Semiparametric Bayesian Inference for Multilevel Repeated Measurement Data

Peter Müller,[1] Fernando A. Quintana[2] and Gary L. Rosner[1]

[1]Department of Biostatistics & Applied Mathematics,

The University of Texas, M. D. Anderson Cancer Center

Houston, TX 77030, U.S.A

[2]Departamento de Estadística, Pontificia Universidad Católica de Chile,

Santiago, CHILE

July 7, 2006

SUMMARY. We discuss inference for data with repeated measurements at multiple levels. The motivating example are data with blood counts from cancer patients undergoing multiple cycles of chemotheraphy, with days nested within cycles. Some inference questions relate to repeated measurements over days within cycle, while other questions are concerned with the dependence across cycles.

When the desired inference relates to both levels of repetition, it becomes important to reflect the data structure in the model. We develop a semi-parametric Bayesian modeling approach, restricting attention to two levels of repeated measurements. For the top level longitudinal sampling model we use random effects to introduce the desired dependence across repeated measurements. We use a non-parametric prior for the random effects

distribution. Inference about dependence across second-level repetition is implemented by the clustering implied in the non-parametric random effects model. Practical use of the model requires that the posterior distribution on the latent random effects be reasonably precise.

KEY WORDS: Bayesian nonparametrics; Dirichlet process; Hierarchical model; Repeated measurement data

1. Introduction

We consider semiparametric Bayesian inference for data with repeated measurements at multiple levels. The motivating data are blood count measurements for chemotherapy patients over multiple courses of chemotherapy. In earlier papers (Müller and Rosner, 1997; Müller et al., 2004), we considered inference for the first course of chemotherapy only. Naturally, such data do not allow inference about changes between cycles. In clinical practice, however, cancer patients receive chemotherapy over multiple courses or cycles of predetermined duration. These courses of therapy typically consist of a period during which the patient receives active drug therapy, followed by a no-drug period to allow the patient to recover for the next round of chemotherapy. Often some aspect of the treatment protocol is intended to mitigate deterioration of the patient's performance across repeated treatment cycles. Inference related to such aspects of the treatment involves a comparison across cycles, which requires modeling of the entire data set, including data from later cycles. In this extended data set, repeated measurements occur at two levels. Each patient receives multiple cycles of chemotherapy, and within each cycle, measurements are recorded over time. Another typical example of this data structure is drug concentration measurements over repeated dosing studies of pharmacokinetics.

A standard parametric approach would base inference on conjugate distributions for the sampling model, hierarchical priors, and random effects distributions. Bayesian inference for such multilevel hierarchical models is reviewed, among many others, in Goldstein, Browne and Rasbash (2002) who also discuss software for commonly used parametric models. Browne,

Draper, Goldstein and Rasbash (2002) compare Bayesian and likelihood-based methods. Heagerty and Zeger (2000) discuss likelihood-based inference for marginal multilevel models. Marginal models regress the marginal means of the outcome on covariates, rather than conditional means given random effects. A recent comprehensive treatment of multilevel models appears in Goldstein (2003).

The proposed semi-parametric Bayesian inference replaces traditional normal random effects distributions with nonparametric Bayesian models. Nonparametric Bayesian random effects distributions in mixed-effects models were first introduced in Bush and MacEachern (1996). Mukhopadhyay and Gelfand (1997) construct a semiparametric Bayesian version of generalized linear models. Applications to longitudinal data models are developed in Kleinman and Ibrahim (1998a), Müller and Rosner (1997) and Walker and Wakefield (1998), among many others. Ishwaran and Takahara (2002) extensively discuss Monte Carlo algorithms for similar semi-parametric longitudinal data models. In particular, they propose clever variations of a sequential importance sampling method known as the Chinese restaurant process. Kleinman and Ibrahim (1998b) extend the approach in Kleinman and Ibrahim (1998a) to allow binary outcomes, using generalized linear models for the top level likelihood. In each of these paper, the authors use variations of Dirichlet process (DP) models to define flexible nonparametric models for an unknown random effects distribution. The DP was introduced as a prior probability model for random probability measures in Ferguson (1973) and Antoniak (1974). See these papers for basic properties of the DP model. A recent review of semiparametric Bayesian inference based on DP models ap-

pears in Müller and Quintana (2004). Related non-Bayesian semi-parametric approaches for longitudinal data are discussed, among others, in Lin and Carroll (2001a) and Lin and Carroll (2001b). These approaches use generalized estimating equations to implement estimation in semiparametric longitudinal data models.

The main novelty in the proposed model is the construction of a nonparametric model for a high dimensional random effects distribution for multilevel repeated measurement data. The proposed approach avoids parameterization of the high dimensional dependence structure. We achieve this by using a mixture of nonparametric models for lower dimensional subvectors. In the application the lower dimensional subvector is the cycle-specific random effects vector θ_{ij} for cycle j and patient i, nested within a higher dimensional patient-specific random effects vector $\theta_i = (\theta_{ij},\ j = 1, \ldots, n_i)$. The mixture model allows us to learn about dependence without having to define a specific parametric structure in a way that is analogous to modeling a multivariate distribution as a mixture of independent kernels. Even if the kernels are independent, the mixture allows us to model essentially arbitrary dependence.

The rest of this article is organized as follows. In section 2, we introduce the proposed sampling model. In section 3, we focus on the next level of the hierarchy by proposing suitable models to represent and allow learning about dependence at the second level of the hierarchy. Section 4 discusses implementation of posterior simulation in the proposed model. Section 5 reports inference for the application that motivated this discussion. A final discussion section concludes the article.

2. First Level Repeated Measurement Model

In our semi-parametric Bayesian model representing repeated measurements at different levels of a hierarchy, the model hierarchy follows the structure of the data. The key elements of the proposed approach are as follows. We consider two nested levels of measurement units, with each level giving rise to a repeated measurement structure. Assume data y_{ijk} are recorded at times k, $k = 1, \ldots, n_{ij}$, for units j, $j = 1, \ldots, n_i$, nested within higher-level units i, $i = 1, \ldots, n$. We will refer to the experimental units i as "subjects" and to experimental units j as "cycles" to simplify the following discussion and remind us of the motivating application. Figure 1 shows the overall structure.

We start by modeling dependence of the repeated measurements within a cycle, $y_{ij} = (y_{ijk},\ k = 1, \ldots, n_{ij})$. We assume $p(y_{ij} \mid \theta_{ij}, \eta)$ to be a nonlinear regression parametrized by cycle-specific random effects θ_{ij}. Here and throughout the following discussion, η are hyperparameters common across subjects i and cycles j. Figure 2 shows typical examples of continuous outcomes y_{ijk}, measured over multiple cycles, with repeated measurements within each cycle.

