# Statistical Methods for Profiling Providers of Medical Care: Issues and Applications

Sharon-Lise T. Normand; Mark E. Glickman; Constantine A. Gatsonis

*Journal of the American Statistical Association*, Vol. 92, No. 439. (Sep., 1997), pp. 803-814.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199709%2992%3A439%3C803%3ASMFPPO%3E2.0.CO%3B2-%23

*Journal of the American Statistical Association* is currently published by American Statistical Association.

# Statistical Methods for Profiling Providers of Medical Care: Issues and Applications

Sharon-Lise T. NORMAND, Mark E. GLICKMAN, and Constantine A. GATSONIS

Recent public debate on costs and effectiveness of health care in the United States has generated a growing emphasis on "profiling" of medical care providers. The process of profiling involves comparing resource use and quality of care among medical providers to a community or a normative standard. This is valuable for targeting quality improvement strategies. For example, hospital profiles may be used to determine whether institutions deviate in important ways in the process of care they deliver. In this article we propose a class of performance indices to profile providers. These indices are based on posterior tail probabilities of relevant model parameters that indicate the degree of poor performance by a provider. We apply our performance indices to profile hospitals on the basis of 30-day mortality rates for a cohort of elderly heart attack patients. The analysis used data from 96 acute care hospitals located in one state and accounted for patient and hospital characteristics using a hierarchical logistic regression model. We used Markov chain Monte Carlo methods to fit the model and to obtain performance indices of interest. In particular, we estimated the posterior probability that mortality at the $i$th hospital is $1-\frac{1}{2}$ times the median mortality rate over all the hospitals in the state. We also calculated the posterior probability that the deviation in average risk-adjusted and "standardized" mortality at the $i$th hospital is "large." We compare the results of evaluating hospitals based on our performance indices to those obtained using conventional measures. With 30-day risk-adjusted mortality rates ranging from 12% to 14%, one-quarter of the hospitals had posterior probabilities that hospital-specific mortality was $1-\frac{1}{2}$ times the median mortality rate greater than 15%. The posterior probability of a large difference between risk-adjusted and standardized mortality rates was less than 6% for three-quarters of the hospitals we examined. Although there were differences in the evaluation of each hospital by the various criteria, one hospital consistently emerged as having the worst performance by all criteria.

KEY WORDS: Acute myocardial infarction; Excess mortality; Gibbs sampler; Hierarchical regression model; Posterior inference; Quality of care.

## 1. INTRODUCTION

Profiling medical care providers on the basis of quality of care and utilization of resources is rapidly becoming a widely used analysis in health care policy and research (Epstein 1995; Green and Wintfeld 1995; Hannan et al. 1994; Kassirer 1994; Landon et al. 1996; McNeil, Pedersen, and Gatsonis 1992; Salem-Schatz, 1994). Although comparative performance measures of health care were proposed as early as 1916 (Codman 1916), their use became widespread only recently. The results of profiling analyses often have far-reaching implications. They are used to generate feedback for health care providers, to design educational and regulatory interventions by institutions and government agencies, to design marketing campaigns by hospitals and managed care organizations, and, ultimately, to select health care providers by individuals and managed care groups. The recent trend of compiling and making available "report cards" for hospitals and individual health care practitioners has brought unprecedented public scrutiny to the practice of medicine. The effects of such scrutiny are undoubtedly complex and will unfold over time. However, the methodology for generating the reports needs more immediate attention (Epstein 1995; Localio et al. 1995).

Profiling is the process of comparing quality of care, use of services, and cost with normative or community standards. For example, hospital readmission rates within 2 weeks of discharge may be compared to a norm based on national rates. The profiling process normally includes a *risk-adjustment* step intended to account for possible differences in patient case mix (Iezzoni 1994; Landon et al. 1996; Salem-Schatz et al. 1994). In addition to a large body of work in medical research, the methodologic aspects of risk-adjustment have been extensively discussed in the literature on observational studies (see Rosenbaum 1995 and references therein). But the essence of profiling analysis lies in developing and implementing performance indices to evaluate medical care providers, such as physicians, hospitals, and care-providing networks. In this article we propose a class of measures for provider performance based on the posterior probability that a provider's patients have an unusually high frequency of adverse events. Our measures are derived from the fit of hierarchical regression models.

A major initiative to evaluate hospital performance in the United States was launched by the Health Care Financing Administration (HCFA) in 1987 with the annual release of hospital-specific data comprising observed and expected mortality rates for Medicare patients. Hospitals observed to have higher-than-expected mortality rates were flagged as institutions with potential quality problems. HCFA derived mortality rates by estimating a patient-level model of mortality for disease-based cohorts using administrative data. The expected hospital-specific mortality rates were calculated by averaging the model-based estimated probabilities of mortality within each hospital over the hospital's patient

Sharon-Lise T. Normand is Assistant Professor of Biostatistics, Department of Health Care Policy, Harvard Medical School, and Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115. Mark E. Glickman is Assistant Professor, Department of Mathematics, Boston University, Boston, MA 02215. Constantine A. Gatsonis is Director, Center for Statistical Sciences, and Associate Professor of Biostatistics, Department of Community Health, Brown University, Providence, RI 02912. This research was supported in part by Agency for Health Care Policy and Research grant RO1-HS07118, and AMI PORT grant R01-HS06341, and by a subcontract with the Health Care Financing Administration. The authors thank Earl Davies of the Harvard Medical School for programming support, and several anonymous referees and an associate editor for helpful comments.


Journal of the American Statistical Association
September 1997, Vol. 92, No. 439, Applications and Case Studies

population. HCFA's approach is typical of many published profiling analyses, and it is mainly for this reason that we discuss it in some detail in this article.

The public release of hospital-specific performance data was suspended in 1994, primarily as a result of the inadequacy of HCFA's administrative databases to provide the necessary detail for case mix adjustment and also the lack of information on patient compliance (Berwick 1990; Kassirer 1994). To remedy the problem, HCFA began a new initiative to carry out streamlined, in-depth data collection on several disease-specific patient cohorts. A subset of this newly collected information forms the dataset analyzed in this article. But our approach is designed to address several methodological concerns about HCFA's approach to profiling beyond the inadequacy of case mix adjustment. First, because of differences in hospital sample size, the precision of the hospital-specific estimates may vary greatly. Large differences between observed and expected mortality rates at hospitals with small sample sizes may be due primarily to sampling variability. Second, hospital practices may induce a strong association among patient outcomes within hospitals even after accounting for patient characteristics. Consequently, the errors associated with the effects of patient covariates on mortality may be underestimated. Third, the HCFA regression model makes no attempt to separate sampling variability from interinstitutional variability. The latter can be partitioned into a *systematic component,* possibly linked to provider characteristics, and a *random component.* Finally, there are concerns about the use of $z$ scores (standardized *expected–observed* mortality) to classify hospitals as aberrant. Such an approach labels a predetermined proportion of hospitals as aberrant even when "excess" mortality can be explained by random error. Thus even if all mortality rates were low and close to each other, a profiling approach such as HCFA's would still classify some hospitals as aberrant. Clearly, algorithms for identifying aberrant providers need to be linked to the outcome under study and to rely on metrics tailored to the needs of the particular profiling analysis. Because it seems prudent to consider several such metrics in a profiling analysis, it would be advantageous to follow an analytic approach that makes it possible to compute and evaluate these metrics within a unified modeling framework.

The statistical literature on methods for profiling providers is relatively limited. Gillis and Hixson (1991) examined the appropriateness of HCFA's quality screening technique using Monte Carlo methods in which the outcome depended on both patient-level and hospital-level characteristics. They defined a high-mortality hospital as one with observed mortality exceeding predicted mortality by more than 1.645 standard deviations. Smith (1994) proposed an analysis of variance (ANOVA) approach to partitioning variation in mortality rates into patient severity, quality of care, and random variation but did not discuss methods for comparing or identify aberrant hospitals. Stukel et al. (1994) developed estimators for standardized summary rates and used them to derive an estimate of excess utilization in the comparison of two areas. More recently, Silber, Rosenbaum, and Ross (1995) examined hospitals with large standardized differences in observed and expected rates of death, adverse events, and death following adverse events for surgical patients. They also evaluated the importance of patient and hospital characteristics as predictors of outcome by estimating the ratio of the variances in the two sets of characteristics. The approach of Silber et al. is valuable in understanding and describing components of variation in the analysis of differences across providers. The importance of incorporating provider characteristics emerges from their analysis, as it did from the earlier work by Gillis and Hixson (1991) and Smith (1994).

Provider profiling and, more generally, the analysis of variations in medical care utilization and outcomes has also been approached using hierarchical regression modeling. The statistical literature on such models is by now extensive (Gilks, Richardson, and Spiegelhalter 1996; Lindley and Smith 1972; Longford 1993; Wong and Mason 1985, 1991). In the area of health care research, Gatsonis, Epstein, Newhouse, Normand, and McNeil (1995) and Gatsonis, Normand, Liu, and Morris (1993) used a hierarchical logistic regression model of the form proposed by Wong and Mason to study variations in angiography rates across states and to evaluate the effects of contextual variables such as geographic location and availability of medical care. The hierarchical model allowed for area-specific coefficients in the patient-level logistic regression model, included area-level covariates, and was fitted via Gibbs sampling. Shwartz et al. (1994) used empirical Bayes methods to rank the amount of random variation across 68 geographic areas in Massachusetts, but did not include patient- or area-level characteristics in the analysis and did not develop a framework for identifying high-variation groups. Of more immediate relevance to hospital profiling and to the methodology discussed in this article is the work of Thomas, Longford, and Rolph (1994), who used a logistic regression model with a random intercept to analyze between-hospital variation in mortality rates. The model allows for variations in overall mortality rate among hospitals but assumes that the effect of patient severity is the same in all hospitals. Thomas et al. developed an empirical Bayes estimator for the difference between adjusted and expected hospital mortality rates and proposed it as an alternative to the estimator used by HCFA. Finally, Goldstein and Spiegelhalter (1996) used hierarchical models with aggregated patient data to realistically account for the uncertainty when comparing institutions.

