7   The Covariance Structure of Mortality Rates in Hospitals

Douglas Staiger

7.1   Introduction

In 1987 the Health Care Financing Administration (HCFA) began publishing standardized mortality scores for each hospital in the country as an indicator of quality of care (Bowen and Roper 1987). Since then, the use of similar patient mortality measures has become widespread, to the point where some insurance plans base hospital reimbursement on such mortality measures (Minnesota Blues 1991; Perry 1989). At the same time, patient mortality has been widely used in studies of the determinants of quality of care in hospitals (Cutler 1995; Garber, Fuchs, and Silverman 1984; Luft, Hughes, and Hunt 1987; McClellan, McNeil, and Newhouse 1994; Staiger and Gaumer 1995) and in studies of patient choice among hospitals (Luft, Hunt, and Maerki 1987; Luft et al. 1990; Staiger 1993). Despite the widespread use of patient mortality as a proxy for quality of care, there is considerable controversy over the statistical reliability and validity of such measures (Hofer and Hayward 1996; Keeler et al. 1992; Krakauer et al. 1992; Luft and Romano 1993; McNeil, Pedersen, and Gatsonis 1992; Park et al. 1990).

A number of questions are of particular interest to both the research and the policy communities. First, how much useful information is there in such inherently noisy measures of quality; for example, how large is the signal-to-noise ratio? A related question is how persistent are these measures of quality of care: Are hospitals with unexpectedly high mortality rates this year likely to have unexpectedly high mortality rates next year, in five years, in ten years? Is

the presumption of high persistence, commonly assumed by both the public and by analysts doing fixed-effect models, consistent with the data? A third question is whether there is a correlation in patient mortality for patients with distinct diagnoses admitted to the same hospital? If so, then combining information from different types of patients may prove to be a useful way of summarizing common hospital-level components of quality of care. A final question of interest is what has happened to the cross-sectional distribution of patient mortality over time; for example, has there been convergence or divergence across hospitals? Have there been any noticeable changes in the variation of these measures in recent years as reimbursement and competitive pressures have grown?

This paper uses annual data from 1974–87 for 492 large hospitals to investigate these questions. I analyze data on standardized mortality rates for Medicare admissions in both specific diagnoses and in aggregate. In addition to presenting simple descriptive evidence on the distribution of the mortality measures, I estimate covariance structures using general method of moment (GMM) methods along the lines of MaCurdy (1982) and Abowd and Card (1989). This method provides a simple and powerful description of the basic features of the data.

The empirical work leads to a number of interesting conclusions. First, 75 to 90 percent (depending on the diagnosis) of the variance in mortality is entirely transitory and can be thought of as independent identically distributed (i.i.d.) measurement error. Second, the remaining nontransitory component of mortality is fairly persistent but is in general not well approximated as a permanent fixed effect. Instead, the nontransitory component is fairly well described as an autoregressive (AR(1)) process with a first-order serial correlation of .8–.95 depending on the diagnosis. Third, the combined data are fit fairly well by a simple three-factor model in which mortality consists of (1) i.i.d. error, (2) a fairly transitory diagnosis-specific component, and (3) a very permanent hospital component that is common across diagnoses. Finally, although there are some interesting changes in the cross-sectional distribution of these mortality measures over time (particularly during the 1970s), there is no obvious evidence that these distributions have tended to converge or diverge over time or changed in any interesting ways during the 1980s.

The key difficulty in interpreting these empirical results is the possibility that much of what we observe in these measures may reflect systematic variation in unobserved patient characteristics rather than quality of care. However, to the extent that this variation reflects quality of care, there are a number of useful lessons to be learned. Clearly, the large i.i.d. component in these mortality measures limits the usefulness of historical measures as indicators of current quality. Similarly, to the extent that this i.i.d. component is measurement error, these mortality measures are a poor choice as independent variables to proxy for quality of care. The presence of an important component of mortality that is serially correlated raises questions about the bias of hospital fixed-effect

models that have been used to analyze mortality. Finally, the three-factor model is consistent with the notion that quality of care depends on both hospital-level infrastructure (e.g., nursing staff, physical plant), which is relatively unchanging, and diagnosis-specific technology (e.g., surgical techniques, medications), which may be a large source of the variation for some diagnoses but disseminates relatively quickly.

The paper proceeds as follows. Section 7.2 describes the data and estimation approach. Section 7.3 presents simple descriptive evidence on how the mortality rate distribution has evolved over time. Section 7.4 estimates the covariance structure for standardized mortality rates. Section 7.5 concludes.

7.2 Data Sources and Methods

7.2.1 Data

The data are derived from a 25 percent random sample of all short-term general hospitals in the continental United States, developed by Abt Associates. Of these hospitals, only those operating continuously from 1974 to 1987 were included in the analysis. Each observation in the data corresponds to a hospital-year from 1974–87, yielding 14 observations per hospital.

The data set contains information on mortality for acute myocardial infarction (AMI) admissions, congestive heart failure (CHF) admissions, and urgent care admissions (an aggregate group that includes AMI and CHF). These three categories are discussed in more detail below. To be included in the sample, a hospital had to have at least one admission in each diagnosis category in every year. This effectively limits the sample to relatively large, urban hospitals. The final sample includes 492 hospitals, with 14 years of data on each hospital.

I use data on 45-day postadmission mortality for a subsample of Medicare patients, chosen on the basis of 59 conditions necessitating urgent admission (see Gaumer, Poggio, and Coelen 1989). These 59 conditions were selected by clinical panels to include cases for which adverse mortality outcomes might reasonably result as the result of care received. These urgent care conditions accounted for just over 12 percent of all Medicare admissions in 1987. Among the urgent care admissions, AMI and CHF are the most frequent diagnoses and account for roughly one-half of the admissions. Mortality measures are available for AMI and CHF separately, and for urgent care diagnoses as a whole.

The mortality data come from a 20 percent sample of Medicare discharge records (the MEDPAR data) combined with social security death records through 1989. For example, the 45-day mortality rate is based on the fraction of urgent care patients admitted during the calendar year that had a date of death within 45 days of admission. Note that these data include all deaths, not just those in the hospital.

Differences in mortality rates across time and across hospitals may reflect

differences in patient mix rather than differences in quality of care. Therefore, I use an expected mortality rate (similar to that used by HCFA in their mortality reporting) for each hospital-year to control for the variation in mortality due to variation in patient mix. The expected mortality rate is based on a reference population of Medicare patients drawn from the MEDPAR data.[1] Mortality rates were computed for the reference population for each of 354 cells defined by diagnosis/procedure (59 groups), gender (2 groups), and age (65–74, 75–84, and over 85). Each study patient was assigned an expected mortality equal to the mean value for the applicable cell. The hospital-year expected mortality rate is the average expected mortality for study patients within each hospital year.

Raw 45-day mortality rates (MR) are transformed into Z-scores by subtracting the expected mortality rate (EMR) and dividing by the estimated standard deviation in mortality. Raw mortality rates display considerable heteroscedasticity, with the variance inversely related to the number of admissions (N). The standard deviation is estimated assuming that mortality is binomially distributed, with the probability of death equal to the expected mortality rate. Thus the mortality Z-score is given by

$$Z = (MR - EMR)/\text{sqrt}[EMR(1 - EMR)/N].$$

This type of Z-score measure is a relatively common method of standardizing across hospitals of different sizes (see, e.g., Luft et al. 1990).

