Quasi-complete separation in random effects of binary response mixed models

Sauter, Rafael ; Held, Leonhard

DOI: https://doi.org/10.1080/00949655.2015.1129539

Quasi-complete Separation in Random Effects of Binary Response Mixed Models

R. Sauter[a*] and L. Held[a]

[a] *Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84, 8001 Zürich*

Clustered observations such as longitudinal data are often analysed with generalized linear mixed models (GLMM). Approximate Bayesian inference for GLMMs with normally distributed random effects can be done using integrated nested Laplace approximations (INLA), which is in general known to yield accurate results. However, INLA is known to be less accurate for GLMMs with binary response. For longitudinal binary response data it is common that patients do not change their health state during the study period. In this case the grouping covariate perfectly predicts a subset of the response, which implies a monotone likelihood with diverging maximum likelihood (ML) estimates for cluster-specific parameters. This is known as quasi-complete separation. In this paper we demonstrate, based on longitudinal data from a randomized clinical trial and two simulations, that the accuracy of INLA decreases with increasing degree of cluster-specific quasi-complete separation. Comparing parameter estimates by INLA, Markov chain Monte Carlo sampling and ML shows that INLA increasingly deviates from the other methods in such a scenario.

Keywords: integrated nested Laplace approximations; Bayesian generalized mixed models; cluster-specific quasi-complete separation

1. Introduction

There has been recent interest in the accuracy of integrated nested Laplace approximations (INLA) [1] for Bayesian inference in binary response mixed models. INLA has been successfully applied to generalized linear mixed models (GLMMs) [2], and a generally high accuracy has been reported. However, for the special case of binary responses, a thorough comparison with Markov chain Monte Carlo (MCMC) sampling has identified larger discrepancies [2]. Here, the relative approximation error, measured as the difference between the marginal posterior mean with MCMC and INLA, and scaled with the (MCMC) posterior standard deviation, was around 30%. These results are in contrast to a more recently published simulation study [3], which reported a high accuracy of INLA.

There is also interest in the accuracy of classical maximum likelihood (ML) estimates in GLMMs with binary responses. ML inference requires numerical integration over the random effects, for which penalized quasi likelihood (PQL) or adaptive Gauss Hermite quadrature (GHQ) are the two most common approaches. In response to the increasing usage of GLMMs in ecology and evolution, an overview of commonly used software packages for GLMMs has been published [4]. A detailed comparison of the estimates obtained by different statistical software packages has identified substantial differences [5], *e. g.* between `PROC NLMIXED` in `SAS` and the function `glmer()` in `R`, although both use adaptive GHQ integration. Also, the accuracy of Bayesian and ML estimation methods has been compared [6], who also consider results with INLA produced in the simulation study by [2].

*Corresponding author. Email: rafael.sauter@uzh.ch

Unfortunately, there is no analytical expression for the approximation error of INLA [1]. A straightforward way to assess INLA's accuracy is a direct comparison with MCMC. Alternatively, the accuracy of INLA in binary response models has been contrasted with the computationally more intensive expectation propagation (EP) algorithm [1] originating from the machine learning literature [7, 8].

There seems to be room for further comparisons of INLA and MCMC in other scenarios than investigated so far. We challenge INLA with a special but still realistic situation, in which not only INLA but also other estimation methods may run into problems. Specifically, we consider a situation, where a covariate is (almost) perfectly classifying the response, known as (quasi) complete separation [9]. For longitudinal data with binary response, cluster-specific (quasi) complete separation may occur if a patient shows no variation in the response, *i. e.* has longitudinal profile $(0, \ldots, 0)$ or $(1, \ldots, 1)$. Based on longitudinal data from a clinical trial on the presence of toenail infections and an additional simulation study, we show that in this case the INLA parameter estimates do not agree with those obtained by MCMC or ML. Further we assess the root mean squared error and bias of the parameter estimates by INLA in a second simulation study. Cluster-specific quasi-complete separation causes a bias for INLA which implies a substantially lower accuracy than reported elsewhere [2, 3, 6].

This paper is organized as follows. We start by reviewing likelihood and Bayesian inference in GLMMs in Section 2. In Section 3, we empirically compare parameter estimates obtained from applying INLA, MCMC and ML to the toenail clinical trial data. Section 4 describes results from two simulation studies with varying degree of cluster-specific quasi-complete separation. We close with some discussion in Section 5.

2. Inference for binary response mixed model

Consider a GLMM for (possibly unbalanced) longitudinal data with binary response $y_{ij} \in \{0, 1\}$ from individuals $i = 1, \ldots, I$ at occasions $j = 1, \ldots, n_i$, linked to times t_{ij} at which the measurements are taken. The total number of observations is $n = \sum_i^I n_i$. The logistic mixed model

$$\text{logit}(\pi_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \mathbf{b}_i$$

assumes that the binary observations y_{ij} are conditionally independent, given the random effects $\mathbf{b}_i$, with success probability $\pi_{ij} = \Pr(y_{ij} = 1 \,|\, \boldsymbol{\beta}, \mathbf{b}_i, \mathbf{D})$. Here $\mathbf{x}_{ij}$ is a vector of length p with explanatory variables and associated fixed effects vector $\boldsymbol{\beta}$. The cluster-specific random effects $\mathbf{b}_i$ are linked to the covariate vector $\mathbf{z}_{ij}$ of length q. The random effects are assumed to follow a multivariate normal distribution, *i. e.* $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$. In the random intercept (RI) model, $q = 1$, $z_{ij} = 1$ and $\mathbf{D}$ is defined by only one hyperparameter $\delta = \sigma_b^2$, the variance of the random intercept. For a random intercept and slope model (RI+RS), $q = 2$, $z_{ij} = (1, t_{ij})^\top$ and the covariance matrix $\mathbf{D}$ consists of three hyperparameters $\boldsymbol{\delta}$, two random effect variances on the diagonal and the corresponding correlation.

2.1. *Likelihood inference*

Likelihood inference is based on the marginal likelihood of the GLMM. The marginal likelihood contribution of individual i is

$$f(\boldsymbol{y}_i \,|\, \boldsymbol{\beta}, \mathbf{D}) = \int \prod_{j=1}^{n_i} f(y_{ij} \,|\, \boldsymbol{\beta}, \mathbf{b}_i) f(\mathbf{b}_i \,|\, \mathbf{D}) d\mathbf{b}_i \tag{1}$$

where $f(\cdot)$ denotes either a probability mass or a density function and $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ is the response vector of the i-th individual.

Corresponding to $\mathbf{y}_i$ the matrices $\mathbf{x}_i$ and $\mathbf{z}_i$ collect the fixed and random effect vectors for all n_i observations and are of dimension $n_i \times p$ and $n_i \times q$. In a linear mixed model, the individual marginal likelihood follows a multivariate normal distribution with mean equal to $\mathbf{x}_i\boldsymbol{\beta}$ and covariance matrix $\mathbf{z}_i\mathbf{D}\mathbf{z}_i^\top + \sigma^2\mathbf{I}_{n_i}$, here $\mathbf{I}_{n_i}$ is the identity matrix of dimension n_i. This is not the case for a GLMM with non-normal response, where numerical integration over the q-dimensional vector $\mathbf{b}_i$ is required to compute (1). This task is usually solved by numerical integration *e. g.* via the Laplace approximation [10]. An alternative approach is based on PQL [11], where bias-corrections are available [12, 13], or the GHQ-approximation, which can be improved by selecting the points, at which the function is evaluated, adaptively [14]. Increasing the number of quadrature points also increases the accuracy of this approximation. With a single quadrature point the GHQ-approximation reduces to the Laplace approximation. In practice, numerical optimization of the marginal likelihood with respect to $\boldsymbol{\beta}$ and $\mathbf{D}$ is performed with random effects fixed at the empirical Bayes estimates $\tilde{\mathbf{b}}_i$. Finding $\tilde{\mathbf{b}}_i$ for fixed $\boldsymbol{\beta}$ and $\mathbf{D}$ is the first step and numerical optimization of the approximated likelihood is the second step, which both are iteratively updated until convergence is reached.

