NBER WORKING PAPER SERIES

THE QUALITY OF HEALTH CARE PROVIDERS

Mark McClellan
Douglas Staiger

Working Paper 7327
http://www.nber.org/papers/w7327

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 1999

We thank Haruko Noguchi, Dhara Shah, and Yu-Chu Shen for outstanding assistance with data management and analysis, and seminar participants at the NBER and various universities for helpful comments. The National Institute on Aging, the Health Care Financing Administration, and the Olin Foundation provided financial support. All errors are our own. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

The Quality of Health Care Providers
Mark McClellan and Douglas Staiger
NBER Working Paper No. 7327
August 1999
JEL No. I10, L15, C33

ABSTRACT

Obtaining better information on the quality of health care providers is one of the most pressing issues in health policy today. In this paper we (1) develop a new method for measuring quality of care that overcomes the key limitations of available quality measures, and (2) apply this method to estimating the quality of hospital care for elderly patients with heart disease. Our approach optimally combines information from all available current and past quality indicators in order to more accurately estimate and forecast each provider's quality level. For patients with heart disease, the method is able to predict and forecast differences in patient outcomes across hospitals remarkably well - far better than existing methods. Our approach also provides an empirical basis for choosing among potential quality indicators. In particular, we find that differences across hospitals in short-term mortality rates following a heart attack, adjusted for patient demographics, are excellent indicators of quality of care: They vary dramatically across hospitals, are persistent over time, are highly correlated with alternative quality indicators, and are highly correlated with mortality rates that adjust more extensively for patient severity. Thus, comparing quality of care across providers may be far more feasible than many now believe.

Mark McClellan
Department of Economics
Stanford University
Palo Alto, CA 94305
and NBER
markmc@texas.stanford.edu

Douglas Staiger
Department of Economics
Dartmouth College
Hanover, NH 03755
and NBER
doug.staiger@dartmouth.edu

1. Introduction

The lack of good information on performance or quality is a core problem in many areas of public policy and evaluation today. The difficulty of developing reliable information on the quality of health care providers for guiding public policies and individual choices is perhaps the most striking example. Many reforms in medical financing and health plan choice have focused on improving competition and efficiency in health care delivery. While these reforms have clearly affected medical prices and expenditures, they have also led to heightened concerns about the quality of medical care. Are managed care plans reducing costs by avoiding providers and services that are more expensive yet worthwhile? How can health plans and providers compete effectively in quality if potential enrollees and patients have little reliable information on quality to use as a basis of their choices? How can providers hope to improve quality if they also have little reliable information?

The reason that these questions are so difficult results from the limited availability of useful information on the quality of health care providers. The quality information problem has many sources. First, measurement is a problem because it is difficult to collect timely and relevant data (often on long-term patient outcomes) for evaluating providers. Even with good data, multidimensionality is a problem. Quality of medical care has many dimensions -- outcomes, processes of care, and others – all of which would ideally be integrated in a quality evaluation. A third obstacle is the noise inherent in any health care quality measure, due to the small sample of patients and large number of factors other than provider quality that influence quality measures for any individual provider. Finally, bias is a problem to the extent that variation in patient treatment or outcomes across providers is the result of systematic differences

in patient mix rather than differences in care. All of these problems have limited the value of explicit information on health care quality, particularly for important health outcomes.

In this paper, we describe and apply a framework for addressing all of these issues in the provision of quality information. We develop measures of major health outcomes for patients treated by different hospitals for their heart disease over time. Our analysis considers a range of serious health outcome measures, including mortality and serious complications that have implications for quality of life, for which the quality evaluation problems are most salient. We adapt vector autoregression (VAR) methods for panel data to estimate the systematic relationship across outcomes and over time, and then use this information to forecast future outcomes and to filter out much of the noise in the observed outcome measures. The basic idea is simple. Any single outcome measure for a health care provider will be a noisy (often very noisy) indicator of that provider's quality. But all of the dimensions of quality that we consider – multiple quality measures, and multiple time periods – are likely to be related to each other, and so can aid the extraction of the signal from any particular measure.

This framework provides the basis for addressing all of the four major problems which have impeded the development and use of quality measures. Our method is computationally feasible even for large datasets with many outcomes and many years. More importantly, this method results in prediction accuracy that is much greater than is possible using existing statistical methods for addressing the signal extraction problem. In addition, it provides a quantitative basis for integrating a large number of dimensions of provider quality, and for limiting the often-substantial costs of data collection efforts to monitor quality. Our method estimates how highly correlated the signals are from alternative quality measures, thereby

providing information on whether some quality measures contain redundant information. For example, we find that outcomes measures based on the limited information in patient claims are highly correlated with measures that control for far more extensive information on patient mix available from patient charts. This result provides an empirical basis for determining whether the collection of additional, potentially expensive information on patient characteristics is worthwhile

.

Our application involves heart disease in elderly Americans. Heart disease is the leading cause of death in the United States, and is clearly a condition for which the quality of medical care provided may have a substantial impact on an individual's health. Between 1970 and 1995, the death rate from heart disease has fallen by more than half, and a number of studies have documented that much of this improvement can be attributed to changes in medical treatment (Goldman, 1984; Hunink et al., 1997; Heidenreich and McClellan, 1998).

The issues we address here are by no means unique to the health care industry. The same problems of measuring quality arise in fields as diverse as airline safety, school test scores, and mutual fund performance. Our approach to addressing the problem of profiling performance is applicable to all of these contexts. In addition, the estimation and forecasting methods we use are in principle applicable in many other settings, from forecasting local area unemployment and wage growth based on survey data to estimating betas for individual stocks based on weekly price data.

In Section 2, we review the issues and previous studies relevant to our analysis. In Section 3, we describe our empirical methods. Section 4 describes the data on heart disease. Section 5 presents our results, and Section 6 concludes.

2. Background

Information on provider quality in any industry is useful for two broad purposes: forecasts and evaluations. Forecast applications are forward-looking, to guide choices about providers that may influence future outcomes. Which hospitals would be the best choices as contractors for a managed care plan? If I have a heart attack, which hospital should I choose? Information on past performance can guide these decisions, but perhaps in a complex way. Relying simply on a provider's performance in the previous year may lead to worse decisions than considering patterns in performance over several years. Evaluation applications are backward-looking, to provide insights into the consequences of alternative policy options or practices in the past. What is the effect of hospital volume or experience on outcomes? What are the consequences of changes in hospital ownership? Do hospitals with greater adherence to standards of care have better outcomes? Policy and practice evaluations generally do not occur at the level of specific providers. But understanding how variations across providers may contribute to the outcome differences resulting from differences in policies or practices may lead to more appropriate inferences about their effects.

Many obstacles hamper the development of reliable forecasts and evaluations. We group these obstacles into four general categories. First, *measurement* is an obstacle: it may be difficult or costly to collect relevant data for evaluating providers in a timely way. For example, the relevant time period for measuring outcomes such as survival after a heart attack may be weeks or longer; until this time has passed, the relevant outcome measures are simply not available. Moreover, obtaining the relevant data may be expensive. For example, collecting

information on survival requires matching an individual's hospitalization records to death records, and collecting information on patient satisfaction requires locating and surveying patients.

Second, the *multidimensionality* of quality magnifies the problems associated with collecting data on any particular measure. Patients are likely to care most about outcomes of care, and relevant outcomes include not only survival but also the occurrence of various complications and functional impairments. Many other factors, such as the processes of care, may also contribute to patients' judgments about their satisfaction with the quality of care received. In addition, the processes of care themselves may be of some interest in evaluating provider quality. For example, the extent to which hospitals apply treatments that have been demonstrated as effective may be useful indicators of quality. Patients and especially health plans may also be interested in resource use and costs of alternative providers. Many clinical reasons suggest that these outcomes are related to each other, but the nature and magnitude of the relationships is generally not obvious. For example, hospitals that perform better in terms of heart attack survival may also do better in avoiding complications. But it is also possible that increased survival is associated with a greater rate of quality of life impairments.

To date, few systematic approaches have been developed to integrate all of these quality measures. Existing provider "report cards" are either *ad hoc*, or rely on clinical considerations rather than empirical relationships to describe "clusters" of quality measures. Similarly, the empirical literature in this area generally focuses on either single measures chosen on *a priori* grounds, or *ad hoc* aggregates across measures or years.

A third obstacle is the substantial amount of *noise* associated with virtually all important measures of provider quality. Most serious health problems, such as heart attacks, are relatively infrequent; many hospitals may treat only a few dozen patients or fewer over an entire year. And most major outcomes, such as long-term mortality, will be influenced by an enormous number of factors other than the quality of the provider. As a result, it is unlikely to be feasible to assess any single outcome measure for a particular provider with a useful degree of precision.

In response to this concern, the National Center for Quality Assurance (NCQA), one of the leading organizations in the development of quality assessment measures, has focused on the development of measures for which 411 cases per provider can reasonably be collected in a specific time period (e.g., six months or a year). As a result, current NCQA measures focus primarily on preventive care, for which the relevant denominator is the entire population treated by a provider or plan, or on outcomes for very common (and less serious) illnesses. None of the current or proposed NCQA quality measures involve outcomes for conditions serious enough to require hospital care.

A second response to the noise problem has been the development of hierarchichal Bayesian models of patient outcomes. The goal of this approach is to estimate posterior distributions for key provider-specific parameters that influence patient outcomes. Unfortunately, the complexity of this approach has limited its application to single outcomes and fairly small samples (see for example Normand, Glickman and Gatsonis, 1997). As we discuss in section 3, our approach is closely related to these empirical Bayes methods but represents a substantial departure in how we manage the complexity of the estimation problem so as to incorporate more outcomes and larger samples.

The final obstacle to measuring provider quality is the possibility of bias due to systematic differences across providers in the patients they treat (case-mix). The debate over the bias of claims-based measured has been extensive, but the empirical evidence for the existence of a bias is mixed (Landon et. al, 1996; Krakauer et. al, 1992; Park et. al, 1990). One response to this problem has been to collect more detailed information on patient condition at the time of admission from patient charts. However chart review is costly enough that it is unclear whether this is a feasible long-term solution. For example, HCFA collected chart information on all Medicare hospital admissions for cardiac conditions in 1994-95 but the cost was around $100 per case.[1] To the extent that such chart data provides the best case-mix adjustment possible, outcome measures based on such data provide a "gold standard" to which other measures may be compared. In section 5, we provide new evidence on the relationship between such chart-based measures and the more commonly available claims-based measures.

[1] Jeffrey Newman, personal communication.

3. Empirical Methods

We now describe our method for multivariate signal extraction using multiple measures of hospital quality, including information from multiple years, multiple diagnoses, and multiple outcomes. We begin by providing some notation, and laying out the basic goals of our empirical work. We then describe our estimation method, first for the case in which we only use quality measures to form predictions, and then in the more general case where we also use other hospital characteristics. Finally, we discuss how our estimator is related to empirical Bayes estimators.

A. Notation and Model

Suppose we observe data for a sample of hospitals (j=1,...N) on multiple dimensions of quality (k=1,...,K) over many years (t=1,...,T). For simplicity of notation we will assume that each hospital has data for all years, and within each year has data on all outcomes, although the methods can easily be extended to cases of missing observations. We are interested in the hospital-specific effect μ from a patient-level equation of the form:

$$(1) \qquad Y^k_{ijt} = \mu^k_{jt} + X_{ijt}\,\Pi^k_t + \omega^k_{ijt} \quad ,$$

where Y is the quality measure (e.g. death within 30 days of admission), X is a set of patient characteristics (e.g. age, gender and comorbidities), ω is the error term, and i indexes the individual patient. Thus, μ^k_{jt} measures the true quality difference in dimension k across hospitals j in year t, controlling for patient characteristics X.

We do not observe the true hospital-specific effects directly, but rather observe estimates of these hospital-specific effects from a patient level regression run separately by year

for each quality measure. In other words, for each hospital we observe a vector of K noisy hospital quality measures for T years. Let M_j be a 1xTK vector of observed quality measures for hospital j, adjusted for differences in X using patient-level regressions.[2] Then

(2) $M_j = \mu_j + \varepsilon_j$

where μ_j is a 1xTK vector of the true hospital effects for hospital j, and ε_j is the estimation error (which is mean zero and uncorrelated with μ_j). Note that the variance of ε_j can be estimated from the patient-level regressions, since this is simply the variance of the regression estimates M_j. In particular, $E(\varepsilon_{jt}'\varepsilon_{jt})=\Omega_{jt}$ and $E(\varepsilon_{jt}'\varepsilon_{js})=0$ for $t \neq s$, where Ω_{jt} is the covariance matrix of the effect estimates for hospital j in year t.