7.2.2   Methods

Modeling and estimating the error structure for a set of variables with panel data is conceptually straightforward. I follow an approach similar to that used by MaCurdy (1982) and Abowd and Card (1989), who use panel data on hours and earnings to decompose the error structure into permanent and transitory components. Since these methods are fairly well known, I will not go into detail on them here.

The basic approach is to choose an error structure for Z_{it} that depends on a $K \times 1$ vector (Θ) of unknown parameters, derive the implied covariance matrix ($\Omega(\Theta)$) for $Z = [Z_{74}, Z_{75}, \ldots, Z_{87}]$, and estimate the unknown parameters by fitting the implied covariance matrix to the actual covariance matrix. For example, if the mortality Z-score is composed of a hospital fixed effect plus i.i.d. noise, then the covariance matrix of Z takes the simple form

$$\Omega_{t,t} = \Theta_0 + \Theta_1$$

$$\Omega_{t,t-k} = \Theta_0 \qquad \text{if } k > 0,$$

---

1. Unfortunately, the reference population changes in 1979. For the years 1974–78, the reference population is all MEDPAR admissions in those years. For the remaining years, the reference population is all MEDPAR admissions during 1979–83. Consequently, expected mortality rates (averaged over the whole sample) take a discrete jump (downward) in 1979. I have rescaled expected mortality rates after 1979 so that, on average, 1978 and 1979 expected mortality rates are equal.

where Θ_1 represents the variance of the i.i.d. measurement error (which only appears in the variances) and Θ_0 represents the variance of the fixed effect (which determines all the off-diagonal covariances).
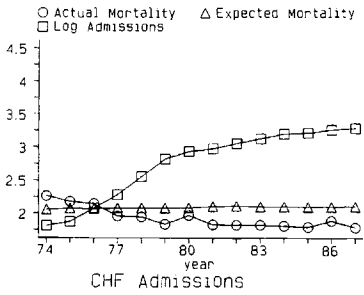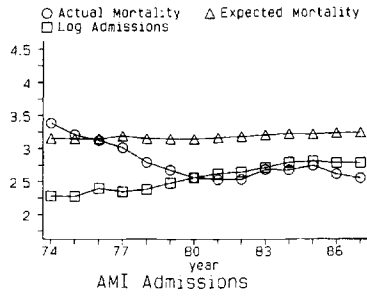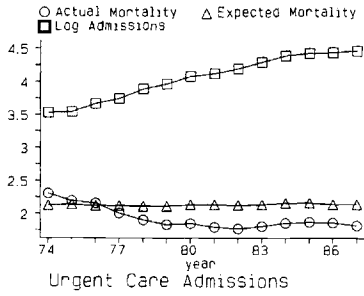
GMM estimates of Θ minimize an optimally weighted sum of squared deviations between the actual and theoretical covariance matrices. Let M be a $J \times 1$ vector of the nonredundant elements of the sample covariance matrix (so if Z is $N \times L$, the sample covariance matrix $Z'Z/N$ has $L \cdot (L + 1)/2$ distinct elements), and let $m(\Theta)$ be a $J \times 1$ vector of the corresponding theoretical moments. Then Θ_{GMM} minimizes the statistic $Q(\Theta) = N \cdot [M - m(\Theta)]' \, V^{-1} [M - m(\Theta)]$, where V is the sample covariance matrix of M (fourth moments of Z). The statistic Q evaluated at the GMM estimate is distributed as chi-squared with $J - K$ degrees of freedom and therefore provides a goodness-of-fit test for the model. Furthermore, parameter restrictions are easily tested using the statistic $L = Q_R - Q$, where Q_R and Q are the goodness-of-fit statistics from the restricted and unrestricted models. Under the null corresponding to the restricted model, L is distributed as chi-squared with degrees of freedom equal to the difference in the degrees of freedom of Q_R and Q.

The GMM framework provides a natural method of testing how increasingly restrictive models fit the data. I first test whether the data can be fit well with a stationary covariance structure, that is, a structure in which $\text{Cov}(Z_t, Z_{t-k})$ only depends on k. I then test even more restrictive error structures against the stationary but otherwise unconstrained model.
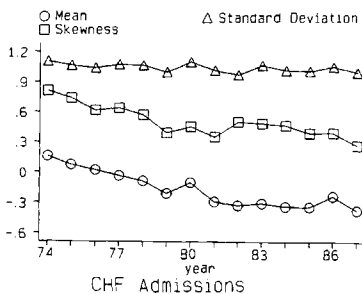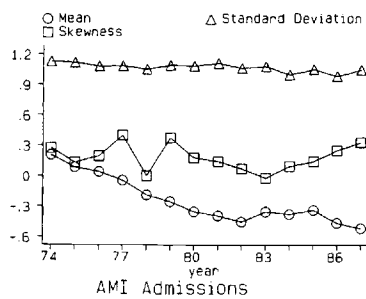
7.3   Descriptive Evidence

Figures 7.1–7.3 summarize some of the basic facts and trends in 45-day mortality rates for urgent care, AMI, and CHF admissions. Figure 7.1 plots trends in the raw data used to construct the mortality Z-scores from 1974 to 1987. The variables have been scaled to better fit on one graph: the mortality rate and expected mortality rate are given as deaths per 10 admits, while the number of admissions has been logged. Actual mortality rates fell for all three admission categories from 1974 until around 1980 and then flattened out. For example, AMI 45-day mortality rates fell from over 33 percent in 1974 to just over 25 percent in 1980. Expected mortality over this time period was surprisingly flat, suggesting that the decline in observed mortality was due to true improvements in quality of care rather than changing patient mix. At the same time that mortality was trending downward, patient volume was trending upward. For example, CHF admissions roughly tripled between 1974 and 1980.

Figure 7.2 plots trends in the first three moments (mean, variance, skewness) of the mortality Z-score measures. These trends highlight important changes in the distribution of the Z-score over time and provide some evidence on whether hospitals are converging or diverging over time. The trend in the mean of the Z-score parallels the trend in actual mortality, falling until 1980 and then remaining relatively flat for all diagnoses. In contrast, the variance of the Z-score is relatively stable, perhaps trending downward slightly for all diagnoses. The

4.5
4
3.5
3
2.5
2

74   77   80   83   86
             year
    Urgent Care Admissions

4.5
4
3.5
3
2.5
2

74   77   80   83   86
             year
       AMI Admissions

4.5
4
3.5
3
2.5
2

74   77   80   83   86
             year
    CHF Admissions

Fig. 7.1   Trends in actual and expected mortality per 10 admissions, and in the log of admissions, 1974–87