We define dependence within each cycle by assuming that observations arise according to some underlying mean function plus independent residuals

$$y_{ijk} = f(t_{ijk};\ \theta_{ij}) + e_{ijk}. \tag{1}$$

Here f is a nonlinear regression with parameters θ_{ij}, and e_{ijk} are assumed i.i.d. $N(0, \sigma^2)$ normal errors. Marginalizing with respect to θ_{ij}, model (1) defines a dependent probability model for y_{ij}. This use of random effects to introduce dependence in models for repeated measurement data is common

practice. The choice of $f(\cdot;\ \theta)$ is problem-specific. In the implementation reported later, we use a piecewise linear-linear-logistic function. In the absence of more specific information, we suggest the use of generic smoothing functions, such as spline functions (Denison et al., 2002).

3. Second-Level Repeated Measurement Model

3.1 *A Semi-parametric Random Effects Model*

We introduce a dependent random effects distribution on $\theta_i = (\theta_{i1}, \ldots, \theta_{in_i})$ to induce dependence across cycles. We will proceed with the most general approach, leaving the nature of the dependence unconstrained. We achieve this by considering a non-parametric prior for the joint distribution $p(\theta_{i1}, \ldots, \theta_{in_i})$.

We first introduce a parametric model, constructed to be comparable to the non-parametric model, to clarify structure, and for later reference. Let $\eta = (m, B, S, \sigma^2)$ denote a set of hyperparameters. We use i, j and k to index patients $i = 1, \ldots, n$, cycles $j = 1, \ldots, n_i$ and observations $k = 1, \ldots, n_{ij}$.

$$\begin{array}{lrcl} \text{Cycle effects:} & p(\mu_j \mid \eta) & = & N(m, B) \\ \text{Random effects:} & p(\theta_{ij} \mid \mu, \eta) & = & N(\mu_j, S), \\ \text{Data:} & p(y_{ij} \mid \theta_{ij}, \eta) & = & \prod_k p(y_{ijk} \mid \theta_{ij}, \sigma^2), \end{array} \tag{2}$$

We assume independence across cycles at all levels. This implies that also posterior inference on θ_{ij} and posterior predictive inference is independent across cycles. We could modify (2) to include a dependent prior $p(\mu_1, \mu_2, \ldots \mid \eta)$ to allow for dependence. However, even for moderate dimension of θ_{ij} this is not practical. Instead, we proceed with a semi-parametric extension that implements learning about dependence through essentially a mixture of

independent models as in (2).

We generalize the random effects distribution for $\theta_i = (\theta_{i1}, \ldots, \theta_{in_i})$ to a mixture of normal model. Let $N(x;\ m, S)$ indicate a normal distributed random variable x with moments (m, S). We assume

$$p(\theta_i \mid G, S) = \int \prod_j N(\theta_{ij};\ \mu_{ij}, S)\, \mathrm{d}G(\mu_i), \tag{3}$$

with a non-parametric prior on the mixing measure G for the latent normal means $\mu_i = (\mu_{i1}, \ldots, \mu_{in_i})$. As usual in mixture models, posterior inference proceeds with an equivalent hierarchical model:

$$\theta_{ij} \sim N(\mu_{ij}, S) \text{ and } \mu_i \sim G. \tag{4}$$

Substituting a common value $\mu_i = \mu_0$ across all patients i, i.e., a point mass $G(\mu) = I(\mu = \mu_0)$, shows the correspondence with the parametric model. In Section 3.2 we will introduce a prior for a discrete random measure G. Denote with μ_h^o the point masses of G. Implicit in the model for G will be an independent $N(m, B)$ prior for the subvectors μ_{hj}^o corresponding to cycles, with independence across h and j. In other words, we generalize (2) by a mixture of independent models. The mixture allows learning about dependence in much the same way as a kernel density estimate with independent bivariate kernels can be used to estimate a dependent bivariate distribution.

3.2 *The Random Probability Measure G*

The probability model for G is the main mechanism for learning about dependence across cycles. We use a Dirichlet process (DP) prior. We write $DP(M, G^\star)$ for a DP model with base measure $G^\star$ and total mass parameter

M. We complete model (1) and (4) with

$$G \sim DP(M, G^\star) \tag{5}$$

See, for example, MacEachern and Müller (2000) for a review of DP mixture models as in (4).

Besides technical convenience and computational simplicity, the main reason for our choice is the nature of the predictive inference that is implied by the DP model. Assume patients $i = 1, \ldots, n$ have been observed. The prior predictive $p(\theta_{n+1} \mid \theta_1, \ldots, \theta_n)$ for patient $n+1$ is of the following type. With some probability, θ_{n+1} is similar to one of the previously recorded patients. And with the remaining probability, θ_{n+1} is generated from a baseline distribution $G^\star$ defined below. The notion of "similarity"is formalized by assuming a positive prior probability for a tie of the latent variables μ_i. Let $k \leq n$ denote the number of unique values among $\mu_1, \ldots, \mu_n$ and denote such values by $\mu_1^*, \ldots, \mu_k^*$. Let m_h, $h = 1, \ldots, k$, denote the number of latent variables μ_i equal to μ_h^*, and let $w_h = m_h/(M+n)$ and $w_{k+1} = M/(M+n)$. The DP prior on G implies

$$p(\mu_{n+1} \mid \mu_1, \ldots, \mu_n) = \begin{cases} \mu_{n+1} = \mu_h^*, & \text{with prob. } w_h(m_1, \ldots, m_k),\ h = 1, \ldots, k \\ G^\star & \text{with prob. } w_{k+1}(m_1, \ldots, m_k). \end{cases} \tag{6}$$

The predictive distribution for μ_{n+1} is a mixture of the empirical distribution of the already observed values and the base measure $G^\star$. The predictive rule (6) is attractive in many applications. For example, consider the application to the multi-cycle hematologic counts. The model implies that with some probability the response for the new patient replicates one the previous pa-

tient responses (up to residual variation), and with the remaining probability the response is generated from an underlying base measure $G^\star$.

4. Posterior Inference

4.1 *Base Measure, Kernel and Regression*

For the base measure $G^\star$ we use the same factorization as in (3),

$$G^\star(\mu_i) = \prod_{j=1}^{n_i} p(\mu_{ij} \mid \eta) = \prod N(\mu_{ij};\ m, B) \tag{7}$$

The advantage of this choice of base measure is that hyperparameters η that define $G^\star$ only need to be defined for the random vector μ_{ij} instead of the higher dimensional vector μ_i. Using a base measure with conditional independence across cycles, any inference about dependence across cycles for a future patient arises from the data-driven clustering of the imputed μ_i vectors. Clustering over locations allows modeling dependence in much the same way as a mixture of bivariate standard normals kernel can approximate any bivariate distribution, with arbitrary variance-covariance matrix, in a bivariate kernel density estimate.