This article presents an approach to profiling providers on the basis of posterior tail probabilities of model parameters that can be interpreted as indicators of provider performance. Providers may be individual physicians, groups of physicians, hospitals, health plans, counties, states, or other meaningful units. The proposed measures can be used to compare each unit to an absolute (external) or relative (internal) standard. The measures are constructed on the basis of a multilevel hierarchical regression model that permits the analyst to incorporate both patient-level and provider-level characteristics. Because the performance indicators are posterior tail probabilities of underlying model parameters, they can be computed directly from the simulated values drawn from the posterior distribution.

We apply our analytic framework to profiling a set of hospitals on the basis of rates of 30-day mortality for patients treated in hospitals for acute myocardial infarction (AMI). In this analysis, risk adjustment is carried out by including a severity index as a patient-level covariate in a hospital-specific logistic regression model. We consider several possible models for describing interhospital variability, including the random intercept model used by Thomas et al. (1994), and evaluate their implications on hospital profiling. Section 2 presents a general framework for profiling providers; Sections 3 and 4 apply our methods to profiling of 96 hospitals on the basis of the mortality rates for a cohort of 3,196 Medicare patients discharged with a principal diagnosis of AMI. Section 5 summarizes our methods and discusses the implications of our proposed framework.

## 2. AN ANALYTIC FRAMEWORK FOR PROFILING PROVIDERS

### 2.1 Modeling Variations Among Patients and Providers

Assume that outcome data are collected on a sample of patients treated by $I$ providers. For each provider, the $(L+1)$-dimensional vector $\mathbf{w}_i = (w_{0i}, w_{1i}, \ldots, w_{Li})$ represents provider $i$ characteristics, $Y_{ij}$ represents the outcome for patient $j$ treated by provider $i$, and the $(T+1)$-dimensional vector $\mathbf{x}_{ij} = (x_{0i}, x_{1ij}, \ldots, x_{Tij})$ represents patient $j$ characteristics, excluding provider characteristics which are defined herein.

*Stage I (patient-level, within-provider model).* Let

$$Y_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\phi}, \mathbf{x}_{ij} \overset{\text{indep.}}{\sim} f(Y_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\phi}, \mathbf{x}_{ij}), \tag{1}$$

where $\boldsymbol{\theta}_i$ is a vector of provider-specific parameters and $\boldsymbol{\phi}$ is a vector of parameters common to all providers. In many situations $\boldsymbol{\theta}_i$ will be a vector of regression coefficients specific to each provider, so that the components of $\boldsymbol{\theta}_i$ represent the effects of patient characteristics on outcome for the $i$th provider.

*Stage II (between-providers model).* The vector $\boldsymbol{\theta}_i$ of provider-specific parameters is modeled as a function of provider characteristics, $\mathbf{w}_i$, and is assumed to follow a distribution parameterized by a vector of hyperparameters, $\boldsymbol{\alpha}$,

$$\boldsymbol{\theta}_i|\boldsymbol{\alpha}, \boldsymbol{\phi}, \mathbf{w}_i \overset{\text{indep.}}{\sim} g(\boldsymbol{\theta}_i|\boldsymbol{\alpha}, \mathbf{w}_i);\ \boldsymbol{\phi}|\boldsymbol{\alpha} \sim h(\boldsymbol{\phi}|\boldsymbol{\alpha}). \tag{2}$$

It is assumed that $\boldsymbol{\theta}_i$ is independent of $\boldsymbol{\phi}$ given $\boldsymbol{\alpha}$ and $\mathbf{w}_i$. The hyperparameter vector, $\boldsymbol{\alpha}$, may contain a set of regression coefficients relating the unobserved Stage I provider parameters, $\boldsymbol{\theta}_i$, to the provider covariates, $\mathbf{w}_i$.

Finally, a prior, $\pi(\boldsymbol{\alpha})$, is assumed for the hyperparameters $\boldsymbol{\alpha}$. We denote the full set of parameters by $\boldsymbol{\Lambda} = \{\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\theta}_i;\ i = 1, 2, \ldots, I\}$ and denote the observed data by $\mathbf{y} = \{y_{ij};\ j = 1, \ldots, n_i;\ i = 1, \ldots, I\}$. The foregoing class of hierarchical models is fairly general, incorporating, among others, the mixed models of Laird and Ware (1982); the Bayesian linear models of Lindley and Smith (1972); the hierarchical logistic regression model of Wong and Mason (1985); semiparametric models, such as a hierarchical version of Cox regression, and models with heavy tails, such as $t$ distributions with low degrees of freedom. Extensions to more stages in the hierarchy are straightforward to incorporate.

The motivation for including the provider characteristics, $\mathbf{w}_i$, in the Stage II model is based on both technical and subject matter considerations. From a technical standpoint, the issue is one of correct specification of the model, on the basis of which performance indices are computed and predictions made. Consideration of nonexchangeable models is a natural step in the process of fitting a hierarchical model that adequately explains the observed variations among providers. Given the substantial lack of precision in estimates derived from traditional profiling analyses, the need for careful modeling is particularly acute. From a subject matter standpoint, a growing body of empirical evidence suggests that provider characteristics are important predictors of patient outcomes. For example, evidence of a relation between hospital characteristics and patient mortality has been provided by several researchers, including Brennan et al. (1991), Hartz et al. (1989), Kuhn, Hartz, Gottlieb, and Rimm (1991), Kuhn, Hartz, Krakauer, Bailey, and Rimm (1994), and McNeil et al. (1992). In a given profiling analysis, the specific provider characteristics would be selected on the basis of published research.

### 2.2 Measures of Absolute and Relative Performance

To introduce our performance indices, define the expected outcome by the $i$th provider, adjusting for patient mix to be

$$\mu_i^A = \frac{1}{n_i}\sum_{j=1}^{n_i} E(Y_{ij}|\mathbf{x}_{ij}, \mathbf{w}_i, \boldsymbol{\Lambda}) = \frac{1}{n_i}\sum_{i=1}^{n_i} E(Y_{ij}|\mathbf{x}_{ij}, \boldsymbol{\theta}_i, \boldsymbol{\phi}), \tag{3}$$

where the last equality holds because the sampling distribution, $f(Y_{ij}|\mathbf{x}_{ij}, \mathbf{w}_i, \boldsymbol{\Lambda})$, depends only on $\boldsymbol{\theta}_i$ and $\boldsymbol{\phi}$.

We define the *standardized outcome* for the $i$th provider as the expected outcome if provider $i$'s patients are treated at a reference provider. This standardized outcome is obtained by averaging the expected outcome over the provider-specific parameters

$$\begin{aligned}\mu_i^S &= \frac{1}{n_i}\sum_{j=1}^{n_i}\int E(Y_{ij}|\mathbf{x}_{ij}, \mathbf{w}_i, \boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\theta}_i,)g(\boldsymbol{\theta}_i|\boldsymbol{\alpha}, \mathbf{w}_i)\, d\boldsymbol{\theta}_i \\ &= \frac{1}{n_i}\sum_{j=1}^{n_i} E(Y_{ij}|\mathbf{x}_{ij}, \boldsymbol{\phi}, \boldsymbol{\alpha}, \mathbf{w}_i).\end{aligned} \tag{4}$$

If $\boldsymbol{\theta}_i$ represents a vector of provider-specific regression coefficients and $\boldsymbol{\alpha}$ represents a matrix of regression coefficients linking provider characteristics, $\mathbf{w}_i$, to the provider-specific effects, $\boldsymbol{\theta}_i$, then the quantity defined in (4) is the expected provider-specific outcome if the matrix of standardized regression coefficients are applied to the $n_i$ patient characteristics at the $i$th hospital. In contrast, (3) represents provider $i$'s adjusted outcome when applying its *own* vector of effects to its patient population.

We now define a first performance index based on the $\mu_i^{\mathrm{A}}$ and $\mu_i^{\mathrm{S}}$. Intuitively, a provider's performance is poor if the posterior probability that $\mu_i^{\mathrm{A}} - \mu_i^{\mathrm{S}}$ being bigger than some benchmark value is "large." This motivates the following performance index. Let $\boldsymbol{\mu}^{\mathrm{A-S}} = \{\mu_1^{\mathrm{A}} - \mu_1^{\mathrm{S}}, \mu_2^{\mathrm{A}} - \mu_2^{\mathrm{S}}, \ldots, \mu_I^{\mathrm{A}} - \mu_I^{\mathrm{S}}\}$ denote the vector of deviations between adjusted and standardized provider outcomes. Let $H(\cdot)$ denote a "benchmark" function of $\boldsymbol{\mu}^{\mathrm{A-S}}$. Then

$$P_i^{\mathrm{A-S}} = P(\mu_i^{\mathrm{A}} - \mu_i^{\mathrm{S}} > H(\boldsymbol{\mu}^{\mathrm{A-S}})|\mathbf{y}) \tag{5}$$

is the posterior probability that the difference between adjusted and standardized outcomes for provider $i$ is larger than some relevant function of the deviations across the sample of providers. For example, $H$ could be the sum of the interquartile range and the 75th percentile, which, for normally distributed populations, would result in the 97.8 percentile; for nonnormal distributions, this value would correspond to a different percentile. The function $H$ could also be a constant function, in which case the performance index $P_i^{\mathrm{A-S}}$ would have the interpretation of measuring provider $i$'s performance relative to an absolute standard.