2.2. *Bayesian inference*

A Bayesian GLMM is a hierarchical model with three stages. The first stage is a model $f(\mathbf{y} \,|\, \boldsymbol{\theta})$ for the observed data $\mathbf{y}$, given the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \mathbf{b}_1^\top, \ldots, \mathbf{b}_I^\top)^\top$. The second stage $f(\boldsymbol{\theta} \,|\, \boldsymbol{\delta})$ is the distribution of $\boldsymbol{\theta}$, given unknown hyperparameters $\boldsymbol{\delta}$. For a GLMM the distribution $f(\boldsymbol{\theta} \,|\, \boldsymbol{\delta})$ is assumed to be Gaussian, such that the GLMM can be described as a Gaussian Markov random field (GMRF) with precision matrix $\mathbf{Q}(\boldsymbol{\delta}) = \mathbf{D}(\boldsymbol{\delta})^{-1}$ [15]. The GMRF is controlled by a relatively small number of hyperparameters $\boldsymbol{\delta}$. The corresponding prior distribution $f(\boldsymbol{\delta})$ is the third stage of the formulation. In GLMMs, the hyperparameters $\boldsymbol{\delta}$ describe the covariance structure of the random effects. The posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ is

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\delta} \,|\, \mathbf{y}) &\propto f(\boldsymbol{\delta}) f(\boldsymbol{\theta} \,|\, \boldsymbol{\delta}) \prod_{i=1}^{I} f(\mathbf{y}_i \,|\, \boldsymbol{\theta}, \boldsymbol{\delta}) \\ &\propto f(\boldsymbol{\delta}) \,|\, \mathbf{Q}(\boldsymbol{\delta}) \,|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{Q}(\boldsymbol{\delta})\boldsymbol{\theta} + \sum_{i=1}^{n} \log f(\mathbf{y}_i \,|\, \boldsymbol{\theta}, \boldsymbol{\delta}) \right\} \end{aligned}$$

and one of the major goals is to calculate the marginal posterior distribution of the k^{th} component of $\boldsymbol{\theta}$:

$$f(\theta_k \,|\, \mathbf{y}) = \int_{\boldsymbol{\delta}} \int_{\boldsymbol{\theta}_{-k}} f(\boldsymbol{\theta}, \boldsymbol{\delta} \,|\, \mathbf{y}) d\boldsymbol{\theta}_{-k} d\boldsymbol{\delta} = \int_{\boldsymbol{\delta}} f(\theta_k \,|\, \boldsymbol{\delta}, \mathbf{y}) f(\boldsymbol{\delta} \,|\, \mathbf{y}) d\boldsymbol{\delta}, \tag{2}$$

here $\boldsymbol{\theta}_{-k}$ denotes all components of $\boldsymbol{\theta}$ except the k^{th} one. Usually MCMC sampling is used to generate samples from $f(\theta_k \,|\, \mathbf{y})$. A binary response GLMM may require advanced sampling algorithms such as block updating [16, 17]. The computationally less intensive INLA approach [1] approximates the marginal posterior distributions by first applying a Laplace approximation [10] to the posterior distribution of $\boldsymbol{\delta}$ and a second Laplace approximation to the posterior of the components of $\boldsymbol{\theta}$ for selected values of $\boldsymbol{\delta}$. INLA uses numerical integration over the hyperparameters to finally obtain the marginal posterior distributions $f(\theta_k \,|\, \mathbf{y})$ of all components of $\boldsymbol{\theta}$. Three different approximation strategies to the first component $f(\theta_k \,|\, \boldsymbol{\delta}, \mathbf{y})$ in equation (2) are discussed in [1]: the first is the least accurate and uses a Gaussian approximation, the second is more precise and computationally more intensive and applies a Laplace approximation while the third is intermediate in accuracy and computing time and uses a simplified Laplace approximation. For all computations involving INLA, we used the intermediate simplified Laplace approximation strategy.

Bayesian inference requires specification of a prior distribution for $f(\boldsymbol{\beta})$ and $f(\boldsymbol{\delta})$. A common approach, also employed in this paper, are independent normal distributions with large variance, *e. g.* $1/\sigma_\beta^2 = 0.0001$, for each component of $\boldsymbol{\beta}$. In the RI model, we follow the approach by [2] and use an inverse gamma $\text{IG}(a_1, a_2)$ prior [18] for the variance σ_b^2. Integration over the hyperparameter for a normal distributed $f(b_i \,|\, \sigma_b^2)$ results in a marginal $t(0, a_2/a_1, 2a_1)$ distribution [18]. For this marginal t distribution a range is defined, which covers the odds ratio $\exp(b_i)$ with a probability of 95%. The values $a_1 = 0.5$ and $a_2 = 0.00802$ for the inverse Gamma prior $f(\sigma_b^2)$ are derived from the relationship between the marginal t distribution $f(b_i)$ and the assumed range for $\exp(b_i)$, which is $[0.2, 5]$ in this case. The same derivation with a range of $[0.1, 10]$ for $\exp(b_i)$ was used by [2, 3]. As discussed by [2] the same approach to determine an informative prior can be extended to the RI+RS model. In the RI+RS model, the covariance matrix $\mathbf{D}$ is assumed to follow an inverse Wishart $\text{IWi}(r, \mathbf{R})$ distribution [18], where $r = 5$ and $\mathbf{R}$ is a diagonal matrix with entries equal to 1.34.

2.3. *Quasi-complete separation*

Fitting a logistic regression model is problematic if a covariate perfectly predicts the response. Such a covariate implies that the ML estimate will be infinite as the likelihood is increasing monotonically. Although a perfect predictor is desirable, one would rarely accept such an extreme estimate based on a finite sample. The problem of divergent ML estimates for such a data configuration is defined as complete separation [9]. A weaker form is quasi-complete separation which occurs if the covariate predicts a subset of the response vector perfectly. Quasi-complete separation leads to infinite ML estimates for the covariate almost perfectly predicting the response but not for additional covariates, if present, which explain the remaining variation in the response.

In the particular case of a binary covariate, which completely separates the response, the corresponding 2×2 table has no off-diagonal entries. For a quasi-complete separation only one of the off-diagonal entries would be zero. A continuous covariate implies complete separation if *e. g.* for all negative values the response is one and for all positive values the response is zero. Quasi-complete separation is present if additionally for covariate values equal to zero the response is either one or zero.

Divergent ML estimates caused by complete separation in generalized linear models (GLMs) may be addressed by a penalized likelihood approach [19]. The suggested penalization depends on the inverse Fisher information matrix and is related to Jeffreys' invariant prior [19]. For a logistic regression with a completely separating binary covariate this approach corresponds to adding 1/2 to each cell of the 2×2 table. While

removing the small sample bias the penalized likelihood approach yields consistent estimates [19] and there exist different approaches to improve the coverage probability of the corresponding confidence intervals [20, 21].

Addressing quasi-complete separation in a logistic regression model with random intercept is discussed by [22]. However, complete separation may now be not only fixed covariate-specific but also be cluster-specific, affecting the random effects $\mathbf{b}_i$. More specifically, if the grouping covariate, which defines the random effect clusters, is separating the response, we encounter a cluster-specific complete separation for the random intercepts. For a logistic mixed model this occurs if all components of $\mathbf{y}_i$ are either equal to one or equal to zero. We have a cluster-specific quasi-complete separation if this occurs only for some i but not all I clusters.

The assumption $\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{D})$ penalizes deviations of $\mathbf{b}_i$ away from zero which in the case of cluster-specific quasi-complete separation stabilizes the marginal likelihood such that the estimates for $\mathbf{b}_i$ are defined. But the penalization decreases if the covariance matrix of the random effects $\mathbf{D}$ increases such that the parameter estimates $\mathbf{b}_i$ are not treated different from the fixed effects if $\mathbf{D}^{-1} \to \mathbf{0}$. Thus in the extreme case of cluster-specific complete separation, the ML estimates for $\mathbf{b}_i$ will not be defined, as the penalization term vanishes with the random effects variance going to infinity. For a random intercept plus random slope model the penalization term may be increased through the random effects correlation, if only one of the two random effect covariates is causing quasi-complete separation.

