Hierarchical Generalized Linear Models in the Analysis of Variations in Health Care Utilization

Michael J. DANIELS and Constantine GATSONIS

In recent years many studies have reported large differences in the use of medical treatments and procedures across geographic regions, hospitals, and individual health care providers. Beyond reporting on the extent of observed variations, these studies examine the role of contributing factors including patient, regional, and provider characteristics. In addition, they may assess the relation between health care processes and outcomes, such as patient mortality, morbidity, and functioning. Studies of variations in health care utilization and outcomes involve the analysis of multilevel clustered data; for example, data on patients clustered by hospital and/or geographic region. The goals of the analysis include the estimation of cluster-specific adjusted responses, covariate effects, and components of variance. The analytic strategy needs to account for correlations induced by clustering and to handle the presence of large variations in cluster size. In this article we formulate a broad class of hierarchical generalized linear models (HGLMs) and discuss their applications to the analysis of health care utilization data. The models can incorporate covariates at each level of the hierarchical data structure, can account for greater variation than what is allowed by the variance in a one-parameter exponential family, and permit the use of heavy-tailed distributions for the random effects. We develop a Bayesian approach to fitting HGLMs using Markov chain Monte Carlo methods and discuss several methods for model checking. The HGLM analysis is presented in the context of two examples of applications to the study of variations in the utilization of medical procedures for elderly Medicare beneficiaries who sustained a heart attack. The first example involves the analysis of clustered longitudinal data with binomial responses and examines geographic and temporal trends in the utilization of coronary angiography across the United States during the 4-year period 1987–1990. The second example involves the analysis of multilevel, clustered data with Poisson responses and examines hospital variations in the utilization of coronary artery bypass graft surgery in 1990. The HGLM analysis incorporates state-level and hospital-level covariates and makes it possible to estimate covariate effects and cluster-specific rates of utilization for both hospitals and states.

KEY WORDS: Health services research; Markov chain Monte Carlo; Multilevel model.

1. INTRODUCTION

The presence of substantial variations in the delivery and outcome of health care has been documented extensively in recent years (Dartmouth Medical School, Center for Evaluative Clinical Sciences 1996; Paul-Shaheen, Clark, and Williams 1987; Wennberg and Gittelsohn 1982). Researchers in this area report often large differences in the use of medical care and its outcomes across geographic regions, hospitals, and individual health care providers. The major aim of the analysis is to measure the magnitude of the variations and to assess the role of contributing factors, including patient, regional, and provider characteristics (Diehr 1984; Diehr, Cain, Connell, and Volinn 1990; Gatsonis, Epstein, Newhouse, Normand, and McNeil 1995). In subsequent steps, researchers study the effects of variations in health care use on patient outcomes; they examine the relationship between measures of process, which define regional or hospital practice patterns, and measures of outcome, such as patient mortality, morbidity, functioning, and satisfaction with care. In this formulation, the analysis of variations is focused on comparing the performance of health care providers, such as hospitals, practice groups, or individual physicians, and is commonly referred to as *provider profiling* (Goldstein and Spiegelhalter 1996; Normand, Glickman, and Gatsonis 1997a; Normand, Glickman, and Ryan 1997b; McNeil, Pedersen, and Gatsonis 1992).

Data on health care use and outcomes have a multilevel structure, usually with patients at the first level and physicians, hospitals, and geographic regions forming the upper-level clusters. Cluster size varies substantially at each level of the hierarchy, and covariates are often available measuring, for example, disease severity and comorbidity for individual patients, and location, size, and organizational characteristics for hospitals. A key analytic goal is to provide cluster-specific estimates of a particular response, such as the rate of utilization of a procedure by hospital or geographic region, adjusted by patient characteristics. Another key goal is to derive estimates of covariate effects, such as differences in health care utilization between patients of different gender or race and practice differences between urban and rural hospitals.

Hierarchical regression modeling provides a general analytic approach that can accomplish these goals. The main purpose of our article is to discuss the application of a broad class of such models, called *hierarchical generalized linear models* (HGLMs). The response variable in these models is distributed according to a one-parameter exponential family, such as binomial, Poisson, exponential, and Gaussian. Covariates can be discrete or continuous. Models of this type have previously been widely used in the analysis of longitudinal data (Gilks, Wang, Yvonnet, and Coursaget 1993; Laird and Ware 1982; Lindstrom and Bates 1988; Longford 1987). They have also been developed and used

Michael J. Daniels is Assistant Professor, Department of Statistics, Iowa State University, Ames, IA 50011 (E-mail: *mdaniels@iastate.edu*). Constantine Gatsonis is Associate Professor and Director, Center for Statistical Sciences, Brown University, Providence, RI 02912. This work was supported in part by a Howard Hughes Medical Institute predoctoral fellowship. The authors thank their colleagues at the Harvard Acute Myocardial Infarction Patient Outcomes Research Team (PORT) for access to the datasets used in this article, the Department of Health Care Policy at Harvard University for supporting the early stages of this work, and Joel Greenhouse and Robert Kass of Carnegie Mellon University for helpful comments. The authors also thank the editor, an associate editor, and three referees for their helpful comments.

in the analysis of effects of contextual factors in the social sciences (Wong and Mason 1985, 1991), in education (Bryk and Raudenbush 1992; Goldstein 1995), in quality control (Natarajan, Ghosh, and Maiti 1998), and in the analysis of spatial data and small area estimation (Breslow and Clayton 1993; Ghosh, Natarajan, Stroud, and Carlin 1998). Applications to the analysis of health care data are relatively recent and have concentrated primarily on binary or multinomial response variables (Calvin and Sedransk 1991; Daniels and Gatsonis 1997; Gatsonis et al. 1995; Gatsonis, Normand, Liu, and Morris 1993; Kahn and Raftery 1996; Malec, Sedransk, Moriarity, and LeClere 1997; Normand et al. 1997a; Normand et al. 1997b). Models for other types of response variables were recently used by Goldstein and Spiegelhalter (1996).

We illustrate the analysis with two examples drawn from a study of variations in the use of cardiac procedures for elderly Medicare beneficiaries who sustained a heart attack. The data for the study were derived from Medicare records compiled by the Health Care Financing Administration. The first example examines geographic and temporal trends in the use of coronary angiography during the 4-year period 1987–1990. Patients were classified by state of residence, and yearly rates of angiography were computed for each state. The goal of the analysis was to examine regional patterns of angiography use in this patient cohort and to assess their temporal evolution (see also Pashos, Newhouse, and McNeil 1993). The second example examines differences in the use of coronary artery by-pass graft (CABG) surgery across hospitals and geographic regions of the United States during a single year, 1990. For this analysis, patients were grouped by the hospital to which they were initially admitted for their heart attack, and hospitals were then grouped by state. In addition to examining geographic practice patterns, the goal of our analysis was to determine whether hospital size and teaching status were related to the use of CABG.

The methodological focus of this article is on fully Bayesian approaches to model fitting and checking for a general class of HGLMs, in which the regression coefficients at each level of the hierarchical structure can be modeled as functions of higher-level covariates. For example, the slope of a patient-level covariate is allowed to vary as a function of hospital characteristics. By assuming a conjugate prior for the second level of the hierarchy, these models make it possible to account for extra variation between clusters and to allow a wider variety of marginal variance structures for the data. The choice of distribution for the random coefficients (effects) is made adaptively among the members of a class of priors, which includes heavy-tailed distributions. The mixed models discussed by Breslow and Clayton (1993) and Zeger and Karim (1991) can be viewed as special cases of this general class of HGLMs, under appropriate assumptions on the exchangeability structure, the conjugate prior, and the distribution of the random effects.

We use Markov chain Monte Carlo (MCMC) simulation of the posterior distribution of the parameters to fit the HGLMs. In particular, we use the Gibbs sampler (Smith and Roberts, 1993) with a Metropolis step (Smith and Roberts 1993). This fully Bayesian approach provides a more realistic account of the uncertainty in the model than is generally possible to obtain from empirical Bayes and other approximations. We discuss several approaches to model selection and checking, including methods based on the posterior predictive distribution and the marginal distribution of the data (Albert and Chib, 1997; Gelfand, Dey, and Chang 1995; Gelman, Meng, and Stern 1996; Verdinelli and Wasserman 1995; Weiss 1996).

Many types of HGLMs can be fitted using the BUGS software (Spiegelhalter, Best, Gilks, and Inskip 1996). However, special programs must be developed for a number of models, because BUGS currently includes only Pareto and inverse gamma priors for variance components. Analytic approximations to the posterior distribution have also been used in some relatively simple HGLMs (Albert 1988; Christiansen and Morris 1997; Kass and Steffey 1989). However, such approximations are not readily available for the more elaborate models typically needed in applications of the kind we are discussing. Non-Bayesian approaches to fitting classes of HGLMs include the use of generalized estimating equations (GEE) (Zeger, Liang, and Albert 1988), the generalized linear mixed model (GLMM) methodology of Breslow and Clayton (1993), and the recently published quasi-likelihood method of Lee and Nelder (1996). We present comparisons of our results to those derived from the SAS subroutine GLIMMIX (Khattree and Naik 1995) and from the statistical package HLM (Bryk, Raudenbush, and Congdon 1998), which fit GLMMs using penalized likelihood methods.

The HGLM approach provides a unified modeling framework for estimating cluster-specific quantities of interest, covariate effects, and components of variance. The models make it possible to pool information across clusters to derive more precise estimates of case-specific and cluster-specific parameters. In addition, the application of HGLMs make it possible to account for correlations in the data and to derive standard errors of estimates, which are more realistic than those obtained by methods ignoring such correlations. The use of MCMC enhances considerably the analyst's flexibility with respect to the complexity of the models and the range of functions of the parameters that can be examined. In Section 2 we present the general form of the models and discuss MCMC methods for model fitting. In Sections 3 and 4 we discuss specific applications to the analysis of variations in the utilization of cardiac procedures for elderly Medicare patients, who had a heart attack. In each example we provide a detailed discussion of the approaches used to fit and check the models under consideration. We present concluding remarks in Section 5.