Our problem is how to use M_j to predict μ_j. More specifically, we wish to create a linear combination of each hospital's observed measures in such a way that it minimizes the mean squared error of our predictions. In other words, we would like to run the following hypothetical regressions:

(3) $\mu^k_{jt} = M_j \beta^k_{jt} + \nu^k_{jt}$

but cannot do this directly, since μ is unobserved and the optimal β will vary by hospital and year.

Equation 3 highlights the key problem in predicting hospital quality. The problem is analogous to classical measurement error: the regressor M^k_{jt} is a noisy estimate of the dependent variable, and therefore should not have a coefficient of one in this hypothetical regression. In other words, the estimated hospital-specific intercepts are not optimal predictors

[2] That is, M_j is the estimated hospital effect from a regression of Y on X with hospital fixed effects included. Since we are only interested in relative rankings, and not the absolute level of the intercept, we construct M so that it is mean zero in each year.

of the true hospital-specific intercepts in terms of minimizing mean squared error. As is usually the case with measurement error, we can improve the predictions in equation 3 by attenuating the coefficient towards zero, and this attenuation should be greatest for hospitals with imprecisely-estimated effects. This is the basic shrinkage or "smoothing" technique that has been applied in the empirical Bayes literature (e.g., Morris, 1983). Moreover, if the true hospital-specific effects from other quality equations (other years, other types of patients) are correlated with the effect we are trying to predict, then using their estimated values can further improve prediction.

B. Estimation

While we cannot estimate equation (3) directly, it *is* possible to estimate the parameters for this hypothetical regression. The minimum mean squared error linear predictor is given by $E(\mu_j| M_j) = M_j \beta$, where $\beta = [E(M_j'M_j)]^{-1}E(M_j'\mu_j)$. This best linear predictor depends on two moment matrices:

(4.1) $E(M_j'M_j) = E(\mu_j'\mu_j) + E(\varepsilon_j'\varepsilon_j)$

(4.2) $E(M_j'\mu_j) = E(\mu_j'\mu_j)$

We *can* estimate these required moment matrices directly as follows:

1) We estimate $E(\varepsilon_j'\varepsilon_j)$ using the patient-level OLS estimate of the covariance matrix for the parameter estimates M_j. Call this estimate S_j. Note that S_j varies across hospitals.
2) We estimate $E(\mu_j'\mu_j)$ by noting that $E(M_j'M_j - S_j) = E(\mu_j'\mu_j)$. If we assume that $E(\mu_j'\mu_j)$ is the same for all hospitals, then it can be estimated by the sample average of $M_j'M_j - S_j$. It is

easy to relax the assumption that $E(\mu_j'\mu_j)$ is the same for all hospitals by calculating $M_j'M_j - S_j$ for subgroups of hospitals.

With estimates of $E(\mu_j'\mu_j)$ and $E(\varepsilon_j'\varepsilon_j)$, we can form least squares estimates of the parameters in equation 3 which minimize the mean squared error. Analogous to simple regression, our prediction of a hospital's true effect is given by:

$$(5) \qquad \hat{m}_j = M_j\, E(M_j'M_j)^{-1}\, E(M_j'\mu_j) = M_j\, [E(\mu_j'\mu_j) + E(\varepsilon_j'\varepsilon_j)]^{-1}\, E(\mu_j'\mu_j)$$

where we use estimates of $E(\mu_j'\mu_j)$ and $E(\varepsilon_j'\varepsilon_j)$ in place of their true values. We can use the estimated moments to calculate other statistics of interest as well, such as the standard error of the prediction and the R-squared for equation 3, based on the usual least squares formulas.

Equation 5 in combination with estimates of the required moment matrices provides the basis for our estimates of hospital quality. Such estimates of hospital quality have a number of attractive properties. First, they incorporate information in a systematic way from many outcome measures and many years into the predictions of any one outcome. Moreover, our estimates of hospital quality are optimal linear predictors for a mean squared error criterion. Finally, these estimates are far simpler to construct than those derived from existing Bayesian approaches (e.g Normand, Glickman and Gatsonis, 1997). The patient-level regressions are somewhat computationally intensive, but can be performed with standard software for estimating fixed effect models, and the required moment matrices for the hospital-level estimates can be estimated in seconds even for large samples of hospitals. We will refer to estimates based on equation (5) as "filtered" estimates, since the key advantage of such estimates is that they optimally filter out the estimation error in the observed quality measures.

In practice, there are a number of reasons to impose more structure on $E(\mu_j'\mu_j)$. First, in order to provide out-of-sample forecasts of these quality measures in future years, some structure on the time-series behavior of these measures is required. Moreover, if the assumed time-series structure is correct, it will improve the precision of the estimated moments (and thus of the estimated effects) by limiting the number of parameters that need to be estimated. Finally, the correlation in quality measures over time and across outcomes is of direct interest, and will be easier to interpret to the extent it can be adequately summarized by a simple time-series model.

Therefore, we assume that each hospital's quality measures change over time according to a vector autoregressive (VAR) model. The VAR model has been applied successfully in other time series and panel data contexts, when the goal is to create a flexible model for forecasting and summarizing the data (Watson, 1994; Holtz-Eakin, Newey and Rosen, 1988). The VAR model is a generalization of the usual autoregressive model, and assumes that each hospital's quality measures in a given year depend on the hospital's quality measures in past years plus a contemporaneous shock that may be correlated across quality measures. In most of what follows, we assume a non-stationary first-order VAR for μ_{jt} (1xK), where:

(6) $\mu_{jt} = \mu_{j,t-1}\Phi + u_{jt}$, with $V(u_{jt}) = \Sigma$ and $V(\mu_{j1}) = \Gamma$.

Thus, we need estimates of the lag coefficients (Φ), the variance matrix of the innovations (Σ) and the initial variance conditions (Γ), where Σ and Γ are symmetric KxK matrices of parameters and Φ is a general KxK matrix of parameters.

The VAR structure implies that $E(M_j'M_j - S_j) = E(\mu_j'\mu_j) = f(\Phi,\Sigma,\Gamma)$. Thus, the VAR parameters can be estimated by Optimal Minimum Distance (OMD) methods (Chamberlain,

1984), i.e. by choosing the VAR parameters so that the theoretical moment matrix, $f(\Phi,\Sigma,\Gamma)$, is as close as possible to the corresponding sample moments from the sample average of $M_j'M_j - S_j$. More specifically, let d_j be a vector of the non-redundant (lower triangular) elements of $M_j'M_j - S_j$, and let δ be a vector of the corresponding moments from the true moment matrix, so that $\delta=g(\Phi,\Sigma,\Gamma)$. Then the OMD estimates of (Φ,Σ,Γ) minimize the following OMD objective function:

$$(7) \qquad q = N\left[\bar{d} - g(\Phi,\Sigma,\Gamma)\right]' V^{-1}\left[\bar{d} - g(\Phi,\Sigma,\Gamma)\right]$$

where V is the sample covariance matrix for d_j, and $\bar{d}$ is the sample mean of d_j. If the VAR model is correct, the value of the objective function, q, will be distributed χ^2 (p) where p is the degree of over-identification (the difference between the number of elements in d and the number of parameters being estimated). Thus, q provides a goodness of fit statistic that indicates how well the VAR model fits the actual covariances in the data.

C. Incorporating known hospital characteristics into the estimates

Thus far, we have assumed that the only data available are quality measures based on patient outcomes. More generally, one would expect that easily observable provider characteristics, such as patient volume or teaching status, might provide additional information about quality. For example, if low patient volume is associated with poor patient outcomes, then small hospitals would be expected to have poor outcomes even when there is little information in their observed outcomes measures. Our approach is easily extended to allow for such a systematic relationship between hospital characteristics and patient outcomes.

Suppose that for each hospital we have a set of hospital characteristics Z_j, where Z_j is 1xL. Note that Z_j may include time-specific characteristics (e.g. patient volume in year t) as individual elements. Then equation 2 can be generalized so that hospital-specific quality is a linear function of observable hospital characteristics and a random effect:

(2′) $M_j = \mu^*_j + \varepsilon_j$, where $\mu^*_j = Z_j \alpha + \mu_j$ and $E(Z_j'\mu_j) = 0$.

As the sample of hospitals grows, the parameter α can be consistently estimated with a weighted least squares regression of M on Z (using estimates of the estimation error in M to form weights), so the conditional mean of the hospital-specific effects ($Z_j\alpha$) is known asymptotically. Our approach can be used to generate estimates of the remaining random effect by replacing M_j with (M_j - $Z_j\alpha$) in the proceeding discussion. The estimated random effect is then added to the conditional mean ($Z_j\alpha$) to form the final estimate for each hospital.

D. Relationship to empirical Bayes estimators and Kalman filters

Our approach is closely related to the literature on empirical Bayes estimation (see Morris, 1983). The relationship is seen most clearly by assuming normality and rewriting equations 2′ and 6 in terms of the distributions for M and μ:

(8) $M_j \mid \mu_j \sim N(\mu_j, S_j)$

(9) $\mu_j \mid \alpha, \Phi, \Sigma, \Gamma \sim N(Z_j\alpha, f(\Phi,\Sigma,\Gamma))$

We take S_j as known, and estimate the parameters (α, Φ, Σ, Γ) from the marginal distribution:

(10) $M_j \mid \alpha, \Phi, \Sigma, \Gamma \sim N(Z_j\alpha, S_j + f(\Phi,\Sigma,\Gamma))$

With normality assumptions for the distributions of (8) through (10), our OMD estimates of the parameters are asymptotically equivalent to quasi-maximum likelihood estimates

(Chamberlain, 1984), but are computationally much more efficient. Our proposed estimator for μ is equivalent to the posterior mean of μ, i.e. $\hat{m}_j = E(\mu_j \mid M_j, \alpha, \Phi, \Sigma, \Gamma)$, where we replace the unknown parameters with consistent estimates. If the normality assumptions are correct, and if the parameters were known, then this Bayes estimator would be the optimal choice for any symmetric loss function (Morris, 1983). Moreover, the posterior distribution for μ would also be normal so that the estimate ($\hat{m}_j$) along with its standard error could be used to form posterior probabilities, for example to calculate the probability that a hospital's effect lies above some value.

Formally, equations (8)-(10) are similar to the statistical assumptions used by Normand, Glickman and Gatsonis (1997) with three key differences. First, Normand, Glickman, and Gatsonis allow for a hospital-specific slope parameter, which multiplies a univariate index of a patient's risk. Thus, the hospital-specific component of patient mortality may vary unidimensionally by type of patient, rather than being a simple intercept shift as in our model. Second, they estimate a much simpler structure for (9), in that they do not allow for multiple quality measures that are correlated. Finally, they work with patient level distributions and thereby avoid making a normality assumption in (8). Since equation (8) is characterizing the distribution for regression intercepts that typically involve a large number of patients, the normality assumption does not seem unreasonable as an approximation. Thus, our model appears to maintain many of the attractive aspects of the hierarchical Bayes approach, while dramatically simplifying the complexity of the estimation.

Finally, note that Equations 2′ and 6 are a linear state-space representation for M. The predictions and forecasts we propose are similar to those used in the time-series literature on state-space models using the Kalman filter (see Hamilton, 1994). We are able to exploit panel data for a large number of health care providers and quality measures to estimate the model's parameters, and therefore avoid many of the technical and computational issues that are at the heart of time-series literature.

4. Data

Our application of these methods involves the quality of care for heart disease in the elderly. We consider two broad types of heart disease patients: patients with heart attack (acute myocardial infarction, AMI) or with ischemic heart disease (IHD). A heart attack is an acute blockage of an artery that provides blood to the heart muscle; it is a major health event that almost always results in hospitalization. Ischemic heart disease hospitalizations involve similar symptoms but somewhat less severe illness, characterized by inadequate blood flow to the heart that does not actually cause death of heart muscle. For this condition, the hospital treatment is intended to assure that a heart attack has not occurred, and especially to try to improve blood flow and reduce heart workload to prevent future heart attacks and recurrent symptoms such as chest pain or breathing problems.

