Statistics
in Medicine WILEY

A class of generalized linear mixed models adjusted for marginal interpretability

Jeffrey J. Gory | Peter F. Craigmile | Steven N. MacEachern

Department of Statistics, The Ohio State University, Columbus, Ohio, USA

Correspondence
Peter F. Craigmile, Department of Statistics, The Ohio State University, 404 Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210, USA.
Email: pfc@stat.osu.edu

Two popular approaches for relating correlated measurements of a non-Gaussian response variable to a set of predictors are to fit a *marginal model* using generalized estimating equations and to fit a *generalized linear mixed model* (GLMM) by introducing latent random variables. The first approach is effective for parameter estimation, but leaves one without a formal model for the data with which to assess quality of fit or make individual-level predictions for future observations. The second approach overcomes these deficiencies, but leads to parameter estimates that must be interpreted conditional on the latent variables. To obtain marginal summaries, one needs to evaluate an analytically intractable integral or use attenuation factors as an approximation. Further, we note an unpalatable implication of the standard GLMM. To resolve these issues, we turn to a class of marginally interpretable GLMMs that lead to parameter estimates with a marginal interpretation while maintaining the desirable statistical properties of a conditionally specified model and avoiding problematic implications. We establish the form of these models under the most commonly used link functions and address computational issues. For logistic mixed effects models, we introduce an accurate and efficient method for evaluating the logistic-normal integral.

KEYWORDS
conditional model, logistic-normal integral, marginal model

1 | INTRODUCTION

The generalized linear mixed model (GLMM) is commonly used to model dependent non-Gaussian responses.[1-3] The GLMM is a *conditional model* that relates a linear combination of p predictors $\mathbf{x}$ and q random effects $\mathbf{U}$ to a response Y through a *link function* g, traditionally via

$$E(Y|\mathbf{U} = \mathbf{u}) = h(\mathbf{x}^T\boldsymbol{\beta} + \mathbf{d}^T\mathbf{u}). \tag{1}$$

Here $\boldsymbol{\beta}$ is a p-vector of fixed effect parameters, $\mathbf{d}$ is a q-vector of covariates, and h is the inverse link function. The GLMM is completed by assuming a distributional form for the random effects $\mathbf{U}$ and the response Y given $\mathbf{U}$.

An alternative for dependent non-Gaussian responses is the *marginal model*, in which the marginal mean is defined via $E(Y) = h(\mathbf{x}^T\boldsymbol{\beta})$ and a covariance structure, usually with no distributional form, is explicitly assumed for the data. Estimation of $\boldsymbol{\beta}$, typically using *generalized estimating equations* (GEE), can be robust to a misspecification of the

covariance structure.[4-6] A marginal model allows for direct inference of *marginal* or *population-averaged* effects rather than conditional effects, but is typically not a fully specified generative model, making model assessment and individual-level prediction difficult, although not impossible.[7,8]

The interpretation of parameters in conditional and marginal models generally differs when the link function *g* is nonlinear.[9-13] Authors have attempted to address this discrepancy by finding a proportional relationship between the marginal and conditional parameters, through the use of a *bridge random effects distribution*[14-16] or approximately using *attenuation factors*.[9,10]

The choice between a marginal and conditional model often hinges on the goal of the analysis. Traditional advice is that if interest focuses exclusively on population-level effects, use a marginal model; if interest focuses on individual-level effects, use a conditional model. We believe that both marginal and conditional summaries are of interest in most analyses, and that a full analysis requires summaries and inferences at both levels. Recently, Hedeker et al[17] aggregate the results of a conditional analysis for binary outcomes across individuals to provide marginal inference. The authors use Gauss-Hermite quadrature to estimate population-averaged probabilities, which are then converted into population logits. This approach allows for the calculation of standard errors and can be extended to other link functions for binary data.

Another issue has received less attention. Kim and Kim[18] provide an example of a conditionally specified GLMM with normal random effects and a Bernoulli response where one can, *without any replication*, consistently estimate the variance of the random effects distribution. We find this phenomenon unpalatable. We refer to this example of false consistency[19] as the *Kim Paradox* and note that it occurs for a range of GLMMs. Hedeker et al's[17] strategy does not remedy this problem.

To produce a cohesive and complete analysis that enables individual-level predictions and diagnostics as well as marginal inference while avoiding the Kim Paradox, we turn to the marginal multilevel models of Heagerty[20] and Heagerty and Zeger.[21] Briefly, these marginally interpretable GLMMs satisfy the following definition.

Definition 1. A GLMM is marginally interpretable if and only if for all $\mathbf{x}^T\boldsymbol{\beta}$

$$\int E(Y|\mathbf{U}=\mathbf{u})f_{\mathbf{U}}(\mathbf{u})d\mathbf{u} = E(Y), \tag{2}$$

where $f_{\mathbf{U}}$ is the joint density of the random effects.

We supplement Heagerty[20] with a definition of the random effects distribution that leads to an expanded and novel parameterization that aids in understanding random effects in GLMMs.

The computational tractability of these models is of concern because parameter estimation for GLMMs can be difficult due to intractable integrals—this often leads to approximate inference.[22] Another possibility is to estimate the parameters using the so-called *h*-likelihood of Lee and Nelder[23] which does not require integration over the random effect, but also is not the same as the likelihood function for the data. Another alternative is to adopt a Bayesian framework and employ Markov chain Monte Carlo (MCMC) to sample from the posterior.[24-26] The adjustment in our proposed model can easily be incorporated into an MCMC algorithm and can be calculated efficiently. We develop a novel approach for evaluating the logistic-normal integral that allows for accurate and efficient calculation of the adjustment for a model with a logit link and normal random effects.

The next section contains a definition of a class of marginally interpretable GLMMs through an adjustment in the conditional model and introduces notation. Section 3 characterizes the properties of these models for the standard link functions and a variety of response distributions and random effect distributions. We discuss the interpretation of random effects in Section 4. Section 5 outlines an efficient Bayesian approach to fitting marginally interpretable GLMMs, including an algorithm for efficiently computing the logistic-normal integrals in logistic GLMMs. An example involving a teratological experiment on rats is given in Section 6. Section 7 contains a discussion of possible avenues for further research. Proofs, simulations, and other technical details appear in the Supplement.

2 | DEFINING A MARGINALLY INTERPRETABLE GLMM VIA AN ADJUSTMENT

Consider the following GLMM for a sample of responses Y_i ($i=1,\ldots,N$), a *p*-vector of predictors $\mathbf{x}_i$, and a *q*-vector of random effects $\mathbf{U}_i$. To produce a marginally interpretable GLMM we let the conditional mean be

TABLE 1 Form and existence of the adjustment for common link functions

Link function	Distribution of U_i	Form of $d_i^T a_i$	Closed form?
identity	mean exists and equals zero	zero	yes
log	exponential tails	independent of $x_i^T \beta$	yes
probit	Gaussian	linear in $x_i^T \beta$	yes
	non-Gaussian	nonlinear in $x_i^T \beta$	no
logit	bridge distribution	linear in $x_i^T \beta$	yes
	most other distributions	nonlinear in $x_i^T \beta$	no
complementary	bridge distribution	linear in $x_i^T \beta$	yes
log-log	most other distributions	nonlinear in $x_i^T \beta$	no
square root	restrictions on domain	nonlinear in $x_i^T \beta$	yes
reciprocal	$E\{1/(x_i^T \beta + d_i^T U_i)\}$ exists	see text	see text

$$E(Y_i | \mathbf{U}_i = \mathbf{u}) = h(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u} + \mathbf{d}_i^T \mathbf{a}_i), \quad i = 1, \dots, N, \tag{3}$$

where the $\mathbf{U}_i$ have joint density given by $f_{\mathbf{U}}(\mathbf{U}_i)$ and, conditional on the $\mathbf{U}_i$, the Y_i are mutually independent with density $f_{Y|\mathbf{U}}(Y_i | \mathbf{U}_i)$ for each i. Following Heagerty,[20] we preserve the mean for Y_i when the random effects are marginalized. To do so, we introduce an *adjustment* $\mathbf{d}_i^T \mathbf{a}_i$:

Definition 2. The *adjustment* $\mathbf{d}_i^T \mathbf{a}_i$, when it exists, is defined by

$$E(Y_i) = h(\mathbf{x}_i^T \boldsymbol{\beta}) = \int h(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u} + \mathbf{d}_i^T \mathbf{a}_i) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u}, \tag{4}$$

for all $\mathbf{x}_i^T \boldsymbol{\beta}$.