2. HIERARCHICAL GENERALIZED LINEAR MODELS

2.1 The General Model

HGLMs are applicable to situations in which a response Y is observed on cases grouped in clusters. Let Y_{ij} denote the response for case $j = 1, \ldots, J_i$, in cluster $i = 1, \ldots, N$.

The data include a covariate vector $\mathbf{X}_{ij}$ associated with the ijth case and a covariate vector $\mathbf{Z}_i$ associated with the ith cluster. If the distribution of Y_{ij} follows a one-parameter exponential family, then a four-level hierarchical model can be written as follows.

Level I (Within-Case Variation). The responses Y_{ij} follow a one-parameter exponential family and are conditionally independent given θ_{ij}. Formally,

$$Y_{ij}|\theta_{ij} \sim f(y_{ij}|\theta_{ij}) = a(y_{ij}) \exp((y_{ij}h(\theta_{ij}) - \psi(\theta_{ij}))). \quad (1)$$

Level II (Within-Cluster Variation). Within each cluster, the θ_{ij}'s are conditionally independent given a cluster-specific parameter vector $\boldsymbol{\alpha}_i$ of dimension $q \times 1$ and scale parameter δ_i ($\delta_i > 0$). The conditional distribution of the θ_{ij} is conjugate to Level I:

$$\theta_{ij}|\boldsymbol{\alpha}_i, \delta_i \sim g(\theta_{ij}; k_1, k_2) \propto \exp(k_1(\boldsymbol{\alpha}_i, \delta_i)h(\theta_{ij}) + k_2(\boldsymbol{\alpha}_i, \delta_i)\psi(\theta_{ij})). \quad (2)$$

The means of the individual θ_{ij} are connected to individual-level covariates through a link function, l, so that $l(E(\theta_{ij}|\boldsymbol{\alpha}_i, \delta_i)) = \mathbf{X}'_{ij}\boldsymbol{\alpha}_i$. (Note that longitudinal data with individual, possibly time-varying covariates can also be handled in this setup.)

The hierarchical structure of levels I and II allows for additional variation in the observed responses beyond that specified by the one-parameter exponential family variance function. It is also possible to combine levels I and II by integrating out the θ_{ij}'s to obtain the distribution of the response conditional on the cluster-specific parameters, $\boldsymbol{\alpha}_i$ and δ_i. For example, a beta-binomial distribution will be specified for binomial data.

Level III (Between-Cluster Variation). The variation between clusters is separated into a systematic and a random component. The former is modeled via regression equations linking cluster-specific parameters to cluster-level covariates. The latter is expressed via distributional assumptions on the error terms. In particular, we assume that

$$\boldsymbol{\alpha}_i, \boldsymbol{\gamma}, \mathbf{D} \sim t_v(G_i^T\boldsymbol{\gamma}, \mathbf{D}), \quad (3)$$

where $t_v(a, b)$ denotes a multivariate t-distribution with v df, location parameter a, and scale parameter (matrix) $\mathbf{b}$. G_i is formed from the Kronecker product $\mathbf{Z}_i \otimes I_{q\times q}$ with $\mathbf{Z}_i$ a $k \times 1$ vector of cluster-level covariates and $\boldsymbol{\gamma}$ a $qk \times 1$ vector of coefficients. Marginally, each $\boldsymbol{\alpha}_{iw}$ $(w = 1, \ldots, q)$ will be a linear function of cluster-level covariates plus the corresponding error term. To construct a prior on the scale parameters δ_i, note that the conditional posterior mean of the θ_{ij} can be written as

$$E[\theta_{ij}|\boldsymbol{\alpha}_i, \delta_i, y_{ij}] = S(\delta_i) * \hat{\theta}_{ij}(y_{ij}) + (1 - S(\delta_i)) * l^{-1}(\mathbf{X}'_{ij}\boldsymbol{\alpha}_i), \quad (4)$$

where l is the link function, $\hat{\theta}_{ij}(y_{ij})$ is some function of the y_{ij}, and $S(\delta_i)$ is a shrinkage parameter, constrained to the interval (0, 1). A "vague" prior on δ_i can be constructed by assuming a uniform distribution on the shrinkage parameter, $S(\delta_i)$, and making the appropriate transformation to obtain $\pi(\delta_i) = dS^{-1}(\delta_i)/d\delta_i$. For further discussion of this prior, see Section 3.2.

Level IV. To complete the Bayesian specification of the multilevel model, we assume proper priors on all remaining hyperparameters. In particular, we assume $\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{\gamma}^*, A)$, with $\boldsymbol{\gamma}^*$ a prior estimate of $\boldsymbol{\gamma}$ and $A^{-1} = \sum_{i=1}^{N} G_i S^{-1} G'_i$. If S is chosen to be approximately equal to $\mathbf{D}$, then the additional information from this prior can be thought of as equivalent to adding one cluster to the dataset (Kass and Wasserman 1995). We also assume $\mathbf{D}^{-1} \sim W_q((qS)^{-1}, q)$, where $W_q((qS)^{-1}, q)$ is a q-dimensional Wishart distribution with q df (see, e.g., O'Hagan 1994).

Finally, we assume a uniform prior on a finite interval for the degrees of freedom v of the t distribution at level III. This interval was chosen to be [1, 100] in the analysis of the examples in Sections 3 and 4. If the interval is large and the data contain little information on the degrees of freedom, then an analysis with a flat prior on v would be similar to assuming a multivariate normal distribution in (3). To give more weight to heavy-tailed distributions in (3), a uniform (0, 1) prior on $1/v$ or the prior suggested by Besag, Green, Higdon, and Mengersen (1995) could be used.

2.2 Model Fitting

The structure of HGLMs makes them natural candidates for the use of MCMC algorithms to simulate the posterior distribution of the parameters. In particular, observations from the full conditionals of $\mathbf{D}$ and $\boldsymbol{\gamma}$ can be generated directly, because these distributions have familiar forms (Wishart and multivariate normal). Observations from the conditional of θ need not be generated, because the use of a conjugate prior at level II allows these parameters to be integrated out. However, observations from the full conditionals for $\boldsymbol{\alpha}_i, \delta_i$, and v cannot be generated directly. We used the random-walk Metropolis–Hastings algorithm (Smith and Roberts 1993) with a normal candidate density for $\log(\delta_i)$, and $\log(v - 1)$. The log transformations were used, because δ_i is constrained to the nonnegative real line and v to the finite interval (1, 100].

To simulate values from the full conditional distributions of $\boldsymbol{\alpha}_i$, we used a version of the Metropolis–Hastings algorithm with a multivariate normal candidate density as suggested by Bennet, Racine-Poon, and Wakefield (1996). We first assumed a normal approximation to the maximum likelihood estimate (MLE) of the $\boldsymbol{\alpha}_i$, conditional on the current value of the δ_i (having integrated out the θ). Formally, $\hat{\boldsymbol{\alpha}}_i|\delta_i \sim \mathrm{N}(\boldsymbol{\alpha}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ is the inverse of the observed information matrix. Using this assumption, the full conditional of $\boldsymbol{\alpha}_i$ is a multivariate normal distribution, with mean equal to a weighted combination of the MLE and the current value of the prior mean. Formally, $\boldsymbol{\alpha}_i|\delta_i, \boldsymbol{\gamma}, \mathbf{D}, y \sim \mathrm{N}(\boldsymbol{\alpha}_i^*, \mathbf{V}_i)$, where $\boldsymbol{\alpha}_i^* = (\mathbf{D}^{-1} + \boldsymbol{\Sigma}_i^{-1})^{-1}(\mathbf{D}^{-1}\hat{\boldsymbol{\alpha}}_i + \boldsymbol{\Sigma}_i^{-1}G_i^T\boldsymbol{\gamma})$ and $\mathbf{V}_i = (\mathbf{D} + \boldsymbol{\Sigma}_i)^{-1}$. To ensure that we sampled from the correct conditional distribution, we included a Metropolis step with acceptance proba-

bility $a(\boldsymbol{\alpha}_i^{(0)}, \boldsymbol{\alpha}_i) = \min(1, \pi(\boldsymbol{\alpha}_i)/\hat{\pi}(\boldsymbol{\alpha}_i)/\pi(\boldsymbol{\alpha}_i^{(0)})/\hat{\pi}(\boldsymbol{\alpha}_i^{(0)}))$, where π denotes the true conditional distribution of the $\boldsymbol{\alpha}_i$, $\hat{\pi}$ denotes the approximate full conditional distribution of the $\boldsymbol{\alpha}_i$, and $\boldsymbol{\alpha}_i^{(0)}$ denotes the current value (Tierney 1994).

The computations involving the multivariate t distributions in level III of the model were simplified by making use of the fact that the t distribution can be expressed as a gamma mixture of normals. This formulation introduces an additional scale parameter τ_i, such that, if $\boldsymbol{\alpha}_i \sim t_v(\mu, \lambda)$, then $\boldsymbol{\alpha}_i|\tau_i \sim \mathrm{N}(\mu, \lambda/\tau_i)$ with $\tau_i \sim \Gamma(v/2, v/2)$ and $E[\tau_i] = 1$. Thus the full conditional of τ_i will follow a gamma distribution (Liu 1996).

Initial values for the individual-level and cluster-level covariates were obtained by first fitting individual GLM regressions for each cluster and then regressing these coefficients on the cluster-level covariates. We used the centering parameterization for the model as recommended by Gelfand, Sahu, and Carlin (1995) to improve model mixing and reduce correlations between parameters. Convergence of the chain was monitored using univariate and multivariate versions of the approach of Gelman and Rubin (1992). In particular, we ran m parallel strings, with half of the strings started at overdispersed maximum likelihood values and the remaining ones started near the grand mean. We then used plots of the average of the $\log|\mathbf{D}|$ for the overdispersed and the underdispersed strings to assess convergence. These plots monitor the overall convergence of the set of parameters $(\boldsymbol{\alpha}_i, \gamma, \mathbf{D})$. In addition, we examined the convergence of the parameters individually through computation of the Gelman and Rubin statistic and plots of the strings. Plots of the average of each parameter over parallel strings and the running average for each parameter were also helpful in assessing convergence (Cowles and Carlin 1995).