Depending on the degree of cluster-specific quasi-complete separation, *i. e.* the proportion of clusters with constant response, convergence problems will arise in the numerical optimization algorithms described in Section 2.1. Also depending on how many clusters are perfectly predicted by the grouping covariate, the normal assumption for the random effects may be inappropriate. Indeed, random effect estimates tend to have extreme values in the presence of cluster-specific quasi-complete separation.

Bayesian inference for GLMMs addresses the complete separation problem in random effects by an additional, possibly informative prior $f(\boldsymbol{\delta})$ [23]. The prior distribution $f(\boldsymbol{\delta})$ needs to be proper [24], so the posterior $f(\boldsymbol{\theta}, \boldsymbol{\delta} \,|\, \mathbf{y})$ will also be proper. Nevertheless, even for Bayesian inference, numerical problems may arise with increasing degree of quasi-complete separation.

3. INLA vs. MCMC for toenail infection data

The data considered in this section are the result from a randomized, double-blinded clinical trial comparing two oral treatments for toenail infections [25–27]. The data are available on `http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-9876/homepage/50_3.htm`. The primary response was the degree of onycholysis, *i. e.* the degree of separation of the nail plate from the nail-bed. The response was classified into absent, mild, moderate or severe onycholysis and was further aggregated to a binary response with either absent or mild (0, not severe) or moderate to severe (1, severe) degree [26, 27].

Follow-up visits were planned to take place 1, 2, 3, 6, 9 and 12 months after baseline. However, the actual times t_{ij} of follow-up visits varied around the foreseen schedule and some patients have less than 6 follow-up measurements due to drop out. For the following analysis 5 patients with no follow-up measurements have been removed such that the dataset consists of 1903 observations from 289 individuals. There are 160 patients who stay always in the not severe state throughout the observation period and 14 patients who remain always in the severe state, while all remaining 115 patients change their

disease state at least once. Time since baseline was centred at the overall mean in order to improve the mixing of the MCMC algorithm [23]. The fixed effects for all models consist of an intercept, the treatment effect, the centred time since baseline in months and the interaction for centred time and treatment, *i. e.* $\mathbf{x}_{ij} = (1, \text{trt}_i, t_{ij}, t_{ij} \times \text{trt}_i)^\top$. The toenail infection data is analysed with a binary response RI ($z_{ij} = 1$) and with a RI+RS ($\mathbf{z}_{ij} = (1, t_{ij})^\top$) model. The RI model has only one hyperparameter which is the random intercept variance σ_b^2. For the RI+RS model the random effect covariance matrix $\mathbf{D}$ is defined by three hyperparameters: the variance for the random intercept $\sigma_{b_1}^2$, the variance for the random slope $\sigma_{b_2}^2$ and the correlation parameter between the two variances ρ.

INLA is implemented in a software package and an `R`-interface is available on `http://www.r-inla.org/`. We used the `r-inla` version built on 14. July 2014. All MCMC sampling was done with `JAGS` [28] through the `R`-interface `R2jags` and the `R`-package `coda` [29]. For binary or binomial response data, `JAGS` uses the algorithms proposed by [17] and [30]. Still we used a relatively large number of 500'000 MCMC iterations with 20'000 additional burnin iterations and thinning of 200 in both models, the RI and RI+RS model, to reach convergence and to ensure a negligible Monte Carlo error of the parameter estimates. ML estimation of the models was undertaken with the `R` package `lme4` [31], version 0.999999-2. We did not use the latest `lme4` version because it restricts the maximal number of quadrature points in the GHQ-approximation to 25 for the RI model and to 1 for the RI+RS model. In the RI model we use 20 quadrature points for the one-dimensional integration and the results are the same as with the current `lme4` version, whereas we use 50 quadrature points for the two-dimensional integration over the joint random effects distribution in the RI+RS model. All computations were done with `R` version 2.15.3 (2013-03-01). See Appendix A for more details about the influence of the number of quadrature points for the toenail data models.

We compare the ML estimates with the marginal posterior means for all components of $\boldsymbol{\beta}$, while fixing the hyperparameters $\boldsymbol{\delta}$ for the Bayesian methods at the ML estimates. Under an uninformative prior for the components of $\boldsymbol{\beta}$ and without any uncertainty in the hyperparameters, the posterior means should be very close to the ML estimates. The only difference between the ML estimate and the mean of the marginal posterior distribution of a fixed effect is the integration over both, random and the remaining fixed effects, while the marginal likelihood only integrates over random effects. Alternatively, the (joint) posterior mode could be used, but this is not the standard output for `r-inla`, which is approximating the marginal posteriors. Anyhow, posterior means and modes will coincide to a reasonable accuracy, since the posterior of $\boldsymbol{\beta}$ is known to be asymptotically Gaussian.

3.1. *Differences in the parameter estimates*

The estimated marginal posterior densities of $\boldsymbol{\beta}$ for both the RI and the RI+RS model are shown in Figure 1. Both are obtained using a fully Bayesian approach with hyperpriors for $\boldsymbol{\delta}$ as described in Section 2.2. Each histogram is based on the MCMC samples provided by `JAGS` and the lines show the corresponding marginal posterior density estimate produced by `r-inla`. For the RI model in the upper row of Figure 1 we see that MCMC and INLA agree rather well for all fixed effects, except for the time covariate, where there is a slight shift towards zero for the posterior by INLA compared to the MCMC histogram. In the lower row of Figure 1 we see more substantial differences between INLA and MCMC for the treatment effect and the interaction between time and treatment.

Figure 2 shows the posterior distributions of the same fixed effects and the same models as in Figure 1 but now with hyperparameters fixed at values which were determined

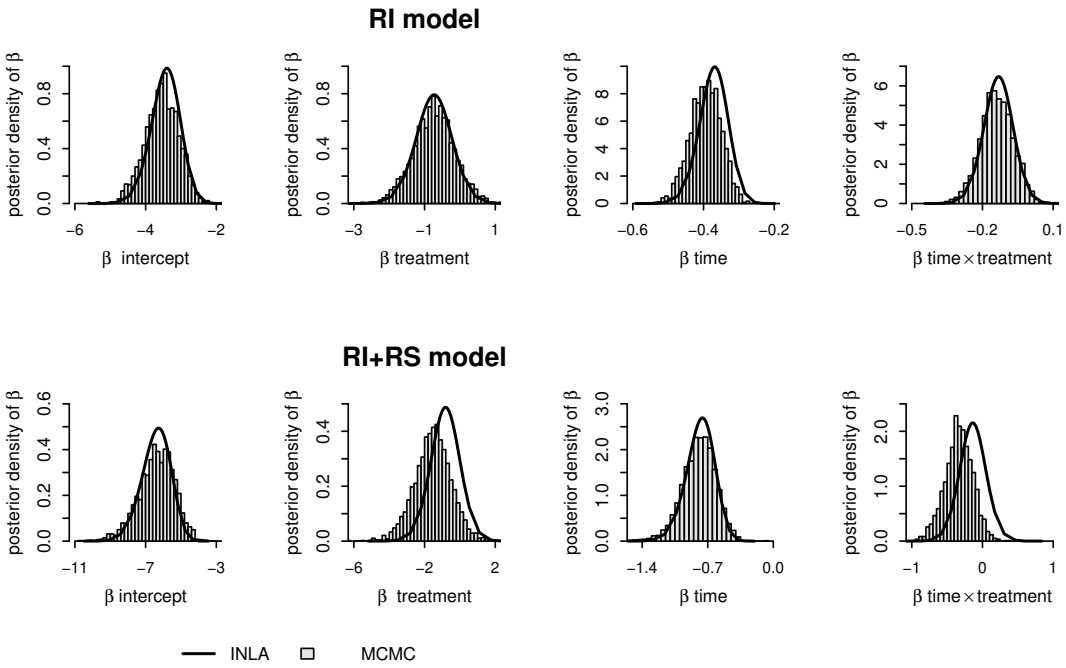

Figure 1. Marginal posterior distributions of fixed effects $\boldsymbol{\beta}$ with MCMC (histogram) and INLA (line).