The following result holds (the proof follows naturally from Equation (2)).

Proposition 1. *A GLMM, as defined above, that satisfies Equation (4) is a marginally interpretable GLMM.*

3 | FUNCTIONAL FORMS OF THE ADJUSTMENT

The form of the adjustment $\mathbf{d}_i^T \mathbf{a}_i$ in a GLMM that satisfies Equation (4) is determined by the choice of link function and random effects distribution. The specific value of the adjustment typically depends on $\mathbf{x}_i^T \boldsymbol{\beta}$. This section describes the form of the adjustment for commonly used link functions. The results are summarized in Table 1. Additional details are provided in the Supplement.

3.1 | Identity link

Proposition 2. *For a model with an identity link, Proposition 2 holds as long as* $E(\mathbf{U}_i) = \mathbf{0}$. *Thus, a standard linear mixed effects model containing random effects with mean zero is a marginally interpretable GLMM.*

3.2 | Log link

Consider a GLMM with a log link function. In this case, $\mathbf{d}_i^T \mathbf{a}_i$ is defined by

$$\exp(\mathbf{x}_i^T \boldsymbol{\beta}) = \int \exp(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u} + \mathbf{d}_i^T \mathbf{a}_i) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u}. \tag{5}$$

Solving Equation (5) for $\mathbf{d}_i^T \mathbf{a}_i$ leads to the following proposition and corollary.

Proposition 3. *A GLMM with a log link satisfying Equations (3) and (4) is marginally interpretable if and only if $\mathbf{d}_i^T \mathbf{a}_i = -\log\{M_{\mathbf{U}}(\mathbf{d}_i)\}$, where $M_{\mathbf{U}}(\mathbf{d}_i) = \mathrm{E}\{\exp(\mathbf{d}_i^T \mathbf{U}_i)\}$ is the moment-generating function of $\mathbf{U}_i$ evaluated at $\mathbf{d}_i$.*

Corollary 1. *For a GLMM with a log link, an adjustment $\mathbf{d}_i^T \mathbf{a}_i$ that makes the model marginally interpretable exists if and only if $M_{\mathbf{U}}(\mathbf{d}_i)$ exists.*

These results constrain the set of possible random effects distributions that can be used with this model to those with exponential or lighter tails. Consequently, the t-distribution is not a valid random effects distribution for a marginally interpretable GLMM with a log link.

The situation simplifies for a simple random intercept model where $U_i \sim \mathrm{N}(0, \sigma^2)$ and $d_i = 1$ for all $i = 1, \ldots, N$. The adjustment $d_i a_i$ is merely a_i.

Corollary 2. *A model of the form $\mathrm{E}(Y_i | U_i = u) = \exp(\mathbf{x}_i^T \boldsymbol{\beta} + u + a_i)$ for which $U_i \sim \mathrm{N}(0, \sigma^2)$ is marginally interpretable if and only if $a_i = -\sigma^2/2$ for all i.*

In this situation, the adjustment depends only on the random effects variance σ^2 and is independent of $\mathbf{x}_i^T \boldsymbol{\beta}$. The adjustment is an additive offset on the log scale that decreases the conditional mean $\mathrm{E}(Y_i | U_i)$ by the same amount for $i = 1, \ldots, N$.

3.3 | Links with bounded domain

Several common link functions, including the probit, logit, and complementary log-log, are defined only on a bounded subset of the real line. In turn, the range of the corresponding inverse link function h is constrained to a bounded interval. For such links, the next theorem shows that one can *always* construct a model to be marginally interpretable when using a link function defined only on a bounded interval.

Theorem 1. *Consider a model of the form given in Equation (3) with inverse link function $h : \mathbb{R} \to [\ell, u]$. Suppose h is increasing and continuous, with $h(\eta) \to \ell$ as $\eta \to -\infty$ and $h(\eta) \to u$ as $\eta \to \infty$. Then an adjustment $\mathbf{d}_i^T \mathbf{a}_i$ that satisfies Equation (4) exists for any choice of random effects distribution.*

3.3.1 | Probit link

Let $g = \Phi^{-1}$ and $h = \Phi$, where Φ is the cumulative distribution function of a standard normal. The range of the inverse link function h is the bounded interval $(0, 1)$. Assuming normal random effects the adjustment has a closed form.

Proposition 4. *For $h = \Phi$ and $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$, a model of the form given by Equations (3) and (4) is marginally interpretable if and only if $\mathbf{d}_i^T \mathbf{a}_i = \{(1 + \mathbf{d}_i^T \boldsymbol{\Sigma} \mathbf{d}_i)^{1/2} - 1\}\mathbf{x}_i^T \boldsymbol{\beta}$.*

3.3.2 | Logit link

The link function $g(\mu) = \log\{\mu/(1 - \mu)\}$ is known as both the *logit link* and the *logistic link*. The inverse link is $h(\eta) = 1/(1 + \exp(-\eta))$ and the adjustment $\mathbf{d}_i^T \mathbf{a}_i$ is defined via

$$\frac{1}{1 + e^{-\mathbf{x}_i^T \boldsymbol{\beta}}} = \int \frac{1}{1 + e^{-(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u} + \mathbf{d}_i^T \mathbf{a}_i)}} f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u}. \tag{6}$$

The range of the inverse link function h is the bounded interval $(0, 1)$, and so the adjustment exists for any choice of random effects distribution. For most choices of random effects distribution the integral on the right-hand side of Equation (6) is analytically intractable and there is no closed-form solution for $\mathbf{d}_i^T \mathbf{a}_i$. One exception is the *bridge distribution* derived by Wang and Louis.[14] Provided the model contains just a single random intercept, the bridge distribution leads to a closed-form solution for $\mathbf{d}_i^T \mathbf{a}_i$ that is linear as a function of $\mathbf{x}_i^T \boldsymbol{\beta}$.