3. EXAMPLE 1: LONGITUDINAL TRENDS IN CARDIAC PROCEDURE UTILIZATION

3.1 The Problem

The data for this example were derived from Medicare use and administrative files compiled by the Health Care Financing Administration. Separate patient cohorts were constructed for each year from 1987 to 1990. Each cohort consisted of elderly Medicare beneficiaries who had sustained a heart attack, or acute myocardial infarction (AMI), in the particular year and no prior hospital admission for AMI during the preceding year. The process of selecting the final cohorts involved several further exclusion criteria, resulting in a total of approximately 790,000 patients in all four years (see Pashos et al. 1993 for details).

The aim of the project was to examine the geographic and temporal trends in the use of medical procedures for elderly AMI patients across the country during this time period and to examine possible links to outcomes, notably patient mortality. Subject matter findings from the project have been published in the medical literature (Gatsonis et al. 1995; McClellan, McNeil, and Newhouse 1994; Pashos et al. 1993; Udverhalyi et al. 1992). Here, we focus on coronary angiography, which is a key diagnostic procedure for deciding on further treatment for AMI patients.

To study geographic variations, patients were grouped by state of residence. For each state and year the observed rate of angiography was computed as the proportion of the state's residents who underwent angiography within 90 days from the hospital admission for their AMI. Because previous studies of geographic variation have shown regional differences in the use of procedures for AMI patients, we used census region (West, Midwest, South, Northeast) as a state-level covariate. In addition, the rate of availability of coronary angiography in 1987 was used as a state-level proxy for the availability of medical procedures to AMI patients. For a given state, the rate of availability of angiography was defined as the proportion of state residents in the 1987 cohort who had their index admission to a hospital that could perform coronary angiography (Gatsonis et al. 1995).

3.2 Models

A binomial assumption appeared appropriate for the state rates of angiography. An exploratory analysis using generalized linear models gave evidence of overdispersion, since the Pearson X^2 statistic (McCullagh and Nelder 1989, p. 34) was quite large for most states. A linear time trend seemed reasonable as the logit of the angiography rates increased fairly linearly with time. Figure 1 shows the time trend of the logit of the pooled regional rates.

In the notation of Section 2.1, the response variable Y_{ij} denotes the observed count of coronary angiographies performed on patients in the ith state $(i = 1, \ldots, 51)$ during the year $1986 + j$ $(j = 1, \ldots, 4)$. We considered HGLMs with the following structure.

- Level I (variation of annual counts within states). The number of angiograms Y_{ij} was assumed to have a binomial distribution with mean $(n_{ij}\theta_{ij})$, where θ_{ij} denotes the angiography rate (per patient in this cohort) and n_{ij} denotes the number of AMI patients in state i during year $1986 + j$.
- Level II (longitudinal variation within states). We assumed that the angiography rates followed a beta distribution with mean m_{ij} and dispersion parameter δ_i. The mean of the beta distribution was expressed as a function of occasion-specific covariates, in this case expressing time (see also Kahn and Raftery 1996). Formally, we assumed that $\mathrm{logit}(m_{ij}) = (1, j-1)(\boldsymbol{\alpha}_{i0}, \boldsymbol{\alpha}_{i1})'$ and $\theta_{ij}|\boldsymbol{\alpha}_i, \delta_i \sim \mathrm{beta}(\delta_i m_{ij}(\boldsymbol{\alpha}_i), \delta_i(1 - m_{ij}(\boldsymbol{\alpha}_i)))$.

The dispersion parameter, δ_i, can be interpreted as a measure of the information contained in the beta prior, expressed in the metric of sample size relative to n_{ij}, the sample size in the binomial likelihood. The prior on δ_i was of the form $\delta_i \sim \pi(\delta_i) = n_{0i}/(n_{0i} + \delta_i)^2$. This prior has median n_{0i} and is rather dispersed (Christiansen and Morris 1997). The constant n_{0i} was set at the minimum of the n_{ij} for the ith state. Other choices for the constants are possible to either encourage or discourage shrinkage. We examine sensitivity to the choice of constants later.

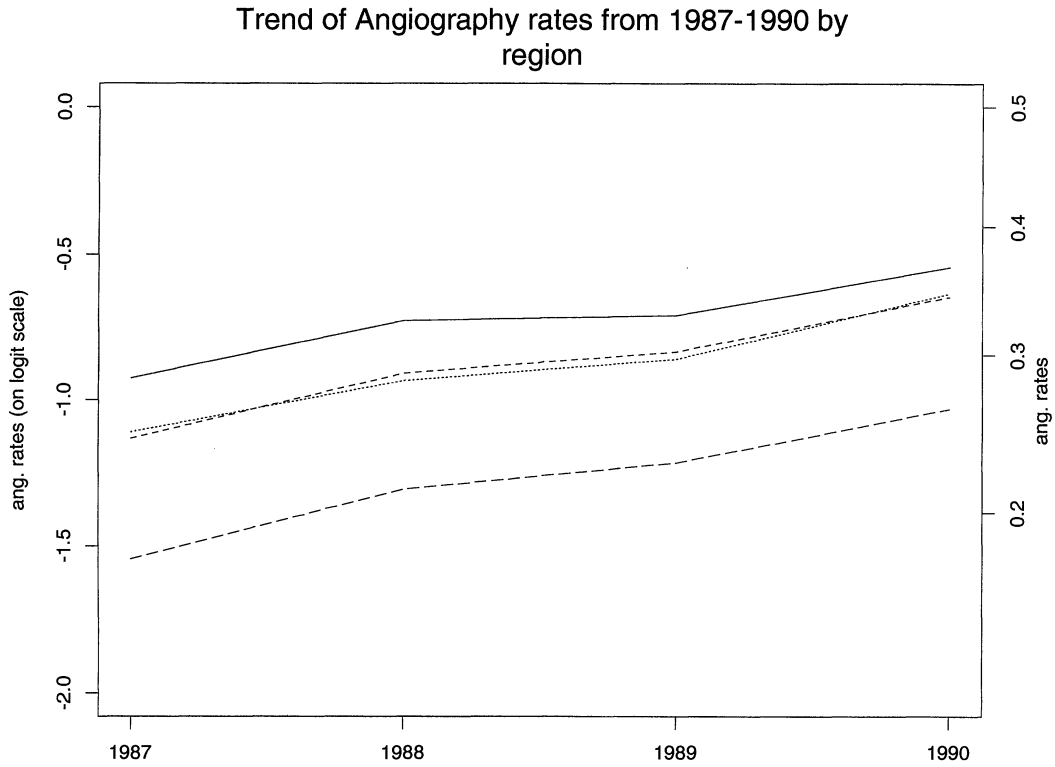

Figure 1. Regional Trends in CABG and Angiography Rates. ———, West; · · ·, Midwest; - - -, South; — — —, Northeast.

- Level III (variation between states). The state-specific coefficients were linked to the state-level covariates by a linear model of the form $\boldsymbol{\alpha}_i|\gamma, \mathbf{D} \sim t_v(G_i'\gamma, \mathbf{D})$, with G_i as defined in Section 2.1 and $\mathbf{Z}_i = (I(\text{West}), I(\text{Midwest}), I(\text{South}), I(\text{Northeast})$, coronary angiography availability at the ith state). Regional indicators were binary (0–1) variables, and availability took values in the interval [0, 1]. The fourth level of the models was specified as in Section 2.1.

3.3 Analysis

3.3.1 Gibbs Sampler. The Gibbs sampler appeared to have converged after 100 iterations. After inspecting the autocorrelation of the simulated values of the state-specific coefficients $\boldsymbol{\alpha}_i$, we decided to sample every fifth observation for posterior estimation. We ran four strings, each of length 1,350, giving us a posterior sample of size 1,000. The average acceptance probability of the Metropolis–Hastings algorithm for the $\boldsymbol{\alpha}_i$'s was nearly 95%, indicating that the normal approximations to the full conditionals of the $\boldsymbol{\alpha}_i$'s were quite accurate. The acceptance probabilities for v and δ_i ranged from 30% to 50% for the random-walk Metropolis–Hastings algorithm. The posterior distribution for the degrees of freedom at the level III distribution was centered around 50; however, the data provided little information on the degrees of freedom. The point estimates for the various parameters reported herein are based on posterior sample means. The 95% credible intervals were computed using the 2.5% and 97.5% quantiles of the samples from the posterior distribution generated by the Gibbs sampler.

3.3.2 Geographic and Temporal Trends. Estimated rates of coronary angiography for the 1987 (baseline) cohort of elderly patients with AMI ranged from a maximum of .37 [95% credible interval (.34, .41)] in Nevada to a minimum of .14 (.12, .16) in Rhode Island. The national median rate was .24 (Fig. 2). By region, the 1987 rates of angiography were highest in the West (.28), followed by the Midwest and South (.24) and then the Northeast (.19). Figure 2 clearly demonstrates the regional trends. In addition, there was a significant positive association between availability of angiography and rate of angiography [$\gamma_5 = .96(.56, 1.38)$].

Angiography use in this patient cohort increased during the 4-year period 1987–1990 in each state of the country. By 1990, the maximum estimated rate of angiography was .43 (.40, .47), again in Nevada, and the minimum rate was .21 (.19, .24), again in Rhode Island. The national median rate in 1990 was .34. As in previous years, the Northeast tended to have lower rates than the rest of the country. The estimated slope of the increase in angiography (modeled as an odds ratio) ranged from a maximum of 1.22 (1.15, 1.27) in Virginia to a minimum of 1.08 (1.01, 1.15) in Montana, with a median state-specific odds ratio of 1.16. In addition, the correlation between the baseline angiography rate and the magnitude of the time trend was negative, $(\hat{D}_{12}/(\sqrt{\hat{D}_{11}\hat{D}_{22}}) = -.47)$, suggesting that states with higher 1987 rates had less steep slopes.