The performance index for the $i$th provider, using the definitions of expected and standardized outcomes in (3) and (4), is conditional on its vector of provider characteristics, $\mathbf{w}_i$. Therefore, the difference $\mu_i^{\mathrm{A}} - \mu_i^{\mathrm{S}}$ is the comparison between the average expected outcome at provider $i$ and the average expected outcome at a pooled collection of *similar* providers. Although this index provides information to evaluate how extreme certain providers are relative to providers of similar characteristics, it may not be as informative when comparing providers with different covariate vectors $\mathbf{w}_i$. This is because the magnitudes of the differences between the $\mu_i^{\mathrm{A}}$ and $\mu_i^{\mathrm{S}}$ for such providers are being measured against different reference groups. Several modifications to this performance index thus may be considered. One modification is to define a fixed set of provider characteristics, $\mathbf{w}^*$, which would serve as the reference set of provider characteristics in the performance index. In this case, $\mathbf{w}_i$ would be replaced with $\mathbf{w}^*$ in (4), but the model in (2) would not change. This approach may be useful in comparing average outcomes at a single type of provider while still adjusting for patient mix.

A second performance index can be motivated in the following manner. Rather than compare average adjusted and standardized outcomes by provider $i$, we can imagine a particular patient whose treatment would potentially be given by provider $i$, and consider the probability of an adverse outcome for this specific patient. More formally, let $\mathbf{x}^* = (x_1, x_2, \ldots, x_T)$ denote a fixed vector of patient covariates and let $Y_i^* = E(Y|\mathbf{x}^*, \boldsymbol{\theta}_i, \phi)$ represent the expected outcome for this patient treated by provider $i$. Define

$$P_i^* = P(Y_i^* > K(\mathbf{Y}^*)|\mathbf{y}) \tag{6}$$

where $\mathbf{Y}^* = \{Y_1^*, Y_2^*, \ldots, Y_I^*\}$ and $K(\cdot)$ is some specified function of $\mathbf{Y}^*$. $P_i^*$ is the posterior probability that the adjusted outcome for a patient described by $\mathbf{x}^*$ at provider $i$ is "large" compared to the adjusted outcomes for similar patients at all of the providers in the sample. For example, if $K(y)$ is the median of $y$, then $P_i^*$ is the probability that the adjusted outcome at provider $i$ is larger than the median adjusted outcome across all providers. When a benchmark value, $\tau$, is available, then provider's $i$ performance can be estimated by $P(Y_i^* > \tau)$.

Choice of performance indices in a given situation depends on the goals of the particular analysis. If one assumes that the future distribution of patients has the same characteristics across providers, as in the observed data, then provider performances can be summarized by comparing the $P_i^{\mathrm{A-S}}$. But if instead we are interested in comparing provider performances for particular types of patients (e.g., relatively healthy patients, or patients in critical condition), then comparing providers based on $P_i^*$ for several different choices of covariate vectors may be more appropriate. We next demonstrate the proposed performance indices in the context of profiling hospitals using 30-day mortality rates.

## 3. PROFILING MORTALITY RATES FOR ACUTE MYOCARDIAL INFARCTION PATIENTS

### 3.1 The Study

As part of the restructuring of HCFA's quality assurance methods, the U.S. government is collecting detailed clinical, socio-demographic, and administrative data for Medicare patients discharged with a principal diagnosis of acute myocardial infarction (AMI) from hospitals in the United States. The pilot phase of this data collection effort, known as the Cooperative Cardiovascular Project (CCP), involved abstracting medical records for patients discharged from hospitals located in Alabama, Connecticut, Iowa, and Wisconsin from June 1992–May 1993 (Ellerbeck et al. 1995; Normand, Glickman, Sharma, and McNeil 1996). AMI was chosen because, despite triggering a vast amount of medical care utilization, it remains a particularly fatal disease in the elderly. For example, in 1990 mortality among Medicare AMI patients was 23% at 30 days after infarction, climbing to 36% at 1 year after infarction (Pashos, Newhouse, and McNeil 1993). Moreover, recent research has shown that rates of medical and surgical interventions and rates of mortality in Medicare AMI patients vary greatly across geographic areas, hospitals, and demographic groups (Gatsonis et al. 1995; McClellan, McNeil, and Newhouse 1995; Pashos et al. 1994). Such variability is notable in a cohort of patients with uniform insurance coverage and relatively homogeneous disease status.

The initial study cohort consisted of 3,269 patients hospitalized at 122 hospitals in one of the four CCP pilot states. After excluding 26 hospitals that treated fewer than 5 AMI patients, we arrived at a final cohort of 3,169 patients across 96 hospitals. The number of AMI patients per hospital ranged from 5–274, with a mean of 33. The outcome of interest was mortality within 30 days of hospital admission. The overall 30-day mortality rate in the sample was 20%. The observed hospital-specific mortality rates ranged from 0% (5 hospitals) to 67% (1 hospital), with a mean of 22% (Table 1). Patient severity at admission was quantified by an index used in previous analyses of the full CCP cohort (see the Appendix). As detailed in the Appendix,

*Table 1. Patient and Hospital Characteristics in the Study Cohort*

| | 25th percentile | Median | Mean | 75th percentile |
|---|---|---|---|---|
| Observed Mortality | | | | |
| Across hospitals | .14 | .22 | .22 | .29 |
| Admission severity | | | | |
| Across patients | −2.47 | −1.80 | −1.65 | − .99 |
| Across hospitals | −1.47 | −1.49 | −1.47 | −1.22 |

| Hospital characteristics | % of patients | % of Hospitals |
|---|---|---|
| Rural (vs. urban) | 54 | 76 |
| Nonacademic (vs. academic) | 79 | 88 |
| Number of beds | | |
| ≤100 (small) | 29 | 64 |
| 101–299 (medium) | 27 | 21 |
| ≥300 (large) | 44 | 15 |

NOTE: Based on 3,196 Medicare beneficiaries age 65 years and older discharged with a principal diagnosis of AMI between June 1, 1992, and May 5, 1993 from one of 96 hospitals. The admission severity index quantifies the burden of illness on entry to the index hospital.

the severity index was constructed on the basis of 34 patient characteristics on admission. Approximately one-half of the patients had a predicted risk of death less than 14% $= 100\{1 + \exp(-\text{median severity})\}^{-1}$, with one-quarter of the patients having mortality risk greater than 27% (Table 1). Patient severity at admission ranged from −2.27 (or 9%) to −.29 (or 46%) across the 96 hospitals. In Table 1 rural hospitals are those not located in a standard Metropolitan Statistical Area, and nonacademic hospitals have no medical residents. Hospitals with fewer than 101 beds were classified as small; those with 101–299 beds, as medium-sized; and those with 300 or more beds, as large. More than three quarters of the hospitals were located in rural areas, and most were nonacademic.

### 3.2 Hierarchical Logistic Regression Model

The variation in mortality rates among hospitals was modeled via a three-level hierarchical regression model of the type proposed by Wong and Mason (1985). At the first level (within-hospital), the probability of death within 30 days was modeled by the hospital-specific logistic regression,

$$\text{logit}(P(Y_{ij} = 1)) = \beta_{0i} + \beta_{1i}\,(\text{severity}_{ij} - \overline{\text{severity}}), \quad (7)$$

where $Y_{ij}$ is the binary indicator of death within 30 days of admission for patient $j$ at hospital $i$, $\text{severity}_{ij}$ is the value of the severity index for that patient, and $\overline{\text{severity}} = -1.65$. We considered two forms for the second level of the hierarchical structure (between-hospitals), an exchangeable model of the form

$$\text{level IIa:}\quad \boldsymbol{\beta}_i = \begin{pmatrix} \beta_{0i} \\ \beta 1i \end{pmatrix} \sim N_2\left(\begin{pmatrix} \gamma_{00} \\ \gamma_{10} \end{pmatrix}, D\right), \quad (8)$$

and a nonexchangeable model of the form

$$\text{stage IIb:}\quad \boldsymbol{\beta}_i \sim N_2(\Gamma \mathbf{w}_i', D),$$

where

$$\Gamma = \begin{pmatrix} \gamma_{00} & \gamma_{01} & \gamma_{02} & \gamma_{03} & \gamma_{04} \\ \gamma_{10} & \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} \end{pmatrix}, \quad (9)$$

and $\mathbf{w}_i = (\text{rural}_i, \text{nonacademic}_i, \text{small}_i, \text{medium}_i)$. The components of $\mathbf{w}_i$ are indicator variables that are 1 if the condition is true and 0 otherwise. Vague prior distributions were assumed for $\Gamma$ and $D$ at the final level of the model.