FIGURE 1 Plot of the adjustment a as a function of the fixed portion of the model κ for various values of σ in a model for which $E(Y|U=u) = h(\kappa + u + a)$, h is the inverse logit function, and $U \sim N(0, \sigma^2)$
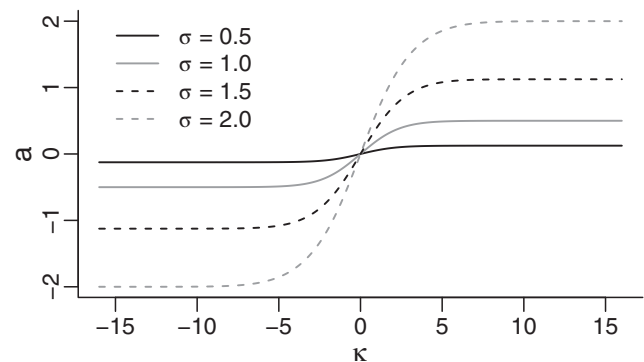
It is more common to assume that $f_{\mathbf{U}}$ is a normal density. Section 5.1 provides details on how to compute the adjustment in this situation accurately and efficiently. The direction and magnitude of the adjustment depend on $\mathbf{x}_i^T\boldsymbol{\beta}$. The adjustment $\mathbf{d}_i^T\mathbf{a}_i$ is typically a nonlinear function of $\mathbf{x}_i^T\boldsymbol{\beta}$. This is illustrated in Figure 1 for the case of a single normal random intercept $U_i \sim N(0, \sigma^2)$. In light of the forthcoming Proposition 8, the same picture would apply for q normal random effects $\mathbf{U}_i \sim N_q(\mathbf{0}, \boldsymbol{\Sigma})$ if we were to replace σ^2 with $\mathbf{d}_i^T\boldsymbol{\Sigma}\mathbf{d}_i$. The direction of the adjustment is driven by the convexity of the inverse link $h(\eta)$ – it is convex for $\eta < 0$ and concave for $\eta > 0$. Hence, the adjustment is negative when $\mathbf{x}_i^T\boldsymbol{\beta} < 0$ and positive when $\mathbf{x}_i^T\boldsymbol{\beta} > 0$. It is also evident from Figure 1 that the magnitude of $\mathbf{d}_i^T\mathbf{a}_i$ is increasing in both σ^2 and $|\mathbf{x}_i^T\boldsymbol{\beta}|$. For very large $\mathbf{x}_i^T\boldsymbol{\beta}$, we have the following:

Proposition 5. *For $h = \text{logit}^{-1}$ and $\mathbf{U}_i \sim N_q(\mathbf{0}, \boldsymbol{\Sigma})$, the value of $\mathbf{d}_i^T\mathbf{a}_i$ that allows a model of the form given by Equation (3) to satisfy Equation (4) converges to $\frac{1}{2}\mathbf{d}_i^T\boldsymbol{\Sigma}\mathbf{d}_i \times \text{sign}(\mathbf{x}_i^T\boldsymbol{\beta})$ as $|\mathbf{x}_i^T\boldsymbol{\beta}| \to \infty$.*

3.3.3 | Complementary log-log link

Let $g(\mu) = \log\{-\log(1-\mu)\}$ denote the complementary log-log link function with inverse link function $h(\eta) = 1 - \exp\{-\exp(\eta)\}$. Again the range of the inverse link function h is the bounded interval $(0,1)$, and so the adjustment exists for any choice of random effects distribution. For this link function, the adjustment $\mathbf{d}_i^T\mathbf{a}$ is defined via

$$1 - \exp\{-\exp(\mathbf{x}_i^T\boldsymbol{\beta})\} = \int [1 - \exp\{-\exp(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{d}_i^T\mathbf{u} + \mathbf{d}_i^T\mathbf{a}_i)\}]f_{\mathbf{U}}(\mathbf{u})d\mathbf{u},$$

or equivalently

$$\exp\{-\exp(\mathbf{x}_i^T\boldsymbol{\beta})\} = \int \exp\{-\exp(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{d}_i^T\mathbf{u} + \mathbf{d}_i^T\mathbf{a}_i)\}f_{\mathbf{U}}(\mathbf{u})d\mathbf{u}. \tag{7}$$

The right-hand side of Equation (7) is is analytically intractable and the adjustment does not have a closed form except for the case of a log-positive stable random effects *bridge distribution*.[14]

3.4 | Link functions with other range restrictions

A number of common link functions map into a proper subset of the real line and therefore require conditions on $\mathbf{x}_i^T\boldsymbol{\beta}$ to ensure that the model is defined. For examples, the square root transformation is typically defined to have nonnegative range, and no real number has a reciprocal of zero. Additive random effects with support on the entire real line could lead to problems in models with these link functions because $\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{d}_i^T\mathbf{U}_i$ can fall outside the domain of the inverse link function h. Thus, special care must be taken with these link functions, as described below.

3.4.1 | Square root link

Consider a GLMM with link function $g(\mu) = \mu^{1/2}$ and inverse link function $h(\eta) = \eta^2$. For such a model, one typically includes the restriction that $\mathbf{x}_i^T \boldsymbol{\beta} \geq 0$. Including the adjustment, we adopt the restriction that $\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{a}_i \geq 0$. The adjustment is defined such that

$$(\mathbf{x}_i^T \boldsymbol{\beta})^2 = \int (\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u} + \mathbf{d}_i^T \mathbf{a}_i)^2 f_\mathbf{U}(\mathbf{u}) d\mathbf{u}. \tag{8}$$

If we assume $\mathrm{E}(\mathbf{U}_i) = \mathbf{0}$, then Equation (8) reduces to

$$(\mathbf{x}_i^T \boldsymbol{\beta})^2 = (\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{a}_i)^2 + \mathrm{Var}(\mathbf{d}_i^T \mathbf{U}_i), \tag{9}$$

which is quadratic in $\mathbf{d}_i^T \mathbf{a}_i$ and leads to the following result:

Proposition 6. *For $h(\eta) = \eta^2$ and $\mathrm{E}(\mathbf{U}_i) = \mathbf{0}$, a model of the form given by (3) and (4) subject to the restriction that $\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{a}_i \geq 0$ is a marginally interpretable model if and only if $\mathbf{d}_i^T \mathbf{a}_i = -\mathbf{x}_i^T \boldsymbol{\beta} + \{(\mathbf{x}_i^T \boldsymbol{\beta})^2 - \mathrm{Var}(\mathbf{d}_i^T \mathbf{U}_i)\}^{1/2}$.*

Thus, for a model with a square root link, $\mathbf{d}_i^T \mathbf{a}_i$ is a nonlinear function of $\mathbf{x}_i^T \boldsymbol{\beta}$ and is only defined when $\mathbf{x}_i^T \boldsymbol{\beta} \geq \{\mathrm{Var}(\mathbf{d}_i^T \mathbf{U}_i)\}^{1/2}$. If the random effects variance is too large, the model cannot be fit.

3.4.2 | Reciprocal link

Consider a model for which $g(\mu) = h(\mu) = 1/\mu$. For a fixed effects model with this link, one typically includes the restriction that $\mathbf{x}_i^T \boldsymbol{\beta} > 0$. When a random intercept U_i is included in the model, the fact that h tends to infinity as its argument approaches zero forces us to also include restrictions on the distribution of U_i. In particular, we want a model with

$$\frac{1}{\mathbf{x}_i^T \boldsymbol{\beta}} = \int \frac{1}{\mathbf{x}_i^T \boldsymbol{\beta} + u} f_U(u) du. \tag{10}$$

Therefore, f_U must be defined such that the integral in Equation (10) exists. Given this restriction, we move away from models of the form given in (3). Rather than adjusting the location of the random effects based on each individual's observed covariates, we instead choose to alter the shape of the distribution of the random effect based on the observed covariates.

One distribution for U_i that allows us to satisfy Equation (10) is a shifted gamma distribution. Specifically, let $\mathbf{x}_i^T \boldsymbol{\beta} + U_i$ follow a gamma distribution with shape parameter α_i and rate parameter β_i so that $\mathrm{E}(U_i) = \alpha_i \beta_i - \mathbf{x}_i^T \boldsymbol{\beta}$. Then the integral in Equation (10) is equal to $\{\beta_i(\alpha_i - 1)\}^{-1}$, and we have $\mathbf{x}_i^T \boldsymbol{\beta} = \beta_i(\alpha_i - 1)$. By placing additional conditions on α_i and β_i, one can determine the appropriate gamma distribution for U_i for each $\mathbf{x}_i^T \boldsymbol{\beta}$. Alternatively, one could let $\mathbf{x}_i^T \boldsymbol{\beta} + U_i$ follow an inverse gamma distribution with parameters α_i and β_i, and be constrained by the relationship $\mathbf{x}_i^T \boldsymbol{\beta} = (\alpha_i \beta_i)^{-1}$. In either case it is the shape, not the location, of the random effects distribution that varies with $\mathbf{x}_i^T \boldsymbol{\beta}$ in this marginally interpretable model.