3.3.3 Discussion. Although coronary angiography use increased in each state of the country during this time period, substantial regional differences persisted. This finding is consistent with the conclusions of a large body of research documenting the existence of practice patterns

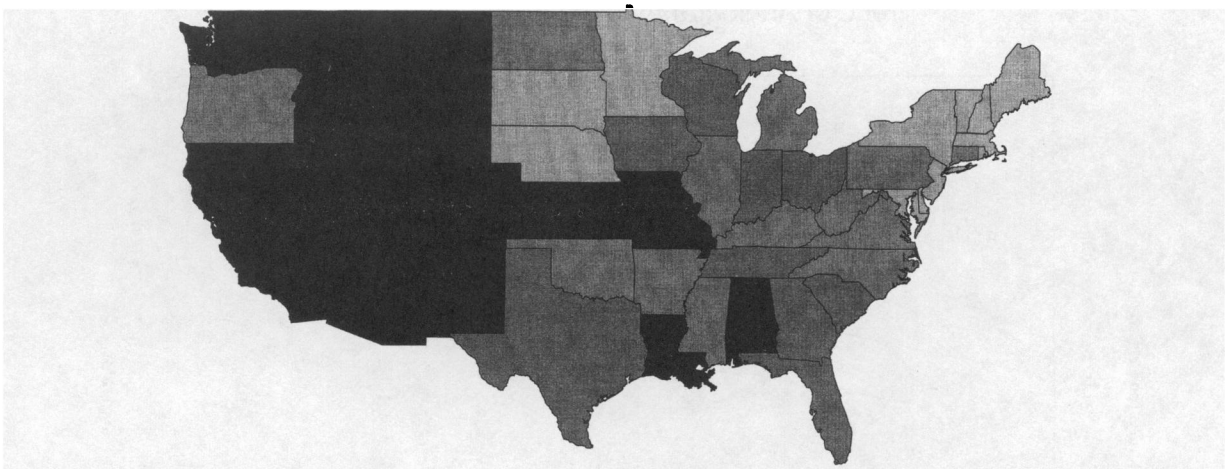

Figure 2. 1987 Fitted Angiography Rates by State. , <.20; , .20–.24; , .24–.28; , >.28.

across regions defined by geopolitical or market criteria (Dartmouth Medical School, Center for the Evaluative Clinical Sciences 1996). An important correlate of utilization was the availability of the procedure at the hospital of a patient's index admission. Although a causal link cannot be determined from this association, the finding suggests the need for further investigation into the relationship between health care market characteristics and health care use. In this analysis we used the 1987 data for procedure availability. An alternative would have been to derive state rates of the availability of coronary angiography for each of the four years and to include them as a time-varying covariate. However, such data were not at our disposal.

Our analysis was based on aggregate data for each state and did not incorporate patient-level characteristics. This was done primarily for reasons of parsimony for the purposes of this example. A more complete analysis could incorporate individual patient data via a logistic model. (For details of such an analysis of the 1987 cohort using hierarchical logistic regression see Gatsonis et al. 1995.) The extension from the cross-sectional (single year) to the longitudinal setting would be carried out along the lines of the model presented in this section.

The assessment of the consequences of geographic and temporal trends in utilization of angiography documented in this analysis calls for a rather complex set of investigations that go well beyond the scope of our article. A key difficulty is that the impact on patient outcome of a diagnostic procedure, such as angiography, is mediated by treatment strategies. It has been shown, for example, that overall mortality in this patient cohort decreased during the period 1987–1990 (Pashos et al. 1993). However, the use of revascularization procedures increased substantially during this time period, as did the use of thrombolytic therapy. In addition to confounding by treatment, a study of the implications of utilization of angiography will also need to account for possible selection effects. An interesting set of results in this regard was provided by McClellan et al. (1994) who used an instrumental variable approach to examine the relation of angiography to patient survival.

3.3.4 Model Selection. Subject matter information and computational considerations are key elements in the construction of HGLMs for particular datasets. The process may require comparisons of alternative models and is often aimed at selecting a parsimonious model to be used for developing final estimates of cluster-specific quantities and covariate effects.

To arrive at a final model for the longitudinal utilization data we used posterior credible intervals to estimate the magnitude of the effect of particular state-level covariates and computed Bayes factors to compare alternative models. We also assessed the need for the level II prior (overdispersion), using the posterior predictive distribution of the data.

The components of γ measuring regional differences, overall time trend, and overall effect of availability of angiography had 95% posterior credible regions that did not include 0. However, the corresponding interval for the parameter γ_{10}, which measures the effect of availability on the time trend, included 0. The Bayes factor (BF) for the hypothesis that $\gamma_{10} = 0$ was BF $= 8.3$, indicating sufficient evidence for the simpler model with no availability effect on the time trend (Kass and Raftery 1995). On the basis of this information, the availability of angiography was removed from the level III equation for α_{i1}.

The derivation of BFs in HGLMs can be computationally demanding because of the large number of parameters and the multi-level structure of the model. This complexity would make it difficult to use importance sampling from the prior or to implement the Laplace approximations of DiCiccio, Kass, Raftery, and Wasserman (1995). We used the approach of Verdinelli and Wasserman (1995), which is appropriate for comparing nested models and proceeds as follows. If $\boldsymbol{\theta} = (\boldsymbol{\omega}, \psi)$ is the parameter vector, H_0: $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ is the null hypothesis, H_a: $\boldsymbol{\omega} \neq \boldsymbol{\omega}_0$ is the alternative hypothesis, P_0 is the prior under H_0, and P is the prior under H_a, the BF can be written as BF $= p(\boldsymbol{\omega}_0|y)E[p_0(\psi)/p(\psi, \boldsymbol{\omega}_0)]$, with the expectation taken with respect to $p(\psi|\boldsymbol{\omega}_0, y)$. The marginal prior p_0 is the prior distribution of ψ under the null hypothesis and $p(\psi, \omega)$ denotes the joint prior distribution of ψ and $\boldsymbol{\omega}$ under the full model. We used the fact that $p(\boldsymbol{\omega}_0|y, \psi_i)$ can be obtained in closed form and

estimated $p(w_0|y)$ by $\hat{p}(\boldsymbol{\omega}_0|y) = (1/N)\sum_{i=1}^{N} p(\boldsymbol{\omega}_0|y, \psi_i)$ (Gelfand and Smith 1990). To compute the correction factor, $E[p_0(\psi)/p(\psi, \boldsymbol{\omega}_0)]$, the Gibbs sampler was rerun with $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, and the correction factor was computed by averaging the ratio of priors over iterations.

We examined the necessity of level II in the prior structure (modeling overdispersion) using discrepancy statistics and coverage probabilities computed from the posterior predictive distribution of the data. Alternative approaches would be to compute Bayes factors; to derive the posterior probability that $p(\delta_i > K|y)$, where K is some large value for which the overdispersion is essentially 0 and δ_i is the dispersion parameter from the level II prior; or to use criteria based on the predictive distribution of future data given the current data (Gelfand and Ghosh 1998; Laud and Ibrahim 1995).

To assess the model's goodness of fit, we used the posterior predictive checking approach of Gelman et al. (1996). We computed statistics of the form $\chi^2(\tilde{y}, \boldsymbol{\alpha}, \delta) = \sum_{i=1}^{n}\sum_{j=1}^{J_i}(\tilde{y}_{ij} - k_{ij}l(x_{ij}\boldsymbol{\alpha}^p))^2/\mathbf{V}(\boldsymbol{\alpha}_i^p, \delta_i^p)$, where $\mathbf{V}(\boldsymbol{\alpha}_i^p, \delta_i^p) = \text{var}(y_{ij}|\boldsymbol{\alpha}_i^p, \delta_i^p)$ denotes the marginal variance of the data having integrated out the $\boldsymbol{\theta}$'s; k_{ij} denotes the exposure or offset for Poisson data (e_{ij} in Sec. 4.2) or the sample size for binomial data (n_{ij} in this example); $(\boldsymbol{\alpha}_i^p, \delta_i^p)$ denotes an observation from the joint posterior distribution of $\boldsymbol{\alpha}_i$ and δ_i; $\tilde{y}$ denotes the observations from the posterior predictive distribution of y; and l denotes the link function defined earlier. If $\tilde{y}$ represents the predicted values of y, then the resulting statistic is denoted by χ^2_{pred}. If $\tilde{y}$ represents the observed values of y, then the resulting statistic is denoted by χ^2_{obs}. Intuitively, χ^2_{pred} measures the variation predicted by the model and χ^2_{obs} measures the variation observed in the data. By repeated sampling, a p value can be defined as the number of times that χ^2_{pred} exceeds χ^2_{obs}. This p value is a useful indicator of whether the variation in the data is consistent with the variation predicted by the model. Extreme p values (near 0 or 1) would indicate inconsistency, with p values near 1 indicating that the variation in the data is considerably smaller than the model prediction and p values near 0 indicating the opposite. However, precise rules for choosing cutoff values for p are not available, and calibration of p values across models is difficult to carry out.

The estimated p value for the longitudinal model without the level II prior was 0, indicating that the variation predicted by the simple model is substantially smaller than the variation in the data. However, when the level II prior was included in the model, the p value was .385, indicating that accounting for overdispersion is needed. Note that if this p value had been near 1, then the conclusion would have been that the variability predicted by the overdispersion model is far greater than the variability in the data. As we show in the second example, the discrepancy statistic of Gelman et al. (1996) can also be used to check for overdispersion in each cluster.

In addition to calculating discrepancy statistics, we also assessed the necessity for modeling overdispersion by computing coverage probabilities of statistics computed from the posterior predictive distribution. In particular, if $y_{il}, l = 1, 2, \ldots, N$, denotes a sample of size N from the posterior predictive distribution of the ith cluster and $\boldsymbol{\Sigma}_i$ denotes the sample covariance matrix, then we defined the statistic, $\mathbf{D}_{il}^{\text{pred}}$, as $\mathbf{D}_{il}^{\text{pred}} = (y_{il} - \bar{y}_i)^T\boldsymbol{\Sigma}_i^{-1}(y_{il} - \bar{y}_i)$. We also defined $\mathbf{D}_i^{\text{obs}}$ by replacing y_{il} with the observed value y_i^{obs} of the vector of responses for cluster i in the formula for $\mathbf{D}_{il}^{\text{pred}}$. A measure of model adequacy can be obtained by counting the fraction of clusters in which $\mathbf{D}_i^{\text{obs}}$ is contained between the 2.5th and 97.5th percentiles of the empirical distribution of $\mathbf{D}_i^{\text{pred}}$ (computed from the above sample of size N). In the analysis of the longitudinal data, the model with overdispersion resulted in coverage probability of ~98%, which is close to 95%. However, the simpler model without overdispersion resulted in coverage probability of ~82%, which is substantially below the 95% level.