We used longitudinal Medicare claims data to identify approximately 220,000 elderly beneficiaries per year with new occurrences of AMI and approximately 360,000 patients per year with new occurrences of IHD. We also used hospital claims, merged with death records, to develop one-year outcome information on mortality and heart-related complications

(rehospitalizations for heart failure, and recurrent AMI or IHD) for all U.S. elderly patients hospitalized with new occurrences of each of these conditions between 1984 and 1994. Mortality outcomes are based on whether a patient died within a given time (e.g. 30 days) of the initial admission. Complication outcomes are based on whether a patient was rehospitalized between 30 and 365 days following the initial admission. Rehospitalizations within 30 days are not counted because they are likely to reflect continuing treatment of the initial event, rather than treatment of complications. The construction and content of these outcome data are described in more detail elsewhere (e.g., McClellan, McNeil, and Newhouse, 1994; Kessler and McClellan; McClellan and Newhouse, 1997; McClellan and Noguchi, 1998a,b). To construct hospital-specific quality measures using these outcome data, we grouped patients according to their hospital of initial admission for heart disease treatment.[3]

Our sample is based on data from 1984 to 1994, and includes 3954 U.S. hospitals that had at least three admissions for AMI and IHD in all years of the 1984-1994 period. Table 1 provides some summary statistics on elderly heart disease patients in the first and last year of this sample. Mortality is substantial for both conditions – one-year AMI mortality is 28 percent and one-year IHD mortality is 11 percent in 1984 – and improved markedly over time. Similarly, complications occurred in many cases, and also declined to some extent over time (heart failure has increased slightly over time; IHD and AMI recurrences have declined over time). Despite the large number of patients, most hospitals treated relatively few cases. The

[3] Many patients are transferred or readmitted for care at other hospitals following their initial admission; for some patients, such transfers may even occur on the same day. Thus, the bulk of care actually received by some patients may have taken place at a hospital other than the hospital of their original admission. From the standpoint of guiding provider choices, however, quality assessment from the perspective of the initial hospital choice seems most appropriate. It is also least subject to selection biases.

average hospital in this population provided the initial care for 50-60 AMI patients per year, and for approximately 80 IHD patients per year in 1984 and 95 IHD patients per year in 1994. This relatively small sample size, coupled with the large number of factors that may influence heart disease mortality, illustrates why signal extraction for particular quality measures is a difficult problem.

We used these patient-level data to construct hospital-level fixed effect measures M_j for each outcome in each year, by estimating patient-level linear regressions for each outcome that included fully-interacted demographic covariates (five-year age groups, gender, black or nonblack race, urban or rural residence) and hospital effects. These adjusted M_j estimates provide the basis for the VAR analyses described in the next section. It is worth noting that the claims data do not include reliable information on comorbidity and severity, so that our adjustment methods will not remove biases in "case mix" that are not correlated with patient demographics. Because these forms of heart disease are urgent health problems, patients are highly likely to go quickly to nearby facilities for care, so that the magnitude of selection biases will be relatively small. We provide further evidence on the bias question in our results and discussion below, through detailed analysis of the relationships among outcome measures as well as integration of detailed chart-review data into our estimation methods.

5. Results

In this section we report the results of applying the methodology described in Section 2 to Medicare patient outcomes data. The results are divided into two sub-sections. The first sub-section presents estimates of the VAR parameters for various models using 9 years of data

from 1984 to 1992. These parameter estimates are of direct interest because they provide insight into fundamental relationships among the multiple dimensions of hospital quality: how strongly are hospital outcomes correlated over time and across different measures? The second sub-section presents evidence on the properties of the filtered estimates of hospital-specific quality. The key issue is whether these filtered estimates extract enough of the signal in hospital-specific outcomes measures to be of practical use. That is, are the measures precise enough to allow informative comparisons across individual hospitals, and do they provide accurate forecasts of hospital outcomes in 1993 and 1994?

A. Vector Autoregression (VAR) estimates

Table 2 illustrates our approach with estimates of the VAR parameters for two basic models. Each column reports parameter estimates for a separate bivariate VAR(1) model (see equation 6), estimated by OMD (see section 3) for two different sets of quality measures for each hospital. The first column contains estimates for a model with 30-day AMI mortality effects (DTH30) and 365-day AMI complication effects (CMP365). The next column is for a model of 30-day AMI mortality and 90-day IHD mortality (IHD_DTH90). We use 90-day mortality for IHD because 30-day mortality is quite low in IHD and very noisy (see Table 1 and further discussion below). Each column reports the initial variance and correlation of the effects in 1984 (Γ), the variance and correlation for the innovations to each effect (Σ), and the lag coefficients (Φ). Recall that each effect depends on lags of both effects, so that for the model in column one, we have:

$$E[\, m_t^{DTH30} \,] = \Phi_{11}\, m_{t-1}^{DTH30} + \Phi_{21}, \quad {}_{t-1}^{CMP365} \quad \text{and} \quad E[\, m_t^{CMP365} \,] = \Phi_{22}\, m_{t-1}^{DTH30} + \Phi_{12}\, m_{t-1}^{CMP365}.$$

The parameter estimates for the model of DTH30 and CMP365 suggest that both dimensions of hospital quality are quite persistent, with coefficients on their own lags of 0.887 for DTH30 and 0.973 for CMP365. There is much more true (signal) variation in hospital quality for 30-day AMI mortality. The variance of DTH30 in 1984 is .00175 and the variance of the innovation to DTH30 is 0.00036; this corresponds to a standard deviation in 30-day AMI mortality rates across hospitals of over 4 percentage points, and a standard deviation for the annual innovations of nearly 2 percentage points. The variation in CMP365 is smaller, corresponding to a standard deviation of just over two-and-a-half percentage points in 1984 and a standard deviation of innovations under 1 percentage point.

A further notable result from the VAR model of DTH30 and CMP365 is that the two measures are negatively correlated. *A priori*, a positive correlation might be expected, because higher values represent worse outcomes. The negative correlation probably reflects the fact that the "marginal" patients who survive if treated by higher quality hospitals are likely to have relatively poor heart function. Thus, hospitals that have worse mortality performance have better rates of subsequent complications, since fewer severely-ill patients survive to develop complications. This negative correlation and the fact that there is very little variation in CMP365 to begin with suggest that CMP365 may be a poor measure of hospital quality, at least when considered in isolation. The negative correlation also suggests that true quality differences and not patient selection are responsible for most of the variation in our hospital quality measures. If healthier heart patients led to low mortality at a hospital, we would also expect complication rates at the hospital to be lower, not higher.

Results from a model of 30-day mortality for AMI and 90-day mortality for IHD admissions are reported in the second column of Table 2. Compared to AMI outcomes at hospitals, 90-day mortality for IHD is not very persistent, with a coefficient on its own lag of only 0.606. As was the case with CMP365, the true variation in IHD 90-day mortality across hospitals is much lower than for AMI, both in 1984 and in terms of the innovations. In contrast to the CMP365 results, however, we estimate a positive correlation in the AMI and IHD mortality rates both in 1984 and in the innovations. Thus, hospitals with low AMI mortality also have low IHD mortality -- as would be expected if high-quality hospitals tend to produce superior outcomes across many patient groups.

In the bottom panel of Table 2, we report p-values for a set of specification tests. The first row reports the general goodness-of-fit test, which tests whether the VAR model provides an adequate fit of the data. Both models fit the data reasonably well, with p-values from the GMM goodness-of-fit test of around 3%. If the two measures in the VAR are not independent, then combining information from both measures will improve prediction. As seen in the second row of the bottom panel of Table 2, independence of the two measures is strongly rejected.[4] Finally, in the last row of Table 2, we test for the presence of a second lag in the VAR model (a VAR(2)). We can formally reject the VAR(1) specification in favor of the VAR(2) specification in one of the models (AMI and IHD mortality), and in this specification the VAR(2) model also performs better on the goodness-of-fit test (p-value=0.24 for the VAR(2)). Nevertheless, we will continue to focus on the VAR(1) specification because this specification is

[4] The test of independence is a Wald test (4 d. f.) of the joint hypothesis that the correlation in 1984 is zero, the correlation in the innovations is zero, and the cross lag terms (Φ_{12},Φ_{21}) are zero.

more parsimonious, easier to interpret, and fits the data reasonably well. Below, we compare the VAR(1) and VAR(2) specifications in terms of their forecasting ability.

The length of follow-up on the mortality measures used in Table 2 (30-day for AMI and 90-day for IHD) was chosen on *a priori* grounds; most studies of AMI outcomes focus on 30-day mortality. However, the optimal length of follow-up is an empirical question: is there a time (e.g. 30 days) after which there is no substantial change in true mortality differences across hospitals, that is, just added noise in outcomes? In Table 3, we explore this question for both AMI (the first two columns) and IHD (the second two columns). Each column of the table reports estimates of bivariate VAR parameters from models of a short-term and a long-term mortality measure. To make the estimates easier to interpret, we report the VAR coefficients in terms of (1) the effect for short-term mortality, and (2) the difference between the effects for long-term and short-term mortality. In this way, the parameter estimates for the second outcome refer directly to the changes in hospital-associated mortality between the short-term and the long-term. For AMI, we report estimates for models of 7-day mortality with the change from 7-day to 30-day mortality, and for models of 30-day mortality with the change from 30-day to 365-day mortality. For IHD, we report similar estimates using 90-day mortality as the intermediate measure.

For AMI, the estimates in Table 3 imply that essentially all of the variation in outcomes across hospitals arises in short-term mortality. We estimate that the variance in 7-day mortality effects in 1984 is 0.00153, while the variance in the change from 7-day to 30-day is only 0.00005 (first column). Similarly, much more variation arises in the first 30 days than between day 30 and 365 (second column). The variance in the innovations is much larger for 7-day

mortality than for the changes in mortality after 7-days. To the extent there are changes in mortality after 30 days, they appear to be negatively correlated with 30-day mortality. Thus, nearly all of the differences across hospitals in terms of AMI mortality appear within the first week, and over long horizons these differences may shrink. These estimates suggest that hospital performance in the first week of AMI care is the principal determinant of long-term outcome differences. From a clinical standpoint, this empirical finding seems to reflect the fact that many of the critical medical interventions in AMI care take place within the first days of care, and have their impact on survival at that time. From an evaluation standpoint, the empirical finding suggests that assessing outcomes soon after AMI is sufficient for detecting the vast majority of mortality-related quality differences across hospitals.

Mortality for IHD exhibits a different pattern. For IHD, there is very little mortality variation at 7 days, with much of the variation emerging in both the 7 to 90 day and 90 to 360 day periods. In 1984, 7-day mortality was positively correlated with 7-90 day mortality (0.792) and 90-day mortality was positively correlated with 90-365 day mortality (0.135). Thus, hospitals with low initial mortality tended to have low mortality in subsequent months, although this pattern is diminished by the negative correlation in the innovations after 90 days. Once again, our empirical findings reflect the clinical care for ischemic heart disease. While IHD symptoms are an urgent problem, the treatments for IHD have more of a long-term focus on preventing the progression of blockages and avoiding future heart attacks. Thus, hospital quality may have little impact on short-term outcomes, yet still substantially influence longer-term complications and death.

The findings in Table 3 have important policy implications for at least two reasons. First, they suggest that the first few days of treatment for AMI patients is the key period in determining outcome differences across hospitals, but not for IHD patients. Thus, attempts to improve patient outcomes should focus on treatment decisions in the first days after AMI, whereas decisions with a longer-term focus and later treatment decisions are more important for providing high-quality IHD care. A second implication is that a mortality-based quality measure for AMI (but not IHD) should involve short-term mortality rather than longer-term mortality, since short-term mortality rates capture nearly all of the signal variation present in longer term AMI mortality but tend to have less noise.

A practical advantage of our VAR method is that it provides a systematic basis for choosing among outcome measures that may be equally valid on *a priori* grounds. In particular, the VAR parameters provide estimates of how much signal variance there is in the original hospital effect data. We can use these estimates of the signal variance, in combination with estimates of the amount of estimation error in each measure (S_j), to examine the signal-to-noise ratio for any quality measure.