1.2
.9
.6
.3
0
-.3
-.6

74   77   80   83   86
             year
    Urgent Care Admissions

1.2
.9
.6
.3
0
-.3
-.6

74   77   80   83   86
             year
       AMI Admissions

1.2
.9
.6
.3
0
-.3
-.6

74   77   80   83   86
             year
    CHF Admissions
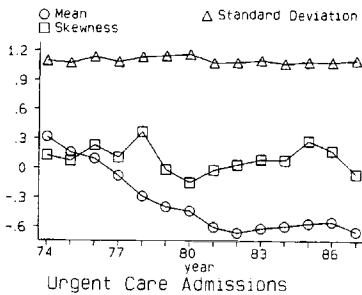
Fig. 7.2   Trends in mortality Z-score moments, 1974–87
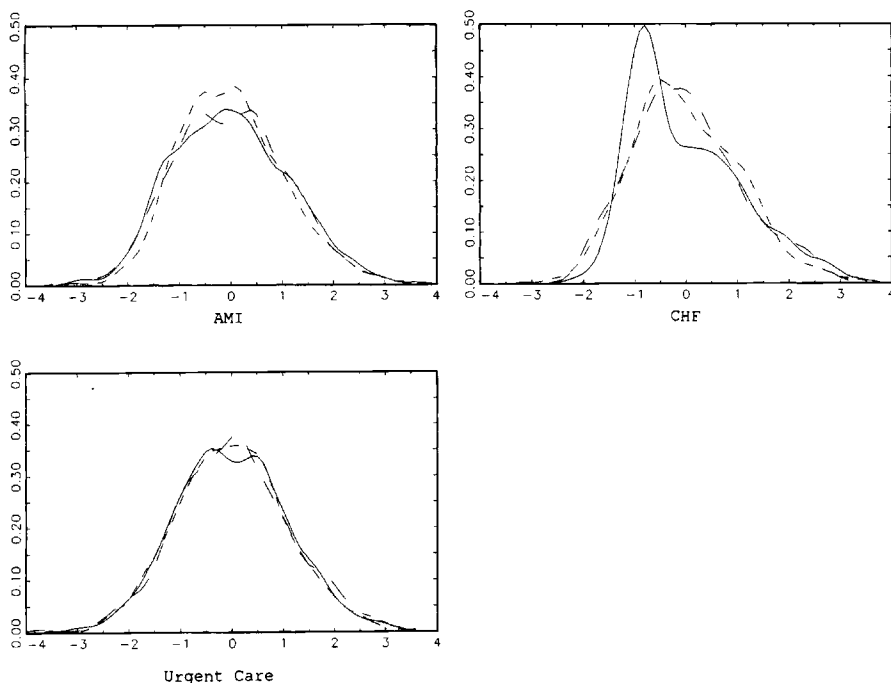
Fig. 7.3 **Empirical p.d.f.'s of mortality Z-score: 1974–75 (*solid line*), 1980–81 (*long dashes*), and 1986–87 (*short dashes*)**

stability of the variance over time, in spite of a large rise in the number of patients on which these measures are based, may seem surprising. However, recall that the Z-score rescales by an estimate of the standard deviation in mortality and therefore should not be overly sensitive to changes in the number of patients. Finally, the skewness of the Z-score distribution has fallen over time, particularly for CHF between 1974 and 1979.

The trends in the moments of the Z-score raise two issues. First, there is clearly a mean shift in mortality over time, and it will be important to remove year effects from this measure. More important, even after allowing for a mean shift, the distribution of the Z-score does not appear to be stationary. Thus a stationary covariance structure may fit the data poorly, particularly in the early years.

Figure 7.3 plots kernel estimates (using a Gaussian kernel) of the partial distribution functions (p.d.f.'s) as a further way of inspecting changes in the distribution of the Z-scores over time. Each panel plots p.d.f. estimates pooling two years of data for three distinct time periods: 1974–75, 1980–81, and 1986–87. Each year's data were demeaned in order to remove the time effects. The changes in the distributions between 1974–75 and 1980–81 are quite striking. The distribution of AMI mortality is relatively fat tailed and flat topped in 1974–75, while the distribution in later years is generally tighter and seems to

exhibit some bimodality. The CHF distribution has a large spike of low-mortality hospitals in 1974–75 with a long tail of higher mortality. By 1980–81 this spike has largely disappeared, and the distribution looks more symmetric.

Overall, a number of interesting facts emerge from figures 7.1–7.3. First, there is no obvious evidence of convergence in the cross-sectional distribution of mortality. Neither is there any striking evidence that this distribution has changed much during the 1980s in response to the changing reimbursement and market conditions facing hospitals. In fact, since some time around 1979–81 the distribution has been remarkably stable in both shape and location. In contrast, over the mid- to late 1970s the mortality distributions changed dramatically: average mortality fell, as did the spread and skewness of the distribution of standardized mortality rates across hospitals.

7.4   Estimates of the Covariance Structure

I now turn to more formal estimates of the covariance structure of the Z-score for mortality in each of the diagnosis categories. The data that provide the basis for these estimates are contained in the sample covariance matrix for each diagnosis category. Since there are 14 years of data, there are $14 \cdot 15/2 = 105$ distinct moments and potential degrees of freedom. I also consider the covariance between AMI mortality and CHF mortality. This matrix is not symmetric (e.g., $\text{Cov}(Z_{AMI,74}, Z_{CHF,75})$ is not equal to $\text{Cov}(Z_{AMI,75}, Z_{CHF,74})$) and therefore provides $14 \cdot 14 = 196$ total degrees of freedom.

Table 7.1 summarizes the results of a series of model specification tests. The first row contains the GMM goodness-of-fit statistic testing whether the data are consistent with a stationary covariance matrix (i.e., a symmetric covariance matrix with constant values along each diagonal). As might be expected from the changes in distributions that were apparent in figure 7.3, only urgent care mortality is fit well by a stationary covariance matrix. The AMI mortality data reject stationarity at the 2 percent level, while CHF and the AMI-CHF covariance overwhelmingly reject stationarity. From direct inspection of the covariance matrices, it is apparent that most of the poor fit is associated with differences between the 1974–78 period and later years.

Despite the poor fit of the unconstrained stationary model, the resulting estimates still provide information about the general covariance pattern found in the data. Figure 7.4 graphs the estimated autocovariances and corresponding 95 percent confidence intervals. These covariograms are the basic data that any stationary model of the error structure is trying to fit. The sharp drop-off in covariance at the first lag for all three diagnoses, combined with the lack of any large contemporaneous correlation between AMI and CHF mortality, suggests significant i.i.d. measurement error in mortality. The magnitude of the decline suggests that as much as 90 percent of the variance in mortality is i.i.d. error. The gradual decline in covariance at higher lags for urgent care and AMI suggests some kind of an AR process—an MA process would be more transitory,

Table 7.1   **Model Specification Tests for Covariance Structure of Mortality Z-Score, 1974–87 (based on GMM goodness-of-fit statistic)**