3.3.5 Goodness of Fit of Final Model. The final model used to derive the results reported in Section 3.3.2 had levels I and II as specified in Section 3.2; level III contained the regression of the state-specific intercepts on region and angiography availability and of the state-specific slopes on region. After selecting the final model, we made global assessments of fit as well as assessments of fit for individual observations and clusters. To assess global model fit, we adapted the methods proposed by Jacquier, Polson, and Rossi (1994) and Weiss (1996).

In the Weiss approach, for each run of the sampler we computed the statistic $\xi(\phi) = \sum_{i=1}^{N} I\{(\boldsymbol{\alpha}_i - G_i^T\boldsymbol{\gamma})^T\tau_i\mathbf{D}^{-1}(\boldsymbol{\alpha}_i - G_i^T\boldsymbol{\gamma}) > \chi^2(q, 1-\phi)\}$, where $1-\phi$ denotes the desired quantile of the chi-squared distribution, I denotes an indicator function, and τ_i denotes the scale parameter in the representation of the multivariate t distribution as a gamma mixture of normals. The statistic $\xi(\phi)$ counts the number of outliers among the values of the state-specific coefficients $\boldsymbol{\alpha}_i$ and has a binomial(N, ϕ) a priori distribution. A global test of model fit can be obtained by comparing this binomial distribution to the posterior distribution of $\xi(\phi)$. As discussed by Weiss (1996), this approach is capable of responding to many types of potential model failures. An inspection of plots of expected and observed distributions of the number of outliers among the 51 states [Fig. 3, (a) and (b), $\phi = .05$] indicates a reasonably good overall fit.

In the approach of Jacquier et al. (1994), we computed posterior quantiles of the τ_i's and compared them to the prior quantiles specified by the prior on τ_i, which was gamma$(v/2, v/2)$. In this comparison, v was set fixed at its posterior mean of about 50. The comparison did not reveal the presence of any outliers (not shown in the figures).

In addition to checking global model fit, we also examined the relative fit of individual data points and clusters (states in this example). The goal of this investigation was to determine if particular individual points or clusters had poor fit relative to others. We used a predictive cross-validation approach in this part of the analysis. We refitted the final model 51 times, leaving out one cluster (state) at a time and using the posterior distribution of the cluster-level pa-

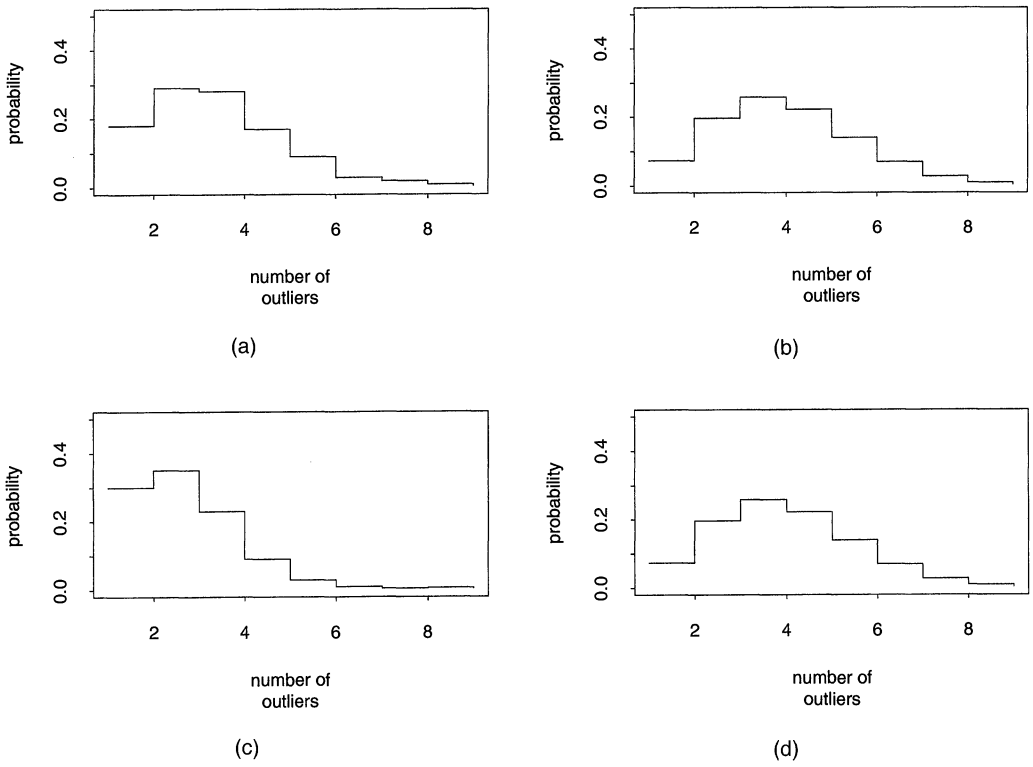

Figure 3. Prior Predictive Plots: (a) and (b) Example I, Observed Under the Model; (c) and (d) Example II, Expected Based on Prior.

rameters given the included clusters to simulate predictive quantities for the removed cluster. An alternative approach would be to use importance sampling and resampling (Rubin 1988).

Using the output from the cross-validation runs, we computed two statistics that are useful in assessing model fit at the individual observation level. First, we computed $p^*_{ij} = P(y^{\text{pred}}_{ij} < y^{\text{obs}}_{ij})$ (Gelfand et al. 1995), which provides information on the degree of consistent overestimation or underestimation of the observed responses. Second, we computed "standardized" residuals, $r^*_{ij} = ((y^{\text{obs}}_{ij} - y^{\text{center}}_{ij})/\sqrt{\text{var}(y^{\text{pred}}_{ij})})$, where y^{center}_{ij} is a measure of the center of the predictive distribution for the ijth observation. We used the mean of the predictive distribution for y_{ij} as the center and the variance as a measure of the spread. These standardized residuals can be used to compare model fit for individual observations. Relative fit for clusters can be checked by computing the Mahalanobis distance measure $(y^{\text{obs}}_i - \bar{y}_i)\mathbf{V}_i^{-1}(y^{\text{obs}}_i - \bar{y}_i)$, where $\mathbf{V}_i$ is the posterior predictive covariance matrix.

An inspection of the Mahalanobis distance measures for each cluster [Fig. 4(a)], the estimates of $p^*_{ij} = P(y^{\text{pred}}_{ij} < y^{\text{obs}}_{ij})$ [Fig. 4(b)], and the standardized residuals for each time point [Fig. 4(c)] points to some states with relatively worse fit than the rest. These states tended to have either very high or low bypass graft rates relative to the overall regional rates and included Kansas in the Midwest, Montana in the West, and Alabama in the South. Overall, the median of the p^*_{ij} was .49, indicating no consistent overestimation or underestimation. Based on examination of the standardized residuals for individual observations, it appears that 95% of the rates were predicted to within two standard deviations of their observed values.

To assess relative fit within clusters, we adapted the approach of Albert and Chib (1997), who suggested using a mixture model to help detect outlying rates within clusters. According to this approach, level II of the binomial HGLM would be replaced by the mixture density $\pi(\theta_{ij}|\boldsymbol{\alpha}_i, \delta_i) = (1 - p)\text{beta}(\delta_i m_{ij}, \delta_i(1 - m_{ij})) + p\,\text{beta}(K^{-1}\delta_i m_{ij}, K^{-1}\delta_i(1 - m_{ij}))$, where p denotes the prior probability of an outlier and K^{-1} denotes the multiplier of the variance for the outlier component of the mixture. Using this mixture density reflects the belief that the jth component in the ith cluster is outlying. Small values of p and $K \geq 2$ would imply that an outlier is unlikely to occur. In this example we set $K = 3$ and $p = .01$.

To assess outliers, define A_{ij} to be the scalar multiple of the variance corresponding to the rate θ_{ij}. A priori, this parameter is equal to 1 with probability $1-p$ and K with probability p. However, if the posterior probability of $A_{ij} = K$ is above some threshold (e.g., 50%), this would indicate that θ_{ij} is an outlying rate for cluster i. A modified Gibbs sampler was used to fit the above model (see Albert and Chib 1997 for details). No within-cluster outliers were detected on the basis of these calculations.

In summary, the final model reported here seems to fit the angiography data quite well. We note that the results were also quite insensitive to the choice of the constant for the prior on the state-specific variance parameter δ_i.