Figure 1 plots estimates of the ratio of signal variance to total (signal plus noise) variance in the original (observed) hospital effects against the number of admissions on which the measure was based. The signal ratio rises with sample size, as the variance of the estimation error declines. The estimates shown are for 1992, and are based on the VAR models from Tables 2 and 3. Not surprisingly, AMI mortality measures perform the best in terms of signal ratios because of the relatively large variance across hospitals in the true effects. For AMI, the signal ratio for 7-day mortality is higher than for 30-day mortality, and much higher than for

365-day mortality. All three measures have roughly the same signal, but 7-day mortality has less estimation error because the overall mortality rate is lower at 7 days; longer-term outcomes largely add pure noise. The signal ratio for IHD mortality shows the opposite pattern, with 7-day IHD mortality having the lowest signal ratio and 365-day IHD mortality having the highest signal ratio – but still well below AMI mortality measures. Thus, it appears that short-term mortality measures are most useful for AMI, while long-term mortality measures are more useful for IHD. Finally, the signal ratios for AMI complications are comparable to 1-year IHD mortality, and roughly half as large as the signal ratio in 7-day AMI mortality for a hospital with 100 admissions.

Figure 1 also highlights how little signal there is in outcome measures for most hospitals. In 1992 over half the hospitals in our sample admitted fewer than 40 AMI patients, and for these hospitals the signal ratio even in their AMI mortality measures is under one-third. Thus, for the majority of hospitals, individual patient outcome measures for a single year provide information on quality that is crude at best.

The dynamics of patient outcomes differ substantially between small and large hospitals. Table 4 contains VAR estimates from models of mortality for AMI and IHD, estimated on the whole sample and then separately for low- and high-volume groups. We split the sample roughly in half, with "high volume" hospitals having at least 25 AMI admissions in every year, and "low volume" hospitals having at least one year with less than 25 AMI admissions. The most striking difference between these two groups is that there is much greater variance in mortality effects in the small hospitals. The variance of both AMI and IHD mortality is 2-3 times larger in the small hospitals, and the innovation variance (the variance of year-to-year

changes in mortality) is more than 3 times larger in the small hospitals. These results are consistent with the notion that there is less discipline on product quality when quality is difficult to observe. In particular, given the difficulty in observing a small hospital's true effect from the patient outcomes data, small hospitals with high mortality might continue to attract patients (and, therefore, survive) and might find it more difficult to detect and correct quality problems as they occur. It is also consistent with quality at small hospitals being more dependent on idiosyncratic changes in heart disease treatment, such as the arrival or departure of an individual physicians or other specific innovation, that would be expected to have relatively smaller effects at larger hospitals.

An important question is whether these patterns which we observe in the VAR estimates are the result of quality differences that vary across providers, or differences in patient mix which persist and are correlated across outcomes. Table 5 presents evidence on this question. The table reports estimates of the correlation between the AMI outcome measures used in the previous tables, which are based on claims data and adjusted for demographic differences across hospitals, and alternative measures of the same outcomes, which are based on detailed medical chart-review data and adjusted for a far more extensive list of patient comorbidities and severity. The data for this comparison come from HCFA's CCP project, which collected information from patient charts for all Medicare patients admitted for AMI during an 8-month period in 1994-95 measures (see the McClellan and Noguchi, 1998b, for more details on the data and variables).

The comparison of claim- and chart-based data suggests that there is very little bias in the claims-based measures. In the first column of table 5, we report the correlation of the chart-

based and claims-based measures. For all of the measures, we estimate a correlation of chart- and claims-based measures of over 0.8, while the correlation for the complications measure and for short term mortality are over 0.95. The second column of table 5 shows the estimated slope parameters from the hypothetical regressions $\mu_{chart} = \beta\mu_{claim}$,[5] which suggest that the claim-based measures introduce some measurement error relative to the chart-based measures. The parameter estimates are below one for all outcome measures, indicating that the claims-based measures tend to overstate differences among providers. Together, these results suggest that some hospitals would be less likely to appear as "outliers" in terms of absolute deviations from a mortality standard. However, the claims-based measures are conveying very similar information about provider quality to that obtained using chart-based data, particularly for the AMI outcome measures that provide the most accurate signals about quality variations.

Overall, five substantive findings emerge from the VAR estimates contained in Tables 2-5. First, there is a substantial amount of correlation in outcomes over time and across measures. Outcomes for AMI are particularly persistent over time, while mortality outcomes appear to be positively correlated between AMI and IHD. In addition, we find that there is a substantial amount of variation across hospitals (especially small hospitals) in outcomes, particularly for AMI mortality. Nevertheless, commonly used risk-adjusted outcome measures have quite low signal ratios for most hospitals. A third substantive finding is that most of the differences across hospitals in AMI mortality emerge within the first week following admission, while differences emerge more gradually for IHD mortality. A fourth finding is that there is little evidence of substantial bias in our claims-based outcome measures. Finally, the VAR model performs

[5] With only one period of chart-review data, we cannot estimate a VAR.

reasonably well in fitting and summarizing key features of the data. We provide further evidence on this issue in the sub-sections that follow.

B. Properties of the filtered estimates

A main goal of this paper is to develop outcomes-based measures of quality of care that are of practical use at the level of individual providers. In this sub-section, we present evidence on the performance of our filtered estimates as indicators of hospital quality. We begin with simple plots comparing the filtered measures to more conventional outcomes-based estimates of hospital quality. We then turn to a more systematic evaluation of the filtered estimates ability to predict (in sample) and forecast (out of sample) variation in the true effects.

Figure 2 plots the observed (unfiltered) data for four hospitals: a small hospital (upper left), a large hospital (lower right), and two midsize hospitals. These hospitals are not a random sample, but rather chosen to represent a wide range of possibilities. Each panel in the figure plots data for a single hospital from 1984 through 1992. Each line plots the estimated hospital-specific effect from a linear probability model (estimated separately by year) that controls for age, race, gender and urban status. A value of 0.04 means that the hospital's mortality was 4 percentage points above the average hospital in that year, with negative values indicating lower mortality than average. Note that these are *absolute* mortality differences: if the national average is 19%, an estimate of 0.04 indicates mortality of 23%. The solid line in each panel plots the observed effects for death within 30 days of admission for AMI admissions. The dashed line is for 90-day mortality among IHD admissions. Estimates such as these are commonly referred to as standardized, or risk-adjusted, mortality rates.

There are two striking features of Figure 2. First, there is considerable variation in hospital outcome estimates both across hospitals and over time, particularly for AMI. For example, the 30-day AMI estimates range from over 15 percentage points below average (in the bottom two panels) to more than 20 percentage points above average (the upper right panel), with one hospital's mortality dropping over 30 percentage points between 1984 and 1990. These differences are particularly large considering that the average hospital's 30-day AMI mortality was 19 percent in 1984. A second striking feature of Figure 2 is the large year-to-year variation, with jumps of 5-10 percentage points not unusual.

Both of these features of figure 2 may reflect the fact that these hospital effects are not very precisely estimated. Figures 3a and 3b plot the AMI and IHD effects separately and add 95% confidence intervals around these estimates. For both AMI and IHD, the confidence intervals on these estimates are quite large. For example, the confidence interval on AMI mortality for the midsize hospital in the lower left panel of figure 3a almost always includes 0 (the national average), despite the fact that its estimated mortality ranges from 4 to 15 percentage points below average. Thus, a clear limitation of these standardized mortality rates is their lack of precision.

Filtered estimates are meant to address this lack of precision. Figures 4a and 4b overlay the filtered estimates (long dashes) and 95% confidence intervals (short dashes) on the plot of the unfiltered estimates (solid line). The filtered estimates in these figures are based on equation (5), incorporating all of each hospital's data from 1984 through 1992 into each years estimate. Estimates of $E(\mu'\mu)$ are derived from the VAR parameter estimates for AMI and IHD mortality reported in Table 2.

It is immediately apparent that the confidence intervals for the filtered estimates are much tighter. The intervals range from +/- 4% at the largest hospital to +/- 6% at the smallest hospital for AMI mortality and about half that for IHD mortality. The tighter intervals are of practical importance in allowing us to interpret these estimates. For example, the midsize hospital in the lower left panel of Figure 4a has AMI mortality that is clearly better than average based on filtered estimates. Another feature of the filtered data is that the estimates move smoothly from year to year. Thus, although it is still a large decline, the decline in AMI mortality for the midsize hospital in the upper right panel of Figure 4a is estimated to be less than half as large with filtered estimates as compared to unfiltered estimates. Finally, one can see the clear tendency of the filtered estimates to "shrink" towards average for the smallest hospital, with the filtered AMI estimates tending on average to be closer to zero than the unfiltered estimates.

Table 6 provides a more systematic evaluation of the ability of the filtered estimates to predict variation in the true hospital effects. The goal of the filtered estimates is to minimize the mean square error of this prediction. If true effects (μ) were observed, a natural metric for evaluating these predictions would be the sample R-squared:

$$(11) \qquad R^2 = 1 - \left(\sum_{j=1}^{N} \hat{u}_j^2 \right) \Big/ \left(\sum_{j=1}^{N} m_j^2 \right),$$

where $\hat{u} = m - \hat{m}$ is the prediction error. Since μ is not observed, we must construct an estimate of this R-squared. For in-sample predictions, we construct an estimate using our estimate of $E(\mu_j' \mu_j)$ for the denominator, and our estimate of $E\left[\left(m_j - \hat{m}_j \right)' \left(m_j - \hat{m}_j \right) \right]$ for the terms in the numerator (where this can be estimated from the estimated moment matrices in

equations 4.1-4.2). Finally, we report a weighted R-squared (weighting by the number of admissions in each hospital). Without weighting, the R-squared of our predictions tend to be smaller, but not dramatically so.

The expected R-squared of in-sample predictions is reported in Table 6 for selected years and outcomes. Each panel in the table reports the results based on different pairs of outcomes. The first panel provides the prediction results based on a VAR model of 30-day AMI mortality and 365-day AMI complications. The remaining panel reports results for a model of 30-day AMI and 90-day IHD mortality. We report the prediction R-squared for each outcome in two representative years, 1988 and 1992. Each column reports the R-squared for predictions using different amounts of data. Predictions in the first column use all years of data for both outcomes and therefore should be most accurate. The second column forms predictions with all years but does not use data on the other outcome. The next two columns form predictions using only the three most recent years of data. The final two columns form predictions using only data for the given year – so that the last column, which does not use data on the other outcome, is a simple shrinkage estimator based on a single year of data.

The filtered estimates are able to predict remarkably well. When we use all the data to form predictions (column one), the prediction R-squared ranges from a low of 0.51 for 90-day IHD mortality in 1988 to a high of 0.73 for 30-day AMI mortality in 1988. In other words, in 1988 the filtered estimates capture 73% of the true variation across hospitals in 30-day AMI mortality. Not surprisingly, the filtered estimates' ability to predict is directly related to the signal ratio in the original data, with AMI mortality being predicted most accurately. In general, the

prediction accuracy declines somewhat in 1992 because there is less data in the surrounding years to rely on for prediction.

The remaining columns of Table 6 report the impact on prediction accuracy from using a limited set of outcome measures in forming the filtered estimate; nine years of data are unlikely to be available in many applications. Provided at least several years of data are available, prediction accuracy is not much affected when prediction is based only on data observed for the same outcome (e.g. only use the 30-day AMI mortality data to predict the 30-day AMI mortality effects). With many years of data on a 30-day AMI mortality, there is not much to be learned about this outcome from other outcomes. Prediction accuracy remains high (R-squared of at least 0.45) with 3 years of data, as seen in columns 3 and 4 of table 6. In particular, prediction accuracy in 1992 (the last year of data being used) is not much reduced: Even with all 9 years of data, the 1992 estimates cannot use future years and so already rely heavily on the most recent years of data. However, when prediction relies only on data from a single year, the prediction R-squared falls considerably, especially when only using data on a single outcome. Thus, the ability of the filtered estimates to incorporate information from many years, or from many outcomes in a single year, is very important in terms of improving prediction accuracy.

A second advantage of the filtered estimates is that the VAR structure allows for forecasting out of sample. To evaluate the performance of these out of sample forecasts, we estimated models using the 1984-1992 data and construct 1-year (1993) and 2-year (1994) ahead forecasts. The accuracy of these forecasts can be predicted as was done in Table 6, and compared to the actual performance of the forecasts against the observed outcome measures in 1993 and 1994. Since much of the observed variance in outcome measures in 1993-1994 is

estimation error, we construct a modified R-squared of the forecast that estimates the fraction of the *systematic* (true) hospital variation in the outcome measure (M) that was explained, i.e.:

$$(12) \qquad R^2 = 1 - \left(\sum_{j=1}^{N} \left(\hat{u}_j^2 - S_j \right) \right) \Big/ \left(\sum_{j=1}^{N} \left(M_j^2 - S_j \right) \right),$$

where $\hat{u} = M - \hat{m}$ is the forecast error, and S_j is the OLS estimate of the variance of the estimate M_j. This modified R-squared estimates the amount of variance in the true hospital effects that has been forecasted. Note that because these are out-of-sample forecasts, the R-squared can be negative: the forecast can perform worse than a naive forecast in which quality is assumed to be equal to the national average at all hospitals.

