0.0354	0.0350
2005 (expected)	0.0552	0.0504	0.0418	0.0423
2005 (actual)	0.0507	0.0502	0.0421	0.0424

5 Discussion

In this study we have applied a simple, bidirectional smoothing estimator to measure hospital quality, based on the work by McClellan and Staiger (1999). In their study, McClellan and Staiger (1999) suggest their method is able to tackle some of the main limitations inherent in hospital quality measurement, allowing them to create indicators which reduce noise both within individual hospitals and across time, as well as integrate different dimensions of quality within a single estimator. Their paper uses US patient level data for elderly American's suffering from heart disease to create quality indicators at the hospital level. They show that their indicators are better able to reflect the multifaceted nature of hospital performance, appear reliable, and forecast quality remarkably well, better than many existing methods.

Recently, Jones and Spiegelhalter (2012) highlighted a number of directly competing bidirectional smoothing estimators, applying these to health care also. They noted that their preferred estimates relied upon specialised software and even so were associated with high computing time. As they state this raises practical issues over the applicability of such estimators in routine performance monitoring. They also noted the similarity of the class of bidirectional smoothing estimators they assess with those proposed by McClellan and Staiger (1999). Despite the practical advantages over competing methods, as noted for example by Jones and Spiegelhalter (2012), this particular approach has not been applied to evaluate hospital quality outside the USA or to other treatment conditions beyond heart attack. It has been shown to have wider application however, as witnessed in its use to evaluate other public service outcomes; specifically educational outcomes in the USA (Kane et al., 2002). This paper has applied the method to English, patient level data to test the robustness of their approach and its generalizability. The paper is also able to address some of the limitations acknowledged by the McClellan and Staiger (1999) study, arising from gaps in the US data on patient co-morbidity. We thus improve on their measure by specifically incorporating co-morbidity information to create even more robust indicators of hospital quality. Our results suggest that this method might be readily applied to other treatment conditions. Indeed we did apply these methods to a wider range of other conditions (including Stroke, TIA and Congestive Heart Failure) and the measures performed as well as the examples reported here. The methods thus seem to be generalizable to a wide set of treatment conditions and transferable across countries using similar administrative data. Our application of this method to a different setting did identify other issues however, stemming from the smaller sample sizes available in the English hospital sector as compared to the US setting. That said, even for medium sized hospitals

the proposed quality measures perform well. Thus, with this caveat aside, the McClellan and Staiger (1999) approach would therefore appear to reduce noise and strengthen the signal, thus improving the ability to assess hospital quality within the NHS.

To outline the approach we relied on a VAR(1) specification for our smoothed estimator, which was chosen for ease of interpretation and parsimony. The signal variances estimated using the VAR parameters were coupled with the estimation error to construct signal to noise ratios for each outcome measure, in each of the two conditions, for the year 2005. This signal, given a sufficient sample of patients, was strong in the majority of cases. While the number of cases required to get a good signal to noise ratio varied by condition, in most cases it included 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, as the variation in the data will be strongly influenced by specific differences in treatment, such as the presence or absence of an individual physician, which would have relatively small effects in a larger hospital. The other striking result from the signal to noise ratios was that in both cases, short-term mortality performed worst. This is an important finding for policy, which tends to emphasize quality measurement through short-term indicators. For AMI, where treatment variation in the short-term has a major implication for survival, one might expect short-term mortality measures to have the stronger signal. Indeed this the finding was reported by McClellan and Staiger (1999) in the US analysis. It is interesting that this is not the case for the UK, raising questions as to why. A possibility is that the wider range of outcome measures available within the UK data allowed us to incorporate more of the short term variation in mortality and readmission into the other outcome measures thus strengthening their signal. While for Hip Replacement, this possibly reflects the largely elective nature of the latter condition. Moreover a robustness check of the models, shows that information from the other outcome measures is important and in some instances can adjusted the persistence of indicators, confirming the assumption made that many outcomes are co-dependent and should be considered together.

This point is perhaps more strongly noted by the interpretation of the correlation of the residuals for the different indicators as reported in the VAR models. Short- and long-term readmissions have strong positive correlations with each other for AMI, while short-term mortality is correlated with long-term readmissions for Hip Replacement. Interestingly, McClellan and Staiger (1999) also observe this negative correlation for AMI between 30-day mortality and year-long readmissions. Although a positive correlation might have been expected, as noted in the text, if the smoothed mortality rates are improving, then "sicker" patients may be being discharged leading to higher future readmission rates.