3.3.6 Comparison With Results From Other Methods. The computations for the analyses in this article were car-

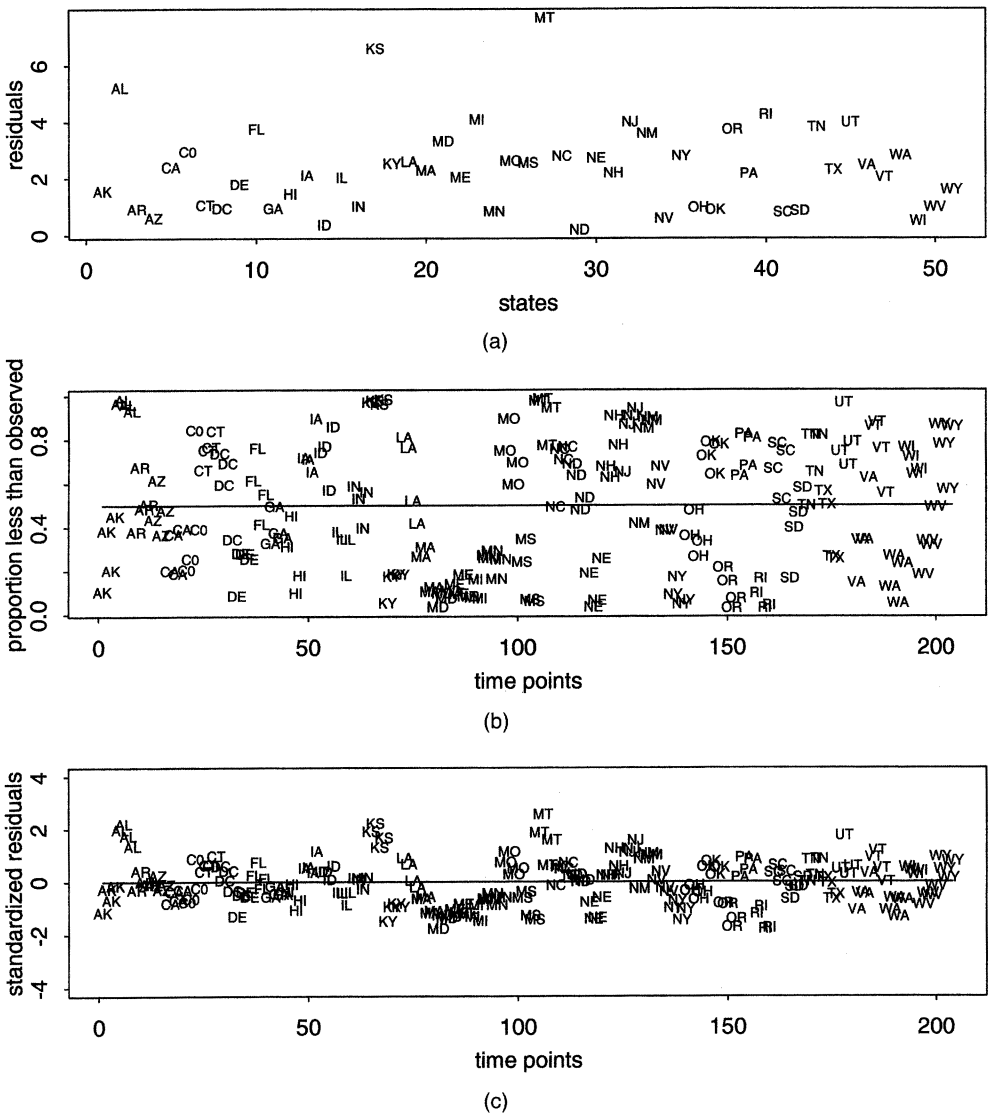

Figure 4. Longitudinal Predictive Plots. (a) Plot of the state specific residuals; (b) plot of P s; (c) standardized residuals for time points.*

ried out using specially developed software written in Fortran. As noted in Section 1, fully Bayesian analyses of some HGLMs can also be implemented using subroutines from BUGS (Spiegelhalter et al. 1996). As the ability of BUGS to handle alternative priors for variances continues to expand, it is likely that it will soon be possible to fit most of the HGLMs discussed in this article using this generally available software. Non-Bayesian subroutines for fitting some classes of HGLMs are available in the software package MLn (Goldstein 1995).

An alternative approach to HGLMs is offered by the class of models known as GLMMs, which can now be estimated using the SAS procedure GLIMMIX (Breslow and Clayton 1993; Khattree and Naik 1995) and the software package HLM (Bryk et al. 1998). In general, the HGLMs allow more flexibility than GLMMs. In particular, HGLMs make it possible to model overdispersion via level II priors, to include mixture distributions to detect potential within-cluster outliers, and to use heavy-tailed distributions for the random effects. GLMM estimates are obtained by maximizing an approximate likelihood such as a penalized likelihood (Lin and Breslow 1996); thus the estimates are not the "true" MLEs of the GLMM. In addition, the HLM software allows only for independent random effects (i.e., a diagonal D matrix). Standard errors and confidence intervals for all estimates can be computationally tedious to obtain. The corresponding computations in the fully Bayesian approach are considerably more streamlined and can provide a more accurate account of the uncertainty in the final estimates.

We used GLIMMIX and HLM to fit a GLMM with a structure similar to the final model of the HGLM analysis. In particular, the GLMM included fixed-effects parameters for regional intercepts and slopes and for angiography availability, and random-effects parameters for the overall intercept and slope. The error distribution was assumed to be binomial, and logit was used as the link function. GLIMMIX and HLM assume a normal distribution on the random effects, which is close to the distribution used in the final HGLM model for this example. For the fixed effects (γ), the estimates and standard errors were quite similar in all three, which is not surprising due to the large overall sample size.

Estimates and standard errors for selected random-effects parameters are shown in Table 1; the standard errors for the GLMM fit with the HLM software are absent from the table as they are not produced as output by the program. The three states were chosen to represent a spectrum of sample sizes, from small (Alaska and Wyoming) to very large (California). The main reason for the differences in the estimates is probably the bias introduced by fitting the GLMM via penalized likelihood and the diagonal $\mathbf{D}$ matrix from fitting the GLMM with HLM; slight differences might be expected from the inclusion of the level II prior in the HGLM. It would be reasonable to expect the standard errors for the GLMM to be generally smaller than those from the HGLM, because of the failure of the GLMM to account for the uncertainty of the covariance matrix $\mathbf{D}$. However, this is not uniformly the case in Table 1, most likely because the GLMM standard errors are obtained by an asymptotic approximation in contrast to HGLM standard errors, which are exact up to Monte Carlo error.

4. EXAMPLE 2: HOSPITAL VARIATIONS IN CARDIAC PROCEDURE UTILIZATION

4.1 The Problem

Numerous studies have documented substantial interhospital variation in the utilization of medical procedures as well as in patient outcomes (see Normand et al. 1997b for relevant references). In our second example we examined differences in the utilization of coronary artery bypass graft surgery (CABG) for elderly heart attack patients treated in hospitals across the country. In addition to geographic variations, we also wanted to examine the effect of hospital characteristics on procedure use. Based on experience from previous studies, we focused our attention to two covariates: hospital size, as measured by the number of beds, and teaching status, classified as "teaching" (major or minor) and "nonteaching."

We limited our attention to data from the 1990 cohort. We excluded hospitals that reported only one patient in this cohort and analyzed the data from the remaining 4,992 hospitals in the 50 states plus the District of Columbia. Most of the hospitals were in the nonteaching category (84%), and only 3% were major teaching hospitals. The median hospital size was 122 beds. The 214,478 patients in our study were distributed unevenly across hospitals. The number of patients ranged from 2 to 381 (interquartile range 11–58). For each hospital, the rate of CABG was computed as the proportion of patients in our cohort who underwent CABG within 90 days from their AMI admission.

Table 1. Comparison of Estimates of Regional-Level Intercepts Using GLIMMIX, HLM, and HGLM (With $\hat{a}_i = \hat{\alpha}_i - G_i^T \hat{\gamma}$)

State	Parameter	GLIMMIX	HLM	HGLM
Alaska	Intercept ($\hat{a}_{1,1}$)	−.11 (.13)	−.20	−.15 (.14)
	Slope ($\hat{a}_{1,2}$)	.01 (.02)	.02	.01 (.03)
California	Intercept ($\hat{a}_{5,1}$)	−.12 (.06)	−.14	−.10 (.05)
	Slope ($\hat{a}_{5,2}$)	.02 (.02)	.03	.02 (.02)
Wyoming	Intercept ($\hat{a}_{51,1}$)	.12 (.10)	.16	.15 (.10)
	Slope ($\hat{a}_{51,2}$)	−.00 (.02)	−.01	−.01 (.03)

The role of patient case-mix as a potential confounder in hospital comparisons is well established in health services research. Based on experience from a variety of studies, it would be reasonable to expect that large teaching hospitals treat patients with more severe illness and thus may have higher rates of cardiac procedures. As a partial adjustment for patient case-mix differences, we created a patient-level comorbidity index (similar to Gatsonis et al. 1995) and averaged over patients by hospital to obtain a hospital-level comorbidity index. The patient-level comorbidity index was created by fitting a patient-level logistic regression model in which the dependent variable was an indicator of whether the patient received CABG within 90 days, and independent variables were age, gender, race, and indicator variables for each comorbid condition with a prevalence of at least .5% in the sample. The value of the index for a patient was the predicted log-odds of CABG based on the patient's characteristics.

4.2 Models

A Poisson assumption for the distribution of the hospital counts of CABG seemed reasonable, because the rates were generally low. As in Example 1, an exploratory analysis using GLMs gave evidence of variation larger than the standard Poisson variance function. A log-transformation of hospital size exhibited a fairly linear relationship with the logarithm of CABG rates. On the basis of these findings, we considered the following model:

- Level I (variation of observed hospital counts): Let Y_{ij} denote the number of CABG operations performed in the jth hospital ($j = 1, \ldots, J_i$) of the ith state ($i = 1, \ldots, 51$). We assume a Poisson distribution for Y_{ij} with mean $(e_{ij}\theta_{ij})$, where θ_{ij} denotes the CABG rate and e_{ij} denotes the corresponding number of cohort members (exposure) in the ijth hospital.
- Level II (variation within states): We considered three specifications for the gamma prior: (A) $\theta_{ij}|\boldsymbol{\alpha}_i, \delta_i \sim \Gamma(\delta_i e_{ij} m_{ij}(\boldsymbol{\alpha}_i), \delta_i e_{ij})$, (B) $\theta_{ij}|\boldsymbol{\alpha}_i, \delta_i \sim \Gamma(\delta_i m_{ij}(\boldsymbol{\alpha}_i), \delta_i)$, and (C) $\theta_{ij}|\boldsymbol{\alpha}_i, \delta_i \sim \Gamma(\delta_i, \delta_i/m_{ij}(\boldsymbol{\alpha}_i))$. In prior specifications A and B, the mean m_{ij} of the CABG rates θ_{ij} is proportional to the standard deviation. In specification C, the mean is proportional to the variance (Christiansen and Morris 1997; Wolfe and Becker 1994). We chose the gamma prior both for its computational attractiveness (conjugate prior for the Poisson distribution) and for the added flexibility in modeling overdispersion with a closed form distribution. The marginal distribution of the data, conditional on $\boldsymbol{\alpha}_i$ and δ_i (the state-specific parameters), is negative binomial in these models. We used prior A for our computations, with $\log m_{ij} = (1, \log(\text{size}), \text{teaching status}, \text{comorbidity.index})\ (\boldsymbol{\alpha}_{i0}, \boldsymbol{\alpha}_{i1}, \boldsymbol{\alpha}_{i2}, \boldsymbol{\alpha}_{i3})'$.