In the foregoing formulation of the model, both coefficients of the logistic regression are allowed to vary among hospitals. A variable intercept would indicate interhospital differences in baseline mortality rates. A variable slope would indicate that the effect of clinical burden (patient severity) on mortality differs across hospitals. For comparative purposes, we also estimated a simple model in which slopes were the same across hospitals, $\beta_1$, and intercepts $\beta_{0i}$ were exchangeable with a $N(\gamma, \sigma^2_{\beta_0})$ distribution. This model (henceforth referred to as the *random-intercept model*) was used in the work by Thomas et al. (1994) discussed in Section 1. To complete the specification, we assumed proper but vague priors for $\beta_1$, $\gamma$, and $\sigma^{-2}_{\beta_0}$.

### 3.3 Model Estimation

Gibbs sampling was used to fit the hierarchical logistic models. The sampler was implemented using the BUGS (Gilks et al. 1996) software for the random-intercept model. A single string with a burn-in of 2,000 iterations and a further 1,250 iterations were used for inference. A specially developed Fortran program using rejection sampling techniques (Zeger and Karim 1991) to draw each $\boldsymbol{\beta}_i$ was used for the models in which both logistic coefficients were permitted to vary. Starting values for the logistic regression coefficients were obtained by fitting a separate logistic regression in each hospital. Hospitals for which the maximum likelihood estimates (MLEs) of $\boldsymbol{\beta}_i$ could not be determined using our Newton–Raphson procedure had starting values imputed from a Bayesian logistic regression analysis on data for hospital $i$ using a normal prior with mean equal to the average of the MLEs for the converged hospitals, and a variance equal to ten times the sample variance of the MLEs. Starting values for $\Gamma$ and $D$ were obtained by calculating the sample average and sample variance of the starting values for the $\boldsymbol{\beta}_i$'s. Five parallel strings of length 600 were simulated. Iterations 301–600 were used for inference. Overdispersed starting values for $D$ were used in the first three strings; the remaining two strings were begun

at the sample variance of the $\{\hat{\boldsymbol{\beta}}_i^{\text{Start}}\}$. Posterior intervals based on the empirical distributions were computed for selected parameters.

Convergence of the Gibbs sampler was assessed according to three criteria. First, the estimated potential scale reduction (PSR) for the Gibbs samples of the elements of $\boldsymbol{\beta}_i$ and $\Gamma$ were examined across the five strings (Gelman and Rubin 1992). Also, a multivariate analog of the PSR statistic was computed by computing the between-strings sums of squares for the overdispersed strings using $S_{\text{high}}^{(k)} = 1/2\sum_{s=1}^{3}(\bar{\boldsymbol{\beta}}_{s\cdot}^{(k)} - \bar{\boldsymbol{\beta}}_{\cdot\cdot}^{(k)})(\bar{\boldsymbol{\beta}}_{s\cdot}^{(k)} - \bar{\boldsymbol{\beta}}_{\cdot\cdot}^{(k)})'$, where $\bar{\boldsymbol{\beta}}_{s\cdot}^{(k)} = \sum_i^I(\boldsymbol{\beta}_{si}^{(k)}/I)$ and $\bar{\boldsymbol{\beta}}_{\cdot\cdot}^{(k)} = \sum_s^{m/2}[\bar{\boldsymbol{\beta}}_{s\cdot}^{(k)}/(m/2)]$, with $k$ indexing iteration, $s$ indexing string, and $i$ indexing hospitals. The between-strings sums of squares for the remaining two strings, $S_{\text{low}}^{(k)}$, were calculated in a similar manner. Because $\ln(|S_{\text{low}}^{(k)}|)$ should increase with increasing $k$ and $\ln(|S_{\text{high}}^{(k)}|)$ should decrease with increasing $k$, convergence is reached when the two quantities are the same. The two determinants were plotted to monitor convergence of the sampler. Finally, an additional string of the sampler was run out to 20,000 iterations, and the final 1,000 draws were examined for convergence using the techniques described earlier. The empirical distributions of the parameters obtained from the long string were compared to that obtained from our shorter strings to determine convergence.

Precision of parameter estimates was assessed by examining the lag-1 autocorrelation among the draws. For a lag-1 autocorrelation among Gibbs draws of $r$ and total number of draws $n$, the effective sample size is approximately $n(1 - r^2)$. The fit of the within-hospital model [(7)] was assessed by examining the residuals from a patient-level logistic regression model. Equation (7) was also fitted separately to several large hospitals, and goodness-of-fit statistics were calculated. Appropriateness of our level II model was judged from boxplots of the posterior draws of the level I parameters plotted against stage II covariates.

Examination of plots of the logarithm of the determinant of the between-strings sums of squares for the overdispersed and underdispersed strings at each iteration supported our assumption of convergence. Comparison of the distribution of the draws of parameters from the single long string (20,000 iterations) to that obtained from the sample of 1,500 based on the shorter strings indicated that the two samples had roughly the same coverage. Finally, the estimated PSR were generally within an acceptable range. The lag-1 autocorrelations for the draws of the $\boldsymbol{\beta}_i$ were found to be negligible. The lag-1 autocorrelations for the draws of the $\Gamma$ were larger, however, resulting in precisions effectively based on 420 independent observations, large enough for our inferences.

### 3.4 Estimation of Hospital Performance Indices

The hospital-specific risk-adjusted mortality rates, $\mu_i^{\text{A}}$, and standardized rates, $\mu_i^{\text{S}}$, are defined as

$$\mu_i^{\text{A}} = \frac{1}{n_i}\sum_{j=1}^{n_i} P(Y_{ij} = 1|\boldsymbol{\beta}_i, \mathbf{x}_{ij}) = \frac{1}{n_i}\sum_{j=1}^{n_i} \text{logit}^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta}_i)$$

and

$$\mu_i^{\text{S}} = \frac{1}{n_i}\sum_{j=1}^{n_i} P(Y_{ij} = 1|\Gamma, \mathbf{x}_{ij}, \mathbf{w}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} \text{logit}^{-1}(\mathbf{x}_{ij}\Gamma\mathbf{w}_i'),$$

and were estimated using

$$\hat{\mu}_i^{\text{A}} = \frac{1}{1{,}500}\sum_{k=1}^{1{,}500}\left\{\frac{1}{n_i}\sum_{j=1}^{n_i}\text{logit}^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta}_i^{(k)})\right\}$$

and

$$\hat{\mu}_i^{\text{S}} = \frac{1}{1{,}500}\sum_{k=1}^{1{,}500}\left\{\frac{1}{n_i}\sum_{j=1}^{n_i}\text{logit}^{-1}(\mathbf{x}_{ij}\Gamma^{(k)}\mathbf{w}_i')\right\}, \quad (10)$$

where $\mathbf{x}_{ij} = (1, \text{severity}_{ij} - \overline{\text{severity}})$, $\boldsymbol{\beta}_i^{(k)}$ is the $k$th draw of the vector of hospital-specific logistic regression coefficients, $\Gamma^{(k)}$ is the $k$th draw of the matrix of reference regression coefficients, and $\mathbf{w}_i = (1, 1)$ in the exchangeable model or $\mathbf{w}_i = (1, \text{rural}_i, \text{nonacademic}_i, \text{small}_i, \text{medium}_i)$ in the nonexchangeable model.

Three types of performance indices were estimated: the probability of excess mortality for the *average* patient, the probability of a large difference between adjusted and standardized mortality, and a $z$ score. The probability of excess mortality for patients of average admission severity, $\mathbf{x}_{ij}^* = (1, 0)$, at each hospital was estimated as

$$\hat{P}_i^* = \frac{1}{1{,}500}\sum_{k=1}^{1{,}500} I(\text{logit}^{-1}(\beta_{0i}^{(k)}) > c \times \hat{\xi}_{.50}^{(k)}), \quad (11)$$

where the indicator function, $I(\cdot)$, is 1 if the condition inside the parenthesis is true and 0 otherwise, $c = 1.5$, and $\hat{\xi}_{.50}^{(k)}$ is the median of $\{(\text{logit}^{-1}(\beta_{0i}^{(k)})); i = 1, 2, \ldots, 96\}$. Because of the dependence among the parameters in the joint posterior distribution of $\{\beta_{0i}; i = 1, 2, \ldots, 96\}$, the median, $\hat{\xi}_{.50}^{(k)}$, depended on the iterations. The value $c = 1.5$ was chosen because physicians felt that deviations this large indicated a potential quality problem.

The posterior probability that the deviation between each adjusted and standardized hospital-specific mortality rate was located in the upper tail of the distribution of the expected differences across all hospitals (Eq. (5)) was estimated as

$$\hat{P}_i^{\text{A}-\text{S}} = \frac{1}{1{,}500}\sum_{k=1}^{1{,}500} I(R_i^{(k)} > H(\mathbf{R}^{(k)})), \quad (12)$$

where

$$R_i^{(k)} = \frac{1}{n_i}\sum_{j=1}^{n_i}\text{logit}^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta}_i^{(k)}) - \frac{1}{n_i}\sum_{j=1}^{n_i}\text{logit}^{-1}(\mathbf{x}_{ij}\Gamma^{(k)}\mathbf{w}_i')$$

and $\mathbf{R}^{(k)} = \{R_i^{(k)}; i = 1, 2, \ldots, 96\}$. We defined $H(\mathbf{R}^{(k)}) = \hat{\xi}_{.50}^{(k)} + 1.5 \times (\hat{\xi}_{.75}^{(k)} - \hat{\xi}_{.25}^{(k)})$ with $\hat{\xi}_q^{(k)}$ the $q$th quantile of $\mathbf{R}^{(k)}$.