As usual in DP mixture models, posterior inference proceeds in the marginal model (6), after analytically integrating out the random measure G. Choosing a conjugate base measure $G^\star$ and kernel $p(\theta_{ij} \mid \mu_{ij}, \eta)$, such as the conjugate normal kernel and base measure used in (4) and (7), further facilitates posterior simulation. See section 4.2 for details.

A minor modification of the model allows us to include cycle-specific covariates. Let x_{ij} denote a vector of covariates for cycle j of patient i. This could, for example, include dose of a treatment in cycle j. A straightforward way to include a regression on x_{ij} is to extend the probability model on θ_{ij} to

a probability model on $\tilde{\theta}_{ij} \equiv (x_{ij}, \theta_{ij})$. The implied conditional distribution $p(\theta_{ij} \mid x_{ij})$ formalizes the desired density estimation for θ as a function of x. This approach is used, for example, in Mallet et al. (1988) and Müller and Rosner (1997).

4.2 *Posterior MCMC*

Posterior simulation in the proposed model is straightforward by Markov chain Monte Carlo simulation (MCMC). See, for example, MacEachern and Müller (2000), Neal (2000), or Jain and Neal (2004) for an explanation of MCMC posterior simulation for DP mixture models.

We briefly explain the main steps in each iteration of the MCMC. An important feature of the model is the conditional independence of the θ_{ij} across cycles j given μ_i and η. This allows us to consider one cycle at a time when updating θ_{ij} in the Gibbs sampler. Updating θ_{ij}, conditional on currently imputed values for μ_i reduces to the problem of posterior simulation in a parametric, non-linear regression with sampling model (1) and prior (4). See the discussion in Section 5.1 for a specific example.

Next, consider updating μ_i conditional on currently imputed values for θ_i, hyperparameters η and $\{\mu_\ell;\ \ell \neq i\}$. We use notation similar to (6), with an additional superindex $^-$ indicating the exclusion of the i-th element, as follows. First, re-order the indices of μ_h^*, such that μ_i is equal to the last of the unique values, i.e., $\mu_i = \mu_k^*$. Let $\{\mu_h^*,\ h = 1, \ldots, k^-\}$ denote the set of unique values among $\{\mu_\ell,\ \ell \neq i\}$, and let $m_h^- = |\{\ell :\ \mu_\ell = \mu_h^*, \ell \neq i\}|$. From this we can find the complete conditional posterior distribution for μ_i. Let

$Q_0 = \int p(y_i \mid \theta)\, \mathrm{d}G^\star(\theta)$ and $q_0 \propto p(y_i \mid \theta)\, G^\star(\theta)$.

$$p(\mu_i \mid \ldots) = \begin{cases} I(\mu_i = \mu_h^*) & \text{w.}pr. \propto m_h^-\, p(y_i \mid \mu_h^*) \\ q_0(\mu_i) & \text{w.}pr. \propto M\, Q_0, \end{cases}$$

$h = 1, \ldots, k^-$. Exploiting the conjugate nature of the base measure $G^\star$ and the kernel $p(\theta_{ij} \mid \mu_{ij}, \eta)$, we can simplify one step further. We can analytically marginalize with respect to μ_i, conditional only on the configuration of ties among the μ_i. Define indicators s_i, $i = 1, \ldots, n$, with $s_i = h$ if $\mu_i = \mu_h^*$ and let $\Gamma_h^- = \{\ell :\ s_\ell = h,\ \ell \neq i\}$ denote the set of indices with common value $\mu_i = \mu_h^*$. Also let $s^- = (s_\ell;\ \ell \neq i)$ and $y^- = (y_\ell;\ \ell \neq i)$. Then we can replace sampling $p(\mu_i \mid \ldots)$ by sampling from

$$p(s_i \mid s^-, y^-, y_i, \eta) = \sum_{h=1}^{k^-} I(s_i = h)\, m_h^- Q_h + I(s_i = k^- + 1)\, M\, Q_0,$$

with $Q_h = \int p(y_i \mid \mu_i = \mu_h^*)\, \mathrm{d}p(\mu_h^* \mid y_\ell,\, \ell \in \Gamma_h^-, \eta)$. This step critically improves mixing of the Markov chain simulation (MacEachern, 1994).

As usual in DP mixture models, we include a transition probability to update μ_h^*, conditional on currently imputed values of all other parameters. As before, $\{\mu_1^*, \ldots, \mu_k^*\}$ are the unique values among $\{\mu_i,\ i = 1, \ldots, n\}$. Let μ_{hj}^* denote the subvector of μ_h^* corresponding to the j-th cycle. For resampling μ_h^*, we condition on s (now again without excluding μ_i) and find $p(\mu_{hj}^* \mid s, \theta, \eta) \propto G^\star(\mu_{hj}^*) \prod_{i \in \Gamma_h} p(\theta_{ij} \mid \mu_{hj}^*)$.

5. Modeling Multiple Cycle Hematologic Data

5.1 *Data and Model*

Modeling patient profiles (e.g., blood counts, drug concentrations, etc.) over multiple treatment cycles requires a hierarchical extension of a basic

one-cycle model. Model (1), (4), (5) and (7) provides such a generalization. Several important inference questions can only be addressed in the context of a joint probability model across multiple cycles. For example, in a typical chemotherapy regimen, some aspects of the proposed treatment are aimed at mitigating deterioration of the patient's overall performance over the course of the treatment. Immunotherapy, growth factors, or other treatments might be considered to ensure reconstitution of blood cell counts after each chemotherapy cycle.

We analyze data from a phase I clinical trial with cancer patients carried out by the Cancer and Leukemia Group B (CALGB), a cooperative group of university hospitals funded by the U.S. National Cancer Institute to conduct studies relating to cancer therapy. The trial, CALGB 8881, was conducted to determine the highest dose of the anti-cancer agent cyclophosphamide (CTX) one can safely deliver every two weeks in an outpatient setting (Lichtman et al., 1993). The drug is known to cause a drop in white blood cell counts (WBC). Therefore, patients also received GM-CSF, a colony stimulating factor given to spur regrowth of blood cells (i.e., for hematologic support). The protocol required fairly extensive monitoring of patient blood counts during treatment cycles. The number of measurements per cycle varied between 4 and 18, with an average of 13. The investigators treated cohorts of patients at different doses of the agents. Six patients each were treated at the following combinations (CTX, GM-CSF) of CTX (in g/m^2) and GM-CSF (in $\mu g/kg$): (1.5, 10), (3.0, 2.5), (3.0, 5.0), (3.0, 10.0) and (6.0, 5.0). Cohorts of 12 and 10 patients, respectively, were treated at dose combinations of (4.5, 5.0) and (4.5, 10.0). Hematologic toxicity was the primary endpoint.

