Statistical Science
2016, Vol. 31, No. 1, 80–95
DOI: 10.1214/15-STS533

Mitigating Bias in Generalized Linear Mixed Models: The Case for Bayesian Nonparametrics

Joseph Antonelli, Lorenzo Trippa and Sebastien Haneuse

Abstract. Generalized linear mixed models are a common statistical tool for the analysis of clustered or longitudinal data where correlation is accounted for through cluster-specific random effects. In practice, the distribution of the random effects is typically taken to be a Normal distribution, although if this does not hold then the model is misspecified and standard estimation/inference may be invalid. An alternative is to perform a so-called nonparametric Bayesian analyses in which one assigns a Dirichlet process (DP) prior to the unknown distribution of the random effects. In this paper we examine operating characteristics for estimation of fixed effects and random effects based on such an analysis under a range of "true" random effects distributions. As part of this we investigate various approaches for selection of the precision parameter of the DP prior. In addition, we illustrate the use of the methods with an analysis of post-operative complications among $n = 18{,}643$ female Medicare beneficiaries who underwent a hysterectomy procedure at $N = 503$ hospitals in the US. Overall, we conclude that using the DP prior in modeling the random effect distribution results in large reductions of bias with little loss of efficiency. While no single choice for the precision parameter will be optimal in all settings, certain strategies such as importance sampling or empirical Bayes can be used to obtain reasonable results in a broad range of data scenarios.

Key words and phrases: Dirichlet process prior, generalized linear mixed models, model misspecification, random effects.

1. INTRODUCTION

When performing regression analyses of clustered or longitudinal data, analysts must account for potential within-cluster correlation to ensure valid inference (McCulloch, 2006; Diggle et al., 2013). For the most part, regression analysis methods for correlated data fall into one of two classes. Marginal methods, such as generalized estimating equations (GEE), account for correlation by using a sandwich estimator for standard errors (Liang and Zeger, 1986). In contrast, mixed effects models add cluster-specific latent terms, referred to as random effects or frailties, to the linear predictor; while within-cluster outcomes are assumed to be conditionally independent given the random effect, correlation is induced marginally (Laird and Ware, 1982). To avoid the curse of dimensionality, where the number of model components increases with the number of clusters, structure is typically placed on the random effects across the population of clusters. In practice, the

Joseph Antonelli is Postdoctoral Fellow, Deparment of Biostatistics, Harvard Chan School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, USA (e-mail: jantonelli@fas.harvard.edu). Lorenzo Trippa is Assistant Professor, Department of Biostatistics, Dana-Farber Cancer Institute, Center for Life Science, 3 Blackfan Circle, Boston, Massachusetts 02115, USA (e-mail: ltrippa@jimmy.harvard.edu). Sebastien Haneuse is Associate Professor, Department of Biostatistics, Harvard Chan School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, USA (e-mail: shaneuse@hsph.harvard.edu).

most common structure adopted is that the random effects are Normally distributed. Regardless of the specific structure adopted, estimation/inference for mixed models is typically likelihood-based; frequentist analyses maximize the induced marginal likelihood, while Bayesian analyses incorporate assumptions on the random effects into a hierarchical model.

That structure is imposed on the random effects is an often-cited drawback of mixed effects models. While methods for model diagnostics have been developed (Verbeke and Molenberghs, 2009; Lange and Ryan, 1989), any given assumption cannot be directly verified and Normality is often criticized as being potentially unrealistic. Furthermore, if the random effects are not Normally distributed and yet are assumed to be, the overarching model is misspecified and likelihood-based estimation is not guaranteed to be consistent (White, 1982). The latter has given rise to a contentious debate literature on the impact, particularly in terms of bias, of misspecification in mixed effects models (Neuhaus, Hauck and Kalbfleisch, 1992; Heagerty and Kurland, 2001; Agresti, Caffo and Ohman-Strickland, 2004; Litière, Alonso and Molenberghs, 2007; Neuhaus, McCulloch and Boylan, 2011; McCulloch and Neuhaus, 2011a, 2011b). This potential for bias has also motivated the development of methods to permit more flexible specifications. Much of the corresponding frequentist literature can be placed into one of two broad categories: methods that estimate/specify the random effects distribution nonparametrically or smoothly (Laird, 1978; Davidian and Gallant, 1993; Zhang and Davidian, 2001; Agresti, Caffo and Ohman-Strickland, 2004) and methods based on flexible families of parametric distributions (Magder and Zeger, 1996; Piepho and McCulloch, 2004; Caffo, An and Rohde, 2007).

The last 20 years has also seen substantial progress in the theoretical development and understanding of so-called Bayesian nonparametric analysis (Dey, Müller and Sinha, 1998; Müller and Quintana, 2004). Furthermore, with recent advances in computing power, and the development and implementation of efficient Markov Chain Monte Carlo algorithms, Bayesian nonparametric methods are becoming more and more practical for everyday analyses (Jara et al., 2011). To the best of our knowledge, while Bayesian nonparametric priors are specifically motivated by the desire to avoid overly restrictive assumptions (Walker and Mallick, 1997), the extent to which misspecification bias in mixed effects models is mitigated by their use has not been examined. In particular, we are unaware of systematic attempts at quantifying operating characteristics (in particular, bias and efficiency) when one uses a nonparametric Bayesian prior, instead of the usual Normal distribution, under various true distributions for the random effects. A crucial question in this context is the extent to which Bayesian analyses using nonparametric priors experience a bias-variance tradeoff: if the truth is that the random effects are indeed Normally distributed, what is the loss of efficiency (if there is one at all) compared to an analysis that correctly adopted a Normal distribution?

In this paper we build on the misspecification literature by considering the use of the Dirichlet process prior as an alternative default choice for the random effects distribution in a logistic generalized linear mixed model (GLMM) for repeated measures binary data. As we elaborate upon, the DP prior is indexed by two parameters: a centering distribution, denoted by G_0, and a precision parameter, denoted by α. In practice, while G_0 is often taken to be a Normal distribution, the specification of α is more challenging. A number of methods have been developed and, toward providing practical guidance to researchers, we investigate the impact of this choice on the operating characteristics. The remainder of this paper is as follows. In Section 2 we provide notation on GLMMs and briefly introduce the DP prior as a specification for the unknown random effects distribution. In Section 3 we review methods for the treatment of α in practice. Sections 4 and 5 provide the main simulation study and a summary of the results, while Section 6 provides a detailed analysis of post-operative complications among $n = 18{,}643$ female Medicare beneficials who underwent a hysterectomy procedure at $N = 503$ hospitals in the US. The paper concludes with a discussion in Section 7.

2. MIXED MODELS FOR BINARY RESPONSE DATA

The emphasis of this paper is on the analysis of clustered, correlated or longitudinal data using generalized linear mixed models. While this class of models is broad, we focus on models for clustered binary response data. Toward this, suppose the observed data consist of N clusters. Let n_i denote the size of the ith cluster and Y_{ij} the binary response of interest for the jth study unit in the ith cluster. Furthermore, let $\mathbf{X}$ and $\mathbf{Z}$ denote mutually exclusive vectors of covariates of lengths p and q, respectively, and consider the model

$$\begin{aligned} Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, b_i &\sim \text{Bernoulli}(\mu_{ij}), \\ g(\mu_{ij}) &= \mathbf{X}_{ij}^T\beta + \mathbf{Z}_{ij}^T b_i, \qquad (1) \\ b_i &\overset{\text{i.i.d.}}{\sim} G, \end{aligned}$$

where $\mu_{ij} = \mathrm{E}[Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, b_i]$, $g(\cdot)$ is a link function, β is a vector of so-called fixed effect regression coefficients and b_i is a vector of so-called random effects. The distribution G characterizes how the b_i vectors vary across the population of clusters and is referred to as the random effects or mixing distribution. In practice, depending on the data application and scientific question, the random effect can take on quite complex structures. In this paper we focus on two relatively common structures. Specifically, in the remainder of this section and in the simulation study of Section 4 we consider the so-called *random intercept* model (e.g., Diggle et al., 2013). In the simulation study we additionally consider the so-called *random intercept/slope* model, commonly used in the analysis of longitudinal data.