Table 7 contains the results of this forecasting exercise. As in Table 6, each panel contains the forecast R-squareds for a different pair of outcomes. For each forecast, we report two R-squared values: the modified R-squared based on evaluating the actual performance (equation 12), and the expected R-squared of the forecast. The expected R-squared of the forecast was constructed using data from 1984-1992 only, and is the prediction from the VAR models of what the actual modified R-squared of the forecast should be. In each column we report the results for a different forecasting method. The first two columns report the results using forecasts based on filtered estimates from a VAR(1) specification and constructed from (1) data on both outcomes, and (2) data on only the same outcome as being forecasted. The next two columns report analogous results for a VAR(2) specification. The remaining three columns consider alternative forecast methods. The fifth column uses the shrinkage estimator from 1992 as a forecast for 1993 and 1994. The sixth column uses the observed effect in 1992

as a forecast. The final column uses the average effect observed between 1984 and 1992 as a forecast.

Forecasts based on filtered estimates perform relatively well across a variety of measures. For example, in the first column, we are able to forecast over 40% of the variation in 30-day AMI mortality effects in 1993 and 1994 based on joint models of 30-day AMI mortality with either 365-day AMI complications or 90-day IHD mortality. The worst forecasting performance is for 90-day IHD mortality, presumably because the IHD effects were not estimated to be very persistent. In each case, the expected R-squared of the forecast is quite close to the actual values, suggesting that these expected R-squared values are an accurate prediction of out-of-sample performance.

In general, forecasts using both outcomes (column one) and using only the same outcome (column two) yield similar results. Data on 30-day AMI mortality seem to improve forecasts slightly for 90-day IHD mortality. This may reflect the lack of persistence in the IHD mortality effects, which makes past values of other outcomes more useful in forecasting. Similarly, forecasts using VAR(2) specifications yield almost identical results to those using VAR(1) specification. Thus, forecast performance does not appear to be sensitive to the lag choice in the VAR.

The alternative forecast methods reported in the last three columns of Table 7 do uniformly worse than the VAR methods in terms of the actual forecast R-squared. Again, this was predicted by our model's estimates of the expected R-squared for these alternative forecasts. In particular, using the 1992 measured effect as a forecast always results in a negative forecast R-squared. The shrinkage (standard Bayesian) estimator from 1992 and the

average outcome from 1984 through 1992 perform similarly. Both generally forecast less than half the variation being forecast by the filtered estimates, and both do particularly poorly at forecasting effects that are not so persistent (e.g. 90-day IHD mortality).

The evidence in Tables 6 and 7 suggests that the filtered estimates are quite accurate, and dominate other forecast methods in terms of prediction accuracy. The accuracy of the filtered estimates may be improved still further by incorporating additional information into the predictions, as discussed in Section 3. The volume of patients treated is one hospital characteristic that has a well-documented association with patient outcomes (e.g., Luft et al., 1990). It is clear in our data that higher volume hospitals have lower mortality. Figure 5 illustrates this relationship by plotting unfiltered and filtered hospital effects for 30-day AMI mortality in 1992 against the number of AMI admissions in 1992. There is a significant (but small) negative association between mortality and volume in each plot, although the relationship is most easily seen using the filtered effects. Thus, information on patient volume should be useful in predicting mortality. Understanding the volume-outcome relationship more precisely may have important policy implications as well.

Table 8 contains estimates of prediction accuracy for 1988, after incorporating the effect of patient volume into the predictions. More specifically, we allow the hospital effect for each outcome in each year to depend linearly on patient volume for that outcome in that year. We assume that the coefficient on patient volume does not change over time. The layout of Table 8 is analogous to that of Table 6, except that now for each outcome we report the prediction R-squared based on predictions including hospital volume. The prediction R-

squared without volume (from Table 6) is also included for comparison. The last column of the table reports the estimated coefficient on patient volume for each outcome.

In general, using patient volume leads to only a slight improvement in prediction accuracy despite the fact that there is a very significant relationship between volume and each outcome. For example, for forecasts that use all years of the same or both outcomes, incorporating volume improves the prediction R-squared by no more than 0.01 in any case. For these forecasts, the hospital-specific effects are being predicted so accurately from the outcomes data that volume adds little new information. For 365-day AMI complications and 90-day IHD mortality, the volume coefficient is too small to contribute much to prediction accuracy: an increase of 100 admissions is associated with a change in these outcomes of less than 0.25 percentage points. Only for AMI mortality, where volume has a relatively large coefficient, does volume meaningfully improve prediction accuracy for some of the weaker forecasts.

These results further emphasize how well the filtered estimates are able to predict. Patient volume has little impact on prediction accuracy because it can only explain a small fraction of the total variation in the true hospital effects. The success of the filtered estimates in predicting the true effects, especially any part that persists over time, is likely to swamp the contribution of other observed hospital characteristics. Thus, the results in Table 8 highlight the value added of the filtered estimates over simply evaluating hospital quality based on hospital characteristics.

6. Conclusion

We have developed and applied a flexible, general, and systematic approach for assessing hospital quality, for use in both evaluation and forecasting. Compared to existing methods for assessing the quality of medical providers, our approach appears to have a number of advantages. It limits measurement costs: the method provides a foundation for identifying particular combinations of quality measures, out of the countless possible measures, that provide a true independent "signal" of important aspects of quality at a hospital. For this reason, it also avoids multidimensionality problems: rather than relying on ad hoc or possibly incorrect *a priori* considerations, it provides a firm empirical basis for the systematic integration of information on quality. Moreover, our method is far less computationally intensive than recently-published Bayesian approaches.

In addition, our method performs far better than alternative approaches for eliminating the noise problem that has plagued research on provider quality in health care. We demonstrate this using outcomes for serious illnesses, including mortality and major disease complications, that should complement the quality measures based on processes of care and less serious outcomes that are under development in many health plans and provider groups today. Our methods are able to extract 40 to 70 percent of the signal variance, and forecast 20 to 50 percent of signal variance one or two years ahead. The methods perform particularly well for AMI outcomes, where the quality of acute hospital care clearly has an important impact on long-term outcomes. We thus demonstrate that it is feasible to collect and integrate outcome information for relatively infrequent conditions, and thereby to evaluate explicitly the quality of care provided for many types of serious illness.

Finally, our results suggest that measures which use much more detailed medical data to account for differences in patient disease severity and comorbidity lead to quite similar predictions regarding provider quality, at least for patients with AMI. However, even if further research demonstrates that better "risk adjustment" does have substantial effects on quality measures, our methods will remain applicable. The "first stage" patient-level regressions to obtain the adjusted hospital measures would require use of more costly, detailed data, but the same VAR framework can be applied. Indeed, our methods can be used to identify when further risk adjustment has a substantial impact on evaluating and forecasting "risk-adjusted" quality, and how measures based on detailed clinical reviews can be integrated optimally with measures based on lower-cost, less-detailed records. Taken together, our research suggests that making reliable, precise predictions about provider quality in health care – and perhaps in many other industries -- may be far more feasible than many now believe.

References

Chamberlain, Gary, "Panel Data" Chapter 22 in *Handbook of Econometrics, Vol. II*, Eds. Z. Griliches and M.D. Intriligator, Elsevier Science, New York, 1984, 1247-1318.

Goldman, Lee, and E.F. Cook, "The decline in ischemic heart disease mortality rates. An analysis of the comparative effects of medical interventions and changes in lifestyle. *Annals of Internal Medicine*. 1984;101(6):825-36.

Hamilton, James, "State-Space Models," Chapter 50 in *Handbook of Econometrics, Vol. IV*, Eds. R.F. Engle and D.L. McFadden, Elsevier Science, New York, 1994, 3039-3080.

Heidenreich, Paul, and McClellan, Mark, "The Benefits of Technological Change in Heart Attack Care: A Literature Review and Synthesis," Stanford University mimeo, 1998.

Holtz-Eakin, Douglas, Whitney Newey and Harvey Rosen, "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56(6), 1988, 1371-95.

Hunink, Maria, Lee Goldman, A.N. Tosteson, et al. The recent decline in mortality from coronary heart disease, 1980-1990. The effect of secular trends in risk factors and treatment. *Journal of the American Medical Association*, 277(7), 1997, 535-42.

Kessler, Daniel, and Mark McClellan, "Do Doctors Practice Defensive Medicine?" *Quarterly Journal of Economics*, May 1996.

Krakauer, H., RC Bailey, KJ Skellan, et al., "Evaluation of the HCFA model for the analysis of mortality following hospitalization," *Health Services Research*, 27(3), 1992, 317-335.

Landon, B., LI Iezzoni, AS Ash, et al., "Judging hospitals by severity-adjusted mortality rates: the case of CABG surgery," *Inquiry*, 33(2), 1996, 155-166.

Luft, Harold S., Deborah W. Garnick, D.H. Mark, et al., 1990, *Hospital Volume, Physician Volume, and Patient Outcomes: Assessing the Evidence*, Ann Arbor, MI: Health Administration Press, 1994.

McClellan , Mark, Barbara McNeil, and Joseph Newhouse, "Does More Intensive Treatment of Acute Myocardial Infarction Reduce Mortality?" *Journal of the American Medical Association*, Sept. 1994.

McClellan, Mark, and Joseph Newhouse, "The Marginal Costs and Benefits of Medical Technology," *Journal of Econometrics*, Spring 1997.

McClellan, Mark and Haruko Noguchi, "Does High-Tech Mean Low Value? Technological Change in Heart Attack Care," *American Economic Review Papers and Proceedings*, May 1998a.

McClellan, Mark and Haruko Noguchi, "Validity and interpretation of treatment effect estimates using observational data: treatment of heart attacks in the elderly," manuscript, Stanford University, February 1998b.

Morris, Carl, "Parametric Empirical Bayes Inference: Theory and Applications," *JASA*, 78(381), 1983, 47-55.

Normand, Sharon-Lise T., Mark E. Glickman and Constantine A. Gatsonis, "Statistical methods for profiling providers of medical care: issues and applications," *JASA*, 92(439), 1997, 803-814.

Park, RE, RH Brook, J Kosecoff, et al., "Explaining variations in hospital death rates: Randomness, severity of illness, quality of care," *JAMA*, 264(4), 1990,484-90.

Watson, Mark, "Vector Autoregressions and Cointegration," Chapter 47 in *Handbook of Econometrics, Vol. IV*, Eds. R.F. Engle and D.L. McFadden, Elsevier Science, New York, 1994, 2843-2915.

</ref_section>

Table 1
Summary statistics for selected years
Means and standard deviations (in parentheses)

	AMI Patients		IHD Patients	
	1984	1994	1984	1994
Number of Admits	53.6 (45.2)	58.5 (56.2)	82.0 (78.8)	95.4 (119.4)
7-day mortality	0.131 (0.085)	0.083 (0.079)	0.010 (0.020)	0.006 (0.019)
30-day mortality	0.187 (0.094)	0.124 (0.095)	0.028 (0.032)	0.016 (0.034)
90-day mortality	0.220 (0.096)	0.149 (0.099)	0.049 (0.044)	0.031 (0.044)
365-day mortality	0.281 (0.099)	0.210 (0.107)	0.108 (0.061)	0.080 (0.065)
365-day complications	0.226 (0.081)	0.182 (0.082)	0.202 (0.077)	0.165 (0.079)

Unweighted means and standard deviations computed from a sample of 3954 hospitals with at least 3 admissions in each year for each diagnosis.
Mortality and complications variables are estimated intercepts from patient-level regressions run separately by year and diagnosis, controlling for age, gender, race, MSA, and rural location. Control variables were demeaned, so that the mean values reported in table represent the average hospital's mortality and complication rates and are not affected by the inclusion of control variables.

Table 2

Estimates of bivariate VAR(1) parameters for hospital-specific effects.
(Standard errors of estimates in parentheses).