	For Covariance Matrix of			
Model Specification	Urgent Care	AMI	CHF	AMI′CHF
1. Unconstrained stationary	104.6	120.26	139.00	296.95
covariance (symmetric)	(.156)	(.022)	(.001)	(.000)
	[91]	[91]	[91]	[182]
Tested against model 1				
2. Fixed effect	8.50	16.77	23.48	11.95
+ i.i.d. error	(.290)	(.019)	(.001)	(.102)
+ ARMA(2,2)	[7]	[7]	[7]	[7]
3. Fixed effect	9.29	23.86	26.32	13.60
+ i.i.d. error	(.410)	(.005)	(.002)	(.137)
+ AR(2)	[9]	[9]	[9]	[9]
4. Fixed effect	10.76	27.92	31.04	13.72
+ i.i.d. error	(.377)	(.002)	(.001)	(.186)
+ AR(1)	[10]	[10]	[10]	[10]
5. i.i.d. error	12.68	37.25	32.13	13.92
+ AR(1)	(.315)	(.000)	(.001)	(.237)
	[11]	[11]	[11]	[11]
6. Fixed effect	77.36	85.90	37.06	15.74
+ i.i.d. error	(.000)	(.000)	(.000)	(.203)
	[12]	[12]	[12]	[12]
7. i.i.d. error	147.6	138.8	84.79	41.08
	(.000)	(.000)	(.000)	(.000)
	[13]	[13]	[13]	[13]

Note: Entries in table are GMM goodness-of-fit statistics. Numbers in parentheses are *p*-values; numbers in brackets are degrees of freedom.

while a fixed effect would be more permanent. In contrast, both CHF and the AMI-CHF covariance look small but fairly persistent, as would be expected with a fixed-effect model.

Although stationarity is not generally supported by the data, it is useful to consider still more restrictive models in order to see whether any simple error structure adequately summarizes the covariogram. Rows 2–7 of table 7.1 test increasingly restrictive models of the error structure against the unconstrained stationary model. I begin with a flexible model that includes a fixed effect, i.i.d. error, and an ARMA(2,2) component. The next two rows restrict the ARMA to be AR(2) and then AR(1). Finally, the last three rows remove the fixed effect or the AR(1) or both (leaving just i.i.d. error).

In general, these goodness-of-fit tests suggest that a simple model of i.i.d. error plus an AR(1) does about as well as any other model in fitting each of the covariance structures. Only for AMI mortality is there a clear preference for adding a fixed effect or an ARMA(2,2). A simple fixed-effect model is
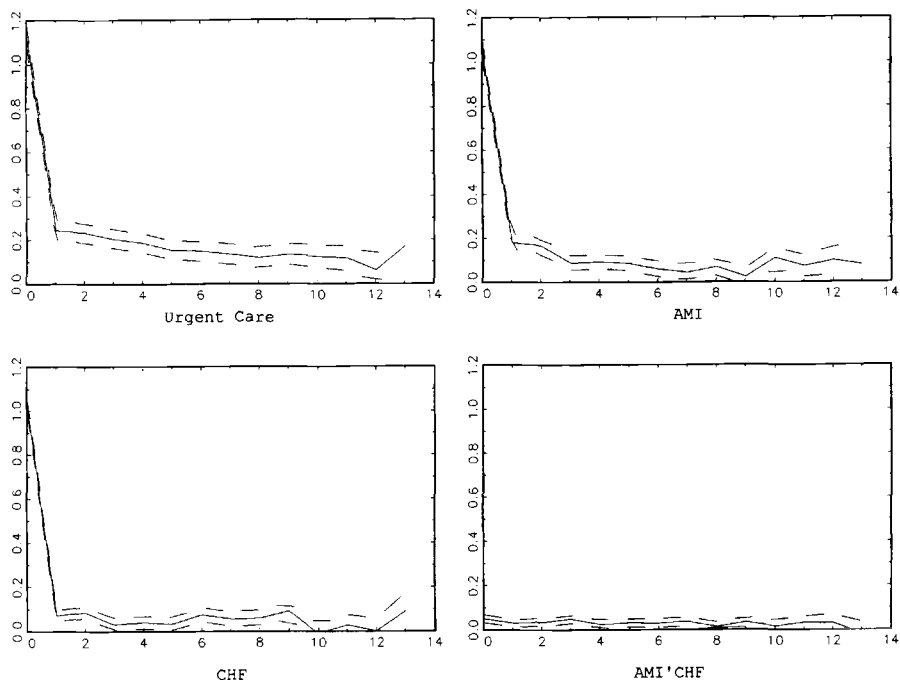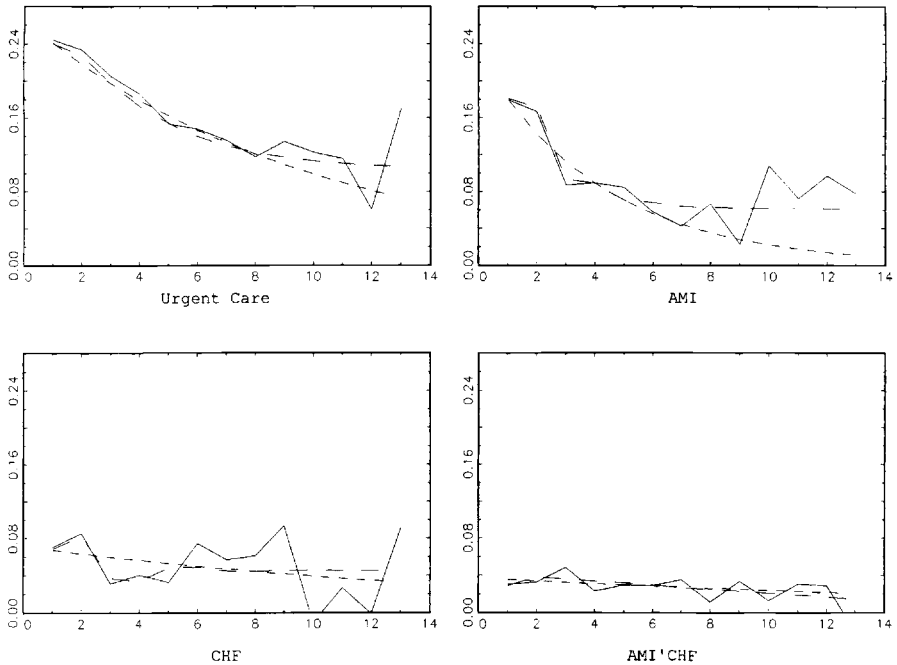
Fig. 7.4   GMM estimates of covariogram (covariance by lag length) for mortality Z-score. Unconstrained stationary covariance and 95 percent confidence interval estimated with data from 1974–87.

soundly rejected for urgent care and AMI. However, for both AMI and CHF, all of the more restrictive models do a poorer job fitting the data than the unconstrained stationary model.

Figure 7.5 illustrates the ability of alternative models to fit the covariogram. These figures graph the estimated covariogram for the unrestricted stationary model (*solid line*), the model with fixed effects, measurement error, and an ARMA(2,2) component (*long dashes*), and a more restrictive model with measurement error and an AR(1) component (*short dashes*). The variance (lag of zero) is estimated equally well by all models and therefore has been left out of these figures to avoid distorting the scale. It is apparent that the more restrictive AR(1) model does a reasonable job fitting the basic features of the data. The more flexible ARMA(2,2) specification is able to pick up some apparent fluctuations in the covariance at short lags while the fixed effect helps to fit the leveling off of the covariogram at longer lags, particularly for AMI mortality.