Finally, HCFA's algorithm was also used to define aberrant hospitals. A logistic regression model was fitted to the entire dataset, and $z$ scores were derived from the standardized difference between observed and expected mortality in each hospital. Specifically, $z_i = n_i(\bar{Y}_i - \bar{p}_i)/\sqrt{\sum_j^{n_i} \hat{p}_{ij}(1-\hat{p}_{ij})}$, where $\hat{p}_{ij} = \text{logit}^{-1}(\mathbf{x}_{ij}\hat{\boldsymbol{\beta}}_{\text{MLE}})$ and $\bar{p}_i = (1/n_i)\sum_{j=1}^{n_i}\hat{p}_{ij}$. Hospitals with $z_i \geq 1.645$ (the top 5%) were classified as aberrant. Note that although more appropriate variance formulas for the standardized difference that account for the correlation between $\bar{y}_i$ and $\bar{p}_i$ are available (see, e.g., Haberman 1976), we chose to be consistent with HCFA's algorithm.

## 4. RESULTS

### 4.1 Interhospital Variation

Table 2 displays the estimated posterior means and standard deviations for the regression parameters using the exchangeable models described in Section 3.2. The estimates of the average baseline mortality rate were similar. In particular, a 95% posterior interval for the average intercept $\gamma_{00}$ derived from the full exchangeable model was $(-1.87, -1.56)$, corresponding to $(.13, .17)$ in the probability scale. The estimates of the average estimated slope in the full model and the overall slope in the random-intercept model were also similar.

There was considerable variation among hospitals in both intercept and slope of the logistic model. Based on the full exchangeable model, the 2.5 and 97.5 percentiles of $(\{\beta_{0i}^{(k)}; i = 1, 2, \ldots, 96\}|\mathbf{y})$ is $(-2.53, -.92)$, indicating a large range in the log-odds of mortality for an average patient randomly selected from one of the 96 hospitals in the sample (see also Fig. 1). The corresponding percentiles for the estimates of the slopes $\beta_{1i}$ were (.62, 1.47), and the coefficient of variation for the slopes was (Table 2) 20% (CV = .21/1.03). We concluded that the data demonstrate substantial variability in the effect of patient severity among the hospitals in our database, and thus a model allowing the logistic model slope to vary across hospitals is preferable to the simple random-intercept model.

Figure 1 displays boxplots of the estimated posterior mean intercepts, $\beta_{0i}$, and the estimated posterior mean slopes, $\beta_{1i}$, stratified by the hospitals characteristics listed in Table 1. There is some evidence that the logistic parameters may not be exchangeable with respect to hospital size, urbanicity, and academic affiliation. Table 2 also displays the estimated level II parameter summaries for the nonexchangeable model [(9)]. In this model, the logistic model intercept, $\gamma_{00}$, represents the log-odds of 30-day mortality across large urban academic hospitals for a patient of average admission severity and has a 95% posterior interval given by $(-2.15, -1.45)$. Hospitals located in rural areas of the state were associated with higher mortality than urban hospitals ($\hat{\gamma}_{01} = .55$) for the average patient. Moreover, some of the variation in the hospital-specific slopes was explained by hospital size with medium-sized hospitals having slightly smaller slopes than large hospitals ($\hat{\gamma}_{14} = -.29$).

### 4.2 Hospital Performance Indices

The risk-adjusted hospital mortality rates, $\hat{\mu}_i^{\text{A}}$, varied

*Table 2. Regression Estimates*

| *Level I parameter* | *Level II parameter* | *Estimated posterior summaries* | | | |
|---|---|---|---|---|---|
| | | *Mean* | *SD* | *Mean/SD* | *Percentiles (2.5, 97.5)* |
| Exchangeable model: Random-intercept model | | | | | |
| $\beta_{0i}$: Intercept | $\gamma$: Intercept | −1.70 | .07 | −24.29 | (−1.85, −1.57) |
| | $\sigma^2_{\beta_0}$: Variance | $(.31)^2$ | .05 | | (.01, .22) |
| $\beta_1$: Severity − $\overline{\text{severity}}$ | | 1.03 | .05 | 20.60 | (.93, 1.13) |
| Exchangeable model: Random-intercept and slope model | | | | | |
| $\beta_{0i}$: Intercept | $\gamma_{00}$: Intercept | −1.72 | .08 | −21.53 | (−1.87, −1.56) |
| $\beta_{1i}$: Severity − $\overline{\text{severity}}$ | $\gamma_{10}$: Intercept | 1.03 | .05 | 19.67 | (.94, 1.15) |
| | | | | *Estimated posterior mean* | |
| | *D*: Variance | | | $\begin{pmatrix} (.42)^2 & -.03 \\ -.03 & (.21)^2 \end{pmatrix}$ | |
| Nonexchangeable model: Random-intercept and slope model | | | | | |
| $\beta_{0i}$: Intercept | $\gamma_{00}$: Intercept | −1.79 | .17 | −10.29 | (−2.15, −1.45) |
| | $\gamma_{01}$: Rural | .55 | .20 | 2.76 | (.15, .93) |
| | $\gamma_{02}$: Non-Academic | −.27 | .27 | −1.24 | (−.71, .14) |
| | $\gamma_{03}$: Small | −.27 | .25 | −1.06 | (−.74, .27) |
| | $\gamma_{04}$: Medium | .29 | .20 | 1.46 | (−.10, .67) |
| $\beta_{1i}$: Severity − $\overline{\text{severity}}$ | $\gamma_{10}$: Intercept | 1.22 | .13 | 9.18 | (.96, 1.52) |
| | $\gamma_{11}$: Rural | .05 | .16 | .33 | (−.27, .36) |
| | $\gamma_{12}$: Nonacademic | −.11 | .17 | −.64 | (−.44, .23) |
| | $\gamma_{13}$: Small | −.08 | .20 | −.39 | (−.50, .28) |
| | $\gamma_{14}$: Medium | −.29 | .15 | −1.88 | (−.58, .01) |
| | | | | *Estimated posterior mean* | |
| | *D*: Variance | | | $\begin{pmatrix} (.35)^2 & -.03 \\ -.03 & (.22)^2 \end{pmatrix}$ | |

NOTE: SD = standard deviation.

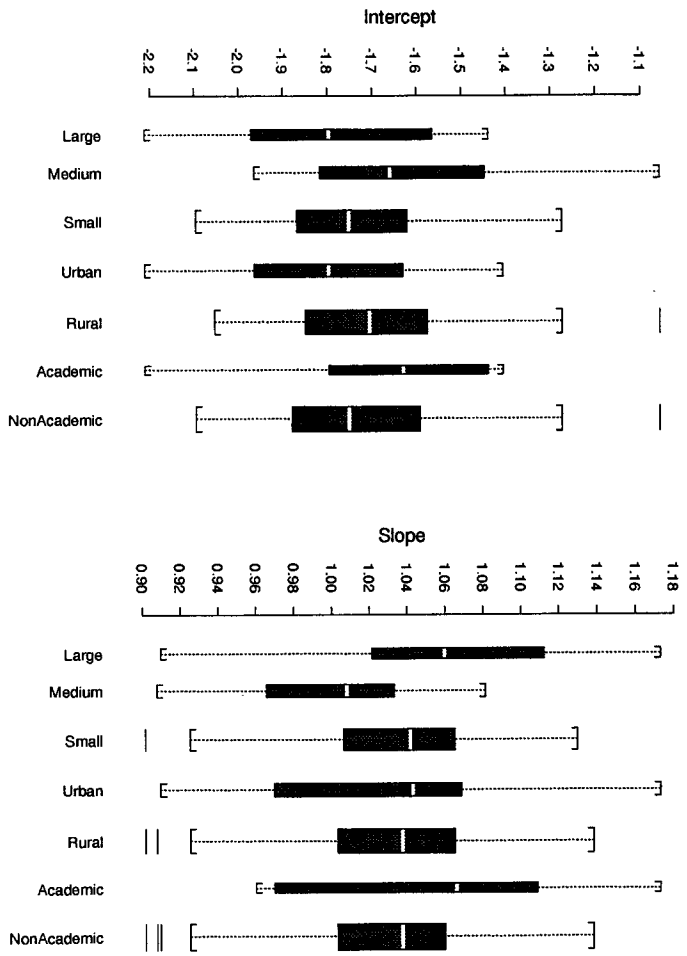


*Figure 1. Estimated Posterior Means for Level I Parameters in the Full Exchangeable Model. Hospital means, calculated as $\{(1/1,500)\sum_{k=1}^{1,500}\beta_{pi}^{(k)};\ i = 1, 2, \ldots, 96;\ p = 1, 2\}$, are stratified by number of beds, geographic location of hospital (rural or urban), and teaching status of the hospital. The width of each boxplot is proportional to the number of hospitals having each particular characteristic.*

from a low of 12% to a high of 44%. Figure 2 displays the observed, $\sum_j(y_{ij}/n_i)$, and adjusted, $\hat{\mu}_i^{\mathrm{A}}$, hospital-specific mortality rates stratified by the geographic location of the hospital. It is clear from the figure that adjustment for severity on admission is substantial. Of particular note is the hospital pictured in the lower right panel of Figure 2 (urban hospitals), whose observed rate is 29% but whose risk-adjusted rate is 37%. This hospital is a medium-sized academic hospital that treated seven patients during the study period; however, the average admission severity of these seven patients was $-.69$ ($\approx$33% on the probability scale). There also appears to be less variability in changes between the observed and adjusted mortality rates for urban hospitals than for rural hospitals.