4 | THE KIM PARADOX AND INTERPRETATION OF RANDOM EFFECTS

Random effects provide a means of introducing dependence and overdispersion into a generalized linear model. In a simple model, this degree of dependence is determined by the distribution of the random effects, most typically by its variance. For example, consider the following hierarchical model:

$$U_i \sim F_{\sigma^2};$$
$$Y_i | U_i = u_i \sim \mathrm{Bernoulli}(u_i),$$

where $i = 1, \ldots, N$ and F_{σ^2} is an arbitrary distribution on $(0, 1)$ with mean μ and variance σ^2. To obtain the marginal distribution for Y_i, we integrate over U_i. Irrespective of σ^2, the resulting marginal distribution for Y_i is Bernoulli(μ). No matter how many of these Bernoullis are collected, there is no replication tied to a single random effect. The data contain no information about σ^2. It is clear that in a simple model such as this, multiple observations tied to a single random effect are required for consistent estimation of σ^2.

Kim and Kim[18] challenged this reasoning with an astonishing result for a conventional Bernoulli GLMM. They showed that the maximum likelihood (ML) estimator is strongly consistent for the random effects variance σ^2, without replication. We call this result the *Kim Paradox* and state it below in a form that, for simplicity of presentation, is not as general as the form presented in Kim and Kim.[18]

The Kim Paradox: *With no replication, one can estimate the random effects variance consistently.* Let parameters β_0, $\beta_1 \neq 0$, and $\tau^2 > 0$ be fixed and known, and let $X_i \sim N(0, \tau^2)$ and $U_i \sim \text{Uniform}(-c, c)$ $(c > 0, i = 1, 2, \ldots)$ be independent sequences of random variables. With h being the inverse logit link, define the conditionally independent sequence $Y_i | X_i = x_i, U_i = u_i \sim \text{Bernoulli}\{h(\beta_0 + \beta_1 x_i + u_i)\}$ for $i = 1, 2, \ldots$. Then $\hat{\sigma}^2$, the ML estimator of the random effects variance σ^2, is consistent.

The Kim Paradox arises from the fact that the marginal mean in a conventional GLMM is distorted by the random effects in such a fashion that there is a one-to-one mapping between $E[Y_i]$ and the random effects variance σ^2. For fixed values of β_0 and β_1, the defining relation $E[Y_i] = E[h(\beta_0 + \beta_1 x_i + U_i)]$ corresponds to a specific value of σ^2. This one-to-one mapping, along with a rich enough set of x_i, ensures that the marginal mean functions are identifiable and consistency of $\hat{\sigma}^2$ follows.

The Kim Paradox is resolved through use of GLMMs of the form given by Equations (3) and (4). The marginal mean in such a model is unaffected by a change in σ^2. Consequently, without replication, the data contain no information about the random effects variance, and one cannot create a consistent estimator of σ^2. This is stated more formally in the following proposition:

Proposition 7. *If $Y_i | U_i$ is Bernoulli-distributed and we have a marginally interpretable GLMM of the form given by Equations (3) and (4) with independently distributed random intercepts U_i, $i = 1, \ldots, N$, then the marginal density of Y_i does not depend in any way on the distribution of U_i.*
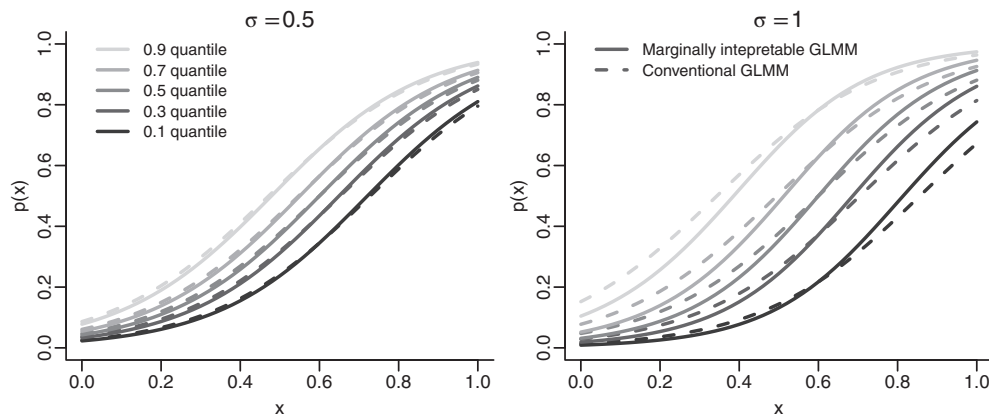
4.1 | Interpretation of random effects

Traditionally in GLMMs, one views each realization of a random effect as a single value that applies to all units in a group of observations sharing that random effect. For example, each random intercept in a conventional random intercepts model corresponds to a shift, and all units with the same random intercept share the same shift. Mathematically, the link-transformed center moves from $\mathbf{x}_i^T \boldsymbol{\beta}$ to $\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{u}$. This common shift applies on the link-transformed scale. When the curvature of the link function is not uniform across the range of the covariates, returning to the data scale differentially distorts these shifts. To preserve the marginal mean of Y_i, it is necessary to allow a different value of the random effect for different units.

This view leads us to the adjustment in Equation (3). This *location shift* of the random effects distribution preserves the marginal mean. We cease to conceptualize each realization of a random effect as a single value shared by all observations in the same group or cluster. Rather, observations sharing the same random effect are viewed as having the same quantile of a location family of distributions.

When the adjustment $\mathbf{d}_i^T \mathbf{a}_i$ depends on $\mathbf{x}_i^T \boldsymbol{\beta}$, the value associated with the random effect for a particular observation depends on the covariates for that observation. Thus, each realization of a random effect represents a set of potential values with the value for a specific observation determined by $\mathbf{x}_i^T \boldsymbol{\beta}$. This allows one to separate systematic variation in the population, captured by $\mathbf{x}_i^T \boldsymbol{\beta}$, from individual-level variation, captured by $\mathbf{d}_i^T \mathbf{U}_i + \mathbf{d}_i^T \mathbf{a}_i$. With this formulation, the individual-level variation still models heterogeneity among clusters and accounts for possible overdispersion, just as it does in a conventional GLMM.

As an illustration, in Figure 2, we compare the individual level variations for marginally interpretable and conventional simple linear logistic GLMMs with a normal random effect. Suppose that

$$Y | \boldsymbol{\beta}^*, U \sim \text{Binomial}\{m, E(p(x) | \boldsymbol{\beta}^*, U)\}.$$

FIGURE 2 For a marginally interpretable simple linear logistic GLMM with a normal random effect, the solid lines denote different quantile functions for the success probability $p(x)$ in terms of the covariate x. The dashed lines denote the quantile functions for a conventional simple linear logistic GLMM. The standard deviation of the random effect is $\sigma = 0.5$ in the left panel and $\sigma = 1$ in the right panel

Then for the marginally interpretable model, letting h denote the inverse logit function, we model the conditional mean as

$$E(p(x)|\boldsymbol{\beta}, U) = h(\beta_0 + \beta_1 x + U + a), \quad (11)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is the vector of fixed effects parameters and the adjustment a is defined by Equation (6). We assume that $U \sim N(0, \sigma^2)$ and allow x to vary from 0 to 1. In the conventional GLMM we remove the adjustment, a, from Equation (11). We let $\beta_0 = -3$ and $\beta_1 = 5$, and calculate the 0.1, 0.3, 0.5, 0.7, and 0.9 quantiles of the success probability $p(x)$ as a function of x for the marginally interpretable GLMM (the solid lines) and the conventual GLMM (the dashed lines). In the left panel, the standard deviation of the random effect is $\sigma = 0.5$ and in the right panel we have $\sigma = 1$.