Table 7.2 provides parameter estimates assuming that mortality is composed of a fixed effect, a stationary AR(1), and stationary i.i.d. error. For example, column (1) estimates that the total variance in urgent care mortality is composed of a large measurement error component (.844) and relatively small

Fig. 7.5   GMM estimates of covariogram (covariance by lag length) for mortality Z-score. Covariance estimated with data from 1974–87: unconstrained (*solid line*), fixed effect + ARMA(2,2) + i.i.d. (*long dashes*), and AR(1) + i.i.d. (*short dashes*).

fixed-effect (.080) and AR(1) (.200) components. The AR(1) is fairly persistent, with a serial correlation of .829. The goodness-of-fit statistic implies that this simple model adequately summarizes the data. Dropping the fixed effect (col. [2]) has little effect on the overall fit of the model but increases the variance and persistence of the AR(1). In other words, the data cannot distinguish between a fixed effect and a somewhat more permanent AR(1). Note that the *t*-tests generally reject the hypothesis that the fixed-effect variance is zero, while chi-squared tests (based on the difference in the goodness-of-fit statistic in models with and without a fixed effect) often do not reject this hypothesis. This is a common feature of GMM models of parameters that are nearly unidentified.

The remainder of the estimates tell a similar story: Measurement error accounts for at least 75 percent of the variance for each diagnosis; the fixed effect accounts for roughly 5 percent of the variance; and a moderately persistent AR(1) accounts for the remainder of the variance, although this component accounts for more of the variance in AMI mortality than in CHF mortality. The covariance between AMI and CHF is quite persistent, making it impossible to

Table 7.2   **Variance Decomposition of Error Structure of 45-Day Mortality Z-score, 1974–87**

Variance Component	Urgent Care		AMI		CHF		AMI′CHF	
	With Fixed Effect (1)	Without Fixed Effect (2)	With Fixed Effect (3)	Without Fixed Effect (4)	With Fixed Effect (5)	Without Fixed Effect (6)	With Fixed Effect (7)	Without Fixed Effect (8)
Variance of i.i.d. error	0.844 (.021)	0.858 (.017)	0.786 (.029)	0.814 (.021)	0.915 (.041)	0.938 (.017)	0.016 (.011)	0.016 (.010)
Variance of fixed effect	0.080 (.040)	–	0.055 (.016)	–	0.044 (.011)	–	−2.14 (1240)	–
Variance of AR(1)	0.200 (.034)	0.266 (.023)	0.208 (.028)	0.227 (.024)	0.051 (.035)	0.071 (.011)	2.18 (1240)	0.037 (.008)
δ	0.829 (.066)	0.906 (.014)	0.648 (.074)	0.792 (.030)	0.576 (.360)	0.941 (.027)	0.999 (.379)	0.957 (.035)
GMM goodness-of-fit statistic	115.39 (.155)	117.31 (.143)	148.18 (.002)	157.51 (.000)	170.04 (.000)	171.13 (.000)	310.66 (.000)	310.87 (.000)
(*p*-value) [d.f.]	[101]	[102]	[101]	[102]	[101]	[102]	[192]	[193]

Notes: Z_{it} = Fixed effect + AR(1) + i.i.d. error. Numbers in parentheses are standard errors of parameter estimates.

identify a fixed effect from a very persistent AR(1). Interestingly, the variance of this persistent effect in the covariance is of roughly the same magnitude as the variance of the fixed effect in the AMI and CHF models. This suggests that AMI and CHF mortality may have a common component that is quite persistent. As a consequence of the failure of stationarity in the data, the goodness-of-fit statistic rejects all the models except urgent care.

The descriptive evidence presented in section 7.3 suggests that the nonstationarity is most evident in the early years of data. Therefore, a simple way to avoid the nonstationarity (and at the same time check on the robustness of the underlying parameter estimates) is to drop the early years from the analysis. Tables 7.3 and 7.4 replicate the results of tables 7.1 and 7.2 using only data for 1979–87. The improvement in the fit of the models is dramatic (see table 7.3). Of course, some of this apparent improvement is due to lower power of these tests with a shorter panel. Still, in striking contrast to table 7.1, stationarity

Table 7.3   **Model Specification Tests for Covariance Structure of Mortality Z-Score, 1979–87 (based on GMM goodness-of-fit statistic)**

	For Covariance Matrix of			
Model Specification	Urgent Care	AMI	CHF	AMI'CHF
1. Unconstrained stationary	31.10	36.30	48.25	67.18
covariance (symmetric)	(.701)	(.455)	(.083)	(.639)
	[36]	[36]	[36]	[72]
Tested against model 1				
2. Fixed effect	3.98	0.55	3.47	3.00
+ i.i.d. error	(.137)	(.758)	(.176)	(.223)
+ ARMA(2,2)	[2]	[2]	[2]	[2]
3. Fixed effect	5.72	5.14	4.11	3.47
+ i.i.d. error	(.221)	(.273)	(.392)	(.482)
+ AR(2)	[4]	[4]	[4]	[4]
4. Fixed effect	7.43	8.10	8.26	3.48
+ i.i.d. error	(.191)	(.151)	(.142)	(.627)
+ AR(1)	[5]	[5]	[5]	[5]
5. i.i.d. error	7.62	8.16	10.08	3.48
+ AR(1)	(.267)	(.227)	(.108)	(.747)
	[6]	[6]	[6]	[6]
6. Fixed effect	32.97	37.38	11.56	3.49
+ i.i.d. error	(.000)	(.000)	(.116)	(.836)
	[7]	[7]	[7]	[7]
7. i.i.d. error	101.6	114.32	54.05	35.05
	(.000)	(.000)	(.000)	(.000)
	[8]	[8]	[8]	[8]

Note: Entries in table are GMM goodness-of-fit statistics. Numbers in parentheses are *p*-values; numbers in brackets are degrees of freedom.

cannot be rejected for any of the covariance matrices. A model of i.i.d. error with an AR(1) cannot be rejected for any of the covariance matrices. For Urgent care and AMI, the fixed-effect model is clearly rejected in favor of the AR(1) model. In contrast, CHF and the AMI-CHF covariance are fit equally well by a fixed-effect model.

Table 7.4 provides parameter estimates of the AR(1) model with and without a fixed effect for the 1979–87 data. Without the fixed effect, these estimates are little changed from those in table 7.2 using the entire panel. The goodness of fit of the models is much improved over table 7.2, as can be seen in the last row of the table. Finally, in contrast to estimates based on the full panel, the fixed-effect component is not particularly well identified.

Overall, the estimates for both the full panel and for the limited 1979–87 data have the same implications. Namely, the mortality data contain significant measurement error, a somewhat transitory serially correlated component, and perhaps a common hospital component that is quite permanent. Table 7.5 investigates whether this simple three-component structure can adequately summarize the joint covariance matrix of the AMI and CHF mortality data. The data that provide the basis for these estimates are contained in the sample covariance matrix for the combined AMI and CHF data (i.e., $Z'Z/N$ where $Z = (Z_{AMI}, Z_{CHF})$). Since there are 14 years of data for each diagnosis group, there are $28 \cdot 29/2 = 406$ distinct moments and potential degrees of freedom. Column (2) of table 7.5 limits the analysis to the 1979–87 data and therefore has $18 \cdot 19/2 = 171$ degrees of freedom.