The estimated posterior probability that mortality at the $i$th hospital was 1-$\frac{1}{2}$ times the median mortality over all 96 hospitals [(11)] ranged from 0 (five hospitals) to 89% (one hospital) (see Table 3). Three-quarters of the hospitals in our sample had an estimated probability of less than 16% for this event. Similarly, the estimated probability that the difference between adjusted and standardized hospital-specific mortality rates, $P_i^{\mathrm{A-S}}$, is large [(12)] ranged from 0 (five hospitals) to 25% (one hospital), with the median estimated probability 2.7%.

Using HCFA's algorithm, nine hospitals were flagged as having potential quality problems (Table 4). There was moderate disagreement among the criteria for classifying hospitals as aberrant. Despite this, regardless of which measure is used ($z_i > 1.65, \hat{P}_i^*, \hat{P}_i^{\mathrm{A-S}}$), hospital 1 is ranked as the worst. This hospital is a rural, medium-sized nonacademic hospital with an observed mortality rate of 35%, an adjusted rate, $\hat{\mu}_1^{\mathrm{A}}$, of 28%, and a standardized mortality rate, $\hat{\mu}_1^{\mathrm{S}}$, of 23%. The average admission severity of patients at hospital 1, $\overline{\text{severity}}_1$, was $-1.87$ (13% on the probability scale), with 25% of the 54 patients having admission severity larger than $-1.19$ (23% on the probability scale).

Because $P^{\mathrm{A-S}}$ represents the probability that the difference between observed and expected mortality is unusually large, this performance index is conceptually closer to inferences based on the $z$ scores than those based on $P^*$. However, $P^{\mathrm{A-S}}$ quantifies how extreme hospitals are relative to *similar* hospitals as defined by hospital location, academic affiliation, and number of beds. Consider hospital 44, which was identified as one of the worst 5% of the hospitals using the HCFA algorithm. The posterior probability that the difference between hospital 44's adjusted and standardized (to small rural nonacademic hospitals) outcome is large was only 5% using our criterion (Table 4: $\hat{P}_{44}^{\mathrm{A-S}} = 5$, rank = 33). There are two reasons for these differences. First, $\hat{P}_{44}^{\mathrm{A-S}}$ estimates the probability of a large difference between observed and expected mortality at hospital 44 compared only to small rural hospitals rather than to all hospitals. Second, because the index $\hat{P}_{44}^{\mathrm{A-S}}$ is based on the difference between the "true" values of the parameters, $\mu_{44}^{\mathrm{A}} - \mu_{44}^{\mathrm{S}}$, and not on a comparison with the observed value, which is based on seven AMI patients, $\mu_{44}^{\mathrm{A}}$ is pulled substantially away from the observed value of $\bar{y}_{44} = .43$ to .20.

Our other performance index, $\hat{P}^*$ estimates the probability that mortality for a specific type of patient (the average) who is treated by a provider $i$ is unusually large even if provider $i$ did not treat such a patient. For example, the probability that mortality is unusually large for such a patient is 71% at hospital 10, yet the probability that the difference between observed and expected mortality is unusual at hospital 10, $\hat{P}_{10}^{\mathrm{A-S}}$, is only 7%. Thus if patients relied on only one index, then the ultimate choice would rule out hospital 10 in one case ($z$-score or $\hat{P}_{10}^{\mathrm{A-S}}$) or not rule it out in the other ($\hat{P}_{10}^*$). This discrepancy may be explained by noting that on average, rural hospitals perform worse than urban hospitals (see Table 2), but hospital 10 is not unusual among rural hospitals. Examination of the posterior probabilities in this study indicates that there are only three hospitals (hospitals 1, 28, and 10) for which there is reasonable confidence ($\hat{P}_i^{\mathrm{A-S}}$ or $P_i^* \geq 0.70$) to suspect a quality problem.

## 5. DISCUSSION

Profiling medical providers is a multifaceted and data-intensive process with significant implications for health care practice, management, and policy. The methodologic

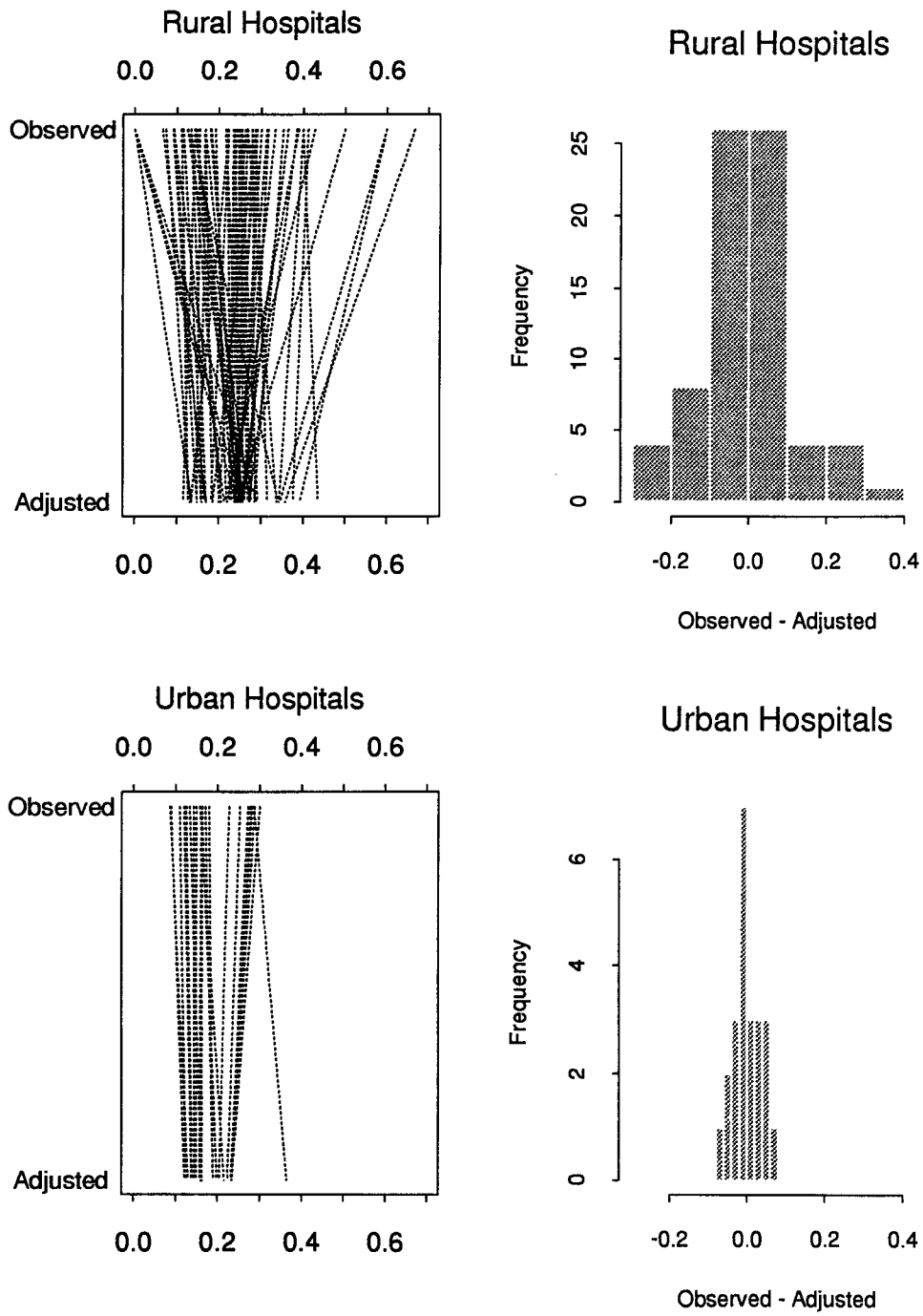


*Figure 2. Observed and Risk-Adjusted Hospital Mortality Rates. Cross-over plots display the observed mortality rate $\bar{y}_i$ (upper horizontal axis) and corresponding risk-adjusted mortality estimates $\hat{\mu}_i^A$ (lower horizontal axis) from the nonexchangeable model. Histograms present the difference, $\bar{y}_i - \hat{\mu}_i^A$.*

issues confronting analysts in this area are not trivial, and simplistic "one-size-fits-all" approaches are not likely to work (Epstein 1995). Major issues include data quality, detail, and availability; choice of performance measures, formulation of statistical analytic strategies; and development of approaches to reporting and interpreting the results of profiling analyses. In this article we focused our attention on performance measures and statistical strategies for deriving reliable estimates.