In Müller and Rosner (1997) and Müller et al. (2004), we reported analyses restricted to data from the first treatment cycle. However, the study data include responses over several cycles for many patients, allowing us to address questions related to changes over cycles. We use the model proposed in Sections 2 and 3 to analyze the full data. The data are WBC in thousands, on a logarithmic scale, $y_{ijk} = \log(\text{WBC}/1000)$, recorded for patient i, cycle j, on day t_{ijk}. The times t_{ijk} are known, and reported as days within cycle. We use a non-linear regression to set up $p(y_{ij} \mid \theta_{ij}, \eta)$. For each patient and cycle, the response $y_{ij} = (y_{ij1}, \ldots, y_{ijn_{ij}})$ follows a typical "bath tub" pattern, starting with an initial base line, followed by a sudden drop in WBC at the beginning of chemotherapy, and eventually a slow S-shaped recovery. In Müller and Rosner (1997) we studied inference for one cycle alone, using a non-linear regression (1) in the form of a piecewise linear and logistic curve. The mean function $f(t;\, \theta)$ is parameterized by a vector of random effects $\theta = (z_1, z_2, z_3, \tau_1, \tau_2, \beta_1)$:

$$f(t;\, \theta) = \begin{cases} z_1 & t < \tau_1 \\ rz_1 + (1-r)g(\theta, \tau_2) & \tau_1 \le t < \tau_2 \\ g(\theta, t) & t \ge \tau_2 \end{cases} \tag{8}$$

where $r = (\tau_2 - t)/(\tau_2 - \tau_1)$ and $g(\theta, t) = z_2 + z_3/[1 + \exp\{2.0 - \beta_1(t - \tau_2)\}]$. The intercept in the logistic regression was fixed at 2.0 after finding in a preliminary data analysis that a variable intercept did not significantly improve the fit. This conclusion is based on comparing the posterior fitted curves $E(f(\cdot;\, \theta_{ij}) \mid data)$ with the observed responses. We did not carry out a formal test of fit.

We use model (8) and assume $\theta_{ij} \sim N(\mu_{ij}, S)$, independently across pa-

tients i and cycles j. Dependence across cycles is introduced by the nonparametric prior $\mu_i \sim G$. Specifying the SSM prior for G, we use the predictive rules corresponding to the DP prior, i.e., $G \sim DP(M, G^\star)$.

We introduce two modifications to the general model to make it suitable for the application. First, we allow a different residual variance for each cycle, i.e, we use $e_{ijk} \sim N(0, \sigma_{ij})$ for the non-linear regression in (1). Second, we add a constraint to the kernel in (3). Let (τ_{1ij}, τ_{2ij}) denote the two elements of θ_{ij} corresponding to the change points τ_1 and τ_2 in (8). We use $p(\theta_{ij} \mid \mu_i, \eta) \propto N(\mu_{ij}, S)\, I(\tau_{1ij} < \tau_{2ij})$.

Finally, we include a regression on covariates x_{ij}. We use the bivariate covariate of the treatment doses of CTX and GM-CSF in cycle j, patient i. Both doses are centered and scaled to zero mean and standard deviation 1.0 using the empirical mean and standard deviation across the $n = 52$ patients. Conditioning on x_{ij}, the mixture of normals for $\tilde{\theta}_{ij} = (x_{ij}, \theta_{ij})$ implies a locally weighted mixture of linear regressions. Compare with Section 4.1.

5.2 *Hyperpriors*

We complete the model with prior specifications for the hyperparameters η. For the residual variance σ_{ij}^2 we assume $\sigma_{ij}^{-2} \sim \mathrm{G}amma(a/2, ab/2)$, parametrized such that $E(\sigma_{ij}^{-2}) = 1/b$, with $a = 10$ and $b = 0.01$. Let $diag(x)$ denote a diagonal matrix with diagonal elements x. For the covariance matrix of the normal kernel in (3), we use $S^{-1} \sim W(q, R^{-1}/q)$ with $q = 25$ degrees of freedom and $R = diag(0.01, 0.01, 0.1, 0.1, 0.1, 0.01, 1, 1)$. The elements of θ_{ij} are arranged such that the first two elements correspond to the covariate x_{ij} and the third through eighth elements corre-

spond to the parameters in the non-linear regression (8), $z_1, z_2, z_3, \tau_1, \tau_2$ and β_1. The base measure $G^\star$ of the DP prior is assumed multivariate normal $G^\star(\mu_{ij}) = N(m, B)$ with a conjugate normal and inverse Wishart hyperprior on the moments. That is, $m \sim N(d, D)$ and $B^{-1} \sim W(c, C^{-1}/c)$ with $c = 25$ and $C = diag(1, 1, 1, 1, 1, .1, 1, 1)$, $D = I_8$, and the hyperprior mean is fixed as the average of single patient maximum likelihood estimates (m.l.e.). Let $\widehat{\theta}_{i1}$ denote the m.l.e. for patient i. We use $d = 1/n \sum \widehat{\theta}_{i1}$. Finally, the total mass parameter is assumed $M \sim \text{G}amma(5, 1)$.

5.3 *Posterior MCMC Simulation*

We implemented posterior MCMC simulation to carry out inference in model (1), (4), (5) and (7), using the described prior and hyperprior choices. The parameters σ^2_{ij}, S, B, m are updated by draws from their complete conditional posterior distributions. All are standard probability models that allow efficient random variate generation. Updating the latent variables μ_i, $i = 1, \ldots, n$ and the total mass parameter M proceeds as described in MacEachern and Müller (2000). Finally, consider updating θ_i. Conditional on μ_i, inference in the model is unchanged from the single-cycle model. Updating the random effects parameters θ_{ij} in a posterior MCMC simulation reduces to a nonlinear regression defined by the sampling model (8) and the normal prior $\theta_{ij} \sim N(\mu_{ij}, S)$. In particular, for the coefficients in θ_{ij} corresponding to the random effects parameters z_1, z_2 and z_3, the complete conditional posterior is available in closed form as the posterior in a normal linear regression model. We use a Metropolis-Hastings random walk proposal to update the elements of θ_{ij} corresponding to τ_1 and τ_2. If the proposed val-

ues violate the order constraint $\tau_1 < \tau_2$, we evaluate the prior probability as zero and reject the proposal.

As starting values for θ_{ij}, we used maximum likelihood estimates based on y_{ij}, substituting average values when too few responses were available for a given cycle. We then ran 200,000 iterations of the described MCMC, discarding the first 100,000 as initial transient, and saving imputed values after every 50-th iteration.

We considered imputed values of the 6-dimensional random effects vectors θ_{ij} for all cycles for the four patients shown in Figure 2 to verify convergence. We used BOA (Smith, 2005) to evaluate convergence diagnostics, and chose the diagnostic proposed in Geweke (1992), using default parameters in BOA. We evaluated the convergence diagnostics for a total number of 60 tracked parameters. The summary of the 60 diagnostics is (min., first quartile, mean, third quart., max.) $= (-1.94, -0.75, 0.01, 0.60, 1.92)$.