For the marginally interpretable model, regardless of the value of σ, it can be shown that adjustment a is equal to zero when $x = 0.6$, becomes positive as x increases above 0.6, and becomes negative as x decreases below 0.6. When $x \neq 0.6$, the magnitude of the adjustment is higher for $\sigma = 1$ than for $\sigma = 0.5$. Thus the individual level effects, as summarized by the quantiles of $p(x)$, are only the same in the marginally interpretable and conventional GLMMs when $x = 0.6$. As x gets closer to 0 or 1, or when the standard deviation of the random effect increases, the quantiles vary greatly between the marginally interpretable and conventional GLMMs. Only in a marginally interpretable GLMM is the population mean of Y as a function of x independent of σ; it is always $\beta_0 + \beta_1 x$. As σ varies in the conventional GLMM, the population mean of Y as a function of x is a nontrivial function of β_0, β_1, and σ. Again this is a motivating reason to use marginally interpretable models in practice.

Another situation where this new formulation of a random effect arises is in a multilevel model where the covariates $\mathbf{x}_i$ differ across individual units in the same group or cluster. For example, students within the same class are liable to have different characteristics. Depending on the choice of link function, the adjustment could be different for different units in the same group. Thus, within a single group, the shift in the mean response associated with the random effect for that group varies with the measured covariates for the individual units sharing that random effect.

5 | COMPUTATIONAL STRATEGIES FOR STATISTICAL INFERENCE

To fit the mean-preserving GLMM, we need to evaluate the integral on the right hand side of Equation (4). The difference between a model of this form and a conventionally specified conditional GLMM is the adjustment $\mathbf{d}_i^T \mathbf{a}_i$. The adjustment typically depends on both $\mathbf{x}_i^T \boldsymbol{\beta}$ and the scale of the random effects distribution. The central tasks for computation are calculation of $\mathbf{d}_i^T \mathbf{a}_i$ and integration over the random effects. These computations are repeated a modest number of times for a classical (likelihood-based) fit and many times for a Bayesian fit via MCMC.

The computational expense added to the algorithm by including the adjustment is driven by how difficult it is to compute $\mathbf{d}_i^T \mathbf{a}_i$. The difficulty varies based on the choice of link function and random effects distribution, and is greater in situations lacking a closed-form solution for $\mathbf{d}_i^T \mathbf{a}_i$. The most common situation without a closed-form solution is a

GLMM with a logit link and normal random effects, which we discuss in Section 5.1. First, we present a result that applies to any GLMM with multivariate normal random effects. The result allows one to simplify computation by reducing a q-dimensional integral to a univariate one.

Proposition 8. *When* $\mathbf{U}_i \sim \mathrm{N}_q(\mathbf{0}, \mathbf{\Sigma})$, *if the q-dimensional integral*

$$\int_{\mathbb{R}^q} h(\kappa + \mathbf{d}^T\mathbf{u} + a)\left(\frac{1}{2\pi}\right)^{\frac{q}{2}} |\mathbf{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{u}^T\mathbf{\Sigma}^{-1}\mathbf{u}\right) d\mathbf{u}$$

exists, then it can be expressed as a univariate integral of the form

$$\int_{\mathbb{R}} h(\kappa + v + a)\frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}v^2\right) dv, \quad \text{with } \tau^2 = \mathbf{d}^T\mathbf{\Sigma}\mathbf{d}.$$

5.1 | Efficient and accurate evaluation of logistic-normal integrals

The logistic-normal integral is analytically intractable, but several numerical approaches exist for evaluating it. Popular approaches include Gauss-Hermite quadrature and adaptive quadrature. The accuracy of Gauss-Hermite quadrature degrades for the logistic-normal integral when the variance of the normal distribution is large.[27] Consequently, specialized algorithms have been developed to evaluate the logistic-normal integral.[27,28]

We have developed a novel and efficient method for calculating $\mathbf{d}_i^T\mathbf{a}_i$ under for the logistic-normal case. Our algorithm for evaluating the logistic-normal integral exploits a recursive formula developed by Pirjol[29] that provides an exact solution on a specifically defined, evenly spaced grid. Pirjol[29] showed that the integral

$$\varphi(\mu, \sigma^2) = \int \frac{1}{1 + e^w} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(w - \mu)^2\right\} dw, \tag{12}$$

satisfies the recursion

$$\varphi(\mu + \sigma^2, \sigma^2) = e^{-\mu - \frac{\sigma^2}{2}}\{1 - \varphi(\mu, \sigma^2)\}, \tag{13}$$

with $\varphi(0, \sigma^2) = 1/2$. We combine this expression with the approximation of Monahan and Stefanski[28] to develop a novel, hybrid approach for approximating the logistic-normal integral. The Supplement describes the algorithm fully and documents its performance.

5.2 | Computing the adjustment

To compute the adjustment in a marginally interpretable GLMM, one must solve Equation (4) for $\mathbf{d}_i^T\mathbf{a}_i$. The adjustment is a deterministic function of $\mathbf{x}_i^T\boldsymbol{\beta}$ and the parameters characterizing $f_{\mathbf{U}}$, which we denote by $\boldsymbol{\alpha}$. When fitting a marginally interpretable GLMM, the adjustment must be included in the calculation of the likelihood. Many model-fitting algorithms are iterative in nature and, at the point in the algorithm at which the adjustment must be included, current estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ will be available to plug into Equation (4) to solve for $\mathbf{d}_i^T\mathbf{a}_i$. When the adjustment can be written in closed form, such as when the model has a probit link and Gaussian random effects or when the model has a log link, computing $\mathbf{d}_i^T\mathbf{a}_i$ is straightforward and has little impact on the amount of computation required to fit the model. When the integral in Equation (4) is analytically intractable and the adjustment does not have a closed form, computing $\mathbf{d}_i^T\mathbf{a}_i$ is considerably more challenging and can add substantial computational expense to fitting the model.

We now introduce an algorithm for efficiently computing $\mathbf{d}_i^T\mathbf{a}_i$ for the common case of a model with a logit link and Gaussian random effects. We present the algorithm for the case of a univariate normal random intercept, but as a consequence of Proposition 8, the strategy described here also applies to a model with multivariate normal random effects. To simplify notation let $\kappa = \mathbf{x}_i^T\boldsymbol{\beta}$ and $a = \mathbf{d}_i^T\mathbf{a}_i$. Further, suppose $h = \mathrm{logit}^{-1}$ and $U_i \sim \mathrm{N}(0, \sigma^2)$. Then Equation (4) reduces to Equation (6) which can be expressed as

$$\int \frac{1}{1 + e^{-w}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\{w - (\kappa + a)\}^2\right] dw.$$

In the notation of (12), this is equivalent to

$$1 - \int \frac{1}{1 + e^{w}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\{w - (\kappa + a)\}^2\right] dw = 1 - \varphi(\kappa + a, \sigma^2).$$

Thus, given κ and σ^2, the equation we must solve for a is

$$h(\kappa) = 1 - \varphi(\kappa + a, \sigma^2). \tag{14}$$

Since the problem is symmetric, we need only consider the case of $\kappa > 0$. When $\kappa < 0$, the adjustment has the same magnitude but opposite sign as if $\kappa = |\kappa|$. Several different techniques, such as binary segmentation or a Newton-Raphson algorithm, can be used to solve Equation (14) for a. Any such technique requires evaluation of $\varphi(\kappa + a^*, \sigma^2)$ for several potential values a^* of the adjustment a. Although we can use the recursion (13) to calculate $\varphi(t\sigma^2, \sigma^2)$ exactly for any integer t, it is unlikely that the desired a will be such that $\kappa + a$ is an integer multiple of σ^2. Thus, we have need of an approximate numerical integration procedure. Since the function $\varphi(\cdot, \cdot)$ is decreasing in its first argument, we can use the recursion (13) to quickly identify an interval of length σ^2 in which $\kappa + a$ must reside. By narrowing our search for the correct value of a to such an interval, we reduce the required number of evaluations of $\varphi(\kappa + a^*, \sigma^2)$.