by ML with `lme4`. The additional red lines now give approximate normal "posterior" distributions based on the ML estimates and the corresponding standard errors. In all plots of Figure 2 we see that the approximate posterior distributions based on the ML es-

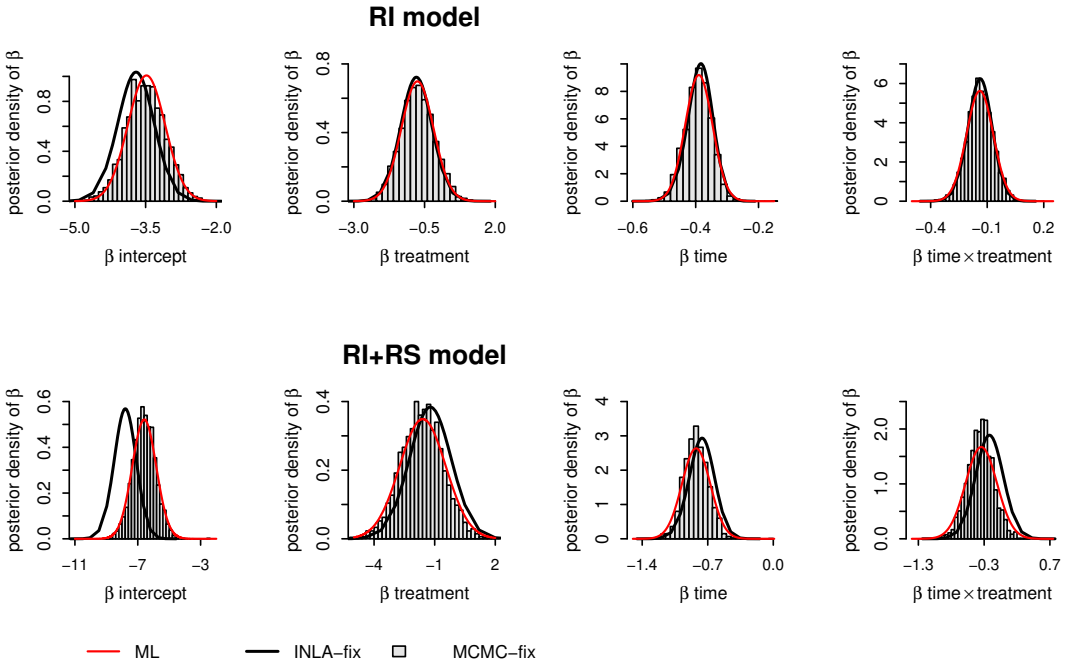

Figure 2. Marginal posterior distributions of $\boldsymbol{\beta}$ by MCMC (histogram), INLA (black line) and based on ML estimates (red line). Hyperparameters values are fixed at corresponding ML estimates.

timates agree well with the MCMC histograms. However, the posterior density estimates provided by INLA exhibit a substantial bias for the intercepts of both the RI and the RI+RS model. There is also some discrepancy for the other fixed effects in the RI+RS model.

The upper half of Table 1 summarizes the differences between the posterior mean estimates obtained with INLA or MCMC and with ML estimates. In the lower half of Table 1, relative differences are given, by scaling the differences from the upper part with the MCMC marginal posterior standard deviation, in the same way as done in the simulation study by [2]. The left part of Table 1 reports differences for the RI, the right half for the RI+RS model. For comparison with the ML estimates, we fixed the hyperparameter of the RI model at the ML estimate σ_b^2=16.04. For the RI+RS model the random intercept variance was fixed at $\sigma_{b_1}^2$ =47.75, the random slope variance at $\sigma_{b_2}^2 = 1.04$ and the correlation at $\rho =$ -0.05.

Table 1. Differences (top) and relative differences (bottom) between parameter estimates obtained with MCMC, INLA and ML for the RI (left) and the RI+RS (right) model. Relative differences are scaled with the MCMC marginal posterior standard deviation. Comparisons of INLA and MCMC with ML are based on hyperparameter values fixed at the corresponding ML estimates.

	RI model			RI+RS model		
	MCMC	ML	ML	MCMC	ML	ML
	INLA	INLA-fix	MCMC-fix	INLA	INLA-fix	MCMC-fix
intercept	-0.073	0.225	0.014	0.101	1.228	0.038
treatment	-0.041	0.038	-0.003	-0.651	-0.362	-0.012
time	-0.024	-0.003	0.003	0.000	-0.056	0.000
time × treatment	-0.006	0.000	-0.001	-0.184	-0.119	0.007
intercept	-0.156	0.557	0.034	0.104	1.692	0.053
treatment	-0.067	0.068	-0.006	-0.661	-0.346	-0.011
time	-0.519	-0.075	0.070	0.001	-0.424	0.000
time × treatment	-0.084	-0.004	-0.018	-0.990	-0.593	0.037

We see especially from the lower part of Table 1 that INLA shows large relative differences compared to MCMC but also to ML. While the relative differences in the RI model are not larger than 0.519, the differences are substantially larger in the RI+RS model with values up to 0.99. Relative differences also increase if we compare INLA with ML, to a maximum of 0.557 for the RI model and 1.692 for the RI+RS model. In contrast, the estimates based on MCMC are much closer to the ML estimates, with a maximum relative difference of 0.07.

The differences shown in the upper half of Table 1 for the RI-model may be considered as acceptable, with a maximal difference of 0.073 on the log-odds ratio scale. However, more substantial discrepancies can be seen for the RI+RS model, in particular for comparisons involving INLA estimates. See Table 1 in Appendix B for the fixed effect estimates of the models presented in Figure 1, 2 and Table 1.

The argument `inla.control` includes several settings which can be modified and which affect the accuracy of the numerical integration of the hyperparameters in `r-inla`. We increased the numerical accuracy and set the step length for the integration to `dz = 0.2` from the default value `dz = 1`, the step length for the gradient calculations to `h = 1e-5` from default `h = 0.01`, the tolerance criteria for the change in the posterior to `tolerance = 1e-6` from default `tolerance = 0.005` and we changed the integration strategy to `int.strategy = "grid"` which uses as default the less accurate central composite design (`int.strategy = "ccd"`). The differences between the posterior distributions shown in Figure 1 and 2 only improved slightly by using these settings, compared to the default ones.

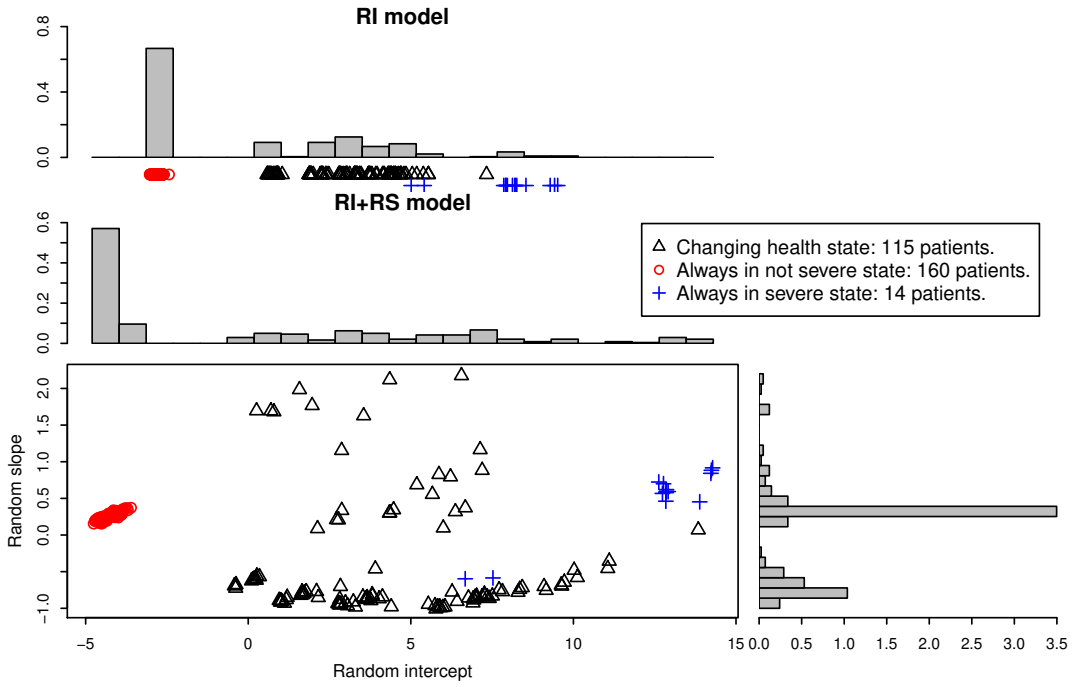

Figure 3. Random effect estimates $\mathbf{b}_i$ for the RI and the RI+RS model. Estimates are marginal posterior means obtained by MCMC.