Because profiling analysis can serve a number of purposes, the choice of performance measures and analytic strategy will have to be customized. Data availability may also play a major role. For example, if a reference standard, such as a national guideline, is available, then it is reasonable to evaluate providers using an absolute cut-off level of performance. Unfortunately, such reference standards are often not available, and comparisons of providers will need to be made using relative measures of performance. For example, we estimated the probability that the risk of death for a patient of average admission severity was $1\text{-}\frac{1}{2}$ times the median mortality rate for similar patients in the sample. It is the information regarding the *actual* magnitude of the

*Table 3. Posterior Probability of Excess Mortality Across 96 Hospitals*

| Index | Minimum | Mean | Percentile 25th | Percentile 50th | Percentile 75th | Maximum |
|---|---|---|---|---|---|---|
| $P_i^{A-S}$ | 0 | 4.2 | .7 | 2.7 | 5.8 | 25.4 |
| $P_i^*$ | 0 | 11.6 | 2 | 3.9 | 15.2 | 89.5 |

NOTE: Table entries are the 100 × the probability of excess mortality, where $\hat{P}_i^{A-S} = P(\hat{\mu}_i^A - \hat{\mu}_i^S > c)$, where $c = \xi_{50}(\hat{\mu}_i^A - \hat{\mu}_i^S) + 1.5 \times$ Interquartile range and $\hat{P}_i^* = P(\text{logit}^{-1}(\hat{\beta}_{0i}) > 1.5 \times \xi_{50}(\text{logit}^{-1}(\hat{\beta}_{0i})))$.

probability associated with this performance index that is important for quality improvement activities. For this reason, we feel that ranks are of limited value (Goldstein and Spiegelhalter 1996) and that efforts need to be directed toward development of indices customized to specific problems.

The performance measures in this article were estimated using a unifying statistical approach based on hierarchical regression modeling. The approach takes into account the hierarchical structure usually present in data for profiling analyses and provides a flexible framework for analyzing a variety of different types of response variables and for incorporating covariates at the various levels of the hierarchical structure. Experience with practical uses of hierarchical modeling is growing rapidly and the computational techniques and software are becoming broadly available (Gilks et al. 1996; Goldstein 1995). As showcased in this article, the hierarchical model can be linked naturally to indices of provider performance and estimates of such indices can be derived in the course of fitting the overall model. A broad variety of performance indices can be accommodated in this framework, and their estimation and evaluation is carried out on the basis of the same underlying statistical model. In addition, hierarchical modeling can be used to address some key technical concerns in profiling analysis, including permitting the impact of patient severity on outcome to vary by provider, adjusting for within-provider correlations, and accounting for differential sample size across providers.

The regression framework presented in this article permits risk adjustment using patient-level data and incorporation of provider characteristics into the analysis. There are methodological difficulties in implementing both of these capabilities of the model. The ability to risk adjust using retrospective data can be hampered by the potential endogeneity of the recorded patient-level covariates and the potential for hidden covariates. With regard to the first factor, it is possible that there are differential error rates associated with the ability or propensity to record information across providers. For example, physicians may comment more frequently in the medical charts at tertiary care hospitals than at small community hospitals. If this is the case, it might be difficult to separate case mix from the provider effect. A second potential problem might arise if an important correlate of outcome is missing from the database. With retrospectively collected data, it is often possible that an important severity measure will be missing from the database, and furthermore, it is likely that the distribution of this unmeasured covariate will vary across providers. As a consequence, the magnitude of the difference between the adjusted and standardized outcome may be exaggerated. Some of these difficulties could be eliminated by putting a prospective data collection system in place.

Finally, the consideration of provider characteristics as possible covariates in the second level of the hierarchical model is dictated by the need to explain as large a fraction as possible of the variability in the observed data. Simple exchangeability across all providers may not be a defensible assumption for many datasets. In such cases more accurate estimates of provider-specific adjusted outcomes will be obtained by inclusion of relevant provider characteristics. Subject matter considerations will play a major role in the choice of covariates and in the interpretation of the results, because choice of the exchangeable model afffects the manner in which the $\theta_i$ are shrunk. For example, if the hospital-specific parameters are exchangeable within rural hospitals and within nonrural hospitals, then one would expect *some* shrinkage toward the prior means

*Table 4. HCFA Highest and Lowest Ranked Hospitals*

| Hospital | No. of AMI patients | No. dead | Hospital location | Academic (Y/N) | Hospital size | Random intercept: HCFA $z_i$ | Random intercept: HCFA Rank | Random intercept: $\hat{P}_i^{A-S}$ (%) | Random intercept: $\hat{P}_i^{A-S}$ Rank | Random intercept and slope: $\hat{P}_i^{A-S}$ (%) | Random intercept and slope: $\hat{P}_i^{A-S}$ Rank | Random intercept and slope: $\hat{P}_i^*$ (%) | Random intercept and slope: $\hat{P}_i^*$ Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 54 | 19 | R | N | M | 3.83 | 1 | 36 | 1 | 25 | 1 | 89 | 1 |
| 28 | 6 | 4 | R | N | M | 2.55 | 2 | 10 | 7 | 15 | 3 | 70 | 3 |
| 2 | 18 | 7 | R | N | S | 2.55 | 3 | 12 | 5 | 19 | 2 | 32 | 9 |
| 10 | 62 | 18 | R | N | M | 2.51 | 4 | 16 | 2 | 7 | 19 | 71 | 2 |
| 90 | 8 | 4 | R | N | S | 2.00 | 5 | 15 | 3 | 13 | 5.5 | 13 | 28 |
| 43 | 27 | 6 | R | N | S | 1.95 | 6 | 3 | 43 | 11 | 8 | 22 | 17 |
| 15 | 81 | 22 | U | Y | L | 1.82 | 7 | 9 | 11 | 5 | 26 | 10 | 31 |
| 44 | 7 | 3 | R | N | S | 1.75 | 8 | 6 | 20 | 5 | 33 | 11 | 30 |
| 95 | 22 | 8 | R | N | S | 1.68 | 9 | 12 | 6 | 14 | 4 | 16 | 21 |
| 29 | 31 | 5 | U | N | S | −1.75 | 93 | 0 | 84.5 | 1 | 74 | 0 | 94 |
| 39 | 6 | 0 | R | N | S | −1.77 | 94 | 2 | 54 | 3 | 48.5 | 2 | 77 |
| 19 | 46 | 4 | U | N | L | −1.80 | 95 | 0 | 90.5 | 0 | 90.5 | 0 | 94 |
| 42 | 70 | 11 | U | Y | L | −2.01 | 96 | 0 | 94.0 | 0 | 94.5 | 0 | 94 |

NOTE: HCFA highest-ranked ($z_i > 1.65$) and lowest-ranked ($z_i < -1.65$) hospitals. The rank of each measure is from worst (1) to best (96). L denotes hospitals with ≥300 beds, M denotes hospitals with 101–299 beds, S denotes hospitals with fewer than 101 beds, R denotes rural hospitals, and U denotes urban hospitals.

*Table A.1 Admission Severity Variables and Weights Comprising the Admission Severity Index*

| $X_p$ | $\hat{\beta}_p$ | $X_p$ | $\hat{\beta}_p$ |
|---|---|---|---|
| Constant | 5.5726 | LV function proxies: | |
| Socio-demographic: | | Cardiac arrest | .9069 |
| (Age—65) | .0681 | Gallop rhythm | −.0310 |
| $(\text{Age}-65)^2$ | −.0010 | Cardiomegaly | −.0094 |
| Admission history: | | Hx CHF | −.1061 |
| Hx cancer | −.1740 | Rales and pulmonary edema | .1520 |
| Admission severity: | | Laboratory results: | |
| Mobility status | | Albumin > 3 (g/dl) | −.4828 |
| Walked independently | −.2740 | Albumin missing | −.4793 |
| Unable to walk | .4700 | $\text{Log}_{10}$[BUN (mg/dl)] | 1.0613 |
| Mobility missing | .3669 | BUN missing | 1.4583 |
| Body mass index ($kg/m^2$) | −.0259 | Creatinine > 2 (mg/dl) | .3279 |
| Body mass missing | −.1525 | Creatinine missing | .1937 |
| Respiration rate breaths/min | | Diagnostic test results: | |
| Respiration (if ≥ 12) | .0429 | Conduction disturbance | .4084 |
| Respiration < 12 | 3.4840 | No EKG (vs EKG reading) | .5050 |
| Respiration missing | 2.2666 | No MI on EKG (vs MI on EKG) | −.1430 |
| Ventricular rate > 100 | .1564 | Anterior MI (vs other MI) | .4384 |
| $\text{Log}_{10}$(MAP) | −4.7101 | Lateral MI (vs other MI) | .2908 |
| MAP missing | −10.1796 | Posterior MI (vs other MI) | .6416 |
| Shock | 1.6194 | Lateral and posterior MI | −.8767 |

NOTE. Hx = history, MAP = mean arterial pressure; BUN = blood urea nitrogen level. Variables indicate the presence of the condition (coded 1 if present and 0 otherwise) with the exception of the following seven continuous covariates, which assume the observed values: age, body mass, respiration rate, MAP, albumin, BUN, and creatinine. The severity index is calculated as $\sum_p \hat{\beta}_p X_{ip}$ for the $i$th patient.

for rural and nonrural hospitals. But the structure of the *between-provider* model is not the only determinant of the resulting shrinkage. In particular, the estimates for hospitals with relatively large numbers of patients generally will be pulled only slightly toward the group mean even if they are quite different from it. In contrast, estimates of hospitals with small numbers of patients are likely to be pulled strongly toward the group mean if they differ substantially from the mean. Substantial shrinkage would be justifiable in such cases, because the raw performance estimates are bound to be imprecise.