This has direct implications for the recent introduction in the UK of a policy to financially penalize high readmission rates. The Hip Replacement model suggests a 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 had a positive association. However, for no condition were all the associations positive, indicating that one should be cautious when interpreting readmission measures in isolation as they may well not be indicative of higher quality. Once again, these findings question the 2011 policy that introduced financial penalties for English NHS hospital readmissions within the same HR.

Overall, these estimates of hospital quality based on four outcome measures for two conditions, an emergency condition (AMI) and an elective condition (Hip Replacement), appear to be acceptable measures and predictors of underlying quality, with straightforward applicability to the English NHS hospital setting. The method itself is less computationally intensive than recently assessed Bayesian approaches and, building on a suggestion by Jones and Spiegelhalter (2012) to evaluate simpler and less intensive measures, we have done precisely that. Although the simplicity of the calculations are a major advantage, a further advantage is that the proposed method provides a systematic approach to assessing noisy hospital quality signals which does not require costly measurement as it relies on routinely collected data. By systematically integrating different dimensions of hospital quality, it also reduces the general criticism that a single measure can not capture the breadth of elements necessary to return an aggregate indicator of quality. Given that the method incorporates correlations across alternative quality signals, identifies and eliminates redundant statistical information, and does so in a straightforward manner there would appear much to recommend this measure over the alternative bidirectional measures assessed by Jones and Spiegelhalter (2012). While recognizing that the aggregation of quality signals will always involve value judgment, especially over the implicit weights employed, the measures presented here are useful precisely because they return a combined signal of quality. They also identify the degree of variation in quality left unexplained after these smoothed shrinkage estimates have removed as much statistical noise as possible. That is, while there will always be a role for interpretation, by incorporating risk-adjustment, smoothing and aggregation of quality dimensions into a single statistical estimate there is, by definition, less noise and therefore more reason to suggest that the measures presented here are more acceptable than the more naive estimators currently used in policy evaluation. Obvious extensions would include investigation into the aggregation of the different measures into a single global measure, or at the very least

a comparison of within hospital quality versus between hospital quality across different treatment conditions using similar methods.

In conclusion, the major advantages of this technique, as illustrated throughout this paper, is their relatively good short-term accuracy and simplicity. Moreover this process can be easily implemented, does not require large amount of historical data, and are relatively low cost. However, this technique does also have disadvantages which are important to note. In particular, any other variables that might influence the forecast, and are not included in the first stage regression will not be accounted for. And, as noted previously the smoothing technique is only valid if we assume that the error terms are random.

References

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.

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.

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*.

Campbell, M. J., R. M. Jacques, J. Fotheringham, R. Maheswaran, and J. Nicholl (2012, March). Developing a summary hospital mortality index: retrospective analysis in english hospitals over five years. *BMJ 344*(mar01 1), e1001–e1001.

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.

Cooper, Z., S. Gibbons, S. Jones, and A. McGuire (2010, January). Does hospital competition save lives? evidence from the english NHS patient choice reforms. http://eprints.lse.ac.uk/28584/.

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.

Dranove, D., D. Kessler, M. McClellan, and M. Satterthwaite (2002, January). Is more information better? the effects of 'Report cards' on health care providers. *National Bureau of Economic Research Working Paper Series No. 8697*.

Fischer, C., H. A. Anema, and N. S. Klazinga (2012). The validity of indicators for assessing quality of care: a review of the european literature on hospital readmission rate. *The European Journal of Public Health 22*(4), 484–491.

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.

Jones, H. E. and D. J. Spiegelhalter (2012). Improved probabilistic prediction of healthcare performance indicators using bidirectional smoothing models. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 175*(3), 729–747.

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.

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.

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. Scholte Op Reimer, H. Van 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.

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.

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.

Normand, S.-L. 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.-L. 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.

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.

Shahian, D. M., R. E. Wolf, L. I. Iezzoni, L. Kirle, and S.-L. T. Normand (2010, December). Variability in the Measurement of Hospital-wide Mortality Rates. *New England Journal of Medicine 363*(26), 2530–2539.

Shen, Y.-C. (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.

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.

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