5.4 *Results*

Posterior inference is summarized in Figures 3 through 5. Let $H(\theta_i) = \int p(\theta_i \mid \mu_i, \eta)\ dG(\mu_i)$ denote the nonparametric mixture model for the random effects distribution. Also, we use Y to denote all observed data. Note that the posterior expectation $E(H \mid Y)$ is identical to the posterior predictive $p(\theta_{n+1} \mid Y)$ for a new subject: $p(\theta_{n+1} \mid Y) = \int p(\theta_{n+1} \mid H, Y)\ dp(H \mid Y) = \int H(\theta_{n+1})\ dp(H \mid Y)$. The high dimensional nature of θ_{ij} makes it impractical to show the estimated random effects distribution itself. Instead we show the implied WBC profile as a relevant summary. Figure 3 shows posterior predictive WBC counts for a future patient, arranged by dose x_{ij} and

cycle j. Each panel shows posterior predictive inference for a different dose of CTX and GM-CSF, assuming a constant dose across all cycles. Within each panel, three curves show posterior predictive mean responses for cycles $j = 1$ through $j = 3$. Each curve shows $E(y_{n+1,jk} \mid Y)$, plotted against $t_{n+1,jk}$. Together, the three curves summarize what was learned about the change of θ_{ij} across cycles. Note how the curve for the third cycle ($j = 3$) deteriorates by failing to achieve the recovery to baseline WBC. Comparing the predicted WBC profiles for high versus low dose of GM-CSF for the same level of CTX confirms that the growth factor worked as intended by the clinicians. The added GM-CSF improves the recovery to baseline for later cycles.

Figure 4a summarizes an important feature of G. Let $p14$ denote the probability of WBC above a critical threshold of 1000 on day 14, i.e., $p14 = p(Y_{n+1,jk} > \log 1000 \mid Y)$ for $t_{n+1,jk} = 14$ (we modeled log WBC). The figure plots $p14$ against cycle, arranged by treatment level x_{n+1} (assuming constant treatment level across all cycles and denoting the common value by x_{n+1}). For each cycle j and treatment level the lines show the marginal posterior predictive probability of a WBC beyond 1000 by day 14. At low to moderate levels of the chemotherapy agent CTX, treatment with high level of the growth factor stimulating factor GM-CSF stops the otherwise expected deterioration across cycles. Even for high CTX, the additional treatment with GM-CSF still mitigates the decline over cycles. Figure 4b plots the posterior predictive minimum WBC (in log 1000) by cycle within doses of the two drugs. Short vertical line segments in both panels indicate pointwise posterior predictive standard deviations. The large posterior predictive uncertainties realistically reflect the range of observed responses in the data.

Compare with Figure 5(b). The numerical uncertainty of the Monte Carlo average is negligible.

Figure 5ab shows another summary of the estimated random effects model $H(\theta_i)$, across cycles, for fixed doses, CTX= $3g/m^2$ and GM-CSF= $5\mu g/kg$. We select two clinically relevant summaries, the nadir WBC (fnadir) and the number of days that WBC is below a critical threshold of 1000 (T_{lo}), to visualize the high dimensional distributions. Both summaries are evaluated for each cycle. For each summary statistic, we show the joint distribution for cycles 1 and 2. The bivariate distributions are shown by plotting 500 random draws. One can recognize distinct clusters in the joint distribution. Figure 5c shows results for the parametric model (2). All hyperparameters were chosen identical as for the semi-parametric model. The parametric model shrinks the random effects to the mean of the estimated unimodal random effects distribution. The shown summaries are highly non-linear functions of the parameters, making it difficult to interpret the shrinkage beyond the fact that the estimated random effects distribution under the parametric model is unimodal and is significantly more peaked. We carried out similar comparisons (not shown) for Figures 2 and 3. The fitted profiles shown in Figure 2 remain almost unchanged under the parametric model. The predictive inference shown in Figure 3, however, changes substantially. The change in Figure 5c relative to Figure 5b shows summaries.

Finally, we carried out sensitivity analyses to investigate the robustness of the reported results with respect to changes in the prior assumptions. Results are summarized in Table 1. The columns of Table 1 report the probability of white blood cell count above a critical threshold in cycle 3,

labeled $P(y_{14} > 0)$, the change in this probability from cycle 2 to 3 (ΔP), and predicted nadir count (FNADIR) in the third cycle, arranged by three different doses of GM-CSF, fixing CTX at 4.5 g/m^2. Reported summaries in Table 1 are with respect to the posterior predictive distribution for a future patient. The three horizontal blocks of the table report changes in the three main layers of the hierarchical model. The first three rows report inference for different choices of the prior expectation $E(\sigma_{ij}^2)$ for the residual variances in (1). The next three lines report inference under different choices of the hyperprior on S. Recall from Section 5.2 that we use a Wishart hyperprior $S^{-1} \sim W(q, R^{-1}/q)$. The rows labeled $r = 0.5$ and $r = 2.0$ rescale the hyperparameter R by 0.5 and 2.0, respectively. The final set of four lines reports inference for different choices of the total mass parameter M, using fixed values of $M = 1$, 5 and 10. The row marked with 5^* reports inference under the G$amma(5, 1)$ hyperprior with $E(M) = 5$ reported in Section 5.2. The rows marked with $E(\sigma_{ij}^2) = 0.01$, $M = 5^*$ and $r = 1.0$ are identical. These are the choices reported in Section 5.2. Under this wide range of reasonable hyperprior parameters, we find the reported features of the posterior inference to be reasonably robust.

6. Conclusion

We have introduced semiparametric Bayesian inference for multi-level repeated measurement data. The nonparametric nature of the model are the random effects distribution for the first level random effects and the probability model for the joint distribution of random effects across second level repetitions.

The main limitation of the proposed approach is the computation-intensive implementation. Important directions of extensions for the proposed model are to different data formats, for example repeated binary data, and to a more structured model for the dependence across cycles. In the proposed model, dependence across cycles is essentially learned by clustering of the imputed random effects vectors for the observed patients. The approach works well for continuous responses with a non-linear regression model (1), assuming the residual variance is small enough to leave little posterior uncertainty for the θ_{ij}. The model is not appropriate for less informative data, for example binary data. Finally, the main reasons for choosing the DP prior were computational ease and the nature of the predictive rule (6). Both apply for a wider class of models known as species sampling models (Pitman, 1996; Ishwaran and James, 2003). Such models could be substituted for the DP model, with only minimal changes in the implementation.

Acknowledgments

Research was supported by NIH/NCI grant 1 R01 CA075981 and by FONDECYT (Chile) grants 1020712 and 7020712.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2**, 1152–1174.