2.1 A Normal Mixing Distribution

Generally, the distribution G is not known. One can, however, impose the assumption that G belongs to some family of distributions and determine the specific member by jointly estimating the parameters that index that family along with the remaining unknown parameters in the model. In practice, and in most software implementations, the default choice for this family is the family of Normal distributions. This gives the logistic-Normal model:

$$\begin{aligned} Y_{ij}|\mathbf{X}_{ij}, b_i &\sim \mathrm{Bernoulli}(\mu_{ij}), \\ \mathrm{logit}(\mu_{ij}) &= \mathbf{X}_{ij}^T\beta + b_i, \\ b_i|\tau^2 &\overset{\text{i.i.d.}}{\sim} G \equiv \mathrm{Normal}(0, \tau^2). \end{aligned}$$

2.2 A Dirichlet Process Mixing Distribution

That variation in the b_i across clusters under a logistic-Normal model is characterized by some specific Normal distribution is often viewed as a strong assumption. In particular, since estimation/inference for GLMMs is usually likelihood based, if the true mixing distribution is not Normal, the adoption of a Normal distribution corresponds to a misspecification of the likelihood with resulting estimates no longer guaranteed to be consistent. To relax this assumption, one could consider a broader class of potential distributions for the unknown G. Within the Bayesian paradigm this is operationalized by postulating a prior distribution for G that (i) specifies the space of distributions that G can take on and (ii) specifies a prior for the distributions in the selected space. While numerous such priors have been proposed in the literature, we consider the Dirichlet process (DP) prior which is indexed by two hyperparameters: G_0 and $\alpha > 0$, referred to as the centering distribution and precision parameter, respectively (Antoniak, 1974; Walker and Mallick, 1997). The corresponding GLMM, referred to here as the logistic-Dirichlet process (logistic-DP) mixing model, can be written as

$$\begin{aligned} Y_{ij}|\mathbf{X}_{ij}, b_i &\sim \mathrm{Bernoulli}(\mu_{ij}), \\ \mathrm{logit}(\mu_{ij}) &= \mathbf{X}_{ij}^T\beta + b_i, \\ b_i|G &\overset{\text{i.i.d.}}{\sim} G, \\ G &\sim \mathrm{DP}(G_0, \alpha). \end{aligned}$$

This model is well established in the literature (Kleinman and Ibrahim, 1998), and we provide details of its specification here. Intuitively, G_0 can be thought of as an a priori "best guess" for G; a common choice is to take G_0 to be a Normal distribution. The precision parameter then controls the extent to which distributions in the space defined by the DP prior differ from G_0. Toward a more intuitive understanding of α, the DP prior literature has often considered the predictive distribution for $\mathbf{b} = (b_1, \dots, b_N)$:

$$\pi(\mathbf{b}|G_0, \alpha) = \int_G \left(\prod_{i=1}^N \pi(b_i|G) \right) \pi(G|G_0, \alpha)\, dG.$$

It is relatively straightforward to show that this joint distribution can be decomposed as

$$\begin{aligned} \pi(\mathbf{b}|G_0, \alpha) = \pi(b_1|G_0, \alpha) \times \pi(b_2|G_0, \alpha, b_1) \times \cdots \\ \cdot \pi(b_N|G_0, \alpha, b_1, \dots, b_{N-1}), \end{aligned}$$

where $b_1|G_0, \alpha \sim G_0$ and the successive conditional distributions are

$$\begin{aligned} &b_i|G_0, \alpha, b_1, \dots, b_{i-1} \\ &\quad\sim \frac{1}{i-1+\alpha} \sum_{k=1}^{i-1} \delta(b_k) + \frac{\alpha}{i-1+\alpha} G_0 \end{aligned} \tag{2}$$

with $\delta(b_k)$ denoting a point mass at b_k. That is, the predictive distribution for b_i is a mixture of the empirical distribution of $(b_1, \dots, b_{i-1})$ and G_0 where the weight is a function of $i - 1$, the number of realizations being conditioned upon, and α. To understand the impact of α, it is useful to consider this expression at two extremes. At one extreme, as $\alpha \longrightarrow \infty$ the joint distribution of $\mathbf{b}$ tends to the product of N independent draws from G_0; if G_0 is taken to be a Normal$(0, \tau^2)$ distribution, then the model reduces to the logistic-Normal model. At the other extreme, as $\alpha \longrightarrow 0$ the joint distribution of $\mathbf{b}$ tends to a point mass with $b_i = b_1$, $i = 2, \dots,$ where b_1 is a random draw from G_0. This common value of b_i can then be absorbed into the (fixed effect) intercept, so that the model reduces to a standard logistic regression without random effects.

The two extremes for α correspond to each cluster having a unique value of b_i from G_0 and all clusters sharing the same value of b_1 (also from G_0). Between these extremes the joint distribution of $\mathbf{b}$ induced by a DP generates a random number of unique values of b. The latter, denoted by K and ranging between 1 and N, is sometimes referred to as the number of clusters generated by the DP (not to be confused with the number of clusters in the observed data, denoted by N). Since K is random, its expectation and variance conditional on the DP and N can be considered and are given by

$$\text{(3)} \qquad \mathrm{E}[\mathrm{K}|G_0, \alpha, N] = \sum_{i=1}^{N} \frac{\alpha}{\alpha + i - 1},$$

$$\text{(4)} \qquad \mathrm{V}[\mathrm{K}|G_0, \alpha, N] = \sum_{i=1}^{N} \frac{\alpha(i-1)}{(\alpha + i - 1)^2}.$$

3. SELECTION OF α

In practice, as mentioned, G_0 is typically taken to be a Normal(0, τ^2) distribution. This is, arguably, a reasonable choice in the sense that fitting a logistic-Normal model is often the default anyway. Unfortunately, because it does not have an intuitive stand-alone interpretation, the choice of α is less clear cut. Furthermore, estimation/inference can be quite sensitive to the specific choice (Escobar, 1994; Dorazio et al., 2008; Dorazio, 2009), particularly important for this paper since our interests lie with the extent to which a DP mixing distribution mitigates bias when the true mixing distribution is not Normal.

A number of strategies for α have been proposed in the literature. Here we provide a synthesis of this literature in the form of two general strategies. The first places a prior on α, while the second estimates α.

3.1 Adopting a Prior for α

When performing Bayesian estimation/inference for a logistic-DP model, a natural approach to the treatment of α is to adopt some prior distribution. A standard choice is a Gamma(ψ_1, ψ_2) distribution, since the full posterior conditional for α is straightforward to sample from (Escobar and West, 1995; Neal, 2000). The choice of specific values for the (ψ_1, ψ_2) hyperparameters can proceed in one of several ways.

After selecting a priori a mean and variance for α one could choose specific values of (ψ_1, ψ_2) based on the moments of a Gamma distribution. That is, make use of the fact that, under this prior, $\mathrm{E}[\alpha|\psi_1, \psi_2] = \psi_1/\psi_2$ and $\mathrm{V}[\alpha|\psi_1, \psi_2] = \psi_1/\psi_2^2$; given a priori values for the mean and variance of α, these expressions can be solved to give the corresponding values of (ψ_1, ψ_2). A second general strategy, proposed by Dorazio (2009), exploits the following representation of the probability mass function for the prior of K that is induced by a Gamma(ψ_1, ψ_2) prior for α and a fixed N:

$$\pi(K|N, \psi_1, \psi_2) = \frac{\psi_2^{\psi_1} S_1(n, K)}{\Gamma(\psi_1)} \cdot \int_0^\infty \frac{\alpha^{K+\psi_1-1} \exp(-\psi_2 \alpha)\Gamma(\alpha)}{\Gamma(\alpha + N)} \, d\alpha,$$

where $S_1(n, K)$ is the unsigned Stirling number of the first kind and $K = 1, \dots, N$. In particular, suppose prior information for K can be directly elicited and is represented by $\pi(K)$. Then consider the Kullback–Leibler divergence between the prior $\pi(K)$ and the induced prior $\pi(K|N, \psi_1, \psi_2)$:

$$D_{\mathrm{KL}}(\psi_1, \psi_2) = \sum_{K=1}^{N} \pi(K) \log\left\{\frac{\pi(K)}{\pi(K|N, \psi_1, \psi_2)}\right\}.$$

Minimization of this quantity with respect to (ψ_1, ψ_2) gives a Gamma prior for α that reflects prior knowledge encoded by $\pi(K)$ but in a computationally convenient form. In practice, in the absence of explicit a priori knowledge regarding K, a uniform prior on $\{1, \dots, N\}$ could be used. In this case, $\pi(K) = 1/N$ and the Kullback–Leibler divergence measure reduces to

$$D_{\mathrm{KL}}(\psi_1, \psi_2) = -\log N - \frac{1}{N} \sum_{K=1}^{N} \log \pi(K|N, \psi_1, \psi_2).$$

If a priori knowledge regarding α is difficult to quantify but elicitation of prior information regarding the number of clusters K is possible, one can build on this approach in conjunction with expressions (3) and (4). In the absence of a priori knowledge on α or the number of clusters K, one could specify a diffuse Gamma distribution that assigns a priori mass to a broad range of values. For example, one option is to choose values of (ψ_1, ψ_2) such that α is centered between 1 and N with a large variance. In the simulations of Sections 4 and 5, we considered a Gamma(1.5, 0.0125) distribution as diffuse; this induces a distribution with a mode of 40 and variance of 9600, and assigns mass across realistic values of α.