Throughout the paper we used the default `simplified.laplace` approximation strategy in `r-inla`. Changing the approximation strategy to `laplace` did not reduce the differences for the posterior distributions as illustrated in the Supplementary Material. Additionally, Figure 2 in the Supplementary Material shows the marginal posterior distributions for the hyperparameters, which are substantially different for INLA compared to ones based on MCMC.

3.2. *Cluster-specific quasi-complete separation*

Table 1 in Section 3.1 clearly indicates that differences between MCMC and INLA, relative to the MCMC standard deviation, exceed the previously reported 30% for binary response GLMMs [2]. A correction in the location of the posterior distribution has been recommended as a possible error-correction [1]. But none of the different approximation strategies did improve the location shift of the marginal posteriors obtained by INLA. The differences between INLA and MCMC got even more pronounced if the time variable was not centred.

A closer look at the random effect estimates, obtained by MCMC, gives some interesting details. Figure 3 gives histograms of the means of the marginal posterior distribution for the random effects. The upper part shows the random effect estimates for the RI and the lower part for the RI+RS model. An additional scatter plot gives the joint distribution of estimated random intercepts and slopes in the RI+RS model. Three clusters can be distinguished: there are 160 patients, who always stay in the non severe state during the observation period (marked with a red circle), 14 patients who stay always in the severe state (marked with a blue cross), while the remaining 115 patients (marked with black triangles) switch their health state at least once. Figure 3 indicates that patients without any variation in the response build clusters and take extreme values for the random effect estimates. As a result, the empirical distribution of the random effect

estimates does not resemble a normal distribution. This hints to a substantial cluster-specific quasi-complete separation problem for the toenail data, as discussed in Section 2.3.

However, there are two patients who are always in the severe response category but their random intercept does not cluster with random effects from the other patients who always stay in the severe state. The reason for the comparably low random intercept is that these two patients are only observed at two, respective three follow-up visits. Thus they are not close to the patients who were observed seven times in the severe response state. Also they were only observed at centered times below zero such that their random slope estimate is negative. On the other hand there is one random effect which is close to the cluster of random effects for patients always being in the same state, although this patient switches the response. This patient was observed at seven occasions but only at the very last observation a moderate infection was declared, such that this profile is very similar to having always a response equal to one.

4. Simulation with varying cluster-specific quasi-complete separation

To assess a possible problem of INLA with cluster-specific quasi-complete separation in more detail, we undertook two simulation studies, with a varying proportion of cluster-specific quasi-complete separation. The first study is based on one simulated dataset only, for which we manipulate the response such that the proportion of patients always having response equal to zero changes. The results are used to examine if the differences between INLA and MCMC respectively ML, discussed in Section 3.1, are persistent or just a random artefact of the toenail data. In the second simulation study we investigate the accuracy of the parameter estimates by INLA by randomly generating replicates of the dataset and assessing the root mean squared error and bias.

In Section 4.1 and 4.2 we simulate balanced datasets with n observations per patient otherwise similar to the toenail infection data. The observed time period ranges from -4.5 to 4.5 and the time differences between follow-up visits are rescaled according to the choice of n. We set the fixed effect for time to 0.8 and for the time treatment interaction to -0.8 while the main effect for treatment is assumed to be zero. The random intercept standard deviation σ_b is set to two.

4.1. *Comparison of estimation methods by changing the response*

For the comparison of the three estimation methods we generated one initial dataset, with $I = 300$ patients observed at baseline and six follow-up visits ($n = 7$). The simulated dataset was guaranteed to have an initial proportion of patients with constant response profile fixed at 5% of all patients. We then successively started to manipulate the response of this dataset and increased the proportion of patients who always remain in state zero by another 5% or 15 patients, until we reached a proportion of 75%. In this way we continually increased the degree of cluster-specific quasi-complete separation in the data.

The plots in Figure 4 show the absolute differences between the parameter estimates by INLA, MCMC and ML. As before, the comparison with ML is based on fixed hyper-parameters. The upper half in Figure 4 shows these differences for a RI and the lower part for a RI+RS model. We see that INLA and MCMC agree well if there is only a low proportion of patients who always remain in the same health state. But with increasing proportion of patients with a constant response profile, the differences between MCMC and INLA increase. If the proportion is 40% or larger, we see substantial discrepancies. The same pattern is visible in the middle panel of Figure 4, where INLA is compared

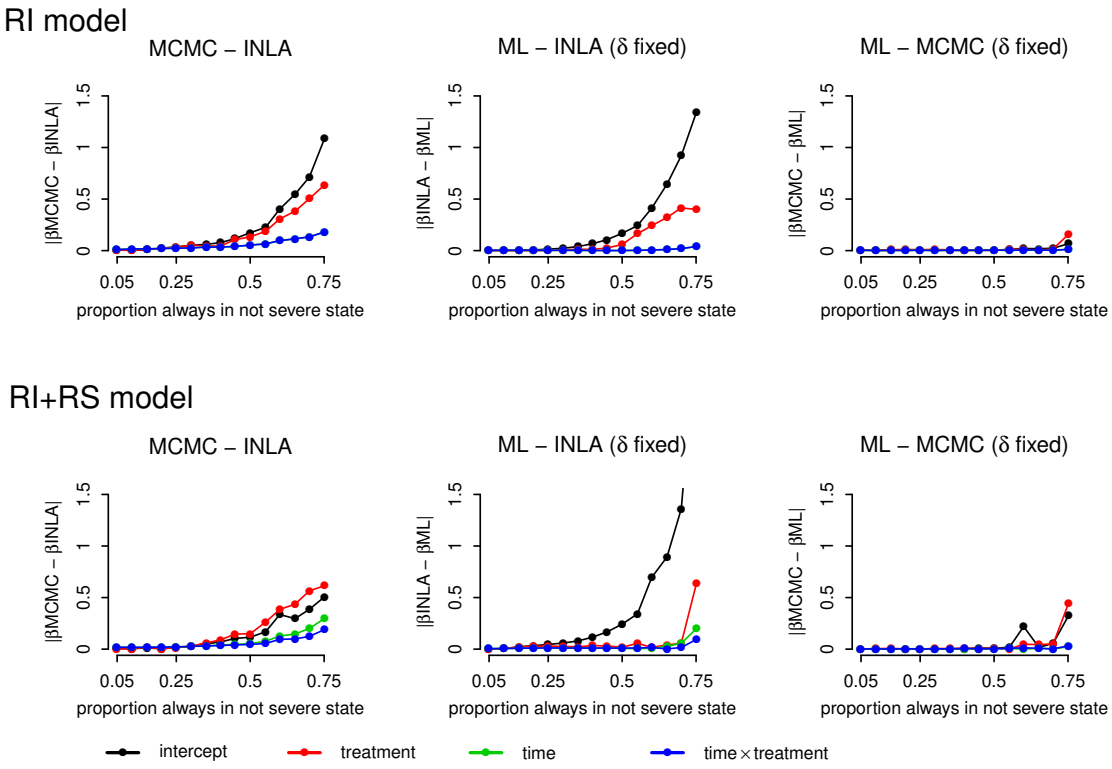

Figure 4. Absolute differences between the marginal posterior means obtained with INLA and MCMC and with ML estimates for simulated data with increasing degree of cluster-specific quasi-complete separation for a RI (upper row) and RI+RS (bottom row) model. Hyperparameters were fixed at the ML estimates for comparisons of INLA and MCMC with ML. The value for the absolute difference of the intercept between ML and INLA (δ fixed) with a proportion of 75% of patients always in the not severe state is 3.1 and not shown in the corresponding plot.

with ML. Here the differences for the fixed intercept seem to increase even more quickly. The last column in Figure 4 compares MCMC with ML which does not show any large differences, even for a large proportion of patients with a constant response profile. The few occasional differences between MCMC and ML for the RI+RS model, shown in the bottom right plot of Figure 4, may be explained by the unstable fixed effect parameter estimates based on `lme4` (see Appendix A). In this simulation we always used 40 quadrature points in the GHQ-approximation.