Data from the full sample once again overwhelmingly reject the stationarity assumption. A simple three-factor model is also soundly rejected against the alternative unconstrained stationary model. The three-factor model assumes that mortality for each diagnosis is composed of (1) a common factor that follows an AR(1) with no measurement error and (2) a diagnosis-specific factor (one for each diagnosis) that consists of i.i.d. measurement error plus an AR(1). Although the model is statistically rejected, it does precisely summarize the key features of the data. There is a very persistent factor that is common to both diagnoses. Added to this is i.i.d. noise for each diagnosis and a relatively transitory AR(1) component with serial correlation of roughly .6 for both diagnoses. Furthermore, the AR(1) component accounts for a much larger share of the variance for AMI mortality than for CHF mortality.

Restricting the sample to 1979–87 dramatically improves the ability of a stationary model to fit the data and improves the fit of the three-factor model as compared to the unconstrained stationary model. Although the goodness-of-fit statistics still reject these models at conventional levels, the rejection is no longer overwhelming. For these data I have estimated the common factor as a fixed effect, since a fixed effect seemed to better fit the AMI-CHF covariance in tables 7.3 and 7.4. The parameter estimates are quite similar to those using the entire panel: there is a persistent common factor, a fairly transitory diagnosis-specific factor, and substantial i.i.d. measurement error.

Table 7.4   Variance Decomposition of Error Structure of 45-Day Mortality Z-score, 1979–87

Variance Component	Urgent Care With Fixed Effect	Urgent Care Without Fixed Effect	AMI With Fixed Effect	AMI Without Fixed Effect	CHF With Fixed Effect	CHF Without Fixed Effect	AMI'CHF With Fixed Effect	AMI'CHF Without Fixed Effect
Variance of i.i.d. error	0.840 (.032)	0.834 (.028)	0.822 (.035)	0.826 (.029)	0.861 (.121)	0.917 (.024)	0.014 (.020)	0.014 (.016)
Variance of fixed effect	−0.224 (1.232)	−	0.016 (.071)	−	0.062 (.017)	−	−0.771 (13652)	−
Variance of AR(1)	0.531 (1.212)	0.313 (.033)	0.234 (.062)	0.246 (.033)	0.086 (.113)	0.090 (.019)	0.822 (13651)	0.051 (.013)
δ	0.945 (.151)	0.893 (.024)	0.778 (.126)	0.802 (.037)	0.397 (.560)	0.939 (.052)	0.999+ (7.105)	0.994 (.057)
GMM goodness-of-fit statistic	38.52 (.581)	38.72 (.616)	44.40 (.330)	44.54 (.369)	56.51 (.054)	58.68 (.045)	70.66 (.681)	70.66 (.710)
(*p*-value) [d.f.]	[41]	[42]	[41]	[42]	[41]	[42]	[77]	[78]

Notes: Z_{it} = Fixed effect + AR(1) + i.i.d. error. Numbers in parentheses are standard errors of parameter estimates.

Table 7.5   **Three-Factor Models for Covariance Structure of AMI and CHF 45-Day Mortality Z-Score**

	1974–87 (1)	1979–87 (2)
1. Goodness-of-fit statistic for unconstrained stationary (symmetric) covariance structure	1456.9 (.000) [364]	199.33 (.002) [144]
2. Goodness-of-fit statistic for three-factor model against model 1	726.53 (.000) [34]	38.49 (.008) [20]
3. Parameter estimates		
A. Common factor		
i. Variance of fixed effect	–	0.048 (.007)
ii. Variance of AR(1)	0.056 (0.004)	–
iii. δ_{common}	0.937 (.009)	–
B. AMI factor		
i. Variance of i.i.d. error	0.730 (.015)	0.775 (.030)
ii. Variance of AR(1)	0.187 (.015)	0.207 (.029)
iii. δ_{AMI}	0.571 (.035)	0.679 (.056)
C. CHF factor		
i. Variance of i.i.d. error	0.779 (.017)	0.862 (.051)
ii. Variance of AR (1)	0.071 (.015)	0.074 (.049)
iii. δ_{CHF}	0.578 (.092)	0.518 (.275)

Note: Numbers in parentheses are *p*-values of goodness-of-fit statistics and standard errors of parameter estimates. Numbers in brackets are degrees of freedom.

The ability of this simple three-factor model to fit the data is seen in figure 7.6. This figure graphs the estimated covariogram for the unconstrained stationary model (*solid line*) against the three-factor model (*dashed line*). As in figure 7.5, the variance (lag of zero) has been left out of these figures to avoid distorting the scale. The three-factor model does a reasonable job of fitting the covariograms. The longest lags of CHF are estimated off of relatively few years but suggest that CHF mortality may be more persistent than one would expect from the three-component model.

Overall, this evidence points to two particularly interesting features of the data. First, the covariance structure is reasonably well approximated by a simple three-factor model. A permanent hospital effect accounts for roughly 4 to 5 percent of the variance in mortality. A more transitory diagnosis-specific effect accounts for another 7 percent of the variation in mortality for CHF and nearly 20 percent of the variation in mortality for AMI. This component has a
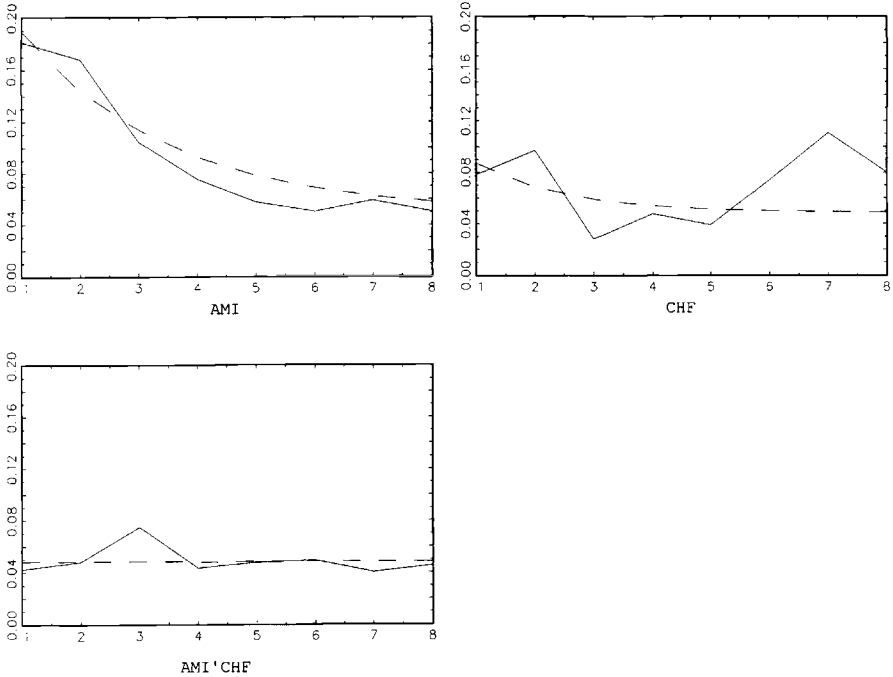
Fig. 7.6   GMM estimates of covariogram (covariance by lag length) jointly estimated for AMI and CHF mortality Z-score. Covariance estimated with data from 1979–87: unconstrained (*solid line*) vs. three-factor model (*dashed line*).

first-order serial correlation in the range of .6, so that diagnosis-specific shocks are mostly dissipated in five years. Finally, the remainder of the variation in mortality appears to be noise. The second interesting feature of the data is the significant nonstationarity in the distribution of mortality between the 1970s and the 1980s.