3.2 Selecting a Value for α

While adopting a prior for α seems natural, the extent to which the data provides meaningful information on either α or K has been questioned (Kyung, Gill and Casella, 2010). An alternative therefore is to perform an analysis which fixes α at some value and to accompany the results with sensitivity analyses. Toward choosing a fixed value of α, if one is able to elicit an a priori best guess of the number of clusters, one could again exploit expression (3) and set α^* to be the value that satisfies $\hat{K} = \sum_{i=1}^{n} \frac{\alpha^*}{\alpha^*+i-1}$. In the absence of a priori knowledge on α or K, one could perform an empirical Bayes analysis of the logistic-DP mixing model that conditions on the MLE of α (Liu, 1996; Dorazio et al., 2008). Practically, this analysis can proceed using the following algorithm:

1. For some starting value of α, draw a sample from the joint posterior for the logistic-DP mixing model including for the unknown number of clusters, K.
2. Calculate the posterior mean of K, denoted $\bar{K}$, and find the value of α that satisfies $\bar{K} = \sum_{i=1}^{N} \frac{\alpha}{\alpha+i-1}$.
3. Repeat steps 1 and 2 and iterate until convergence.

While appealing that the final value of α^* is data-driven (as opposed to being an arbitrary choice), an important drawback is the substantial computational burden. Until convergence, the entire MCMC scheme needs to be rerun each time a new value of α is obtained. In the simulations presented in Sections 4 and 5, for example, this strategy generally required between 20–30 runs of the entire MCMC scheme, although it sometimes required more than 50.

As an alternative we propose a simplified strategy that relies on finding the MLE of α via importance sampling. Specifically, suppose a sample of size R has been drawn from the posterior that corresponds to a Gamma(ψ_1, ψ_2) prior for α. The posterior samples $\{\alpha^{(1)}, \dots, \alpha^{(R)}\}$ can be used to estimate the marginal posterior of α; we denote the corresponding estimate of the marginal posterior density by $\hat{\pi}_g(\alpha|\text{data}, \psi_1, \psi_2)$. Now define the importance weights

$$w(\alpha) = \frac{\pi_u(\alpha|C)}{\pi_g(\alpha|\psi_1, \psi_2)},$$

where $\pi_u(\alpha|C)$ is the density corresponding to a uniform prior for α on the interval $(0, C)$ with $C > N$ to ensure we assign mass to all reasonable values of α, and $\pi_g(\alpha|\psi_1, \psi_2)$ is the density corresponding to a Gamma(ψ_1, ψ_2) prior. An estimate of the marginal posterior density under a uniform prior for α can be obtained via reweighting using the importance weights to give

$$\hat{\pi}_u(\alpha|\text{data}) \propto \hat{\pi}_g(\alpha|\text{data}, \psi_1, \psi_2) * w(\alpha).$$

Since the marginal posterior distribution under a uniform prior is approximately the marginal likelihood for α, one can approximate the MLE for α by taking the mode of $\hat{\pi}_u(\alpha|\text{data})$. Hence, the MLE of α can be obtained with relatively little computational burden.

4. SIMULATION STUDY

As mentioned above, in practice and in most statistical software implementations, the mixing distribution in a GLMM is often taken to be a Normal distribution. Here we present a comprehensive simulation study with the dual goals of: (i) characterizing the bias-variance trade-off associated with using a logistic-DP model, as opposed to a logistic-Normal model; and, (ii) examining the performance of the logistic-DP model under the various strategies for α described in Section 3. Here we describe the framework used to conduct the simulation; the results are presented in Section 5.

4.1 Generating Correlated Data

To generate correlated binary response data, we adapt the setup used by Heagerty and Kurland (2001). Specifically, we used model (1) to generate correlated binary response data consisting of $N = 100$ clusters, each with $n_i = 10$ study units/observations. Throughout we consider two covariates. The first, denoted X_{1ij}, is a binary cluster-specific covariate with $P(X_{1i} = 1) = 0.5$ across all clusters. The second, denoted X_{2ij}, is a within-cluster covariate with the jth unit in the cluster taking on the jth value in the set $\{-4.5, -3.5, \dots, 3.5, 4.5\}$.

4.2 True Mixing Distributions and Fixed Effects

We consider two specifications for the random effects. The first is the random intercept model in which $\mathbf{Z}_{ij} = 1$ so that, building on the generic specification given by model (1), the data are generated according to the model

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T \beta + b_i,$$

with $g(\cdot)$ taken to be the logit function, $\mathbf{X}_{ij} = (1, X_{1ij}, X_{2ij})$, $\beta = (-2, 1, 0.5)$ and b_i a cluster-specific scalar. The second specification is the random intercept/slope model in which $\mathbf{Z}_{ij} = (1, X_{2ij})$, so that the data are generated as

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T \beta + \mathbf{Z}_{ij}^T b_i,$$

with $g(\cdot)$ again taken to be the logit function, $\beta = (-2, 1, 0.5)$ and $b_i = (b_{i,0}, b_{i,2})$ a cluster-specific vector that includes a random intercept and random slope.

For the random intercept model, we consider five scenarios for the "true" mixing distribution, G:

(RI-1) *Normal*: $b_i \sim \text{Normal}(0, \sigma^2)$, with $\sigma = 2$.

(RI-2) *Students' t*: $b_i \sim t_\nu$, where $\nu = \frac{8}{3}$.

(RI-3) *Standardized Gamma*: $b_i = \sigma(a_i - \lambda)/\sqrt{\lambda}$, with $a_i \sim \text{Gamma}(\lambda, 1)$, $\sigma = 2$ and $\lambda = 0.5$.

(RI-4) *Mixture of Normals*: b_i is a mixture of a $\text{Normal}(0, \sigma_0^2)$ and $\text{Normal}(0, \sigma_1^2)$, depending on the cluster-level covariate $X_{1ij} = 0/1$ and $(\sigma_0, \sigma_1) = (\sqrt{7}, 1)$.

(RI-5) *Two-point*: b_i takes on values in $\{-2, 2\}$, with equal probability.

For the random intercept/slope model, we consider two scenarios:

(RIS-1) *Normal*: $b_i \sim \text{MVN}_2(0, \Sigma)$, with $\Sigma = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

(RIS-2) *Standardized log-Normal*: Let $a_i \sim \text{MVN}_2(0, \Sigma)$ with Σ as in (RIS-1). Take b_i to be $\log(a_i)$, standardized to have mean zero and (marginal) variance of 1.0 for each of the two components.

4.3 Analyses

Under each of the seven scenarios for G, we generated $R = 5000$ data sets. For each data set, we performed a series of analyses as follows:

- A standard frequentist analysis based on the logistic-Normal model.
- A Bayesian analysis based on the logistic-Normal model.
- A Bayesian analysis based on the logistic-DP model with fixed α:
  - (i) Set at each of seven values: $\{100, 50, 25, 10, 5, 1, 0.1\}$.
  - (ii) Chosen via empirical Bayes, using the algorithm proposed by Dorazio et al. (2008).
  - (iii) Chosen via the importance sampling approach of Section 3.2.
- A Bayesian analysis based on the logistic-DP model with a $\text{Gamma}(\psi_1, \psi_2)$ prior on α with the following:
  - (i) $(\psi_1, \psi_2) = (1.5, 0.0125)$, chosen as a prior that is diffuse with respect to α (see Section 3.1).
  - (ii) $(\psi_1, \psi_2) = (0.491, 0.004)$, chosen using the Kullback–Leibler divergence of Dorazio (2009) based on a uniform prior for K on $\{1, \ldots, N\}$.

Note, in the Results section below, we refer to the last four of these analyses (i.e., empirical Bayes, importance sampling, a diffuse prior and a prior chosen via the Kullback–Leibler criterion) as "general-purpose" strategies for α in the sense that they are strategies that an analyst could adapt to be specific to their data/analyses.

Throughout, the frequentist analyses were performed using the `glmmML()` function in `R` (Broström and Holmberg, 2012). For Bayesian analyses of the logistic-Normal model, a noninformative flat prior was adopted for the β regression coefficients together with a $\text{Gamma}(0.5, 0.01)$ prior for the precision, τ^{-2}. For Bayesian analyses of the logistic-DP model, we used the `DPglmm()` function in the `DPpackage` package for `R` (Jara et al., 2011). The latter uses a slightly different parameterization of the logistic-DP model, the details of which along with our choice of priors (other than α) are provided in the online Supplementary Materials (Antonelli, Trippa and Haneuse, 2016). For each Bayesian analysis, three independent chains were run. Summaries are based on pooling 5000 posterior samples from each chain, obtained after thinning every fourth sample and removing a 20% burn-in. Convergence of the MCMC schemes was evaluated by calculating the potential scale reduction (PSR) factor (Gelman et al., 2013) across all model parameters.