		Bivariate VAR(1) of DTH30 and:	
		CMP365	IHD_DTH90
Parameter estimates:			
Lag coefficients			
DTH30:	Φ_{11}	0.887	0.920
		(0.012)	(0.014)
	Φ_{21}	-0.017	-0.177
		(0.022)	(0.066)
2nd outcome:	Φ_{22}	0.973	0.606
		(0.020)	(0.044)
	Φ_{12}	0.038	0.020
		(0.009)	(0.007)
Innovations			
Variance of DTH30 innovation		0.00036	0.00034
		(0.00004)	(0.00004)
Variance of 2nd innovation		0.00009	0.00012
		(0.00003)	(0.00002)
Correlation of the innovations		-0.589	0.384
		(0.104)	(0.086)
Initial conditions			
Variance of DTH30 in 1984		0.00175	0.00176
		(0.00012)	(0.00012)
Variance of 2nd outcome in 1984		0.00067	0.00024
		(0.00007)	(0.00004)
Correlation in 1984		-0.483	0.341
		(0.047)	(0.063)
Specification Tests:			
P-value for GMM goodness-of-fit test		0.035	0.029
P-value for test of independence of DTH30 and 2nd outcome		0.000	0.000
P-value for test of restrictions from VAR(2) to VAR(1)		0.341	<0.001

DTH30 are the intercepts for mortality within 30 days among AMI admissions.
CMP365 are the intercepts for readmission with complication between 30 and 365 days among AMI admissions.
IHD_DTH90 are the intercepts for mortality within 90 days among IHD admissions.

Table 3

Estimates of bivariate VAR(1) parameters for hospital-specific effects.
(Standard errors of estimates in parentheses).

Bivariate VAR(1) using: *1st outcome:*	AMI DTH7	AMI DTH30	IHD DTH7	IHD DTH90
2nd outcome:	AMI DTH7-30	AMI DTH30-365	IHD DTH7-90	IHD DTH90-365
Parameter estimates:				
Lag coefficients				
1st outcome: Φ_{11}	0.813 (0.022)	0.898 (0.011)	0.856 (0.056)	0.578 (0.047)
Φ_{21}	0.434 (0.146)	0.137 (0.057)	-0.013 (0.024)	0.187 (0.038)
2nd outcome: Φ_{22}	0.966 (0.082)	0.944 (0.043)	0.351 (0.065)	0.342 (0.049)
Φ_{12}	0.009 (0.011)	0.023 (0.006)	0.416 (0.156)	0.430 (0.062)
Innovations				
Variance of 1st innovation	0.00039 (0.00005)	0.00040 (0.00005)	0.00000 (0.00000)	0.00011 (0.00002)
Variance of 2nd innovation	0.00001 (0.00001)	0.00004 (0.00002)	0.00015 (0.00002)	0.00024 (0.00003)
Correlation of the innovations	-0.302 (0.202)	-0.779 (0.147)	0.295 (0.204)	-0.361 (0.073)
Initial conditions				
Variance of 1st outcome in 1984	0.00153 (0.00014)	0.00173 (0.00013)	0.00002 (0.00000)	0.00021 (0.00004)
Variance of 2nd outcome in 1984	0.00005 (0.00002)	0.00020 (0.00004)	0.00008 (0.00005)	0.00028 (0.00005)
Correlation in 1984	0.332 (0.165)	-0.299 (0.082)	0.792 (0.286)	0.135 (0.129)
Specification Tests:				
P-value for GMM goodness-of-fit test	0.056	0.008	0.020	0.001
P-value for test of independence of the two outcomes	0.000	0.000	0.000	0.000

DTH7, DTH30 and DTH90 are the intercepts for mortality within 7, 30 and 90 days.
DTH7-30 (DTH7-90) are the change in the mortality intercepts between 7 and 30 (90) days.
DTH30-365 (DTH90-365) are the change in mortality intercepts between 30 (90) and 365 days.

Table 4

Comparison of VAR parameters in low and high volume hospitals.
Bivariate VAR(1) for AMI 30-day and IHD 90-day mortality.
(Standard errors of estimates in parentheses).

	Full Sample	Low Volume: <25 admits in at least 1 year	High Volume: ≥25 admits in all years
Parameter estimates:			
Lag coefficients			
AMI_DTH30: Φ_{11}	0.920	0.911	0.912
	(0.014)	(0.020)	(0.014)
Φ_{21}	-0.177	-0.325	-0.055
	(0.066)	(0.095)	(0.056)
IHD_DTH90: Φ_{22}	0.606	0.531	0.813
	(0.044)	(0.061)	(0.028)
Φ_{12}	0.020	0.013	0.005
	(0.007)	(0.010)	(0.006)
Innovations			
Var. of AMI_DTH30 innovation	0.00034	0.00057	0.00016
	(0.00004)	(0.00008)	(0.00002)
Var. of IHD_DTH90 innovation	0.00012	0.00022	0.00003
	(0.00002)	(0.00004)	(0.00000)
Correlation of the innovations	0.384	0.476	0.555
	(0.086)	(0.101)	(0.098)
Initial conditions			
Var. of AMI_DTH30 in 1984	0.00176	0.00235	0.00109
	(0.00012)	(0.00024)	(0.00008)
Var. of IHD_DTH90 in 1984	0.00024	0.00035	0.00013
	(0.00004)	(0.00009)	(0.00002)
Correlation in 1984	0.341	0.330	0.464
	(0.063)	(0.087)	(0.063)
Specification Tests:			
P-value for GMM goodness-of-fit test	0.029	0.033	<0.001
P-value for test of independence of AMI_DTH30 and IHD_DTH90	0.000	0.000	0.000
Sample size:	3954	1943	2011

AMI_DTH30 are the intercepts for mortality within 30 days among AMI admissions.
IHD_DTH90 are the intercepts for mortality within 90 days among IHD admissions.

Table 5

Comparison of claims-based outcomes measures to chart-based outcomes measures.
All comparisons based on data for AMI admissions from the CCP Project, 1994-95.
(Standard errors of estimates in parentheses).

	Correlation of claims-based outcome with chart-based outcome	Estimate of slope coefficient for $\mu_{chart} = \beta\mu_{claim}$
DTH7	0.95 (0.01)	0.81 (0.03)
DTH30	0.91 (0.01)	0.77 (0.03)
DTH365	0.80 (0.04)	0.69 (0.06)
CMP365	0.96 (0.02)	0.91 (0.04)

Estimates computed from a sample of 3622 hospitals.
DTH7 are intercepts for mortality within 7 days.
DTH30 are intercepts for mortality within 30 days.
DTH365 are intercepts for mortality within 365 days.
CMP365 are intercepts for readmission with complication between 30 and 365 days.

Table 6

Summary of estimated prediction accuracy using alternative methods of signal extraction.
All estimates based on VAR(1) models from Table 2.

	Expected R-squared of prediction based on:					
	All nine years of data for:		Three most recent years of data for:		Concurrent year of data only for:	
	Both outcomes	Same outcome	Both outcomes	Same outcome	Both outcomes	Same outcome
A. Based on model of AMI DTH30 and AMI CMP365						
For AMI DTH30:						
1988	0.71	0.71	0.58	0.58	0.42	0.34
1992	0.66	0.66	0.63	0.63	0.48	0.40
For AMI CMP365:						
1988	0.65	0.64	0.46	0.45	0.28	0.20
1992	0.60	0.58	0.51	0.50	0.33	0.23
B. Based on model of AMI DTH30 and IHD DTH90						
For AMI DTH30:						
1988	0.73	0.72	0.60	0.59	0.44	0.35
1992	0.68	0.67	0.64	0.64	0.51	0.40
For IHD DTH90:						
1988	0.51	0.50	0.46	0.45	0.38	0.27
1992	0.52	0.51	0.52	0.51	0.47	0.32

All expected R-squared values refer to a weighted R-squared, with weights proportional to the number of admissions at each hospital.
AMI DTH30 are the intercepts for mortality within 30 days among AMI admissions.
AMI CMP365 are the intercepts for complication within 365 days among AMI admissions.
IHD DTH90 are the intercepts for mortality within 90 days among IHD admissions.

Table 7

Summary of forecast accuracy using alternative forecasting models.
Forecasting 1993 and 1994 values using data from 1984 to 1992.

	Modified R-squared of forecast based on:						
	VAR(1), forecasting with:		VAR(2), forecasting with:		Alternative forecast methods		
	Both outcomes, 84-92 data	Same outcome, 84-92 data	Both outcomes, 84-92 data	Same Outcome, 84-92 data	Shrinkage estimator, 1992 data	Outcome in 1992	Average outcome, 84-92 data
A. Based on model of AMI_DTH30 and AMI_CMP365							
For AMI_DTH30:							
1993 actual	0.46	0.47	0.47	0.47	0.19	< 0	0.30
expected	*0.52*	*0.52*	*0.51*	*0.50*	*0.31*	*< 0*	*0.22*
1994 actual	0.40	0.41	0.40	0.41	0.15	< 0	0.21
expected	*0.42*	*0.41*	*0.41*	*0.40*	*0.23*	*< 0*	*0.10*
For AMI_CMP365:							
1993 actual	0.64	0.63	0.64	0.63	0.27	< 0	0.31
expected	*0.53*	*0.52*	*0.54*	*0.52*	*0.20*	*< 0*	*0.24*
1994 actual	0.57	0.57	0.57	0.57	0.25	< 0	0.21
expected	*0.47*	*0.46*	*0.49*	*0.47*	*0.18*	*< 0*	*0.17*
C. Based on model of AMI DTH30 and IHD DTH90							
For AMI DTH30:							
1993 actual	0.47	0.47	0.48	0.48	0.19	< 0	0.30
expected	*0.55*	*0.55*	*0.52*	*0.52*	*0.32*	*< 0*	*0.28*
1994 actual	0.41	0.41	0.41	0.41	0.15	< 0	0.21
expected	*0.45*	*0.45*	*0.44*	*0.44*	*0.25*	*< 0*	*0.17*
For IHD DTH90:							
1993 actual	0.21	0.19	0.20	0.19	< 0	< 0	< 0
expected	*0.21*	*0.20*	*0.15*	*0.14*	*0.08*	*< 0*	*< 0*
1994 actual	0.19	0.18	0.21	0.20	< 0	< 0	< 0
expected	*0.09*	*0.08*	*0.11*	*0.10*	*< 0*	*< 0*	*< 0*

See notes to Table 6.

Table 8

Comparison of estimated prediction accuracy with and without controlling for patient volume.
All estimates based on VAR(1) models.

	Expected R-squared of prediction based on:				
	All nine years of data for:		Concurrent year of data only for:		Coefficient for volume/100 (s.e.)
	Both outcomes	Same outcome	Both outcomes	Same outcome	
A. Based on model of AMI DTH30 and AMI CMP365					
For AMI DTH30, 1988:					
covariates: None	0.71	0.71	0.42	0.34	
Volume	0.72	0.72	0.50	0.39	-0.0188 (0.0005)
For AMI CMP365, 1988:					
covariates: None	0.65	0.64	0.28	0.20	
Volume	0.65	0.64	0.28	0.21	0.0024 (0.0002)
B. Based on model of AMI DTH30 and IHD DTH90					
For AMI DTH30, 1988:					
covariates: None	0.73	0.72	0.44	0.35	
Volume	0.74	0.73	0.51	0.40	-0.0188 (0.0005)
For IHD DTH90, 1988:					
covariates: None	0.51	0.50	0.38	0.27	
Volume	0.52	0.51	0.39	0.30	-0.0013 (0.0001)

See notes to Table 6.