7.5   Conclusion

The empirical results presented in this paper have a number of implications. For those using these mortality variables as proxies for quality of care, the statistical properties of mortality should raise some concern. The amount of noise in these measures is on the order of 80 to 90 percent of the total variance. In a simple regression using mortality as a right-hand–side proxy for quality, this would lead to an attenuation bias of at least 80 percent, making the estimates of little use. Knowledge of the magnitude of measurement error can allow one to correct for the attenuation bias, so these estimates (or a similar method) might be used to correct for the bias. Even with such a correction, however, such low signal-to-noise ratios severely limit our ability to use such measures to forecast future hospital mortality.

Alternatively, for research that uses mortality as a dependent variable in a panel data setting, these estimates clearly indicate that the empirical model must allow for serially correlated errors. Moreover, the most common approach of adding hospital fixed effects does not appear to be adequate. It remains to be seen whether allowing for a more complicated error structure would significantly affect the conclusions from such studies.

In thinking more generally about the process that determines quality of care in a hospital, the three-factor model may give some insight. An obvious interpretation of the three-factor model is one in which the hospital effect represents general infrastructure such as the nursing staff, physical plant, or skill of the medical staff. These characteristics might be expected to be fairly permanent, and in fact, they represent what one often thinks of when thinking of a top-notch hospital. In contrast, the diagnosis-specific component could be thought of as technological innovations specific to that diagnosis. Casual observation suggests that AMI is a diagnosis that has had more technological innovation over the past 20 years, and this is consistent with the fact that the variance of the diagnosis-specific factor is much larger in AMI than in CHF. On the other hand, such innovations in treatment technology tend to diffuse to other hospitals fairly rapidly, so it is not surprising that this diagnosis-specific component does not persist much beyond five years.

Of course, there are alternative interpretations of the results. For example, the hospital component may reflect permanent differences in the population that each hospital serves, which are not captured by the adjustment for expected mortality. Distinguishing the quality-of-care interpretation from the case-mix interpretation is an important topic of future research.

Finally, the results suggest that there have been important changes in the distribution of patient mortality across hospitals between the 1970s and the 1980s. The reasons for this shift, and the corresponding change in the autocovariance structure of mortality, are unknown. It remains to be seen whether a simple extension of the models considered here can explain this anomaly. One possible explanation is that important technology shocks always begin with flagship hospitals and then diffuse through the remainder of the population. Thus, mortality in "innovative" years might be much more correlated than in average years. This is a topic of future research.

References

Abowd, J., and D. Card. 1989. On the covariance structure of earnings and hours changes. *Econometrica* 57:411–45.
Bowen, O., and W. Roper. 1987. *Medicare hospital mortality information, 1986.* Publication no. 00744. Washington, D.C.: U.S. Department of Health and Human Services, Health Care Financing Administration.

Cutler, D. 1995. The incidence of adverse medical outcomes under prospective payment. *Econometrica* 63:29–50.

Garber, A., V. Fuchs, and J. Silverman. 1984. Case mix, costs and outcomes: Differences between faculty and community services in a university hospital. *New England Journal of Medicine* 310:1231–37.

Gaumer, G., E. Poggio, and C. Coelen. 1989. Effects of state prospective reimbursement programs on hospital mortality. *Medical Care* 27:724–36.

Hofer, T., and R. Hayward. 1996. Identifying poor-quality hospitals: Can hospital mortality rates detect quality problems for medical diagnoses? *Medical Care* 34:737–53.

Keeler, E., L. Rubenstein, K. Kahn, D. Draper, E. Harrison, M. McGinty, W. Rogers, and R. Brook. 1992. Hospital characteristics and quality of care. *JAMA* 268:1709–14.

Krakauer, H., R. C. Bailey, K. Skellan, J. Stewart, A. Hartz, E. Kuhn, and A. Rimm. 1992. Evaluation of the HCFA model for the analysis of mortality following hospitalization. *Health Services Research* 27:317–35.

Luft, H., D. Garnick, D. Mark, D. Peltzman, C. Phibbs, E. Lichtenberg, and S. McPhee. 1990. Does quality influence choice of hospital? *JAMA* 263:2899–906.

Luft, H., S. Hughes, and R. Hunt. 1987. Effects of surgeon volume on quality of care in hospitals. *Medical Care* 25:489–503.

Luft, H., R. Hunt, and S. Maerki. 1987. The volume-outcome relationship: Practice-makes-perfect or selective-referral patterns? *Health Services Research* 22:147–82.

Luft, H., and P. Romano. 1993. Chance, continuity, and change in hospital mortality rates: Coronary artery bypass graft patients in California hospitals, 1983 to 1989. *JAMA* 270:331–37.

MaCurdy, T. 1982. The use of time series processes to model the error structure of earnings in a longitudinal data analysis. *Journal of Econometrics* 18:83–114.

McClellan, M., B. McNeil, and J. Newhouse. 1994. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA* 272:859–66.

McNeil, B., S. Pedersen, and C. Gatsonis. 1992. Current issues in profiling quality of care. *Inquiry* 29:298–307.

Minnesota Blues payment plan is the first to tie reimbursement to outcomes. 1991. *Outcomes Measurement and Management* 2:1–2.

Park, R. E., R. Brook, J. Kosecoff, J. Keesey, L. Rubenstein, E. Keeler, K. Kahn, W. Rogers, and M. Chassin. 1990. Explaining variations in hospital death rates: Randomness, severity of illness, quality of care. *JAMA* 264:484–90.

Perry, L. 1989. Michigan Blues plan initiates payment bonuses, penalties tied to standards of quality. *Medicine and Health* 19:58.

Staiger, D. 1993. Quality of care and patient volume. Unpublished manuscript.

Staiger, D., and G. Gaumer. 1995. Price regulation and patient mortality in hospitals. August. Unpublished manuscript.

Comment   David Meltzer

In recent years, the need to control health care costs and the recognition that much of health care is of uncertain benefit has increased interest in measuring

David Meltzer is assistant professor in the section of general internal medicine, department of economics, and Harris Graduate School of Public Policy at the University of Chicago and a faculty research fellow of the National Bureau of Economic Research.

the quality of health care. This has been evident at multiple levels of the health care system, with efforts to evaluate the quality of care provided by HMOs operating under prepayment, hospitals operating under diagnosis-related groups, and individual practitioners in a variety of settings. In all these areas, as well as in a number of studies to which the paper alludes, mortality has been used as an indicator of quality. Nevertheless, there has been surprisingly little work examining the statistical properties of these mortality variables. Staiger's paper raises important questions about the use of hospital mortality rates as an indicator of quality.