4.4 Operating Characteristics

To investigate the performance of the various analysis strategies across the "true" mixing distributions, we evaluated a number of (traditionally) frequentist operating characteristics. For estimation of the fixed effects we investigated the potential for a bias-variance trade-off by evaluating percent bias for the MLEs in the frequentist analyses and the posterior medians in the Bayesian analyses. We also evaluated the standard deviation of the MLE and posterior medians across the $R = 5000$ simulated data sets to obtain the true variability of our fixed effect estimates.

The imposition of structure on the cluster-specific random effects, typically in the form of a Normal distribution, is well known to result in shrinkage in their corresponding point estimates, particularly in the tails of the distribution. In addition to improving estimation with respect to fixed effects in the GLMM, a motivation for relaxing the Normality assumption for the mixing distribution G is to mitigate this phenomenon and reduce the amount of shrinkage. To investigate this phenomenon, we examined the extent to which a posterior median of a given true random effect value exhibits shrinkage. Specifically, for any given value of b_i we can plot the posterior median vs. the true value to investigate the amount of shrinkage in the predicted values. As a more quantitative measure, we estimated the mean squared error of prediction (MSEP) as

$$\frac{1}{RN}\sum_{r=1}^{R}\sum_{i=1}^{N}(\hat{b}_i^{(r)} - b_i^{(r)})^2, \tag{5}$$

where $b_i^{(r)}$ is the random effect for the ith cluster in the rth simulated data set and $\hat{b}_i^{(r)}$ is the corresponding posterior median.

5. RESULTS

Tables 1–5 and Figure 1 summarize results from the simulations studies. Across each of the mixing distribution scenarios, results for the frequentist analysis based on the logistic-Normal model and the corresponding Bayesian analysis are qualitatively similar and we restrict attention to the latter. In addition, across all "true" random intercept mixing distributions and analysis procedures, estimation of the within-cluster effect, β_2, exhibited low bias (between -1.6% and 1.8%) and posterior uncertainty did not vary across models. As such, when the true mixing distribution is one of the five random intercept specifications, (RI-1)–(RI-5), we therefore restrict presentation and discussion of results to those for the global intercept β_0 and the between-cluster effect, β_1. Complete tables, that include results for frequentist estimation of the logistic-Normal model as well as for β_2 in the random intercept model simulations, are provided in the online Supplementary Materials document (Antonelli, Trippa and Haneuse, 2016).

5.1 Fixed Effects, β: Percent Bias

From Table 1 when the true mixing distribution is either a Normal or Students' t-distribution, an analysis that assumes Normality exhibits little bias (e.g., 1.4% and 0.5% for β_1). Analysis based on the logistic-DP model for these mixing distributions also exhibited little bias unless the value of α was set to be a low number

TABLE 1

Estimated percent bias for fixed effects estimation, across various analyses based on the random intercept model under the five "true" random effect specifications (RI-1)–(RI-5), *described in Section* 4.2. *Percent bias is for the posterior median. Results are based on* $R = 5000$ *simulated data sets, each with* $N = 100$ *and* $n_i = 10$

	Normal		**Students' t**		**Gamma**		**Mixture**		**Two point**	
	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
Logistic-Normal										
Bayesian	0.9	1.4	0.1	0.5	15.2	8.2	−11.4	−30.5	6.0	1.3
Logistic-DP: Fixed α										
$\alpha = 100$	−0.9	−0.4	−0.8	−0.3	13.1	5.1	−10.8	−23.1	9.8	4.4
$\alpha = 50$	−1.1	−0.6	−1.1	−0.4	11.3	4.1	−10.1	−18.9	11.8	5.6
$\alpha = 25$	−1.5	−1.0	−1.5	−0.6	9.0	3.3	−9.7	−15.0	12.5	5.7
$\alpha = 10$	−2.2	−1.9	−2.1	−1.0	6.2	2.2	−9.6	−11.1	10.1	4.9
$\alpha = 5$	−3.0	−3.0	−2.5	−1.5	4.5	1.3	−9.6	−9.1	7.1	4.2
$\alpha = 1$	−4.8	−6.5	−3.7	−3.2	2.5	−1.4	−10.3	−8.3	2.8	2.6
$\alpha = 0.1$	−7.0	−10.8	−5.2	−5.7	1.2	−4.1	−12.0	−12.0	0.7	0.9
EB*	−1.4	−0.7	−1.7	−0.7	5.5	1.8	−9.5	−12.8	3.5	3.1
IS*	−0.7	−1.6	−1.0	−0.9	5.4	0.9	−10.0	−17.4	3.4	3.2
Logistic-DP: Random α**										
Diffuse	−1.0	−0.5	−1.0	−0.3	8.7	3.3	−10.4	−20.2	6.0	3.9
KL*	−1.1	−0.6	−1.2	−0.4	6.6	2.2	−10.2	−17.9	3.9	3.1

*EB = Empirical Bayes; IS = Importance sampling; KL = Kulback–Leibler.
**Diffuse: Gamma(1.5, 0.0125); KL criterion: Gamma(0.491, 0.004).

(i.e., less than 5); under a true Normal mixing distribution, the actual number of clusters is $K = 100$ and low fixed values of α do not reflect this. When the true mixing distribution is a Gamma distribution, we see that a naïve analysis based on the logistic-Normal model yields substantial bias for the intercept β_0 (15.2%) and moderate bias for the between-cluster effect (8.2%). In contrast, each of the logistic-DP analyses outperforms the logistic-Normal analysis with percent bias reduced to approximately 2% when either a fixed α is chosen based on the empirical Bayes approach or on importance sampling or when a Gamma prior is chosen using the Kullback–Leibler approach of Dorazio (2009). The logistic-DP model also dramatically improves upon a logistic-Normal model when the true mixing distribution is a mixture of Normals, although the bias is not completely removed. Bias for estimation of both β_0 and β_1 when the truth is a two-point distribution is smaller under a logistic-DP model than a logistic-Normal model assuming α is chosen to be very small (i.e., consistent with the "true" number of clusters), though bias can be increased if α is too large.

From Table 2, we see that when the true mixing distribution is a bivariate Normal for the random intercepts and slopes, an analysis based on the logistic-Normal model yields small bias. Unless a poor value of α is chosen (i.e., one that is completely inconsistent with the true number of clusters), we see that bias is of essentially the same magnitude under the logistic-DP analysis. Interestingly, when the true mixing distribution is skewed (i.e., RIS-2; the standardized log-Normal distribution), a naïve analyis based on the logistic-Normal model exhibits substantial bias in all three fixed effect parameters (i.e., 12.7% for β_0, 7.5% for β_1 and −25.0% for β_2). For any given strategy regarding α, with the exception of particularly small values, we find large decreases in the bias across all three parameters for all logistic-DP analyses.

Finally, we note that, across the board, the four general-purpose strategies for α perform either optimally or very well relative to the logistic-Normal in both sets of simulations.

5.2 Fixed Effects, β: Uncertainty

Table 3 provides the empirical standard deviation of the sampling distribution of the posterior median for β_0 and β_1 from the random intercept model simulations.

TABLE 2
Estimated percent bias for fixed effects estimation, across analyses based on the random intercept/slope model under specifications (RIS-1; *Normal*) *and* (RIS-2; *Standardized log-Normal*) *described in Section* 4.2. *Percent bias is for the posterior median. Results are based on $R = 5000$ simulated data sets, each with $N = 100$ and $n_i = 10$*

	Normal			Standardized log-Normal		
	β_0	β_1	β_2	β_0	β_1	β_2
Logistic-Normal						
Bayesian	2.4	2.3	3.0	12.7	7.5	−25.0
Logistic-DP: Fixed α						
$\alpha = 100$	2.7	2.8	3.3	12.5	7.7	−20.2
$\alpha = 50$	2.5	2.6	3.1	11.6	6.8	−17.7
$\alpha = 25$	2.1	2.0	2.2	9.9	5.2	−14.7
$\alpha = 10$	0.7	0.5	−0.4	6.7	2.4	−12.2
$\alpha = 5$	−0.9	−1.2	−3.4	4.1	0.5	−13.4
$\alpha = 1$	−5.0	−4.9	−9.0	−0.9	−3.4	−19.6
$\alpha = 0.1$	−9.6	−9.2	−14.1	−5.0	−7.5	−24.3
EB*	2.5	2.5	3.0	4.6	1.0	−14.0
IS*	2.7	2.7	3.4	4.4	0.9	−14.4
Logistic-DP: Random α**						
Diffuse	2.6	2.7	3.2	7.3	3.3	−15.0
KL*	2.5	2.6	3.0	5.3	1.7	−15.1

*EB = Empirical Bayes; IS = Importance sampling; KL = Kulback–Leibler.
**Diffuse: Gamma(1.5, 0.0125); KL criterion: Gamma(0.491, 0.004).