Browne, W. J., Draper, D., Goldstein, H. and Rasbash, J. (2002). Bayesian and likelihood methods for fitting multilevel models with complex level-1

variation. *Computational Statistics and Data Analysis* **39**, 203–225.

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275–285.

Denison, D., Holmes, C., Mallick, B. and Smith, A. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley, New York.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., editors, *Bayesian Statistics 4*, Oxford. Oxford University Press.

Goldstein, H. (2003). *Multilevel Statistical Models*. Kendall's Library of Statistics. Edward Arnold Publishers Ltd.

Goldstein, H., Browne, W. J. and Rasbash, J. (2002). Multilevel modelling of medical data. *Statistics in Medicine* **21**, 3291–3315.

Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science* **15**, 1–19.

Ishwaran, H. and James, L. J. (2003). Generalized weighted chinese restaurant processes for species sampling mixture models. *Statistica Sinica* **13**, 1211–1235.

Ishwaran, H. and Takahara, G. (2002). Independent and identically distributed Monte Carlo algorithms for semiparametric linear mixed models. *Journal of the American Statistical Association* **97**, 1154–1166.

Jain, S. and Neal, R. M. (2004). A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational*

and Graphical Statistics **13**, 158–182.

Kleinman, K. and Ibrahim, J. (1998a). A semi-parametric bayesian approach to the random effects model. *Biometrics* **54**, 921–938.

Kleinman, K. P. and Ibrahim, J. G. (1998b). A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine* **17**, 2579–2596.

Lichtman, S. M., Ratain, M. J., Echo, D. A., Rosner, G., Egorin, M. J., Budman, D. R., Vogelzang, N. J., Norton, L. and Schilsky, R. L. (1993). Phase I trial and granulocyte-macrophage colony-stimulating factor plus high-dose cyclophosphamide given every 2 weeks: a Cancer and Leukemia Group B study. *Journal of the National Cancer Institute* **85**, 1319–1326.

Lin, X. H. and Carroll, R. J. (2001a). Semiparametric regression for clustered data. *Biometrika* **88**, 1179–1185.

Lin, X. H. and Carroll, R. J. (2001b). Semiparametric regression for clustered data using generalized estimating equations. *Journal Of The American Statistical Association* **96**, 1045–1056.

MacEachern, S. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation* **23**, 727–741.

MacEachern, S. N. and Müller, P. (2000). Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichlet Process Mixture Models. In Ruggeri, F. and Ríos-Insua, D., editors, *Robust Bayesian Analysis*, pages 295–316, New York. Springer-Verlag.

Mallet, A., Mentré, F., Gilles, J., Kelman, A., Thomson, A., S.M., Bryson and Whiting, B. (1988). Handling covariates in population pharmacoki-

netics with an application to gentamicin. *Biomedical Measurement Informatics and Control* **2**, 138–146.

Mukhopadhyay, S. and Gelfand, A. E. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association* **92**, 633–639.

Müller, P. and Quintana, F. (2004). Nonparametric Bayesian Data Analysis. *Statistical Science* **19**, 95–110.

Müller, P., Quintana, F. and Rosner, G. (2004). Hierarchical meta-analysis over related non-parametric bayesian models. *Journal of the Royal Statistical Society, Series B (Methodological)* **66**, 735–749.

Müller, P. and Rosner, G. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association* **92**, 1279–1292.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. In Ferguson, T. S., Shapeley, L. S. and MacQueen, J. B., editors, *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, IMS Lecture Notes, pages 245–268. IMS.

Smith, B. J. (2005). Bayesian Output Analysis program (BOA), version 1.1.5. http://www.public-health.uiowa.edu/boa.

Walker, S. and Wakefield, J. (1998). Population models with a nonparametric random coefficient distribution. *Sankhya, Series B* **60**, 196–214.

Table 1

Important inference summaries for low, medium and high choices of three hyperprior parameters. The columns report three important summaries related to Figure 4. See the text for details. The three horizontal blocks report inference when changing hyperprior parameters related to the nonlinear regression sampling model (first block, $E(\sigma^2)$), the prior on the random effects (marked with r), and the nonparametric prior (M). See the text for details.

	$P(y_{14} > 0)$			$\Delta P(y_{14} > 0)$			FNADIR		
	dose GM-CSF								
	2.5	5.0	10.0	2.5	5.0	10.0	2.5	5.0	10.0
$E(\sigma^2)$									
0.01	0.60	0.81	0.97	0.02	−0.07	−0.017	−1.7	−1.3	0.00
0.10	0.64	0.72	0.90	0.00	−0.27	−0.040	−1.6	−1.4	−0.54
1.00	0.62	0.58	0.97	0.04	−0.36	0.000	−1.5	−1.7	−0.34
r									
0.5	0.62	0.72	0.87	−0.10	0.14	0.090	−1.7	−1.1	−0.19
1.0	0.60	0.81	0.97	0.02	−0.07	−0.017	−1.7	−1.3	0.00
2.0	0.84	0.81	0.99	−0.12	−0.12	0.010	−1.6	−1.7	−0.37
M									
1	0.58	0.78	0.97	−0.09	−0.21	−0.010	−1.7	−1.4	−0.33
5	0.59	0.74	0.97	−0.08	−0.25	−0.010	−1.7	−1.5	−0.46
5*	0.60	0.81	0.97	0.02	−0.07	−0.017	−1.7	−1.3	0.00
10	0.56	0.68	0.98	−0.11	−0.30	−0.010	−1.6	−1.3	−0.40

Figure Legends

Figure 1. Model structure. Circles indicate random variables. Arrows indicate conditional dependence. The dashed box and the solid lines (without arrows) show how μ_i is partitioned into subvectors. The sampling model $p(y_{ijk} \mid \theta_{ij})$, the random effects model $p(\theta_{ij},\, j = 1, \ldots, n_i \mid G)$ and the nonparametric prior $p(G \mid \eta)$ are defined in (1), (3) and (5), respectively.

Figure 2. Repeated measurements over time (DAY) and cycles. Each panel shows data for one patient. Within each panel, the curves labeled 1, 2, and 3 show profiles for the first, second and third cycle of chemotherapy (only two cycles are recorded for patients 15 and 17). The curves show posterior estimated fitted profiles. The observed data are indicated by "1","2" or "3" for cycles 1,2 and 3, respectively.

Figure 3. Prediction for future patients treated at different levels of CTX and GM-CSF. For each patient we show the predicted response over the first three cycles as solid, dashed and dotted lines, respectively. CTX levels are 1.5, 3.0 and 4.5 g/m^2(labeled as 1,3 and 4 in the figure). GM-CSF doses are 2.5, 5 and 10 $\mu g/kg$(labeled as 3,5 and 10). Inference is conditional on a baseline of 2.0. Posterior predictive standard deviations are approximately 0.6.