Our algorithm for solving Equation (14) proceeds as follows. We start with $t = 0$ and increment t by one until $1 - \varphi(t\sigma^2, \sigma^2) \leq h(\kappa) < 1 - \varphi((t + 1)\sigma^2, \sigma^2)$. We use t^* to denote the value of t for which this inequality holds, and note that the value of a satisfying Equation (14) must lie in the interval $[t^*\sigma^2 - \kappa, (t^* + 1)\sigma^2 - \kappa)$. We then employ binary segmentation, implemented using the the `uniroot` function in R,[30] to search within this interval for the appropriate value of a. To evaluate $\varphi(\kappa + a^*, \sigma^2)$ for $\kappa + a^* \in [t^*\sigma^2, (t^* + 1)\sigma^2)$ we use the hybrid approach introduced in Section 5.1 and Section S2 of the Supplement that combines the approximation of Monahan and Stefanski[28] with the recursive result of Pirjol.[29] Specifically, we use the Monahan and Stefanski method to compute $\varphi(\kappa + a^* - t^*\sigma^2, \sigma^2)$ and then apply Equation (13) t^* times to obtain $\varphi(\kappa + a^*, \sigma^2)$.

5.3 | Improving MCMC mixing

A challenge in using MCMC to sample from a high-dimensional posterior density is poor mixing. Due to the large number of unknown parameters, the only proposals that tend to get accepted are those representing relatively small steps from the current state of the Markov chain. Consequently, there is substantial autocorrelation in the Markov chain and it is necesssary to run the algorithm for an exceedingly long time to generate a representative sample from the target posterior. One way to improve mixing is to sample the parameters in blocks, but this may not always be enough. In a GLMM, when there are many realizations of the random effects, the latent random variables $\mathbf{U}_i$ ($i = 1, \ldots, N$) can dominate the likelihood and cause very few proposed $\boldsymbol{\beta}^{prop}$ to be accepted.

In a marginally interpretable GLMM, the conditional mean is

$$E(Y_i | \mathbf{U}_i = \mathbf{u}_i) = h(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{d}_i^T \mathbf{u}_i + \mathbf{d}_i^T \mathbf{a}_i), \quad i = 1, \ldots, N,$$

where $\mathbf{d}_i^T \mathbf{a}_i$ is a deterministic function of $\mathbf{x}_i^T \boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, the parameters characterizing the random effects distribution. To preserve the mean, if $\mathbf{x}_i^T \boldsymbol{\beta}$ is larger then $\mathbf{d}_i^T \mathbf{u}_i + \mathbf{d}_i^T \mathbf{a}_i$ must be smaller, and vice versa. Thus, if the current state $\mathbf{U}_i^{(t)}$ of the latent variables is compatible with the current state $\boldsymbol{\beta}^{(t)}$ of the fixed effects parameters, it is difficult to find a proposal $\boldsymbol{\beta}^{prop}$ that yields a greater value of the likelihood given $\mathbf{U}_i^{(t)}$. Thus, $\boldsymbol{\beta}^{prop}$ is rarely accepted, those $\boldsymbol{\beta}^{prop}$ that do get accepted tend to represent small steps from the current state $\boldsymbol{\beta}^{(t)}$, and the chain mixes slowly.

To overcome this problem, we suggest the following solution. Along with each proposed $\boldsymbol{\beta}^{prop}$, simultaneously propose random effects $\mathbf{U}_i^{prop}$ that are consistent with the proposed $\boldsymbol{\beta}^{prop}$, while respecting the effect upon the adjustment. That is, given a proposal $\boldsymbol{\beta}^{prop}$ for the fixed effects parameters, also propose random effects $\mathbf{U}_i^{prop}$ in such a manner that $\mathbf{x}_i^T \boldsymbol{\beta}^{prop} + \mathbf{d}_i^T \mathbf{U}_i^{prop} + \mathbf{d}_i^T \mathbf{a}_i^{prop} = \mathbf{x}_i^T \boldsymbol{\beta}^{(t)} + \mathbf{d}_i^T \mathbf{U}_i^{(t)} + \mathbf{d}_i^T \mathbf{a}_i^{(t)}$. (Here, $\mathbf{d}_i^T \mathbf{a}_i^{prop}$ denotes the value of the adjustment calculated using $\mathbf{x}_i^T \boldsymbol{\beta}^{prop}$ and $\boldsymbol{\alpha}$, and $\mathbf{d}_i^T \mathbf{a}_i^{(t)}$ denotes the value of the adjustment calculated using $\mathbf{x}_i^T \boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\alpha}$.) Then the conditional mean,

$h(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{d}_i^T\mathbf{U}_i + \mathbf{d}_i^T\mathbf{a}_i)$, is unaffected by the proposal and there is no net impact on the likelihood. The decision to accept or reject the proposed $\boldsymbol{\beta}^{prop}$ and $\mathbf{U}_i^{prop}$ is based entirely on the prior distribution for $\boldsymbol{\beta}$ and the random effects distribution assumed for $\mathbf{U}_i$. A formal update of $\mathbf{U}_i$ is still required after this simultaneous update of $\boldsymbol{\beta}$ and $\mathbf{U}_i$, but this strategy improves the acceptance rate for $\boldsymbol{\beta}$, encourages larger moves for $\boldsymbol{\beta}$, and thereby facilitates faster mixing. An example illustrating this approach is given in the Supplement.

6 | RAT TERATOLOGY EXAMPLE

We analyze data from a teratological experiment on rats conducted by Weil.[31] A group of 16 female rats was fed a diet containing a chemical agent during pregnancy and lactation, while another group of 16 female rats was fed a control diet. Counts were made of the number of pups in each litter to survive four days from birth and to survive the 21-day lactation period. Interest lies in the proportion of pups to survive 21 days among those alive after four days. To clearly distinguish between different parameterizations, let $\boldsymbol{\beta}^*$ denote the marginal parameters in a marginally interpretable GLMM and $\boldsymbol{\beta}$ denote the cluster-specific parameters in a conventional GLMM.

For $i = 1, 2$ and $j = 1, \ldots, 16$, we denote the number of pups in litter j receiving treatment i to survive 4 days by m_{ij}, the number of pups to survive 21 days by Y_{ij}, and the proportion of pups to survive 21 days by $p_{ij} = Y_{ij}/m_{ij}$. We include in our model a fixed effect for the treatment ($x_1 = 1$ for the treatment group, $x_2 = -1$ for the control group) and random effects for litter (denoted U_{ij}). Earlier analyses of these data established that there is more between-litter heterogeneity in the treatment group than in the control group.[15,21,32] We therefore allow different random effects variances for the two treatment groups. We assume $U_{ij} \overset{ind}{\sim} N(0, \sigma_i^2)$ and $Y_{ij}|\boldsymbol{\beta}^*, U_{ij} \overset{ind}{\sim} \text{Binomial}\{m_{ij}, E(p_{ij}|\boldsymbol{\beta}^*, U_{ij})\}$, define h as the inverse logit function, and model the conditional mean as

$$E(p_{ij}|\boldsymbol{\beta}^*, U_{ij}) = h(\beta_0^* + \beta_1^* x_i + U_{ij} + a_i),$$

where $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)^T$ is the vector of fixed effects parameters and a_i is the adjustment that ensures the model is marginally interpretable. Since all litters in the same treatment group have the same covariate x_i and the same random effects variance σ_i^2, they also have the same adjustment a_i.