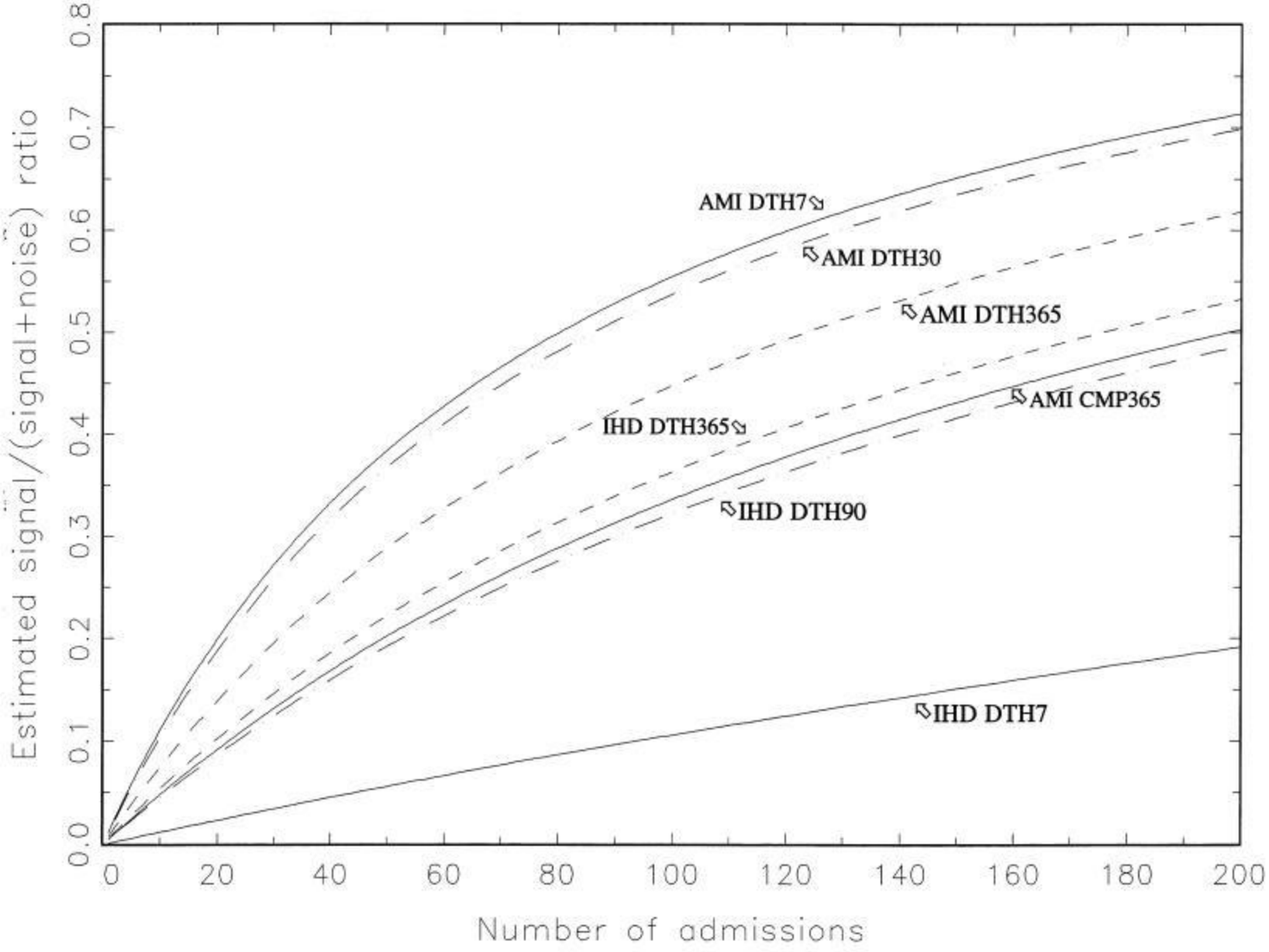

Figure 1. Estimate of signal to noise ratio (signal variance as a percent of total variance) for various outcome measures.

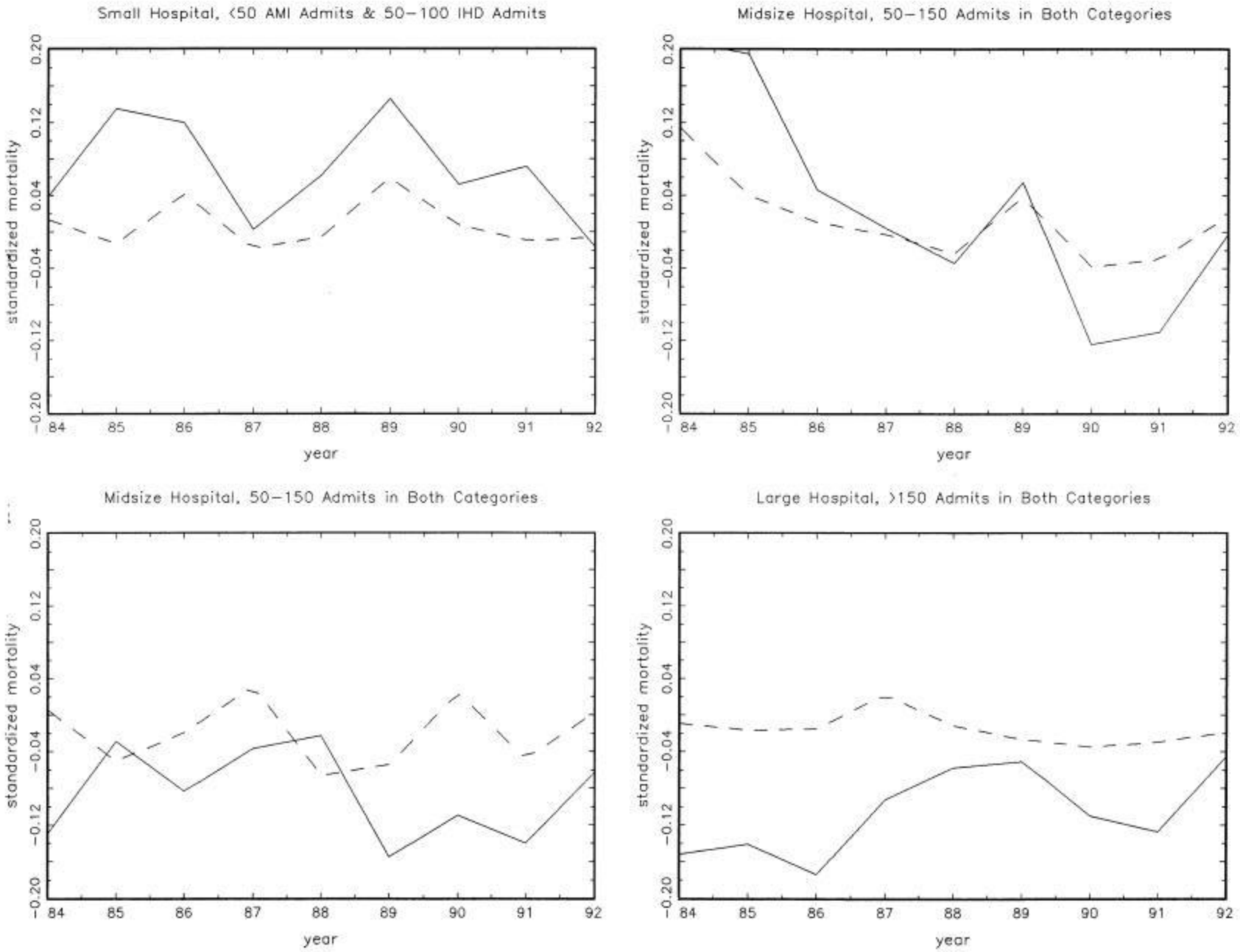

Figure 2. Trends in standardized 30-day mortality rates for AMI (solid line) and 90-day mortality for IHD (dashed line) for four selected hospitals. Rates are based on all Medicare admissions.

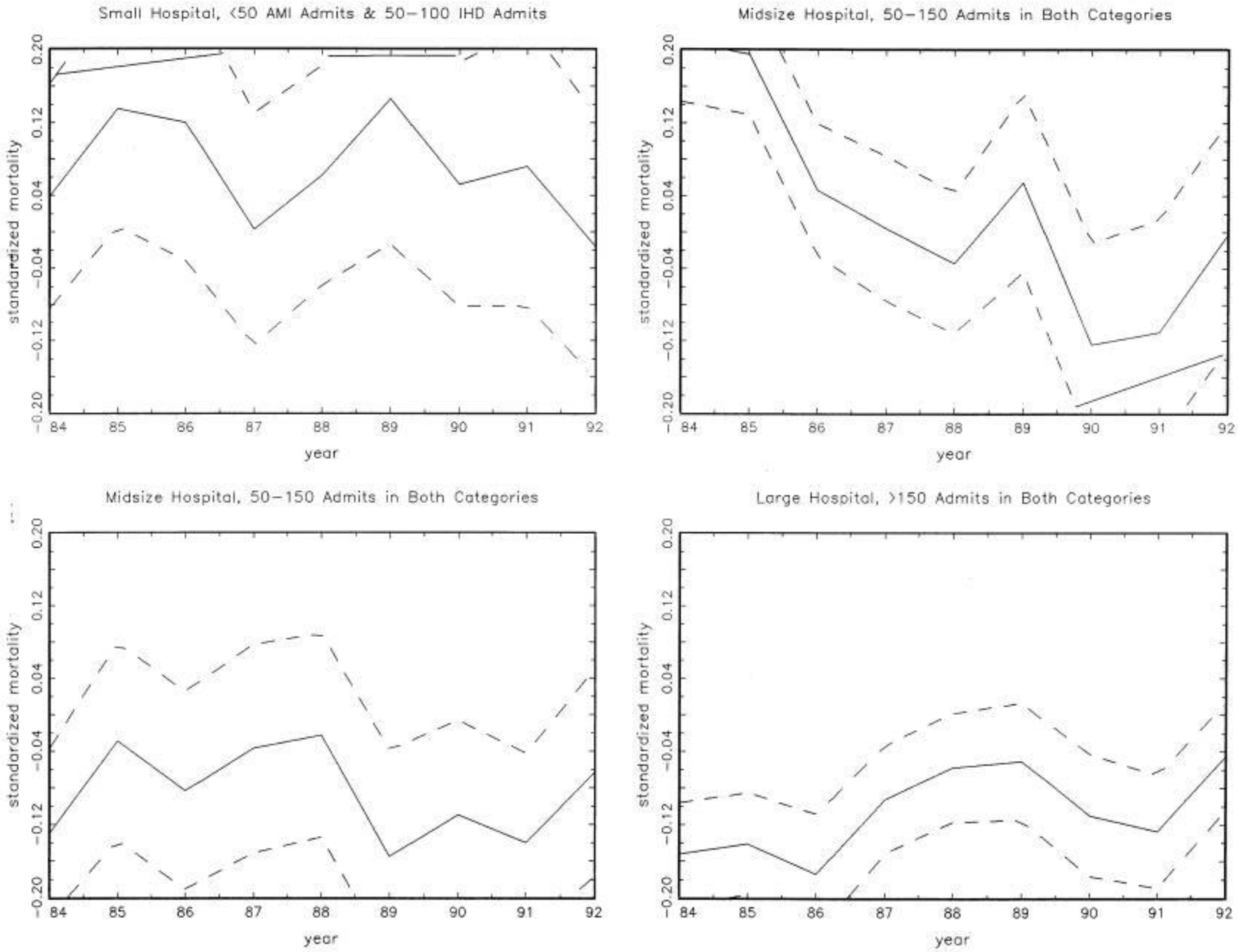

Figure 3a. Trends for AMI admissions: standardized 30-day mortality rates (solid line) and 95% confidence interval (dashed line) for four selected hospitals. Rates are based on all Medicare admissions.