  The prior distribution of the variance parameter δ_i had the form $\delta_i \sim \pi(\delta_i) = e_{0i}/(e_{0i} + \delta_i)^2$. In computations reported here, we set the constant e_{0i} equal to 1. This specification assumes that the variance of the observed counts in hospitals in a given state, conditional on the state-specific parameters, is equal to a constant scale factor multiplied by the standard Poisson variance.

- Level III (variation between states): The state-specific coefficients α_i were linked to the state-level covariates by a model similar to the one in Section 3.2, with the parameters γ representing the effect of state characteristics. The latter were now limited to indicators of Census region.

4.3 Analysis

4.3.1 Gibbs Sampler. The structure and convergence diagnostics of the Gibbs sampler were similar to those reported in Example 1 and will not be elaborated here. The posterior distribution for the degrees of freedom at level III was also similar. As in Example 1, we report here posterior sample means and 95% credible intervals constructed in the basis of the 2.5% and 97.5% quantiles of the sample from the posterior distribution.

4.3.2 Model Selection. The final model in this analysis was somewhat more simplified than the full model described in Section 4.2. We describe the process of model selection before proceeding to the presentation of the subject matter results.

An examination of the 95% credible intervals for the components of γ, measuring regional effects, showed no significant regional differences in the effect of teaching status and size on hospital rates of CABG. In addition, the Bayes factor for testing whether regional differences existed in the effect of both hospital size and teaching status was 1.5. On the basis of this evidence, the final model did not include regional effects for teaching status and hospital size. Thus the vector γ was of length 10 in the final model, with the first four components corresponding to the regional effects on the state-specific intercepts, the fifth and sixth components corresponding to the contrywide mean effect of size and teaching status, and the remaining four components corresponding to regional differences in the effect of comorbidity.

The p value for the discrepancy statistic was close to 1.0, indicating that the model with overdispersion predicted considerably more variation than was present in the data. However, when we fit the model without the level II prior, we obtained a p value of 0, indicating that the reduced model predicted substantial less variation than was present in the data. The extreme p values for both models suggest that a more appropriate model might allow for overdispersion in some states and not in others. The computational burden of carrying out the analysis with the more complex model is substantial, and we did not pursue it for this example. To assess whether it is necessary to model extra variability for a particular state, one could compute appropriate Bayes factors, use the discrepancy statistic for that state, or compute the posterior probability that δ_i is greater than some critical value.

4.3.3 Goodness of Fit of Final Model. Figure 3, (c) and (d), shows the observed and expected probability mass functions (pmf's) from the Weiss statistic for the state-specific effects. The observed pmf seems to be underdispersed with fewer outliers than expected, indicating that the distribution the state-specific effects may have heavier tails than needed.

Figure 5 compares the posterior distribution of the scale parameters τ_i to their prior distribution. The horizontal lines correspond to the 1%, 5%, 25%, 75%, 95%, and 99% quantiles of the prior distribution, whereas the plotted points represent the observed [5%, 25%, 75%, 95%] quantiles from the posterior distribution of the τ_i. For only two of the 51 states does the 95% quantile of the posterior exceed the 95% quantile of the prior, indicating a reasonable model fit. The discrepancy is largest in Maryland. (A vertical line goes through the percentiles for Maryland in Figure 5.)

To assess relative fit of the various clusters (states), we used predictive cross-validation, by dropping out groups of states at a time instead of a single state, to speed up the computation. The median for p^* was .50, indicating no consistent overestimation or underestimation. The standardized residuals were computed as in Example 1; 99% of the hospitals had residuals below 2. Large residuals corresponded to extreme p^* values, and hospitals with the largest residuals tended to have high ($> .2$) CABG rates. Such hospitals represented about 8% of the total number of hospitals in the database.

Finally, using the mixture models suggested by Albert and Chib (1997), we determined that several states had within-cluster outliers. To better account for these outliers, our final model included the mixture distribution for level II for each state as discussed in the goodness-of-fit section for Example I (Sec. 3.3.5).

4.3.4 Variations by State and Hospital Type. Hospital size was an important predictor of CABG utilization. If the effect of hospital size is measured as the percentage increase in CABG rate corresponding to a unit increase in the logarithm of hospital size, the national mean effect of hospital size, $\hat{\gamma}_5$, was 13% (9%, 16%) (95% credible interval). The median state-specific effect of hospital size was also 13%. The greatest effect of hospital size on CABG rate occurred in Virginia, with a rate increase of 19% (7%, 31%) and the smallest effect in Florida, with a rate increase of 9% (0%, 17%).

The evidence on the effect of teaching status was less conclusive. The estimate of $\hat{\gamma}_6$, the national mean effect of teaching status, was 3% $(-4\%, 11\%)$, indicating an insignificant increase in CABG rates between teaching and nonteaching hospitals. The median state-specific effect of teaching status corresponds to a 5% increase in the rate of CABG. However, there was large variation across states in the effect of teaching status. For example, in Indiana the estimated effect of teaching status was a 40% $(-1\%, 98\%)$ increase in CABG rates in teaching over nonteaching hospitals. In Mississippi the estimated effect was a 19% $(-35\%, -2\%)$ decrease in CABG rates in teaching over nonteaching hospitals.

Nationwide, the highest rates of CABG (procedures per patient in this cohort) tended to occur in larger teaching hospitals. The estimated rates of CABG for an average size (approximately 196 beds) nonteaching hospital ranged from a maximum of .12 (.11, .14) in Arkansas to a minimum

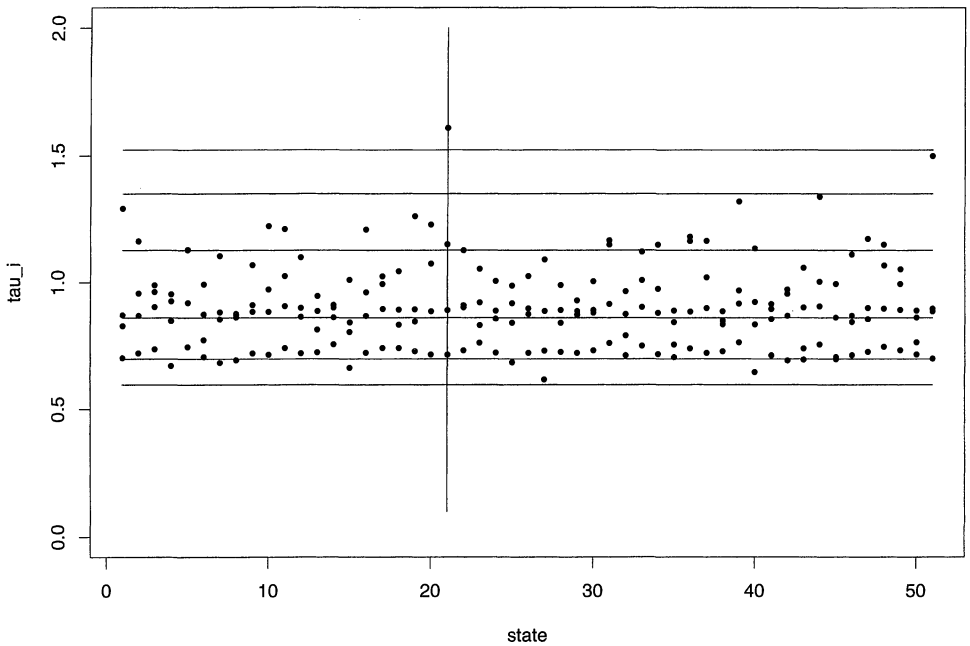

Figure 5. Plot Comparing Prior Distribution for τ_i to the Posterior Distribution.

of .07 (.05, .09) in Rhode Island. The nationwide median rate of CABG for this class of hospitals was .09. The corresponding state-specific median rate for large (270 beds) teaching hospitals was .10 with a maximum of .14 (.10, .19) in Utah and a minimum of .08 (.07, .10) in Indiana.

4.3.5 Regional Variations. The Northeastern region of the country tended to have the lowest CABG rates. For example, the estimated rate of CABG for an average size nonteaching hospital at average comorbidity was highest in the West and Midwest, both with rates of .10 (.09, .11); followed by the South, with a rate of .09 (.08, .10); and then the Northeast, the lowest, with a rate of .08 (.07, .09).

4.3.6 Discussion of CABG Analysis. Studies of hospital quality of care have often reported higher quality in larger than smaller, teaching than non-teaching, and urban than rural hospitals (see, e.g., Keeler et al. 1992). These findings are consonant with the results of our analysis, which suggest that adjusted CABG rates for AMI patients are higher for those admitted to larger hospitals and, to a lesser degree, for those admitted to teaching hospitals. We note that because teaching hospitals tended to be larger than nonteaching hospitals (median number of beds was 400 in teaching hospitals, compared to 100 in nonteaching hospitals), it is difficult to separate the effects of hospital size and teaching status.

Interstate and regional variation in CABG use was substantial, even after hospital characteristics were taken into account. For example, in the Northeast (a low-use region) the median state-specific rate for large (270 beds) teaching hospitals was .10, whereas in the West (a high-use region) the median state-specific rate for small (50 beds) teaching hospitals was very similar, .09. The geographic pattern for the utilization of CABG was similar to that for coronary angiography, suggesting that medical care for elderly AMI patients was generally less aggressive in the Northeast than in other parts of the country. The presence of such geographic and hospital variations raises the question of whether patient outcomes show a corresponding trend. However, such an investigation is well beyond the scope of this example.

The analysis for this example used patient data aggregated at the hospital level. However, as in the first example, a more detailed analysis would proceed by modeling the individual patient response (CABG, yes or no) as a function of patient characteristics, using binary regression. The binary coefficients could then be allowed to vary by hospital and state in the upper levels of the hierarchical model. In a further elaboration, the response on an individual patient could be multinomial, indicating also the type of revascularization procedure (CABG or angioplasty). A hierarchical polytomous regression analysis of interstate variations, without reference to hospital effects, has been reported by Daniels and Gatsonis (1997).