We adopt a Bayesian approach and use MCMC to sample from the posterior distribution of the unknown parameters in our model. Our prior distributions for β_0^*, β_1^*, $\log(\sigma_1^2)$, and $\log(\sigma_2^2)$ are $N(0, 2)$, $N(0, 1)$, $N(-1/2, 1)$, and $N(-1/2, 1)$, respectively. These priors were chosen such that the prior distribution for the expected survival rate for each of the two treatment groups is approximately uniform over the interval $(0, 1)$. To sample from our target posterior we iteratively update blocks of parameters using Metropolis steps. We first update $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)^T$, then $\boldsymbol{\alpha} = (\sigma_1^2, \sigma_2^2)^T$, and finally $\mathbf{U} = (U_{1,1}, \ldots, U_{1,16}, U_{2,1}, \ldots, U_{2,16})^T$. When necessary, we compute the adjustment a_i using the technique described in Section 5. This MCMC algorithm was carried out both with the adjustment included in the model and without it. Each chain was run for 1 010 000 steps, with the first 10 000 steps discarded as burn-in and every 100th step thereafter retained for the final sample. This resulted in 10 000 draws from the posterior distribution for each model. Additional details regarding the algorithm are provided in the Supplement.

We first compared the fits of the marginally interpretable model and an analogous conventional GLMM using the deviance information criterion (DIC).[33,34] The DIC for the marginally interpretable model is 87.7, whereas the DIC for the conventional model it is 99.0. By this metric, the marginally interpretable model fits to the data better. Exploring the model fits further, Table 2 provides posterior means and standard deviations for the parameters in both the marginally interpretable model and the conventional GLMM. Figure 3 displays kernel density estimates based on the posterior samples for the two models. Notably, the tail area above zero for β_1^*, which corresponds to the treatment effect in the marginally interpretable model, is 0.016. This is considerably less than the tail area of 0.088 for β_1 in the conventional GLMM. Thus, many would draw different conclusions about the importance of the treatment effect using the two different models. Indeed, Bayes factors for a test of no treatment effect ($H_0 : \beta_1^* = 0$ or $H_0 : \beta_1 = 0$), computed using the Savage-Dickey density ratio,[35,36] come in at 0.45 for the marginally interpretable model and 1.37 for the conventional GLMM. This confirms the disparity, as the Bayes factor for the conventional model favors the null hypothesis whereas the Bayes factor for the marginally interpretable model does not.

Another approach for investigating the treatment effect is to compare the expected 21-day survival rates between the two treatment groups. For the marginally interpretable model, the expected proportion of rat pups to survive 21

Parameter	Marginally Interpretable GLMM	Conventional GLMM
β_0^*, β_0	1.62 (0.24)	1.92 (0.30)
β_1^*, β_1	−0.49 (0.23)	−0.38 (0.29)
σ_1	1.55 (0.42)	1.58 (0.42)
σ_2	0.74 (0.30)	0.72 (0.28)

TABLE 2 Posterior means of the unknown parameters in the model for the rat teratology data of Weil[31] (with corresponding posterior standard deviations in parentheses)
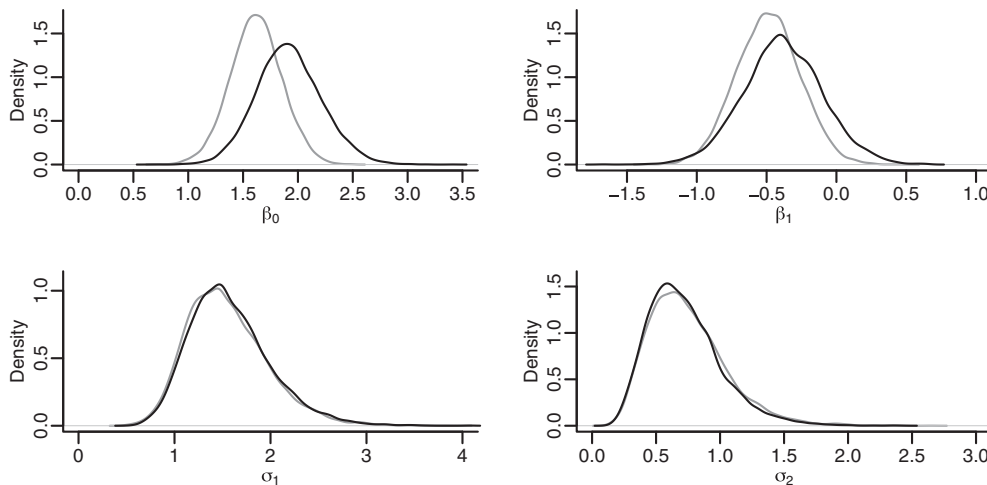
FIGURE 3 Kernel density estimates of the posterior densities for the unknown parameters in the model for the rat teratology data of Weil;[31] estimates obtained from the marginally interpretable model are in gray while those obtained from the conventional GLMM are in black
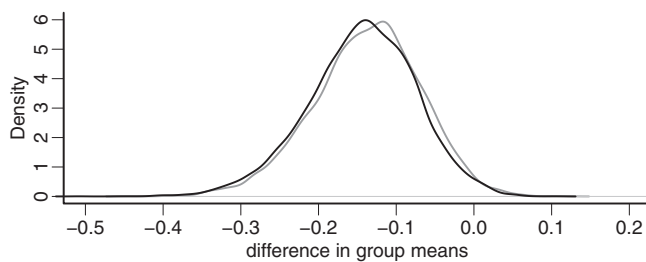
FIGURE 4 Kernel density estimates of the posterior density of the difference in the expected 21-day survival rate between the two treatment groups based on the marginally interpretable model (black) and the conventional GLMM (gray)

days among those alive after 4 days is $E[p|\boldsymbol{\beta}^*, \boldsymbol{\alpha}, x = 1] = h(\beta_0^* + \beta_1^*)$ for the treatment group and $E[p|\boldsymbol{\beta}^*, \boldsymbol{\alpha}, x = -1] = h(\beta_0^* - \beta_1^*)$ for the control group. For the conventional GLMM, the same expectation is $E[p|\boldsymbol{\beta}, \boldsymbol{\alpha}, x = 1] = \int h(\beta_0 + \beta_1 + u)f_U(u)du$ for the treatment group and $E[p|\boldsymbol{\beta}, \boldsymbol{\alpha}, x = -1] = \int h(\beta_0 - \beta_1 + u)f_U(u)du$ for the control group. Note that the parameters $\boldsymbol{\alpha}$ enter this expression through the random effects distribution f_U. Interest lies in whether or not the quantity $E[p|\boldsymbol{\beta}, \boldsymbol{\alpha}, x = 1] - E[p|\boldsymbol{\beta}, \boldsymbol{\alpha}, x = -1]$ (or the same quantity with $\boldsymbol{\beta}^*$ replacing $\boldsymbol{\beta}$) is nonzero. Kernel density estimates of the posterior density for this quantity under the two models are shown in Figure 4. The posterior density estimates are quite similar under the two models, with most of the posterior mass below zero. The tail area above zero for the marginally interpretable model is 0.016, matching the tail area for β_1^*. The tail area above zero for the conventional GLMM is 0.014. This is close to the tail area for the marginally interpretable model, but contrasts sharply with the tail area of 0.088 for β_1, This demonstrates that the marginal parameter β_1^* targets inference on the quantity of interest—the difference between the two group means—whereas the cluster-specific parameter β_1 does not.