TABLE 3

True standard errors for fixed effects estimation, across analyses based on the random intercept model under the five "true" random effect specifications (RI-1)–(RI-5), *described in Section* 4.2. *True standard errors are calculated as the standard deviation of the parameter estimates across* $R = 5000$ *simulated data sets, each with* $N = 100$ *and* $n_i = 10$

	Normal		**Students' *t***		**Gamma**		**Mixture**		**Two point**	
	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
Logistic-Normal										
Bayesian	0.34	0.47	0.27	0.35	0.34	0.45	0.38	0.42	0.42	0.53
Logistic-DP: Fixed α										
$\alpha = 100$	0.33	0.46	0.26	0.34	0.31	0.39	0.38	0.44	0.36	0.35
$\alpha = 50$	0.33	0.46	0.26	0.34	0.30	0.37	0.39	0.47	0.34	0.30
$\alpha = 25$	0.33	0.46	0.26	0.34	0.30	0.36	0.39	0.51	0.33	0.27
$\alpha = 10$	0.33	0.46	0.26	0.34	0.30	0.35	0.40	0.57	0.31	0.25
$\alpha = 5$	0.33	0.47	0.27	0.34	0.30	0.35	0.41	0.60	0.30	0.24
$\alpha = 1$	0.33	0.47	0.27	0.35	0.29	0.34	0.41	0.64	0.28	0.23
$\alpha = 0.1$	0.33	0.47	0.27	0.35	0.29	0.34	0.41	0.66	0.27	0.23
EB*	0.33	0.46	0.26	0.34	0.31	0.35	0.40	0.54	0.29	0.24
IS*	0.34	0.46	0.26	0.34	0.31	0.35	0.40	0.54	0.29	0.24
Logistic-DP: Random α**										
Diffuse	0.33	0.46	0.26	0.34	0.31	0.37	0.39	0.47	0.30	0.24
KL*	0.33	0.46	0.26	0.34	0.31	0.36	0.39	0.50	0.29	0.24

*EB = Empirical Bayes; IS = Importance sampling; KL = Kulback–Leibler.
**Diffuse: Gamma(1.5, 0.0125); KL criterion: Gamma(0.491, 0.004).

Results from the random intercept/slopes simulations are substantively similar and are therefore not shown [see the Supplementary Materials document for additional detail (Antonelli, Trippa and Haneuse, 2016)].

Given the adopted sample sizes (i.e., $N = 100$ and $n_i = 10$), the standard error of the posterior median for β_1 based on a logistic-Normal model when the mixing distribution is truly Normal is approximately 0.47. Interestingly, the corresponding standard error estimates under each of the logistic-DP model analysis are approximately the same; that none of these are larger than 0.47 indicates that there is no loss of efficiency associated with using the logistic-DP model even when the true mixing distribution is a Normal. This result also holds when the true mixing distribution is a Students' t-distribution. When the true mixing distribution is a Gamma distribution, the standard error of the posterior median of β_1 under a (misspecified) logistic-Normal model is approximately 0.45. The corresponding standard errors for the posterior medians under each of the logistic-DP analyses were smaller (between 0.34 and 0.39), indicating a relative improvement in efficiency under the nonparametric model. This result also holds, and is more dramatic, when the true mixing distribution is a two-point distribution with the standard error under a logistic-Normal model approximately 0.53 and under the various logistic-DP model analyses between 0.24 and 0.35. Finally, when the true mixing distribution is a mixture of Normals, the standard errors under a logistic-Normal model are approximately 0.42. Under each of the logistic-DP model analyses the standard error is greater (between 0.44 and 0.66). The increase in variability is difficult to interpret, however, because of the substantial bias that the posterior medians exhibit (between −8.3% and −23.1%; see Table 1).

5.3 Random Effects, b_i: Shrinkage

Figure 1 provides a visual representation of shrinkage in the estimation of the cluster-specific random effects, for select model fits, when the true mixing distribution is a standardized Gamma distribution (i.e., RI-3 in Section 4.2). From the figure, a naïve logistic-Normal model exhibits substantial shrinkage in the estimated random effects (i.e., bias toward zero), particularly in the tails of the distribution. Under each of the displayed logistic-DP models the shrinkage in the tails is mitigated, indicating a superior fit overall. This improvement is likely due to the additional flexibility the DP prior provides in modeling a skewed mixing distribution, a scenario that is ruled out when G is taken to be a Normal distribution. When the true mixing distribution is a two-point distribution (i.e., RI-5

TABLE 4
Shrinkage of estimated random effects under select analyses based on a random intercept model when the "true" mixing distribution is a Two-point distribution (RI-5; *see Section* 4.2). *Estimates are the mean of the posterior medians, taken across* $R = 5000$ *simulated data sets, each with* $N = 100$ *and* $n_i = 10$

	True value	
	$b_i = -2$	$b_i = 2$
Logistic-Normal		
Bayesian	−1.79	1.88
Logistic-DP: Fixed α		
$\alpha = 100$	−1.79	1.99
$\alpha = 50$	−1.82	2.06
$\alpha = 25$	−1.86	2.11
$\alpha = 10$	−1.92	2.10
$\alpha = 5$	−1.96	2.06
$\alpha = 1$	−1.97	1.97
$\alpha = 0.1$	−1.95	1.93
EB*	−1.97	1.99
IS*	−1.97	1.98
Logistic-DP: Random α**		
Diffuse	−1.95	2.03
KL*	−1.96	1.99

*EB = Empirical Bayes; IS = Importance sampling; KL = Kulback–Leibler.
**Diffuse: Gamma(1.5, 0.0125) KL criterion: Gamma(0.491, 0.004).

in Section 4.2), Table 4 shows that the shrinkage exhibited by the logistic-Normal model is almost completely mitigated by the fit of a logistic-DP model, especially for small values of α (i.e., those that are consistent with the "true" number of clusters). Beyond the Gamma and two-point distributions, shrinkage under a logistic-DP model was only minimally improved under the Normal, Students' t and mixture distributions; the corresponding figures are provided in the online Supplementary Materials document (Antonelli, Trippa and Haneuse, 2016). Finally, as with estimation of the fixed effects, each of the four general-purpose strategies for α perform very well in Figure 1 and Table 4.

5.4 Random Effects, b_i: Mean Squared Error of Prediction

Finally, Table 5 provides estimates of MSEP (as defined in Section 4.4) for both the random intercept and random intercept/slope model simulations. For the random intercept model simulations (first five columns) we see that the logistic-DP analysis generally outperforms the logistic-Normal analysis. For example, the MSEP is substantially reduced under both the standardized Gamma and the two-point distributions regardless of the strategy adopted for α. Furthermore, as long as a poor value of α is avoided, MSEP under either the Normal distribution or the Students t-distribution is no worse for the logistic-DP analysis than the logistic-Normal analysis. Settings where the logistic-DP does exhibit worse performance are when the true G is a mixture distribution or when the truth is a Normal or a Students' t-distribution and a poor value of α is chosen.

The last four columns of Table 5 provide MSEP specific to $b_{i,0}$ and $b_{i,2}$ in the random intercept/slope model simulations. For the most part, the same general conclusions are drawn. When the true mixing distribution is a multivariate Normal, the logistic-DP analyses exhibit the same performance as the logistic-Normal analyses as long as a particularly poor value of α is not chosen. When the true mixing distribution is a standardized log-Normal (i.e., skewed), the MSEP based on a logistic-Normal model is estimated to be 0.59 and 0.58 for $b_{i,0}$ and $b_{i,2}$, respectively. The corresponding MSEPs are uniformly lower under each of the logistic-DP analyses, with each of the general-purposes strategies performing well.

6. APPLICATION

To further illustrate the concepts and methods described in Sections 2 and 3, we present an analysis of $n = 18{,}643$ female Medicare beneficiaries who underwent a surgical procedure for the removal of their uterus, a hysterectomy, at one of $N = 503$ hospitals in the U.S between 2009–2012. The patients were identified in the Medicare Inpatient File with the only inclusion/exclusion criteria being that the patient was at least 65 years of age at the time of the procedure and had not transferred to the hospital at which the hysterectomy was performed from some other facility. Furthermore, due to considerations of patient confidentiality, the data available for analyses only includes patients treated in hospitals at which at least 20 hysterectomy procedures had been performed between 2009–2012.