Staiger uses annual data on mortality after Medicare admission for urgent care, congestive heart failure (CHF), and acute myocardial infarction (AMI) for a sample of about 500 hospitals from 1974 through 1987. He investigates trends in the distribution of these rates using descriptive statistics and generalized method of moments estimators of the covariance structure. There are three major findings. First, that 75 to 90 percent of the variation in mortality is entirely transitory consistent with i.i.d. measurement error, suggesting that mortality is likely to be a poor proxy for quality whether used on the right- or left-hand side of a regression. Second, that mortality rates have a fairly persistent component consistent with a first-order autoregressive (AR(1)) process with an autocorrelation of .8–.95, suggesting that hospital effects may not be well approximated by a fixed-effects model. Third, that the combined AMI/CHF data are well fit by a three-component model with an i.i.d. error, a moderately transitory disease-specific component, and a permanent hospital-specific component, which may provide some insight into the determinants of the quality of care in hospitals.

The result that the vast majority of the variability in adjusted mortality rates is consistent with i.i.d. measurement error is the most striking result of the paper and raises serious questions about the value of mortality rates as a measure of quality. If 75 to 90 percent of the variation in mortality is gone in one year, past variations in hospital mortality provide little information on future mortality. Luft and Hunt (1986) also make this point in the context of work on the volume-outcomes relationship. The variability in annual hospital mortality rates is probably particularly large in Staiger's sample of hospitals because inclusion in the sample requires only a single admission in each diagnosis in each year. One could conceivably restrict the use of such measures to only larger hospitals, but quality may often be of greatest concern in smaller hospitals. Even if a few statistically significant outliers could be identified, little insight would be gained into quality in the majority of hospitals. The crudeness of the Health Care Financing Administration's (HCFA's) severity-of-illness adjustments is also worrisome, particularly if it causes hospitals to select patients for treatment by considering their effect on the hospital's mortality statistics rather than their potential to benefit from treatment.

How, then, should we address the need to assess hospital quality? There are

several possibilities. One is to use mortality statistics but improve risk stratification using more detailed clinical data in order to increase the signal-to-noise ratio. It is difficult to know to what extent it may be possible to refine such measures, but a recent study of mortality for CHF suggests that simple clinical measures of severity of illness (blood pressure, respiration rate, EKG changes, and serum sodium on presentation) substantially increase the amount of explained variation in in-hospital mortality compared to age and sex (Chin and Goldman 1996). Another possibility is to measure a broader set of intermediate outcomes (such as postoperative infection or bleeding), which may occur more frequently and therefore exhibit less noise. The issue with that approach is how much weight to put on those outcomes if patients ultimately recover from them. Another strategy is to follow measures of the process of care such as time to thrombolysis or appropriate choice of antibiotics. Like nonfatal adverse outcomes, these have the advantage of occurring more frequently than fatal outcomes and the drawback that they may not necessarily translate into significant outcomes. However, if the connection between process and outcomes is as believed, this approach has the advantage over using outcomes measures that it tells the hospital not only that there is a problem but also what to do about it. This may explain why this approach has been frequently adopted by hospitals in their attempts to improve quality, for example, through the critical pathways approach (Coffey et al. 1992). Interestingly, the Joint Commission on Accreditation of Healthcare Organizations has recently moved toward outcomes-based measures of quality (JCAHO 1995), but the noise inherent in those measures and the potential to game the system by manipulating case mix suggest that monitoring the consequences of those changes will be important.

Despite the large component of variability in these measures, the paper does report a statistically significant persistent component, which may have implications for the use of fixed-effects models with hospital mortality data and provide insights into the behaviors of doctors, hospitals, and the process of technical change. As the paper points out, one needs to be cautious in attributing the persistent component of adjusted mortality rates to quality because there may be persistent differences in the underlying severity of illness across hospitals that are not captured by the crude HCFA severity-of-illness adjustments. One reason that the severity-of-illness adjustments may be misleading is that they are rescaled only once over a 15-year period that contained substantial technical change. Even if patients sorted only on observable differences in severity of illness, technical changes that particularly improved outcomes for sicker or healthier groups would result in persistent deviations of hospitals' outcomes from expected. Future work could address this by using the original Medicare data to perform annual risk adjustment.

Much of the remainder of the paper is devoted to examining whether the error process is stationary, and it finds that stationarity is generally rejected for the full sample, though not for the latter half of the sample. This latter result may

simply reflect the low power of the test for the subsample, and it raises the question of what one is to conclude about a process that is intermittently stationary; but the deeper question is why we should be concerned about stationarity. Presumably, it is for prediction; yet the first half of the paper tells us that the vast majority of variation in mortality rates is transient, so knowing that the error structure is stationary is of little consequence in enhancing our ability to predict.

The paper also examines alternative models of the error structure of mortality rates and suggests that it may be inconsistent with the use of hospital fixed effects. This concern is too frequently neglected by researchers using fixed-effects models. Table 7.1 examines increasingly restrictive models from ARMA(2,2) to i.i.d. measurement error and finds that an AR(1) generally does as well as any other model for CHF and urgent care admissions, and that there is only a weak preference for adding an ARMA(2,2) or fixed effect for AMI. Unfortunately, the test statistics reported test only against the unrestricted stationary model. Since the models are generally nested, it would have been useful to test each additional restriction individually. Table 7.2 does report the effect of adding a fixed effect to the AR(1). However, what one really wants to know is whether one can use a fixed-effect model and do without the AR(1), and the paper does not report the more interesting exercise of assuming a fixed effect and adding an AR(1).

The final section of the paper examines the covariance of the CHF/AMI data in order to try to gain some insight into the process underlying these changes in mortality rates over time. It finds that the joint process is rather well fit by a fairly permanent hospital-specific component, a moderately transient disease-specific component, and an i.i.d. term consistent with sampling error. The paper also provides the interesting interpretation that the hospital-specific component might reflect a hospital's infrastructure—such as personnel and physical plant—while the disease-specific component might reflect technical progress in treating individual diseases. Ideally, one would like to measure these factors—technical innovations, knowledge of those innovations by specific providers, volume of experience for both providers and hospitals, and so forth—and test whether they are related to outcomes.

Though significant insights into the determinants of the quality of care seem unlikely to come from examination of the error structure alone, the finding that the majority of variation in hospital mortality rates appears to reflect random variation is of great consequence for the study of quality. It is there that research in this area seems likely to focus in the future.

References

Chin, Marshall, and Lee Goldman. 1996. Correlates of major complications or death in patients admitted to the hospital with congestive heart failure. *Archives of Internal Medicine* 156 (16): 1814–20.

Coffey, R. J., et al. 1992. An introduction to critical paths. *Quality Management in Health Care* 1:45–54.

Joint Commission on Accreditation of Healthcare Organizations (JCAHO). 1995. *JCAHO 1995 accreditation manual.* Oakbrook Terrace, Ill.: Joint Commission on Accreditation of Healthcare Organizations.

Luft, H., and S. Hunt. 1986. Evaluating individual hospital quality through volume statistics. *JAMA* 255 (20): 2780–84.