7 | CONCLUSION AND DISCUSSION

In this article, we have defined a class of marginally interpretable GLMMs and described the form of the adjustment that appears in these models for several common link functions. Unlike conventional GLMMs, which must be interpreted conditional on the random effects, these marginally interpretable GLMMs preserve the marginal mean even when the link function is nonlinear. Consequently, model parameters can be given a *population-averaged* interpretation. In this

sense, marginally interpretable GLMMs are comparable to marginal models fit via GEE, but unlike a purely marginal model, a marginally interpretable GLMM is a fully specified model with a density for the data that can be used to make individual-level predictions in addition to marginal inferences. We have also provided details regarding how to fit marginally interpretable GLMMs, including a fast and accurate algorithm for computing the logistic-normal integral.

While we have focused on efficient calculation for the logistic-normal case, there is also a question of whether there exist extensions to the probit-normal case where, from Proposition 4, the adjustment is available in closed form. Possible strategies suggested by a reviewer involve investigating methods summarized by Genz and Bretz[37] or adapting the algorithm of Albert and Chib.[38] Albert and Chib use data augmentation to simplify the MCMC algorithm; since the adjustment is available as closed form that only rescales the relationship between the probit of the success probability and β, this algorithm extends simply to the marginally interpretable case. In the future, we will investigate extending our method to more general versions of our marginally interpretable GLMM (not just the logistic-normal or probit-normal cases).

Many of the examples we have provided relate to models with normal random effects, in part because the normal distribution is a common choice for random effects distributions. However, a marginally interpretable GLMM does not require normal random effects and the methods described here apply to a wide array of random effects distributions. One interesting class of random effects distributions consists of mixtures of normal distributions. Mixed models that represent the random effects distribution as a mixture of normals[16,39] allow considerable flexibility in the shape of the random effects distribution.

Although we discuss fitting marginally interpretable GLMMs using Bayesian techniques, these models are also compatible with frequentist techniques. Regardless of the method used, the key is to include the adjustment at the appropriate step in the algorithm. Further, for models with a logit link and normal random effects, any model-fitting technique, be it Bayesian or frequentist, could benefit from our more accurate approach to evaluating the logistic-normal integral. (We demonstrate by simulation in Section S3 of Supplement that combining our method of calculating the adjustment with the commonly used Gauss-Hermite quadrature calculations for the likelihood, allows for successful and efficient recovery of model parameters.)

Another area where the marginally interpretable GLMM can improve inference is in hypothesis testing. When comparing two group means in a GLMM with a nonlinear link function, testing whether there is a difference between the two groups is not necessarily the same as testing whether the group means differ because the different groups may require different adjustments to preserve the marginal means. In other words, failure to account for the impact of the nonlinear link may lead one to test the wrong hypotheses. The importance of the adjustment was demonstrated with the rat teratology data in Section 6. Using the conventional GLMM, testing for a nonzero treatment effect yielded a different result than testing for a difference in the expected survival rates between the two treatment groups. Using the marginally interpretable GLMM avoids such inconsistencies because it makes the appropriate adjustments in the presence of random effects.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Peter F. Craigmile 🄳 https://orcid.org/0000-0002-9212-3950
Steven N. MacEachern 🄳 https://orcid.org/0000-0003-4106-1232

REFERENCES

1. Stiratelli R, Laird N, Ware JH. Random-effects models for serial observations with binary response. *Biometrics*. 1984;40:961-971.
2. Gilmour AR, Anderson RD, Rae AL. The analysis of binomial data by a generalized linear mixed model. *Biometrika*. 1985;72:593-599.
3. McCulloch CE, Searle SR, Neuhaus JM. *Generalized, Linear, and Mixed Models*. 2nd ed. Hoboken, NJ: Wiley; 2008.
4. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13-22.
5. Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*. 1986;42:121-130.
6. Chiou J-M, Müller H-G. Estimated estimating equations: semiparametric inference for clustered and longitudinal data. *J Royal Stat Soc Ser B*. 2005;67:531-553.

7. Barnhart HX, Williamson JM. Goodness-of-fit tests for GEE modeling with binary responses. *Biometrics*. 1998;54:720-729.

8. Wei P. Goodness-of-fit tests for GEE with correlated binary data. *Scand Stat Theory Appl*. 2002;29:101-110.

9. Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*. 1988;44:1049-1060.

10. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int Stat Rev*. 1991;59:25-35.

11. Neuhaus JM, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*. 1993;80:807-815.

12. Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*. 2nd ed. Oxford, UK: Oxford University Press; 2002.

13. Ritz J, Spiegelman D. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Stat Methods Med Res*. 2004;13:309-323.

14. Wang Z, Louis TA. Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika*. 2003;90:765-775.

15. Wang Z, Louis TA. Marginalized binary mixed-effects models with covariate-dependent random effects and likelihood inference. *Biometrics*. 2004;60:884-891.

16. Caffo B, An M-W, Rohde C. Flexible random intercept models for binary outcomes using mixture of normals. *Comput Stat Data Anal*. 2007;51:5220-5235.

17. Hedeker D, du Toit SHC, Demirtas H, Gibbons RD. A note on marginalization of regression parameters from mixed models of binary outcomes. *Biometrics*. 2018;74:354-361.

18. Kim Y, Kim D. Posterior consistency of random effects models for binary data. *J Stat Plann Infer*. 2011;141:3391-3399.

19. MacEachern SN. Nonparametric Bayesian methods: a gentle introduction and overview. *Commun Stat Appl Methods*. 2016;23:445-466.

20. Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*. 1999;55:688-698.

21. Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference. *Stat Sci*. 2000;15:1-19.

22. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc*. 1993;88:9-25.

23. Lee Y, Nelder JA. Hierarchical generalized linear models. *J Royal Stat Soc Ser B*. 1996;58:619-678.

24. Zeger SL, Karim MR. Generalized linear models with random effects: a Gibbs sampling approach. *J Am Stat Assoc*. 1991;86:79-86.

25. Gamerman D. Sampling from the posterior distribution in generalized linear mixed models. *Stat Comput*. 1997;7:57-68.

26. Miglioretti DL, Heagerty PJ. Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics*. 2004;5:381-398.

27. Crouch EAC, Spiegelman D. The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t)exp(-t^2)dt$: application to logistic-normal models. *J Am Stat Assoc*. 1990;85:464-469.

28. Monahan JF, Stefanski LA. Normal scale mixture approximations to F*(z) and computation of the logistic-normal integral. In: Balakrishnan N, ed. *Handbook of the Logistic Distribution*. New York, NY: Marcel Dekker; 1992:529-540.

29. Pirjol D. The logistic-normal integral and its generalizations. *J Comput Appl Math*. 2013;237:460-469.

30. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2018.

31. Weil CS. Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food Cosmet Toxicol*. 1970;8:177-182.

32. Liang K-Y, Hanfelt J. On the use of the quasi-likelihood method in teratological experiments. *Biometrics*. 1994;50:872-880.

33. Spiegelhalter David J, Best Nicola G, Carlin Bradley P, Der Van Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B* 2002;64:583–639.

34. Spiegelhalter David J, Best Nicola G, Carlin Bradley P, Der Van Linde A. The deviance information criterion: 12 years on. *J Royal Stat Soc Ser B* 2014;76:485–493.

35. Dickey JM. The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann Math Stat*. 1971;42:204-223.

36. Verdinelli I, Wasserman L. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J Am Stat Assoc*. 1995;90:614-618.

37. Genz A, Bretz F. *Computation of Multivariate Normal and t Probabilities*. New York, NY: Springer Science & Business Media; 2009.

38. Albert J, Chib S. Bayesian residual analysis for binary response regression models. *Biometrika*. 1995;82:747-769.

39. Magder LS, Zeger SL. A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *J Am Stat Assoc*. 1996;91:1141-1151.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Gory JJ, Craigmile PF, MacEachern SN. A class of generalized linear mixed models adjusted for marginal interpretability. *Statistics in Medicine*. 2021;40:427–440. https://doi.org/10.1002/sim.8782