Of primary scientific interest for this application is the characterization of variation in quality of post-operative care in the U.S. among Medicare beneficiaries undergoing a hysterectomy. Toward this, we consider a binary outcome of whether or not the patient experienced a surgery-specific complication, defined as either an in-hospital death or readmission within 30 days of discharge. These were chosen because they represent key markers of quality of care in the lit-

TABLE 5

Mean squared error of prediction for random effects, across various analyses based on (i) *the random intercept model under the five "true" random intercept specifications* (RI-1)-(RI-5), *and* (ii) *the random intercept/slope model under the two "true" random intercept specifications* (RIS-1) *and* (RIS-2). *Estimates are based on* $R = 5000$ *simulated data sets, each with* $N = 100$ *and* $n_i = 10$

	Random intercept					Random intercept/slope			
						Normal		Standardized log-Normal	
	Normal	Students' t	Gamma	Mixture	Two-point				
	b_i	b_i	b_i	b_i	b_i	$b_{i,0}$	$b_{i,2}$	$b_{i,0}$	$b_{i,2}$
Logistic-Normal									
Bayesian	0.94	2.00	1.25	1.17	0.78	0.59	0.22	0.59	0.58
Logistic-DP: Fixed α									
$\alpha = 100$	0.95	1.99	1.07	1.19	0.62	0.59	0.22	0.54	0.55
$\alpha = 50$	0.95	1.99	1.03	1.21	0.56	0.59	0.22	0.52	0.54
$\alpha = 25$	0.96	1.99	0.99	1.23	0.47	0.59	0.23	0.50	0.51
$\alpha = 10$	0.98	2.01	0.96	1.28	0.34	0.59	0.22	0.46	0.49
$\alpha = 5$	1.01	2.03	0.94	1.32	0.27	0.57	0.22	0.44	0.46
$\alpha = 1$	1.10	2.12	0.98	1.44	0.22	0.60	0.23	0.39	0.41
$\alpha = 0.1$	1.23	2.26	1.07	1.59	0.22	0.76	0.30	0.50	0.54
EB*	0.96	2.00	0.96	1.25	0.23	0.59	0.22	0.43	0.44
IS*	0.96	2.09	0.97	1.25	0.22	0.59	0.22	0.43	0.44
Logistic-DP: Random α**									
Diffuse	0.95	1.99	1.00	1.21	0.26	0.59	0.22	0.46	0.48
KL*	0.95	1.99	0.97	1.23	0.23	0.59	0.22	0.43	0.46

*EB = Empirical Bayes; IS = Importance sampling; KL = Kulback–Leibler.
**Diffuse: Gamma(1.5, 0.0125); KL criterion: Gamma(0.491, 0.004).

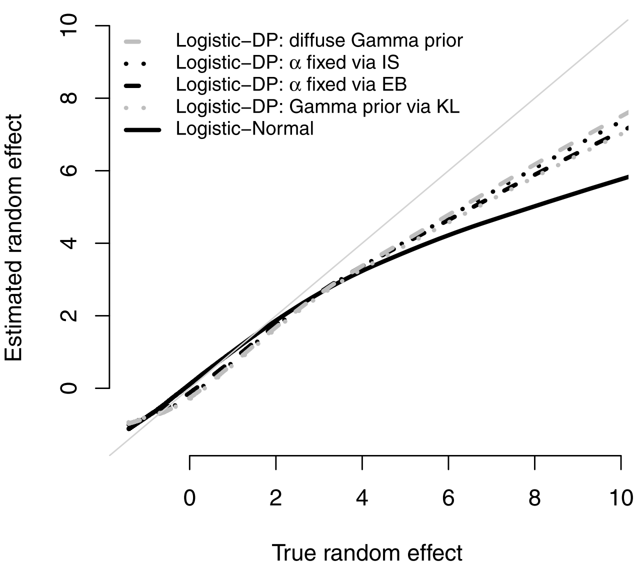

FIG. 1. *Shrinkage of estimated random effects under select analyses when the "true" mixing distribution is a standardized Gamma distribution* (RI-3; *see Section* 4.2). *Estimates are the mean of the posterior medians, taken across* $R = 5000$ *simulated data sets, each with* $N = 100$ *and* $n_i = 10$. *The solid grey line indicates the* 45° *line.*

erature, and also because they do not rely on point-of-access coding (which can be subject to inaccurate and/or idiosyncratic coding practices). When identifying readmissions, attention was restricted to those that were likely to be related to the surgery. Toward this, in collaboration with a group of two-dozen expert reviewers (doctors and surgeons), a list of admission codes was developed that only included those corresponding to likely unplanned admissions and therefore indicative of a negative surgical outcome.

Figure 2 provides a histogram of the $N = 503$ raw hospital-specific complication rates. While the raw rates vary from 0% to a maximum of 12%, there is a large point mass at 0% corresponding to 323 hospitals (64.2%) with the distribution of the remaining rates skewed to the right. As such, while not conclusive, the marginal distribution of the raw rates suggest that a single Normal distribution may be insufficient to capture between-hospital variation in these data. Toward performing an adjusted analysis, wherein differences in the characteristics of the patients across the hospitals are accounted for, we fit a series of GLMMs based on the methods described in Sections 2 and 3. Due to the

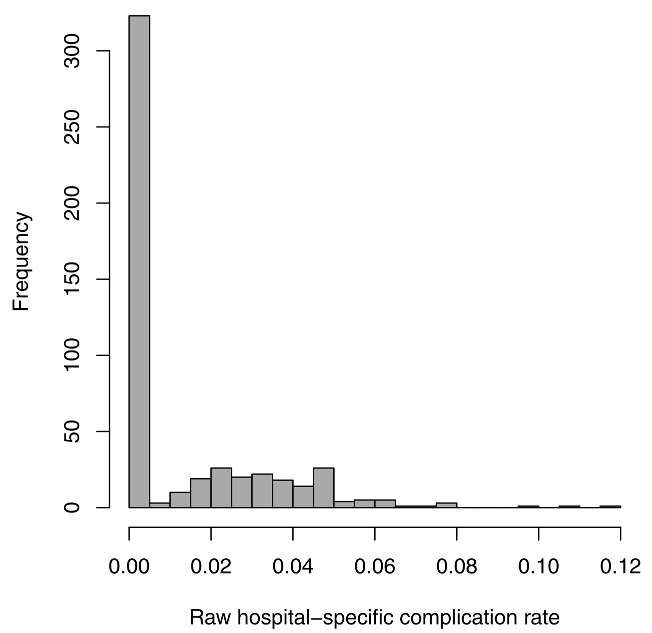

FIG. 2. *Distribution of the raw hospital-specific complication rates for* $n = 18{,}643$ *female Medicare beneficiaries who underwent a hysterectomy at* $N = 503$ *hospitals. The point mass corresponds to* 323 *hospitals* (64.2%) *that had a zero complication rate.*

large sample size and subsequent computational burden, we do not perform the DP analyses based on empirical Bayes or the Kullback–Leibler criterion. In each of these models we included information about the patients' race (white, black, other), age (65–69 years, 70–74 years, 75–79 years, 80–84 years, ≥85 years) and year of surgery (2009, 2010, 2011, 2012) as fixed effects. Given the nature of the potential correlation (i.e., patients clustered within hospitals), we present results based on a random intercept model. For analyses based on a logistic-Normal model, we adopted a diffuse Normal prior for the regression coefficients and a Gamma(0.2, 0.025) prior on the precision of the random effect distribution. For the DP models we also assume noninformative priors on the regression coefficients as well as for the mean of the centering distribution. In addition, we again adopt a Gamma(0.2, 0.025) prior on the precision of the centering distribution. In all cases, three MCMC chains were run of length 25,000, with the first 20% removed as a burn-in and the remaining samples thinned to retain every 10th sample. Convergence was assessed by visually inspecting trace plots as well as ensuring the PSR was below 1.05 for all unknown model components [see the online Supplementary Materials document for details (Antonelli, Trippa and Haneuse, 2016)].