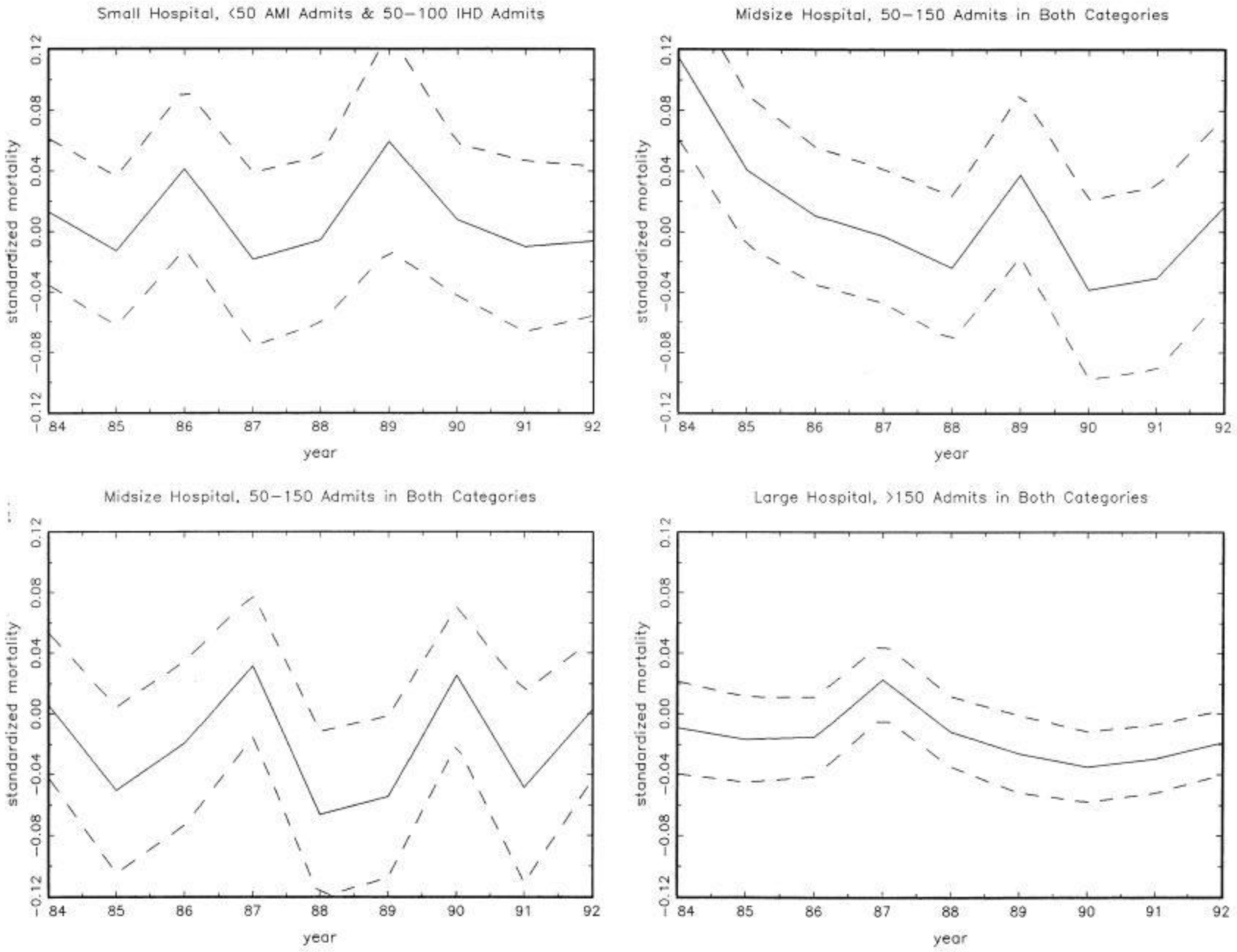

Figure 3b. Trends for IHD admissions: standardized 90-day mortality rates (solid line) and 95% confidence interval (dashed line) for four selected hospitals. Rates are based on all Medicare admissions.

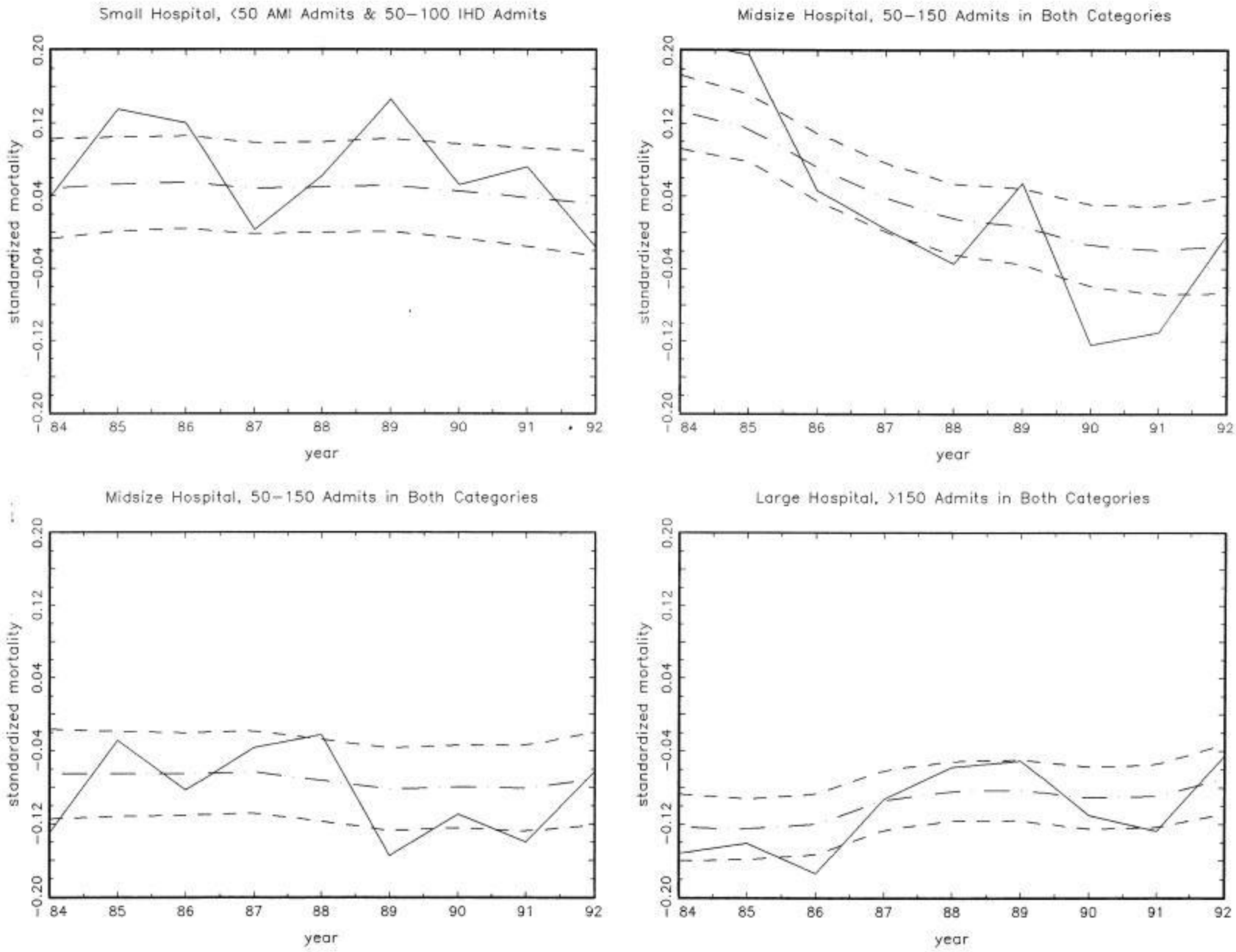

Figure 4a. Trends for AMI admissions: Actual values (solid line), predicted values (long dashes) and 95% confidence interval (short dashes) for standardized 30-day mortality rates. Rates based on all Medicare admissions.

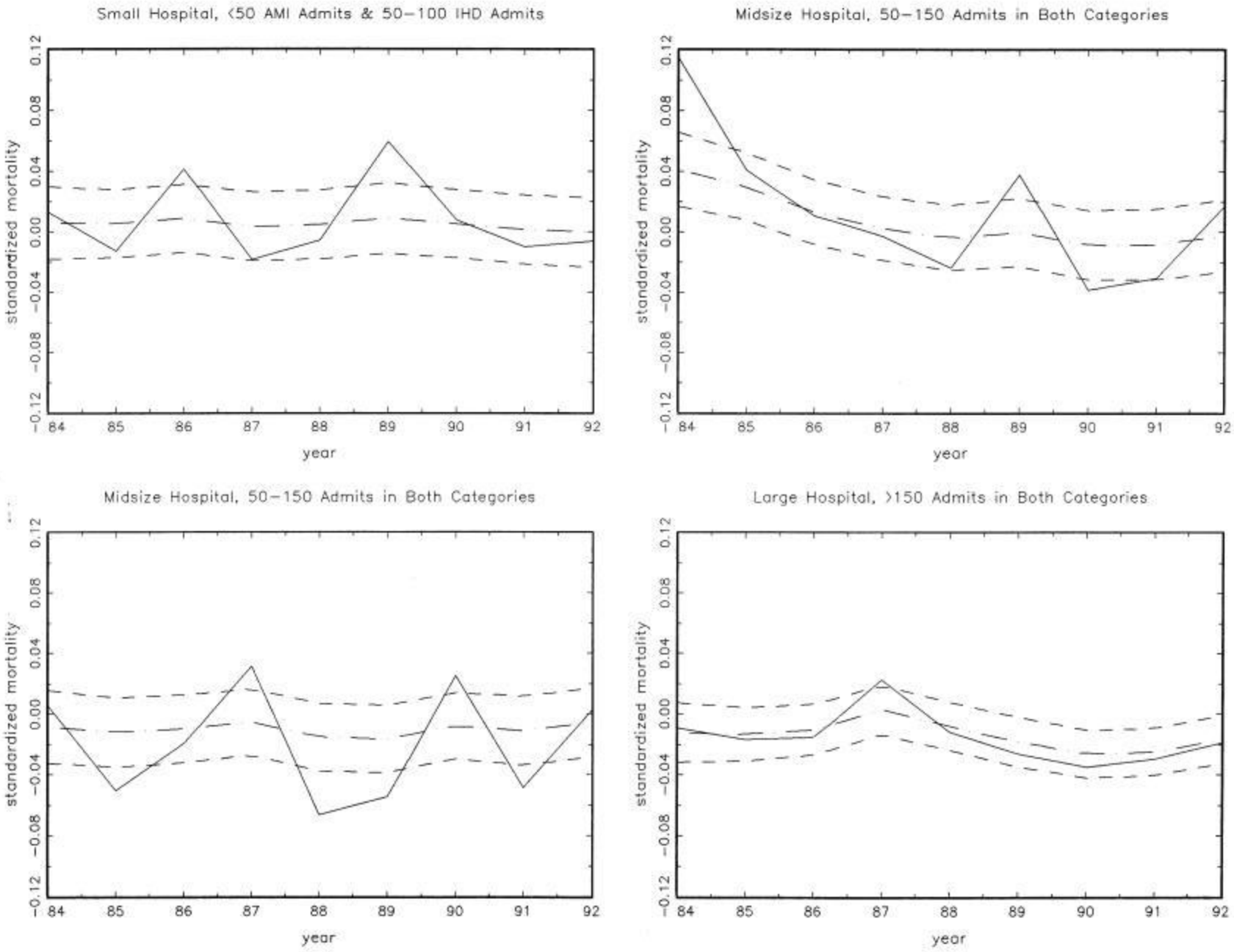

Figure 4b. Trends for IHD admissions: Actual values (solid line), predicted values (long dashes) and 95% confidence interval (short dashes) for standardized 90-day mortality rates. Rates based on all Medicare admissions.

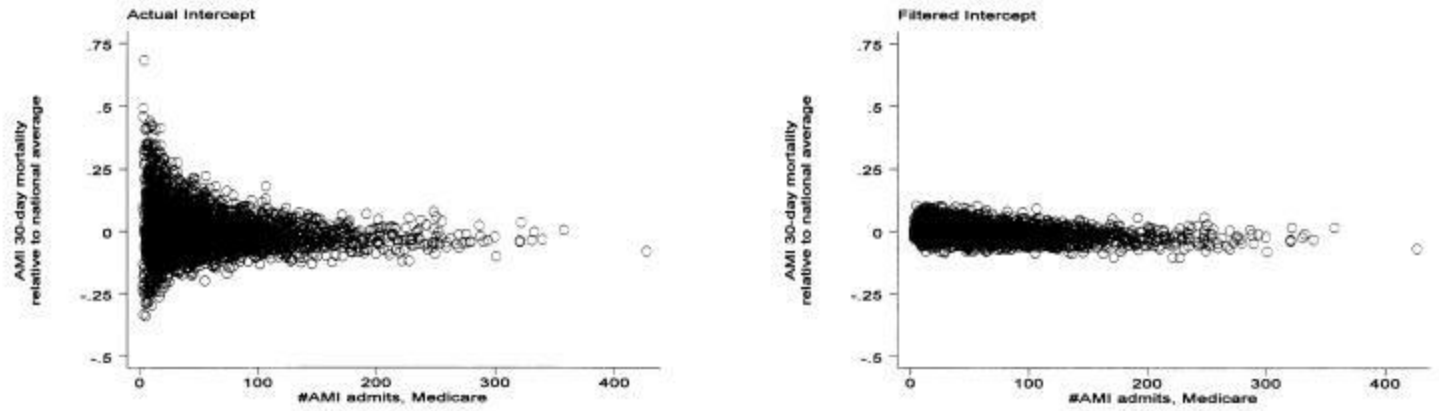

Figure 5. Plots of actual and filtered hospital-specific intercepts against patient volume. Based on 30-day mortality for Medicare AMI admits, 1992.