5. DISCUSSION

The HGLM framework provides a broad and flexible class of models for multilevel clustered data. We have focused our attention on four-level models with linear structures for the regression models and with multivariate t error distributions for the higher levels of the hierarchy. These models can be extended in several directions:

a. Additional levels can be added to the hierarchical structure in order to model more complex clustering. For example, patients may be clustered first by physician or clinic and then by hospital and geographic region or hospital system.

b. Nonlinear regression models can be incorporated into the hierarchical prior structure.

c. The assumption of constant variance across clusters can be relaxed to accommodate situations in which the

variance may depend on cluster characteristics. In the case of the longitudinal analysis, for example, allowing the covariance matrix for the state-specific parameters to depend on covariates, such as region, might help the model to account for some of the more extreme observations.

d. A nonconjugate prior distribution can be chosen for level II to allow for a wider variety of variance functions.
e. A mixture of normals for the random coefficient distribution can be used to account for potential multimodal distributions (Escobar and West 1995).
f. The prior at level II can specify a point mass for some clusters (no overdispersion) and the usual conjugate prior for other clusters as discussed in Example II. The general computational approach via MCMC can be modified to accommodate the more complex models resulting from any of these possible extensions.

The class of HGLMs discussed in this article provide the user with numerous capabilities not available in similar classes of models, notably GLMMs and the HGLMs recently discussed by Lee and Nelder (1996), but still lack the computational simplicity of these approaches. However, the continual development of the BUGS software should allow fitting the full HGLMs in the near future in a standard package without the need for specialized software. Finally, further exploration of the computation of Bayes factors in the HGLM setting is needed, especially Bayes factors to test for heterogeneity ($\mathbf{D} = 0$) and to compare nonnested models, such as those arising in the selection of alternative level II priors.

[Received June 1996. Revised June 1998.]

REFERENCES

Albert, J. H. (1988), "Bayesian Estimation of Poisson Means Using a Hierarchical Log-Linear Model," in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Oxford, U.K.: Clarendon Press, pp. 519–531.

Albert, J. H., and Chib, S. (1996), "Bayesian Tests and Model Diagnostics in Conditionally Independent Hierarchical Models," *Journal of the American Statistical Association*, 92, 916–925.

Bennet, J. E., Racine-Poon, A., and Wakefield, J. C. (1996), "MCMC for Nonlinear Hierarchical Models," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 339–358.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), "Bayesian Computation and Stochastic Systems," *Statistical Science*, 10, 3–41.

Breslow, N., and Clayton, D. (1993), "Approximate Inference in Generalized Linear Models," *Journal of the American Statistical Association*, 88, 9–25.

Bryk, A. S., and Raudenbush, S. W. (1992), *Hierarchical Linear Models: Application and Data Analysis Methods*, Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., and Congdon, R. T. (1998), *Hierarchical Linear Modeling*, Version 4.03, Chicago, IL: Scientific Software International.

Calvin, J. A., and Sedransk, J. (1991), "Bayesian and Frequentist Predictive Inference for the Patterns of Care Studies," *Journal of the American Statistical Association*, 86, 36–48.

Christiansen, C. L., and Morris, C. N. (1997), "Hierarchical Poisson Regression Modeling," *Journal of the American Statistical Association*, 92, 618–632.

Cowles, M. K., and Carlin, B. P. (1996), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, 91, 883–904.

Daniels, M. J., and Gatsonis, C. G. (1997), "Hierarchical Polytomous Regression Models With Applications to Health Services Research," *Statistics in Medicine*, 16, 2311–2326.

Dartmouth Medical School, Center for the Evaluative Clinical Sciences (1996), *The Dartmouth Atlas of Health Care in the United States*, Chicago: American Hospital Publishing.

DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997), "Computing Bayes Factors by Combining Simulation and Asymptotic Approximations," *Journal of the American Statistical Association*, 92, 903–915.

Diehr, P. (1984), "Small Area Statistics: Large Statistical Problems," *American Journal of Public Health*, 74, 313–314.

Diehr, P., Cain, K., Connell, F., and Volinn, E. (1990), "What Is Too Much Variation? The Null Hypothesis in Small Area Analysis," *Health Services Research*, 24, 741–771.

Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Gatsonis, C., Epstein, A., Newhouse, J., Normand, S., and McNeil, B. (1995), "Variations in the Utilization of Coronary Angiography for Elderly Patients With an 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.

Gelfand, A. E., Dey, D. K., and Chang, H. (1995), "Model Determination Using Predictive Distributions With Implementation via Sampling-Based Methods," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. David, and A. F. M. Smith, Oxford, U.K.: Clarendon Press, pp. 147–167.

Gelfand, A. E., and Ghosh, S. K. (1998), "Model Choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, 1–12.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995), "Efficient Parameterizations for Normal Linear Mixed Models," *Biometrika*, 82, 479–488.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., Meng, X-L., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies," *Statistica Sinica*, 6, 733–807.

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

Ghosh, M., Natarajan, K., Stroud, T. W. F., and Carlin, B. P. (1998), "Generalized Linear Models for Small-Area Estimation," *Journal of the American Statistical Association*.

Gilks, W. R., Wang, C. C., Yvonnet, B., and Coursaget, P. (1993), "Random Effect Models for Longitudinal Data Using Gibbs Sampling," *Biometrics*, 49, 441–454.

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

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

Jacquier, E., Polson, N., and Rossi, P. (1994), "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business and Economic Statistics*, 12, 371–389.

Kahn, M. J., and Raftery, A. E. (1996), "Discharge Rates of Medicare Stroke Patients to Skilled Nursing Facilities: Bayesian Logistic Regression With Unobserved Heterogeneity," *Journal of the American Statistical Association*, 91, 29–41.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kass, R. E., and Steffey, D. (1989), "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)," *Journal of the American Statistical Association*, 84, 717–726.

Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

Keeler, E., Rubestein, L. Kahn, K., Draper, D., Harrison, E., McGinty, M., Rogers, W., and Brook, R. (1992), "Hospital Characteristics and Quality of Care," *Journal of the American Medical Association*, 268, 1709–1714.

Khattree, R., and Naik, D. N. (1995), *Applied Multivariate Statistics With*

SAS Software, Cary, NC: SAS Institute, Inc.

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

Laud, P. W., and Ibrahim, J. G. (1995), "Predictive Model Selection," *Journal of the Royal Statistical Society*, Ser. B, 57, 247–262.

Lee, Y., and Nelder, J. A. (1996), "Hierarchical Generalized Linear Models" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 58, 619–678.

Lin, X., and Breslow, N. E. (1996), "Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion," *Journal of the American Statistical Association*, 91, 1007–1016.

Lindstrom, M. J., and Bates, D. M. (1988), "Newton–Raphson and E-M Algorithms for Linear Mixed Effect Models for Repeated Measures Data," *Journal of the American Statistical Association*, 83, 1014–1022.

Liu, C. (1996), "Bayesian Robust Multivariate Linear Regression With Incomplete Data," *Journal of the American Statistical Association*, 91, 1219–1227.

Longford, N. T. (1987), "A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models With Nested Random Effects," *Biometrika*, 74, 817–827.

Malec, D., Sedransk, J., Moriarity, C. L., and LeClere, F. B. (1997), "Small Area Inference for Binary Variables in the National Health Interview Survey," *Journal of the American Statistical Association*, 92, 815–826.

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.

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models*, London: Chapman and Hall.

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

Natarajan, K., Ghosh, M., and Maiti, T. (1998), "Hierarchical Bayes Quality Measurement Plan," unpublished manuscript submitted to *Communications in Statistics: Simulation and Computation.*

Normand, S. L., Glickman, M., and Gatsonis, C. (1997a), "Statistical Methods for Profiling Providers: Issues and Applications," *Journal of the American Statistical Association*, 92, 803–814.

Normand, S. L., Glickman, M. E., and Ryan, T. (1997b), "Modeling Mortality Rates for Elderly Heart Attack Patients: Profiling Hospitals in the Cooperative Cardiovascular Project," in *Case Studies in Bayesian Statistics*, eds: C. Gatsonis, J. Hodges, R. Kass, and N. Singpurwalla, New York: Springer-Verlag, pp. 155–236.

O'Hagan, A. (1994), *Kendall's Advanced Theory of Statistics, Vol. 2B: Bayesian Inference*, New York: Wiley.

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, 1832–1837.

Paul-Shaheen, P., Clark, J. D., and Williams, D. (1987), "Small Area Analysis: A Review and Analysis of the North American Literature," *Journal of Health Politics, Policy and Law*, 14, 741–809.

Rubin, D. B. (1988), "Using the SIR Algorithm to Simulate Posterior Distributions," in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Oxford, U.K.: Clarendon Press, pp. 395–402.

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Ser. B, 55, 3–23.

Spiegelhalter, D. J., Best, N. G., Gilks, W. R., and Inskip, H. (1996), "Hepatitis B: A Case Study in MCMC Methods," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 21–44.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1762.

Udvarhelyi, I. S., Gatsonis, C., Epstein, A., Pashos, C., Newhouse, J. P., and McNeil, B. J. (1992), "Acute Myocardial Infarction in the Medicare Population: Process of Care and Clinical Outcomes," *Journal of the American Medical Association*, 18, 2530–2536.

Verdinelli, I., and Wasserman, L. (1995), "Computing Bayes Factors Using the Savage–Dickey Density Ratio," *Journal of the American Statistical Association*, 90, 614–618.

Weiss, R. E. (1996), "Bayesian Model Checking with Applications to Hierarchical Models," technical report, UCLA School of Public Health, Dept. of Biostatistics.

Wennberg, J., and Gittelsohn, A. (1982), "Variations in Medical Care Among Small Areas," *Scientific American*, 246, 120–134.

Wolfe, R., and Becker, M. (1994), "Modeling Utilization Rates," technical report, University of Michigan, Dept. of Biostatistics.

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

Zeger, S., Liang, K., and Albert, P. (1988), "Models for Longitudinal Data: A Generalized Estimating Equations Approach," *Biometrics*, 44, 1049–1066.