Table 6 and Figure 3 provide a summary of the results for the fixed effects and random effects, respec-

TABLE 6

Posterior summaries for the fixed effects (*log odds and log odds ratio*) *parameters from a series of analyses of complications following a hysterectomy, based on* $n = 18{,}643$ *Medicare beneficiaries treated at* $N = 503$ *hospitals. Shown are posterior medians* (*standard deviations*)

		Race*		**Age, years***				**Year of surgery***		
	Intercept	**Black**	**Other**	**70–74**	**75–79**	**80–84**	**≥85**	**2010**	**2011**	**2012**
Logistic-Normal										
Bayesian	−4.55 (0.15)	0.20 (0.30)	−0.38 (0.31)	0.19 (0.17)	0.64 (0.17)	0.43 (0.23)	1.01 (0.27)	−0.22 (0.18)	−0.12 (0.17)	0.16 (0.18)
Logistic-DP: Fixed α										
$\alpha = 500$	−4.58 (0.15)	0.19 (0.31)	−0.38 (0.32)	0.18 (0.17)	0.65 (0.17)	0.42 (0.23)	0.99 (0.28)	−0.23 (0.17)	−0.12 (0.18)	0.16 (0.17)
$\alpha = 300$	−4.58 (0.16)	0.19 (0.32)	−0.39 (0.33)	0.18 (0.17)	0.65 (0.17)	0.41 (0.23)	0.99 (0.28)	−0.22 (0.18)	−0.11 (0.18)	0.17 (0.18)
$\alpha = 100$	−4.60 (0.16)	0.20 (0.32)	−0.38 (0.32)	0.19 (0.17)	0.65 (0.17)	0.43 (0.23)	1.01 (0.27)	−0.22 (0.17)	−0.11 (0.17)	0.16 (0.18)
$\alpha = 50$	−4.57 (0.16)	0.20 (0.32)	−0.39 (0.32)	0.18 (0.17)	0.64 (0.17)	0.41 (0.23)	0.99 (0.28)	−0.23 (0.17)	−0.11 (0.17)	0.16 (0.17)
$\alpha = 25$	−4.58 (0.16)	0.19 (0.31)	−0.39 (0.31)	0.18 (0.18)	0.64 (0.17)	0.42 (0.23)	1.00 (0.28)	−0.22 (0.18)	−0.12 (0.18)	0.16 (0.17)
$\alpha = 10$	−4.58 (0.16)	0.19 (0.32)	−0.39 (0.32)	0.18 (0.17)	0.64 (0.17)	0.42 (0.23)	1.00 (0.27)	−0.22 (0.17)	−0.11 (0.17)	0.16 (0.17)
$\alpha = 1$	−4.58 (0.53)	0.21 (0.32)	−0.39 (0.32)	0.18 (0.17)	0.65 (0.17)	0.41 (0.23)	0.99 (0.28)	−0.23 (0.18)	−0.12 (0.18)	0.16 (0.18)
IS**	−4.56 (0.16)	0.21 (0.33)	−0.39 (0.32)	0.18 (0.17)	0.64 (0.17)	0.41 (0.23)	0.98 (0.28)	−0.23 (0.18)	−0.13 (0.18)	0.15 (0.18)
Logistic-DP: Random α***										
Diffuse	−4.59 (0.18)	0.21 (0.32)	−0.38 (0.32)	0.19 (0.17)	0.65 (0.17)	0.42 (0.23)	0.99 (0.28)	−0.21 (0.18)	−0.12 (0.17)	0.17 (0.17)

*Reference race group is "White;" referent age group is "65–69 years;" reference year of surgery is "2009."
**IS = Importance sampling.
***Diffuse: Gamma(2, 0.008).

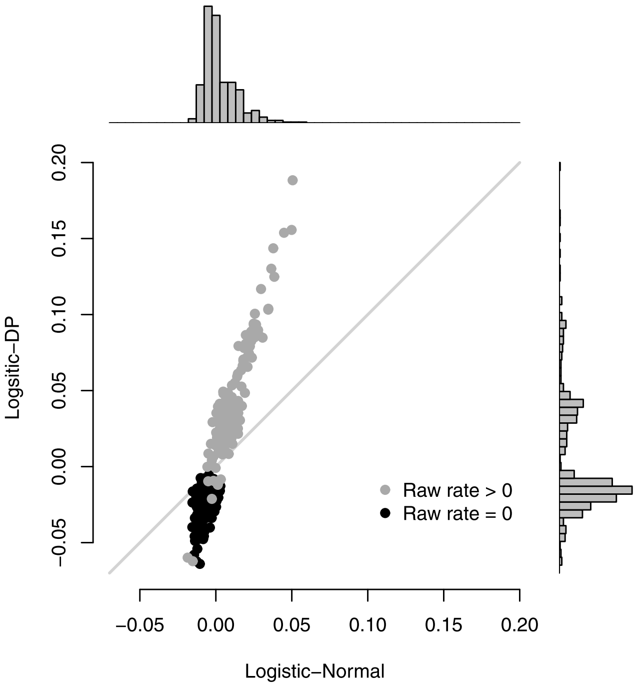

FIG. 3. *Comparison of posterior medians for the hospital-specific random effects from a logistic-Normal random intercept model and a logistic-DP random intercept model applied to* $n = 18{,}643$ *Medicare beneficiaries who underwent a hysterectomy at* $N = 503$ *hospitals.*

tively. From Table 6 one can see that the conclusions one draws for the fixed effects regression parameters are robust across all specifications for the mixing distribution. In particular, both the posterior medians and standard deviations are insensitive to whether or not a Normal mixing distribution is adopted or if a DP prior is adopted. Interestingly, that the posterior standard deviations are insensitive suggests that there was no penalty paid for adopting the more flexible DP prior specification.

Figure 3 provides posterior medians for the $N = 503$ hospital specific random effects (i.e., the b_i random intercepts) for the logistic-Normal model and for the logistic-DP model with α fixed and chosen via importance sampling (see Section 3.2); analogous figures for other logistic-DP analyses were similar and are, therefore, not presented. From the figure, in contrast to the results for the fixed effects, we find that the conclusions one draws regarding the random effects are highly sensitive to specification of the mixing distribution. In particular, one key feature of the comparison is that the posterior medians of the hospital-specific random effects exhibit substantially greater variation under the logistic-DP model than under the logistic-Normal model. This is likely because the structure imposed by the common Normal distribution across the random effects is being informed primarily by the large number of hospitals with an observed raw complication rate of 0%. Such structure is not imposed under the logistic-DP model and the random effects for the hospitals for which the raw rates were greater than 0% (i.e., the grey dots in the figure) are given much greater flexibility to deviate away from the point mass. A second key feature of the comparison is that the marginal distribution of the posterior medians under the logistic-DP model is much more aligned with the distribution of the raw rates in Figure 2. Specifically, those hospitals with a 0% complication rate remain "clustered," while the hospitals with a raw complication rate that is greater than 0% form a distinct mass in the distribution. In contrast, the corresponding marginal distribution under the logisitic-Normal is unimodal again likely due to the structure imposed by adopting a single common Normal distribution for the random effects.

7. DISCUSSION

As researchers consider GLMMs for the analysis of clustered or longitudinal data, Bayesian nonparametric formulations offer an appealing and flexible framework if the mixing distribution is potentially not Normal. While the literature on Bayesian nonparametrics is well established and user-friendly software is readily available, to our knowledge, the extent to which bias due to misspecification of the mixing distribution is mitigated has not been quantified. Overall, our simulation-based results suggest that, in a broad range of settings, the use of a Bayesian nonparametric prior does mitigate bias that arises when the mixing distribution is incorrectly assumed to be a Normal distribution. Our results also indicate that little penalty is paid (in terms of variance) if the mixing distribution is truly a Normal distribution and yet a more flexible specification is adopted. Furthermore, in some settings, specifically the Gamma and two-point distributions, we found that fitting a logistic-DP model yielded results that not only exhibited lower bias but also had improved efficiency. This latter observation is consistent with the theoretical results of Kyung, Gill and Casella (2009) who showed that estimation of fixed effects in a linear mixed effects model often enjoys reduced variance under a Dirichlet random effects model compared to that under a Normal model.

The results from our simulation study also indicate that adopting a Bayesian nonparameteric prior may not completely resolve bias induced by misspecification of the mixing distribution. In this respect, the results based on the true mixing distribution being a mixture of Normals are important. We found that none of the logistic-DP analyses we considered completely removed bias, with the best case being a bias of approximately -9.6% when α was between 5–25. Furthermore, while it is difficult to interpret measures of variability when the estimates are themselves biased, the results of Table 3 show that in some settings uncertainty is increased when one adopts a flexible prior on the unknown mixing distribution.

During the review process, the reviewers made a number of interesting suggestions for discussion and future work. One such direction is a Bayesian treatment of marginalized models and the possible impact of this choice under model misspecification. Marginalized models utilize a marginal likelihood in which the random effects have been integrated out, and these models have been shown to handle misspecification of random effects distributions quite well (Heagerty and Zeger, 2000). Similarly, research has been done on marginalized copula methods and their ability to estimate generalized linear models in the presence of correlated data. O'Brien and Dunson (2004) provide a flexible approach to logistic regression using copulas permitting a wide variety of correlation structures in the data. These marginalized models also have an advantage, as they allow for population-based interpretations of parameters as opposed to interpretations conditional on random effect values in conditional models. Another general direction for future work is the investigation of the relative merits of alternative Bayesian nonparametric prior formulations. Polya trees (Lavine, 1992) and species sampling priors (Lee et al., 2013), for example, can both be viewed as generalizations of the Dirichlet process prior. The latter, in particular, has been used in the context of GLMs and GLMMs (Hanson and Johnson, 2002, Branscum and Hanson, 2008; Trippa, Müller and Johnson, 2011) and is implemented in the `DPpackage` package for `R` (Jara et al., 2011). Whether or not the desirable theoretical aspects of these alternatives result in meaningful practical benefits (i.e., in terms of bias and efficiency) is unclear and an avenue we are currently exploring. Third, while the random intercept model and the random intercept/slopes model are commonly used for binary response data, analysts may encounter any of a broad range of data scenarios including count and ordinal data (Kottas, Müller and Quintana, 2005; Leon-Novelo et al., 2010), varying cluster sizes (e.g., Dunson, Chen and Harry, 2003), and more complex correlation structures seen in spatial (Banerjee, Carlin and Gelfand, 2014) and spatio-temporal models (Waller et al., 1997). Finally, it may be possible to formulate the choice of how α is treated as a question of model choice. Within this framing, it may be possible to exploit formal techniques of model choice such as the deviance information criterion (DIC) or log pseudo marginal likelihood (LPML) to empirically decide which approach is best (Celeux et al., 2006; Geisser and Eddy, 1979; Basu and Chib, 2003).