Figure 4. Clinically relevant summaries of the inference across cycles: probability of $WBC > 1000$ on day 14 (left panel) and estimated nadir WBC count (right panel). The left panel shows the posterior probability of WBC above 1000 on day 14, plotted by treatment and cycle. The right panel shows the minimum WBC (in log 1000) plotted by treatment and cycle. Reported CTX doses are in $g/m^2$and GM-CSF doses are in $\mu g/kg$. The vertical error bars show plus/minus 1/2 pointwise posterior predictive standard deviation. We added a small horizontal offset to each line to avoid overlap.

Figure 5. Estimated $H(\theta)$. We show the bivariate marginals for cycle 1 and 2 for two relevant summaries of θ, for doses CTX=3 g/m^2 and GM-CSF=5 $\mu g/kg$. Panel (a) shows the estimated distribution of T_{lo}, the number of days that WBC is below 1000, for the first two cycles. Panel (b) shows the same for the minimum WBC (in log 1000). Panel (c) shows the same inference as Panel (b) for the parametric model. The distributions are represented by scatterplots of 500 simulated draws. For the integer valued variable T_{lo} we added additional noise to the draws to visualize multiple draws at the same integer pairs. For comparison, the 45 degree line is shown (dashed line).

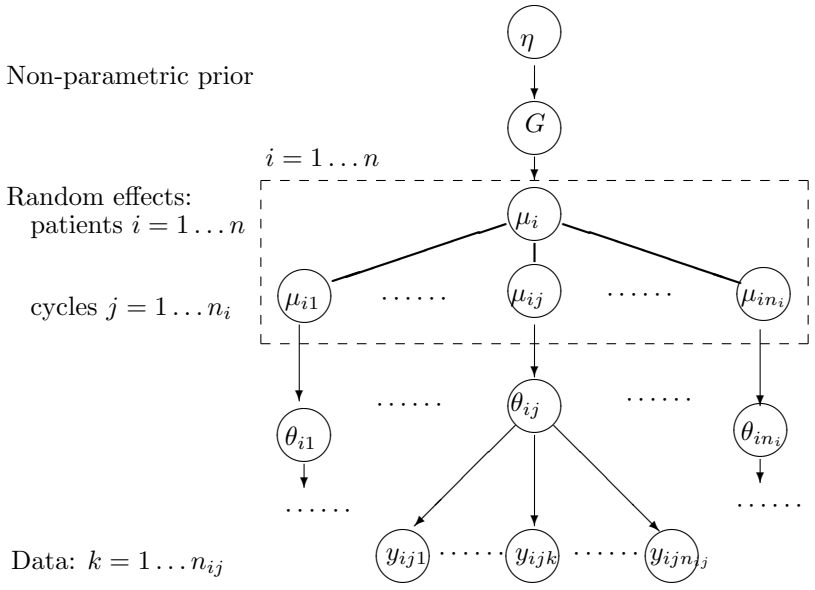

Non-parametric prior
η
G
$i = 1 \ldots n$
Random effects:
patients $i = 1 \ldots n$
μ_i
cycles $j = 1 \ldots n_i$
μ_{i1}
μ_{ij}
μ_{in_i}
θ_{i1}
θ_{ij}
θ_{in_i}
Data: $k = 1 \ldots n_{ij}$
y_{ij1}
y_{ijk}
$y_{ijn_{ij}}$

Figure 1.

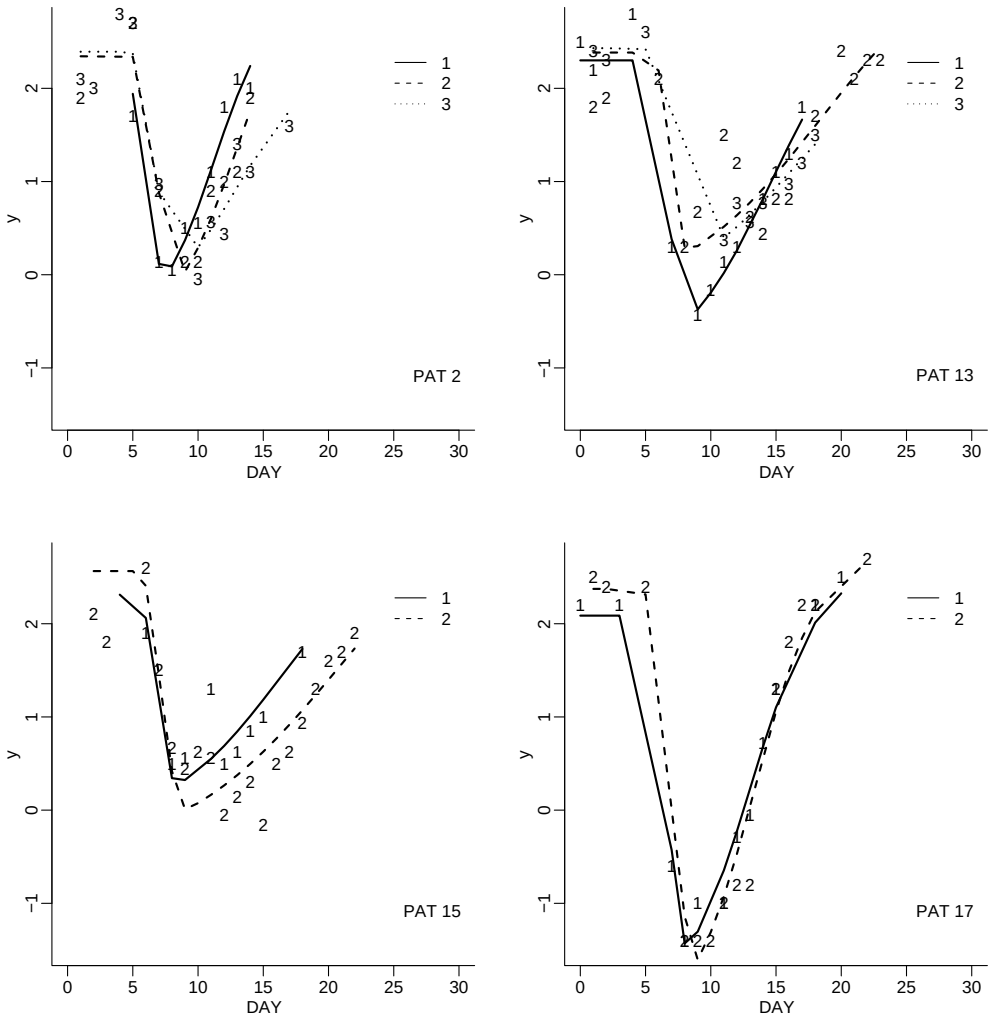


Figure 2.