4.2. *Assessment of root mean squared error and bias*

The comparison of methods in Section 4.1 shows that discrepancies between INLA and the other two methods increase along with increasing degree of cluster-specific quasi-complete separation. Therefore we assess the accuracy of INLA estimates in the following simulation study. In order to keep the scope limited we report results for the random intercept model only. There are three parameters which we allow to vary, the number of patients I, the number of observations per patient n and the fixed intercept. We used four different settings with I equal to 50 or 125 and with n equal to 10 or 25. The fixed intercept is varying from -8.5 to -2 by 0.5 steps such that the proportion of patients always observed in the same state is varying. The fixed intercepts were chosen such that a large range of cluster-specific quasi-complete separation results in the simulated datasets. For a given fixed intercept the proportion of patients always observed in the same state is not necessarily the same across the four different scenarios. Still the range of cluster-specific

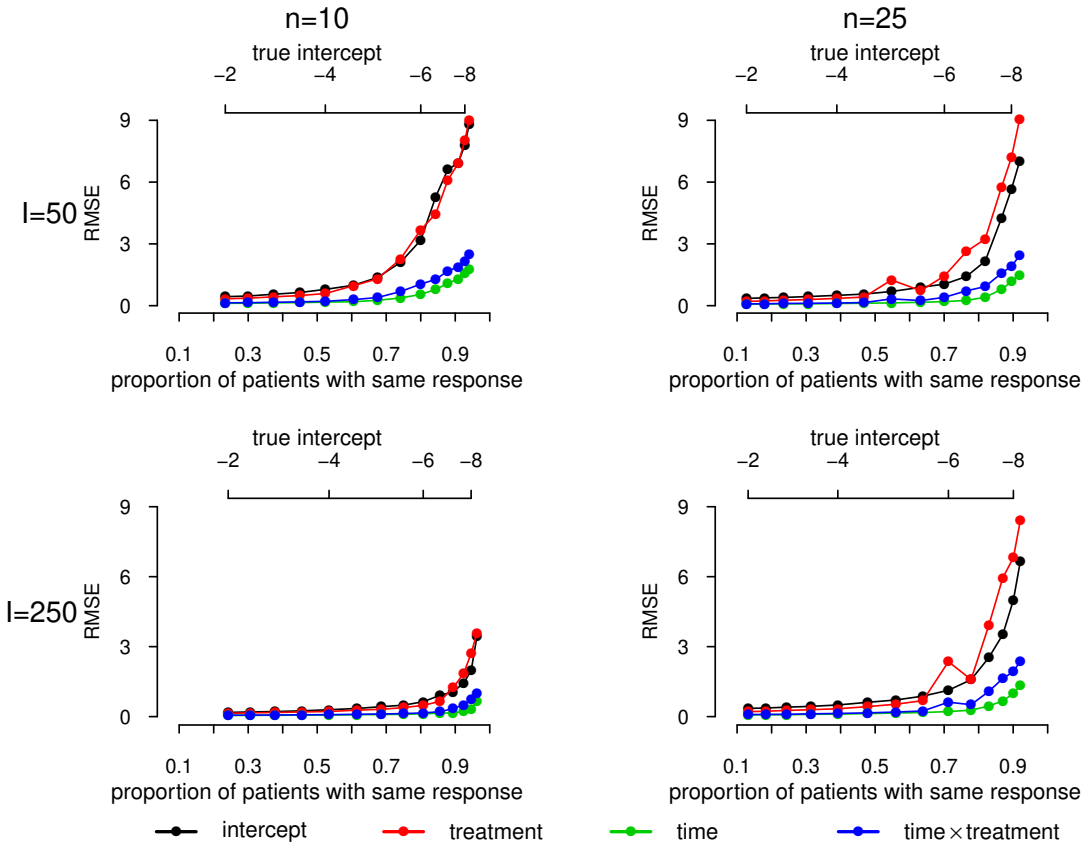

Figure 5. RMSE for marginal posterior means of fixed effects $\boldsymbol{\beta}$ for a RI model based on 1'000 iterations with different number of patients I and number of observations per patient n.

quasi-complete separation reaches from below 20% to above 80% in all four simulations.

For each of the 14 different intercepts and for each of the four scenarios we iteratively simulate 1'000 datasets, resulting in 56'000 `r-inla` calls. We rule out any datasets which include a complete separation of the response given the treatment, which would result in diverging fixed effect estimates and repeat the iteration if this occurs. For the four parameter combinations we report the root mean squared error RMSE$= \sqrt{1/N \sum_{i=1}^{N} (\hat{\beta}_i - \beta)^2}$ in Figure 5 and the bias $1/N \sum_{i=1}^{N} \hat{\beta}_i - \beta$ in Figure 6 based on the marginal posterior means.

Figure 5 shows for all four combinations of I and n an increasing RMSE with increasing proportion of patients always having the same response for all fixed effect estimates. Although the simulation with $n = 10$ and $I = 125$, in the bottom left plot, has a lower RMSE compared to the other three scenarios. The RMSE in all plots of Figure 5 is increasing with increasing quasi-complete separation Figure 6 illustrates that there is also an increasing bias with increasing proportion of quasi-complete separation. Compared to the other three scenarios Figure 6 shows that for $I = 125$ and $n = 10$ the assessed bias is relatively small.

Although the intercept has the largest bias also the other fixed effects are affected increasingly by increasing cluster-specific quasi-complete separation. This is in line with the results for the toenail dataset in Section 3.1 and also with the simulation in Section 4.1, where the estimates for the intercept based on INLA had the largest difference compared to the other two methods.