To conclude, we make the general point that, in the absence of knowledge about the true mixing distribution, no single analysis strategy will work well (or be best in any sense) across all possible true data scenarios. While describing and quantifying operating characteristics can serve to provide guidance, in practice, substantive knowledge about the underlying data generating mechanisms is far more valuable. Nevertheless, we found that adopting a Bayesian nonparametric prior may outperform or do no worse than adopting the current convention of Normal distribution in a broad range of settings. These benefits are not for free, however, and are accompanied by additional complexity and the nontrivial task of specifying α. Toward providing practical guidance on this, however, our experience in conducting a broad range of simulations suggest to us that, combined with appropriate sensitivity analyses, the four "general-purpose" strategies consisting of choosing a specific value via empirical Bayes or importance sampling, or choosing a diffuse prior or one prior chosen via the Kullback–Leibler criterion, will be reasonable in a broad range of practical settings.

ACKNOWLEDGMENTS

We would like to thank the Associate Editor and referees for their thoughtful comments that have improved the original manuscript.

Joseph Antonelli was supported by NIH Grant ES007142. Lorenzo Trippa was supported by the Claudia Adams Barr Program in Innovative Basic Cancer Research. Sebastien Haneuse was supported by NIH Grant R-01 CA181360-01.

SUPPLEMENTARY MATERIAL

Supplement to "Mitigating bias in generalized linear mixed models: The case for Bayesian nonparametrics" (DOI: 10.1214/15-STS533SUPP; .pdf). We include in the supplementary files a detailed description of both the model and prior specification for the Logistic-DP model. We also include extended simulation results that include all parameters from the model and an additional simulation that looks at a larger sample size. Finally, we include convergence diagnostics for all Bayesian models in the Medicare application.

REFERENCES

AGRESTI, A., CAFFO, B. and OHMAN-STRICKLAND, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Comput. Statist. Data Anal.* **47** 639–653. MR2100566

ANTONELLI, J., TRIPPA, L. and HANEUSE, S. (2016). Supplement to "Mitigating bias in generalized linear mixed models: The case for Bayesian nonparametrics." DOI:10.1214/15-STS533SUPP.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, Boca Raton.

BASU, S. and CHIB, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *J. Amer. Statist. Assoc.* **98** 224–235. MR1965688

BRANSCUM, A. J. and HANSON, T. E. (2008). Bayesian nonparametric meta-analysis using Polya tree mixture models. *Biometrics* **64** 825–833. MR2526633

BROSTRÖM, G. and HOLMBERG, H. (2012). glmmML: Generalized linear models with clustering (2011). R package version 0.82-1.

CAFFO, B., AN, M.-W. and ROHDE, C. (2007). Flexible random intercept models for binary outcomes using mixtures of normals. *Comput. Statist. Data Anal.* **51** 5220–5235. MR2370867

CELEUX, G., FORBES, F., ROBERT, C. P. and TITTERINGTON, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Anal.* **1** 651–673 (electronic). MR2282197

DAVIDIAN, M. and GALLANT, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80** 475–488. MR1248015

DEY, D., MÜLLER, P. and SINHA, D., eds. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. *Lecture Notes in Statistics* **133**. Springer, New York. MR1630072

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2013). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR3087136

DORAZIO, R. M. (2009). On selecting a prior for the precision parameter of Dirchlet process mixture models. *J. Statist. Plann. Inference* **139** 3384–3390. MR2538090

DORAZIO, R. M., MUKHERJEE, B., ZHANG, L., GHOSH, M., JELKS, H. L. and JORDAN, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics* **64** 635–644, 670–671. MR2432438

DUNSON, D. B., CHEN, Z. and HARRY, J. (2003). A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* **59** 521–530. MR2004257

ESCOBAR, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89** 268–277. MR1266299

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

GEISSER, S. and EDDY, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** 153–160. MR0529531

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*. CRC Press, Boca Raton.

HANSON, T. and JOHNSON, W. O. (2002). Modeling regression error with a mixture of Polya trees. *J. Amer. Statist. Assoc.* **97** 1020–1033. MR1951256

HEAGERTY, P. J. and KURLAND, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88** 973–985. MR1872214

HEAGERTY, P. J. and ZEGER, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statist. Sci.* **15** 1–26. MR1842235

JARA, A., HANSON, T. E., QUINTANA, F. A., MÜLLER, P. and ROSNER, G. L. (2011). DPpackage: Bayesian non-and semiparametric modelling in R. *J. Stat. Softw.* **40** 1.

KLEINMAN, K. P. and IBRAHIM, J. G. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics* 921–938.

KOTTAS, A., MÜLLER, P. and QUINTANA, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *J. Comput. Graph. Statist.* **14** 610–625. MR2170204

KYUNG, M., GILL, J. and CASELLA, G. (2009). Characterizing the variance improvement in linear Dirichlet random effects models. *Statist. Probab. Lett.* **79** 2343–2350. MR2556367

KYUNG, M., GILL, J. and CASELLA, G. (2010). Estimation in Dirichlet random effects models. *Ann. Statist.* **38** 979–1009. MR2604702

LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73** 805–811. MR0521328

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 963–974.

LANGE, N. and RYAN, L. (1989). Assessing normality in random effects models. *Ann. Statist.* **17** 624–642. MR0994255

LAVINE, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **20** 1222–1235. MR1186248

LEE, J., QUINTANA, F. A., MÜLLER, P. and TRIPPA, L. (2013). Defining predictive probability functions for species sampling models. *Statist. Sci.* **28** 209–222. MR3112406

LEON-NOVELO, L. G., ZHOU, X., BEKELE, B. N. and MÜLLER, P. (2010). Assessing toxicities in a clinical trial: Bayesian inference for ordinal data nested within categories. *Biometrics* **66** 966–974. MR2758233

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

LITIÈRE, S., ALONSO, A. and MOLENBERGHS, G. (2007). Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics* **63** 1038–1044, 1310. MR2414580

LIU, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Ann. Statist.* **24** 911–930. MR1401830

MAGDER, L. S. and ZEGER, S. L. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *J. Amer. Statist. Assoc.* **91** 1141–1151. MR1424614

MCCULLOCH, C. E. (2006). *Generalized Linear Mixed Models*. Wiley Online Library, New York.

MCCULLOCH, C. E. and NEUHAUS, J. M. (2011a). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statist. Sci.* **26** 388–402. MR2917962

MCCULLOCH, C. E. and NEUHAUS, J. M. (2011b). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics* **67** 270–279. MR2898839

MÜLLER, P. and QUINTANA, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19** 95–110. MR2082149

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804

NEUHAUS, J. M., HAUCK, W. W. and KALBFLEISCH, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79** 755–762.

NEUHAUS, J. M., MCCULLOCH, C. E. and BOYLAN, R. (2011). A note on type II error under random effects misspecification in generalized linear mixed models. *Biometrics* **67** 654–660. MR2829095

O'BRIEN, S. M. and DUNSON, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics* **60** 739–746. MR2089450

PIEPHO, H.-P. and MCCULLOCH, C. E. (2004). Transformations in mixed models: Application to risk analysis for a multienvironment trial. *J. Agric. Biol. Environ. Stat.* **9** 123–137.

TRIPPA, L., MÜLLER, P. and JOHNSON, W. (2011). The multivariate beta process and an extension of the Polya tree model. *Biometrika* **98** 17–34. MR2804207

VERBEKE, G. and MOLENBERGHS, G. (2009). *Linear Mixed Models for Longitudinal Data*. Springer, New York. MR2723365

WALKER, S. G. and MALLICK, B. K. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *J. Roy. Statist. Soc. Ser. B* **59** 845–860. MR1483219

WALLER, L. A., CARLIN, B. P., XIA, H. and GELFAND, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *J. Amer. Statist. Assoc.* **92** 607–617.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163

ZHANG, D. and DAVIDIAN, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* **57** 795–802. MR1859815