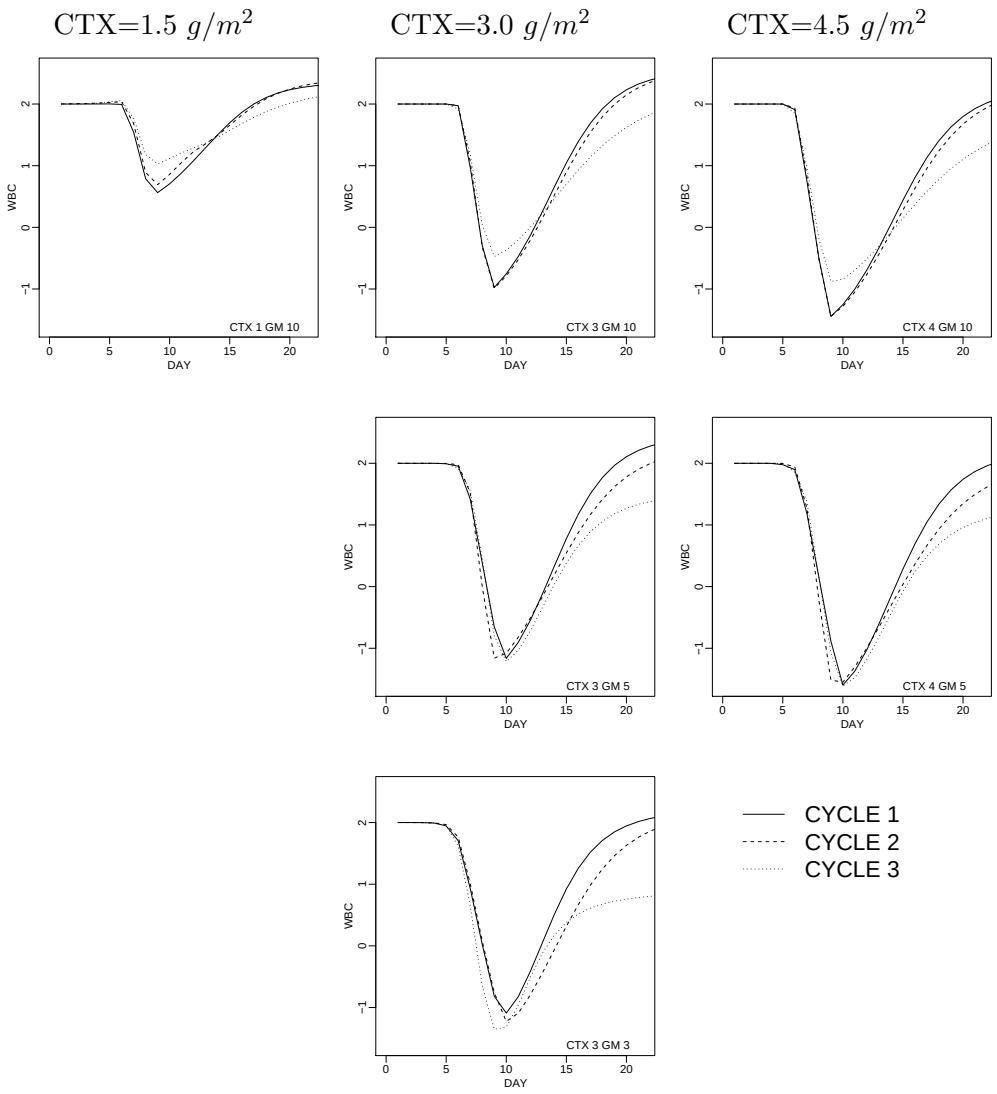
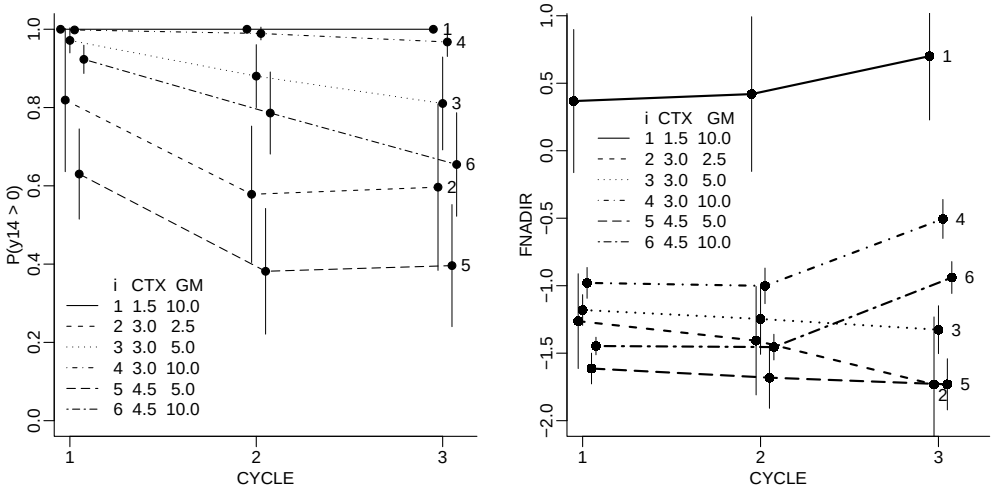
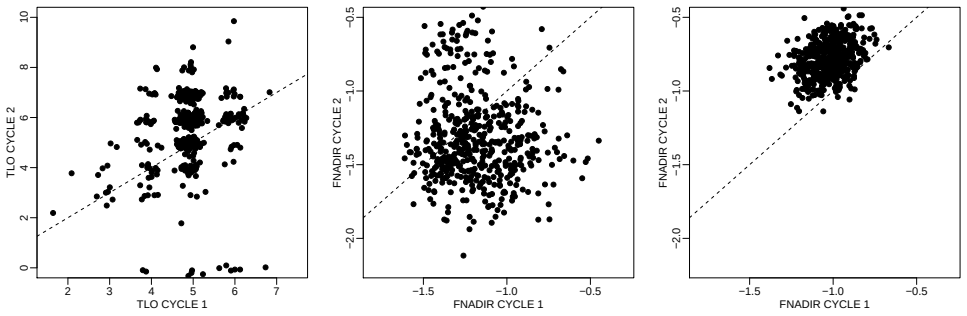

CTX=1.5 g/m^2
CTX=3.0 g/m^2
CTX=4.5 g/m^2
WBC
DAY
CTX 1 GM 10
CTX 3 GM 10
CTX 4 GM 10
CTX 3 GM 5
CTX 4 GM 5
CTX 3 GM 3
CYCLE 1
CYCLE 2
CYCLE 3

Figure 3.



(a) $P(y14 > 0 \mid Y)$ (b) Minimum WBC

Figure 4.



(a) T_{lo} for cycle 1 and 2 (b) Minimum WBC for cycle 1 and 2 (c) Like (b) for the parametric model (2)

Figure 5.