## APPENDIX: DEVELOPMENT OF THE ADMISSION SEVERITY INDEX

A model predicting the log-odds of mortality using covariates measured within the first 24 hours of admission was developed in collaboration with a panel consisting of physicians, health services researchers, representatives of physician specialty societies, and other health care organizations. Covariate information stemming from more than 200 variables was retrospectively abstracted from medical charts and administrative data. A logistic regression model linking 30-day mortality to admission covariates was estimated using a developmental sample of 10,936 AMI Medicare patients discharged from hospitals located in the four CCP pilot states and validated on a sample of 3,645 AMI patients from the same four states using a three-phase procedure. First, stepwise regression models were fitted to subsamples of the developmental cohort, using 20 random starting models for each subsample. Second, the model with the largest likelihood in each subsample was identified, and a backward selection logistic regression procedure was used in the developmental cohort using only those covariates associated with models having the largest likelihood. Third, after assessing model fit, the regression coefficients were reestimated using the full cohort of 14,581 patients. Table A.1 lists the individual severity variables and their estimated regression coefficients that resulted from the final model. Admission severity for the $j$th patient at the $i$th hospital in this article was defined as $\sum_{p=1}^{P} \hat{\beta}_p X_{ijp}$ with $\hat{\beta}_p$ as specified in the Table A.1 (see Normand et al. 1996).

*[Received December 1994. Revised January 1997.]*

## REFERENCES


Anderson, T. W. (1971), *An Introduction to Multivariate Statistics*, New York: Wiley.

Berwick, D. M., and Wald, D. L. (1990), "Hospital Leaders' Opinions of the HCFA Mortality Data," *Journal of the American Medical Association*, 263, 247–249.

Brennan, T. A., Herbert, L. E., Laird, N. M., Lawthers, A., Thorpe, K. E., Leape, L. L., Localio, A. R., Lipsitz, S. R., Newhouse, J. P., Weiler, P. C., and Hiatt, H. H. (1991), "Hospital Characteristics Associated With Adverse Events and Substandard Medical Care," *Journal of the American Medical Association*, 265, 3265–3269.

Codman, E. (1916), "Hospital Standardization," *Surgery, Gynecology, and Obstetrics*, 22, 119–120.

Ellerbeck, E. F., Jencks, S. F., Radford, M. J., Kresowik, T. F., Craig, A. S., Gold, J. A., Krumholz, H. M., and Vogel, R. A. (1995), "Quality of Care for Medicare Patients With Acute Myocardial Infarction: A Four-State Pilot Study From the Cooperative Cardiovascular Project," *Journal of the American Medical Association*, 273, 1509–1514.

Epstein, A. (1995), "Performance Reports on Quality—Prototypes, Problems, and Prospects," *New England Journal of Medicine*, 333, 57–61.

Gatsonis, C. A., Epstein, A. M., Newhouse, J. P., Normand, S. L., and McNeil, B. J. (1995), "Variations in the Utilization of Coronary Angiography for Elderly Patients With Acute Myocardial Infarction: An Analysis Using Hierarchical Logistic Regression," *Medical Care*, 33, 625–642.

Gatsonis, C., Normand, S. L., Liu, C., and Morris, C. (1993), "Geographic Variation of Procedure Utilization: A Hierarchical Model Approach," *Medical Care*, 31, YS54–YS59.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995), *Bayesian Data Analysis*, New York, Chapman and Hall.

Gelman, A., and Rubin, D. (1992), "Inference From Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–472.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Gillis, K., and Hixson, J. (1991), "Efficacy of Statistical Outlier Analysis

for Monitoring Quality of Care," *Journal of Business and Economic Statistics*, 9, 241–252.

Goldstein, H. (1995), *Multilevel Statistical Models*, London: Edward Arnold.

Goldstein, H., and Spiegelhalter, D. J. (1996), "Statistical Aspects of Institutional Performance: League Tables and Their Limitations" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 159, 385–444.

Green, J., and Wintfeld, N. (1995), "Report Cards on Cardiac Surgeons—Assessing New York State's Approach," *New England Journal of Medicine*, 332, 1229–1232.

Haberman, S. (1976), "Generalized Residuals for Log-Linear Models," in *Proceedings of the 9th International Biometric Conference*, Raleigh, NC, pp. 104–122.

Hannan, E., Siu, A., Kumar, D., Kilburn, H., and Chassin, M. R. (1995), "The Decline in Coronary Artery Bypass Graft Surgery Mortality in New York State. The Role of Surgeon Volume," *Journal of the American Medical Association*, 273, 209–213.

Hartz, A. J., Krakauer, H., Kuhn, E. M., Young, M., Jacobsen, S. J., Gay, G., Muentz, L., Katzoff, M., Bailey, R. C., and Rimm, A. A. (1989), "Hospital Characteristics and Mortality Rates," *New England Journal of Medicine*, 321, 1720–1725.

Iezzoni, L. I. (1994), *Risk Adjustment for Measuring Health Care Outcomes*, Ann Arbor, MI: Health Administration Press.

Kassirer, J. P. (1994), "The Use and Abuse of Practice Profiles," *New England Journal of Medicine*, 330, 634–635.

Kuhn, E. M., Hartz, A. J., Gottlieb, M. S., and Rimm, A. A. (1991), "Relationship of Hospital Characteristics and the Results of Peer Review in Six Large States," *Medical Care*, 29, 1028–1038.

Kuhn, E. M., Hartz, A. J., Krakauer, H., Bailey, R. C., and Rimm, A. A. (1994), "The Relationship of Hospital Ownership and Teaching Status to 30- and 180-Day Adjusted Mortality Rates," *Medical Care*, 32, 1098–1108.

Laird, N. M., and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974.

Landon, B., Iezzoni, L., Ash, A. S., Schwartz, M., Daley, J., Hughes, J. S., and Mackiernan, Y. D. (1996), "Judging Hospitals by Severity Adjusted Mortality Rates: The Case of CABG Surgery," *Inquiry*, 33, 155–166.

Lindley, D. V., and Smith, A. F. M. (1972), "Bayes Estimates for the Linear Model" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 1–41.

Localio, A. R., Hamory, B., Sharp, T., Weaver, S., TenHave, T. R., and Landis, R. (1995), "Comparing Hospital Mortality in Adult Patients With Pneumonia: A Case Study of Statistical Methods in a Managed Care Program," *Annals of Internal Medicine*, 122, 125–132.

Longford, N. (1993), *Random Coefficient Models*, Oxford, U.K.: Oxford University Press.

McClellan, M., McNeil, B. J., and Newhouse, J. P. (1994), "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality? Analysis Using Instrumental Variables," *Journal of the American Medical Association*, 272, 859–866.

McNeil, B., Pedersen, S., and Gatsonis, C. (1992), "Current Issues in Profiling Quality of Care," *Inquiry*, 29, 298–307.

Normand, S. L., Glickman, M. E., Sharma, R., and McNeil, B. J. (1996), "Using Admission Characteristics to Predict Short-Term Mortality From Myocardial Infarction in Elderly Patients: Results From the Cooperative Cardiovascular Project," *Journal of the American Medical Association*, 275, 1322–1328.

Pashos, C. L., Newhouse J. P., and McNeil, B. J. (1993), "Temporal Changes in the Care and Outcomes of Elderly Patients With Acute Myocardial Infarction, 1987 Through 1990," *Journal of the American Medical Association*, 270, 1832–1836.

Pashos, C. L., Normand, S. L., Garfinkle, J. B., Newhouse, J. P., Epstein, A. M., and McNeil, B. J. (1994), "Trends in the Use of Drug Therapies in Patients With Acute Myocardial Infarction: 1988–1992," *Journal of the American College of Cardiology*, 23, 1023–1030.

Rosenbaum, P. (1995), *Observational Studies*, New York: Springer-Verlag.

Salem-Schatz, S., Moore, G., Rucker, M., and Pearson, S. (1994), "The Case for Case-Mix Adjustment in Practice Profiling," *Journal of the American Medical Association*, 272, 871–874.

Silber, J. H., Rosenbaum, P. R., and Ross, R. N. (1995), "Comparing the Contributions of Groups of Predictors Which Outcomes Vary With Hospital Rather Than Patient Characteristics," *Journal of the American Statistical Association*, 90, 7–18.

Shwartz, M., Ash, A., Anderson, J., Iezzoni, L. I., Payne, S., and Restuccia, J. D. (1994), "Small Area Variation in Hospitalization Rates: How Much You See Depends on How You Look," *Medical Care*, 32, 189–201.

Stukel, T. A., Glynn, R. J., Fisher, E. S., Sharp, S. M., Lu-Yao, G., and Wennberg, J. E. (1994), "Standardized Rates of Recurrent Outcomes," *Statistics in Medicine*, 13, 1781–1791.

Smith, D. (1994), "Evaluating Risk Adjustment by Partitioning Variation in Hospital Mortality Rates," *Statistics in Medicine*, 13, 1001–1013.

Thomas, N., Longford, N., and Rolph, J. (1994), "Empirical Bayes Methods for Estimating Hospital-Specific Mortality Rates," *Statistics in Medicine*, 13, 889–903.

Wong, G., and Mason, W. (1985), "The Hierarchical Logistic Regression Model for Multilevel Analysis," *Journal of the American Statistical Association*, 80, 513–524.

——— (1991), "Contextually Specific Effects and Other Generalizations of the Hierarchical Linear Model for Comparative Analysis," *Journal of the American Statistical Association*, 86, 487–503.

Zeger, S. L., and Karim, M. R. (1991), "Generalized Linear Models With Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79–86.