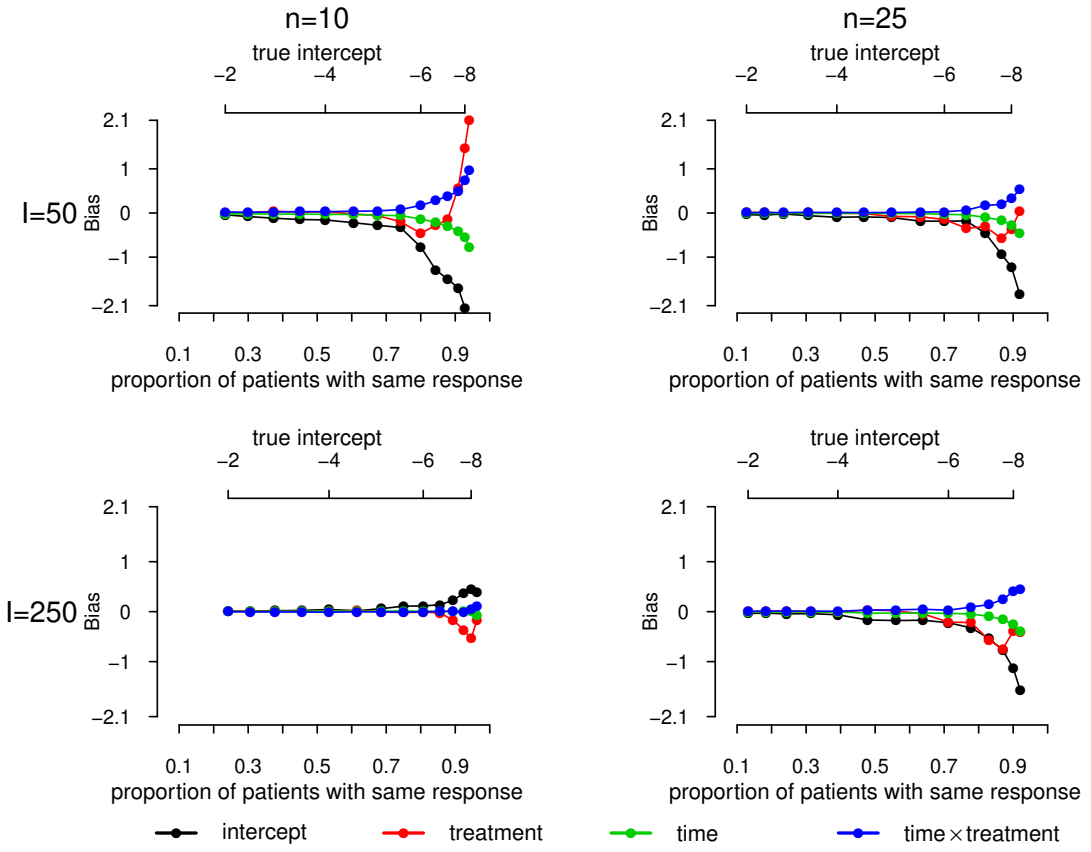

Figure 6. Bias for marginal posterior means of fixed effects $\boldsymbol{\beta}$ for a RI model based on 1'000 iterations with different number of patients I and number of observations per patient n.

5. Discussion

We showed that the approximation error by INLA increases for binary response GLMMs if the data shows a substantial and increasing degree of cluster-specific quasi-complete separation. INLA estimates agree rather well with MCMC and ML, unless the degree of cluster-specific quasi-complete separation is high. The simulation in Section 4.1 disclosed already large discrepancies if the proportion of patients with a constant response is 40%. Differences shown in Figure 4 are in the same range as the ones found for the toenail infection trial in Section 3.1, where 55.4% of the patients always stayed in the not severe state. This large degree of cluster-specific quasi-complete separation causes INLA to fail to produce reliable parameter estimates.

This was confirmed by the simulation study in Section 4.2, which illustrated that the RMSE as well as the bias increases with increasing proportion of patients always being in the same state. Although MCMC sampling is known to converge to the true posterior distribution if the number of samples is large enough, it would require much more computing time to analyse the same number of replicated datasets. However, INLA is much faster than MCMC, required less than 10 seconds per call and thus it was possible to assess the RMSE and bias with `r-inla` based on this rather large number of replicates with modest computational effort.

As illustrated in Appendix A also ML estimation may result in numerical instabilities in such situations and MCMC may request a large number of iterations. However, only INLA shows already at a comparably low degree of cluster-specific quasi-complete sep-

aration a systematic bias. This finding contrasts the results by [2] and [3], who do not investigate this scenario and thus are too optimistic regarding INLA's accuracy.

In the context of Bayesian inference most often critique is directed to the choice of the prior distributions. Usually one would assume that there must be a possibly very informative prior, which helps to stabilize the deteriorating INLA estimates if cluster-specific quasi-complete separation is present. We thus looked at different prior specifications for the hyperparameters. It has been argued [23] that an inverse gamma prior on the random effects variances may result in large sensitivity of parameter estimates. Indeed, the alternative half-normal prior distribution on σ_b [23] shows less prior sensitivity [32]. We therefore investigated if part of the discrepancies between INLA and MCMC are due to the inverse gamma prior in the RI model. Naturally, as consequence of adapting the prior, the parameter estimates changed for the toenail data. However, the differences between INLA and MCMC did not decrease, such that our main findings persisted under the alternative half-normal, and also under more informative prior specifications. Another model modification is to relax the normality assumption for the random effects in the RI model and to use a t-distribution. This model can be considered in `r-inla` [33], but differences still did not decrease substantially.

Alternatively, the non-normal distribution of the random effects, shown in Figure 3, suggests to use a mixture of normal distributions [34, 35]. This formulation has been shown to provide a better fit to the data [36]. However, implementation of such a mixture model in INLA is not straightforward and a combination with an expectation-maximization (EM) type algorithm might be required [37].

Nevertheless there are possibly ways in how this specific problem could be addressed in INLA to improve its performance, *e. g.* in [1] section 6.1 a possible alternative way to approximate the posterior marginals for the hyperparameters based on a Gaussian copula is mentioned. Finally it is important to highlight that (quasi) complete separation in mixed models is not INLA related, but a general problem, for which awareness should be high, indifferently what kind of inference is applied. If encountering cluster-specific quasi-complete separation for a binary response GLMM based on longitudinal data, one could perhaps avoid this by Markov models based on time-dependent transition probabilities $\Pr(y_{ij} = 1 \,|\, y_{i(j-1)}, \boldsymbol{\beta}, \mathbf{b}_i, \mathbf{D})$ instead of $\Pr(y_{ij} = 1 \,|\, \boldsymbol{\beta}, \mathbf{b}_i, \mathbf{D})$ as discussed in [27, 38].

If using INLA for a binary, or even binomial GLMMs, one should always pay some effort in investigating if there is cluster-specific quasi-complete separation present in the data. In practice one should check if the variance for the hyperparameters is large and if there are clusters with very high and very low random effect estimates. These may be valuable hints towards a possible large cluster-specific quasi-complete separation, which requires further investigation, as INLA may under these circumstances provide biased parameter estimates.

Acknowledgements

A lot of support and insights resulted from discussions with Andrea Riebler and Håvard Rue. Comments and suggestions of two anonymous reviewers helped to substantially improve the paper.

References

[1] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society - Series B. 2009 2;71:319–392.

[2] Fong Y, Rue H, Wakefield J. Bayesian inference for generalized linear mixed models. Biostatistics. 2010;11(3):397–412.
[3] Grilli L, Metelli S, Rampichini C. Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models. Journal of Statistical Computation and Simulation. 2014 July;0(0):1–9.
[4] Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution. 2009;24(3):127 – 135.
[5] Zhang H, Lu N, Feng C, Thurston SW, Xia Y, Zhu L, Tu XM. On fitting generalized linear mixed-effects models for binary responses using different statistical packages. Statistics in Medicine. 2011;30(20):2562–2572.
[6] Capanu M, Gönen M, Begg CB. An assessment of estimation methods for generalized linear mixed models with binary outcomes. Statistics in Medicine. 2013;32(26):4550–4566.
[7] Minka T. A family of algorithms for approximate Bayesian inference [dissertation]. MIT; 2001.
[8] Kuss M, Rasmussen CE, Herbrich R. Assessing approximate inference for binary Gaussian process classification. Journal of Machine Learning Research. 2005;6:1679–1704.
[9] Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika. 1984;71(1):1–10.
[10] Tierney L, Kadane JB. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association. 1986;81(393):82–86.
[11] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993;88(421):9–25.
[12] Breslow NE, Lin X. Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika. 1995;82(1):81–91.
[13] Lin X, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. Journal of the American Statistical Association. 1996;91(435):1007–1016.
[14] Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics. 1995;4(1):12–35.
[15] Rue H, Held L. Gaussian Markov Random Fields: Theory and Applications. London: Chapman & Hall/CRC Press; 2005.
[16] Gamerman D. Sampling from the posterior distribution in generalized linear mixed models. Statistics and Computing. 1997 MAR;7(1):57–68.
[17] Holmes CC, Held L. Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Analysis. 2006;1(1):145–168.
[18] Held L, Sabanés Bové D. Applied Statistical Inference; Likelihood and Bayes. Springer; 2014.
[19] Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80(1):27–38.
[20] Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Statistics in Medicine. 2002;21(16):2409–2419.
[21] Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. Statistics in Medicine. 2006;25(24):4216–4226.
[22] Abrahantes JC, Aerts M. A solution to separation for clustered binary data. Statistical Modelling. 2012;12(1):3–27.
[23] Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics. 2008;2(4):1360–1383.
[24] Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis. 2006 09;1(3):515–534.
[25] De Backer M, De Vroey C, Lesaffre E, Scheys I, De Keyser P. Twelve weeks of continuous oral therapy for toenail onychomycosis caused by dermatophytes: A double-blind comparative trial of terbinafine 250 mg/day versus itraconazole 200 mg/day. Journal of the American Academy of Dermatology. 1998;38(5, Supplement 2):S57 – S63.
[26] Lesaffre E, Spiessens B. On the effect of the number of quadrature points in a logistic random effects model: an example. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2001;50(3):325–335.
[27] Molenberghs G, Verbeke G. Models for Discrete Longitudinal Data. Springer; 2005.
[28] Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs

Sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing; 2003.
[29] Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. R News. 2006;6(1):7–11.
[30] Frühwirth-Schnatter S, Frühwirth R, Held L, Rue H. Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. Statistics and Computing. 2009;19(4):479–492.
[31] Bates D, Maechler M, Bolker B. lme4: Linear mixed-effects models using S4 classes.. 2011.
[32] Roos M, Held L. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. Bayesian Analysis. 2011;6(2):259–278.
[33] Martins TG, Rue H. Extending integrated nested Laplace approximation to a class of near-Gaussian latent models. Scandinavian Journal of Statistics. 2014;:online.
[34] Verbeke G, Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. Journal of the American Statistical Association. 1996;91(433):217–221.
[35] Komárek A, Lesaffre E. Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. Computational Statistics & Data Analysis. 2008;52(7):3441 – 3458.
[36] Verbeke G, Molenberghs G. The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models. Biostatistics. 2013;14(3):477–490.
[37] Van De Wiel MA, Leday GG, Pardo L, Rue H, Van Der Vaart AW, Van Wieringen WN. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics. 2013;14(1):113–128.
[38] Diggle PJ, Heagerty PJ, Liang KY, Zeger SL. Analysis of Longitudinal Data. 2nd ed. Oxford Statistical Science Series; Oxford: Oxford University Press; 2003.

Appendix A. ML estimation for toenail data and varying quadrature points

The choice of the number of quadrature points may have an influence on the fixed effect estimates in RI and RI+RS models [26]. Specific implementations may differ in different software packages [5]. We illustrate in Figure A1 that the fixed effect estimates for the toenail data are varying with the number of quadrature points used in the adaptive GHQ-approximation. This confirms the findings by [26] who compared adaptive and non-adaptive GHQ-approximation. Figure A1 shows differences for the fixed effect estimates obtained by `PROC NLMIXED` in `SAS` and `lme4` in `R` confirming the findings by [5] who also state a large difference between the two software implementations. Figure A1 suggests that, to obtain accurate estimates, the RI+RS model needs more quadrature points than the simpler RI model. Strikingly, the fixed effects obtained by `lme4` start to vary again for more than 81 quadrature points. Additionally `lme4` repeatedly produced a warning message resulting from the optimization algorithm `nlminb` which is indicated by small bars at the bottom of the two `glmer` plots. For 82 and 83 quadrature points, indicated with crosses, `glmer` aborted with an error message.

Again the `lme4` version 0.999999-2 was used, as the number of quadrature points is hard coded to a maximum of 25 in later versions. The `R` version 2.15.3 (2013-03-01) was used. If a newer `R` version together with `lme4` version 0.999999-2 is used, convergence criteria and related error and warning messages may be different. For `SAS` we used version 9.3. In contrast to the text above, models shown in Figure A1 are based on the data with uncentred timescale. Due to randomization of the trial and the uncentred timescale the treatment effect was omitted for the models here.

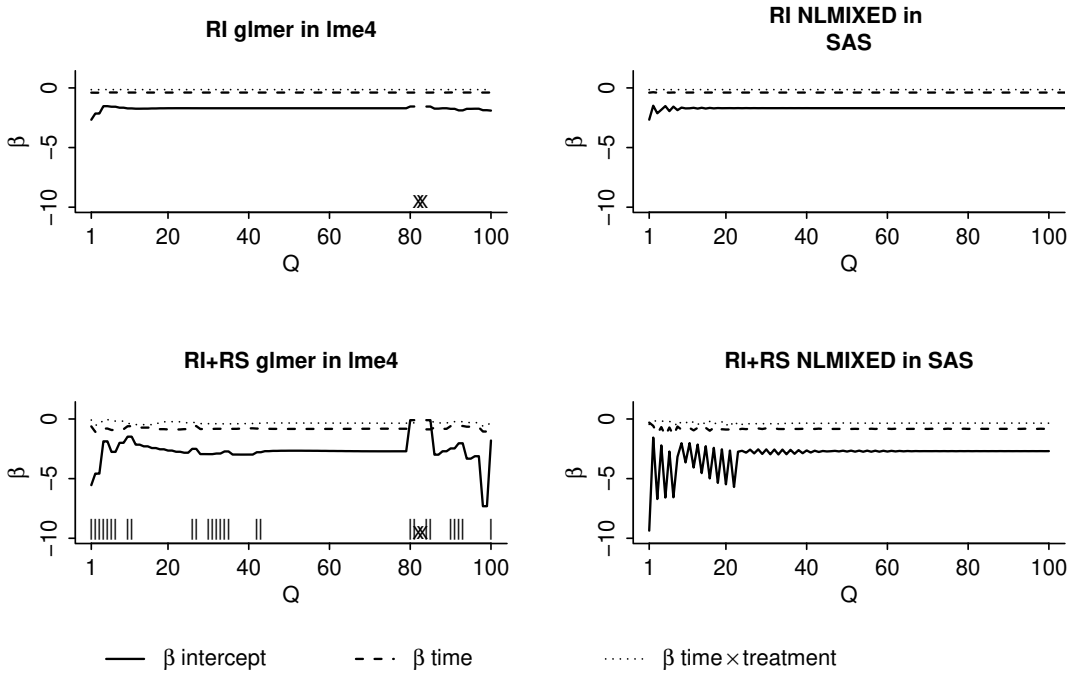

Figure A1. Fixed effects estimates for toenail data with varying number of quadrature points. A x indicates that `glmer` finished with "Error message: Downdated X'X is not positive definite, 1." and | that `glmer` finished with "Warning message:In mer_finalize(ans) : false convergence (8)".

Appendix B. Parameter estimates for RI and RI+RS model for toenail data

Table B1. Fixed effect estimates and hyperparameters by INLA, MCMC and ML estimation for the RI (upper part) and RI+RS model (lower part). For the fixed effects the standard errors are shown in parentheses. For INLA and MCMC the means of the marginal posterior distribution are shown. The last two columns show the results if the hyperparameter values are fixed at the estimates obtained by ML.

	INLA	MCMC	ML	INLA-fix	MCMC-fix
intercept	-3.441	-3.515	-3.482	-3.707	-3.495
	(0.406)	(0.469)	(0.396)	(0.386)	(0.404)
treatment	-0.737	-0.778	-0.753	-0.791	-0.749
	(0.507)	(0.605)	(0.571)	(0.552)	(0.566)
time	-0.372	-0.396	-0.390	-0.387	-0.393
	(0.0404)	(0.0460)	(0.0434)	(0.0402)	(0.0435)
time × treatment	-0.133	-0.139	-0.139	-0.139	-0.138
	(0.0618)	(0.0705)	(0.0709)	(0.0640)	(0.0696)
σ_b^2	12.848	16.858	16.036		
intercept	-6.280	-6.180	-6.588	-7.816	-6.627
	(0.797)	(0.974)	(0.766)	(0.703)	(0.726)
treatment	-0.848	-1.499	-1.593	-1.231	-1.581
	(0.818)	(0.985)	(1.144)	(1.039)	(1.047)
time	-0.761	-0.761	-0.824	-0.768	-0.824
	(0.147)	(0.175)	(0.151)	(0.136)	(0.132)
time × treatment	-0.144	-0.328	-0.344	-0.224	-0.351
	(0.184)	(0.186)	(0.239)	(0.211)	(0.201)
$\sigma_{b_1}^2$	25.7083	42.7380	47.7495		
$\sigma_{b_2}^2$	0.7441	0.9055	1.0356		
ρ	0.0249	-0.0742	-0.0531		