Marginal Correlation in Longitudinal Binary Data Based on Generalized Linear Mixed Models

Marginal correlation in longitudinal binary data based on generalized linear mixed models

Tony Vangeneugden[1,2] **Geert Molenberghs**[2,3] **Annouschka Laenen**[2]
Helena Geys[2,4] **Caroline Beunckens**[2]
Cristina Sotto[2,5]

[1] Tibotec, Johnson & Johnson, Mechelen, Belgium
Email: tvangene@tibbe.jnj.com
[2] Hasselt University, I-BioStat, Diepenbeek, Belgium
[3] Katholieke Universiteit Leuven, I-BioStat, Leuven, Belgium
[4] Janssen Pharmaceutica, Johnson & Johnson, Beerse, Belgium
[5] School of Statistics, University of the Philippines, Diliman, Quezon City, Philippines

April 9, 2009

Abstract

This work aims at investigating marginal correlation within and between longitudinal data sequences. Useful and intuitive approximate expressions are derived based on generalized linear mixed models. Data from four double-blind randomized clinical trials are used to estimate the intra-class coefficient of reliability for a binary response. Additionally, the correlation between such a binary response and a continuous response is derived to evaluate the criterion validity of the binary response variable and the established continuous response variable.

Some Keywords: Binary data; Intraclass correlation; Random effects; Reliability; Variances.

1 Introduction

In applied sciences, one is often confronted with the collection of hierarchical data or repeated measures, in particular longitudinal or clustered data. Methods for continuous such data are centered around the well-developed linear mixed effects model (LMM, Verbeke and Molenberghs 2000); the same is true for software implementation. Drawing from the normal distribution, the LMM allows one to obtain marginal characteristics, such as marginal means, marginal covariate effects, and marginal correlation coefficients, in a very straightforward way. This is because the natural parameters in an LMM have a hierarchical and a marginal interpretation at the same time. Hence, deriving the intraclass correlation (ICC) from a random-intercept LMM is particularly straightforward and coincides

with the correlation from a compound-symmetric structure, the latter being the marginalization of the former. This makes the LMM a flexible tool to study psychometric reliability based on longitudinal data, as in Vangeneugden *et al* (2005). Reliability reflects the amount of error inherent in any measurement and hence, in a general sense, how replication of the administration would give a different result (Streiner and Norman 1995).

While also non-Gaussian outcomes are prominent, model formulation is less straightforward. One distinguishes between marginal and random-effects model families with now no easy relationship between both. An example of the marginal family is generalized estimating equations (GEE, Liang and Zeger 1986), whereas the generalized linear mixed model (GLMM, Breslow and Clayton 1993) is a well-known random-effects model. Whereas GEE is convenient and frequently used, it models the marginal regression function, treating the second and higher-order moments as nuisance. When the correlation is of primary scientific interest, e.g., when determining the ICC or studying reliability, a non-likelihood method like GEE has clear limitations. The GLMM has a full likelihood basis, but fails to produce the marginal correlations in an easy fashion, owing to a non-linear link function, as well as the mean-variance link (Molenberghs and Verbeke 2005, Chapter 16). Due to the flexibility of the GLMM, it is a viable modeling candidate, even when the marginal correlation is of interest. We will show that the derivation of such correlations is generally feasible and derive the intra-class correlation coefficient of *reliability*. Note that, in classical terms, reliability is defined as the variance attributed to the difference among subjects divided by the total variance (Shrout and Fleiss 1979) and therefore takes the form of the intra-class correlation coefficient. We also investigate correlation of this binary response variable with an established continuous, interval-scaled variable in view of the *criterion validity* of the derived response variable and the continuous response variable.

Reliability is an important aspect of any clinical-trial response. Fleiss (1986) states: "The most elegant design of a clinical study will not overcome the damage by unreliable or imprecise measurement." In clinical trials, one typically wants to differentiate among treatments. If reliability is low, the ability to differentiate between the subjects in the different treatment arm decreases. Fleiss lists consequences of *unreliability* and brings up reasons for attenuation of correlation in studies designed to estimate correlation between variables. First, he mentions poor reliability. A second cause is biased sample selection where patients are selected with a minimum level of a certain measurement with

low reliability. Third, an increased sample size for trials with a primary parameter with low reliability causes attenuation since one can then easily show that for a paired t-test, the required sample size becomes $n = n^*/R$ where R denotes the reliability coefficient and n^* is the required sample size for the true score, i.e., the required sample size in the ideal but hypothetical case where the reliability is equal to 100%. Investigators in the mental disorders traditionally have been more concerned with the reliability of their measures than have their colleagues in other medical specialties.

When the trials are finished and reported, it is astonishing how little attention is given to the observed reliability of a certain scale. Here, we propose a framework to study *trial- or population-specific reliability*. Clinical trial data can be used to make progress when studying reliability as well as generalizability in case of interval scaled data (Vangeneugden *et al* 2004, 2005), given one is willing to make a number of assumptions, enabling one to "translate" biomedical data to a parallel measurements setting. The softer an endpoint or the less it has been calibrated, the more crucial psychometric validation becomes. Such analyses focus on variance components rather than treatment differences and can provide insight into scale behavior in (sub)populations: trial-population specific reliability coefficient can be produced and via generalizability testing, sources of variation and their impact on reliability can be studied. Here, a general formula will be derived to handle broad classes of data, with an application to a binary response. The goal is to use clinical trial data at hand and to evaluate reliability of the binary response. The intention is not to replace up-front validity and reliability testing but to stimulate *post hoc* evaluation on the performance of the scale or any other measurement. These methods can also deliver a population-trial specific measure for reliability in case there is a need to confirm earlier reliability testing results; regulatory authorities might question reliability of the scale in the specific trial population.

The *validity* of a questionnaire is defined as the degree to which the questionnaire measures what it purports to measure. This can be performed through the analysis of *content*, *construct*, and *criterion validity*. Content validity can be defined as the extent to which the instrument assesses all the relevant or important content or domains. Criterion validity can be divided into two types: *concurrent validity* and *predictive validity*. With concurrent validity we correlate the measurement with a criterion measure (gold standard), both of which are given at the same time. In predictive validity, the criterion will not be available until some time in the future at which time the true endpoint

is actually observed. This also clearly links validity testing to surrogate marker validation as shown in Alonso *et al* (2002). Of course, while measures of correlation are an important aspect of surrogacy evaluation, there is more to it than this (Baker and Kramer 2003, Burzykowski, Molenberghs, and Buyse 2005). The most commonly used method to assess the validity is by calculation of the Pearson correlation coefficient.

Thus, using concepts of Vangeneugden *et al* (2004, 2005), we will show how correlations can be derived by means of a GLMM, with particular attention to the reliability functions, operationalized by means of the ICC. At the same time, our results apply to settings quite different form psychometric validation, where nevertheless marginal correlations are of interest. It will be clear in what follows that, in the non-Gaussian case, reliability will no longer be constant, excepting special cases. And correlation between concurrently measured response variables will be derived via fitting a joint, bivariate GLMM. Our framework allows for derivation of a correlation coefficient between two response variables of any kind. As an example, the correlation between a binary response and an interval scaled response will be derived to investigate criterion validity between the derived binary response and the more standard continuous response.

In Section 2, the motivating case study is introduced, while methodology is described in Section 3. Section 4 reports simulations, directed at evaluating the quality of the approximation. In Section 5, we will apply the derived formulae to the data introduced above to estimate reliability of a binary response variable. In Section 6, we will extend the methodology to calculate correlation between concurrently measured responses to study criterion validity.

2 Motivating Study

In this section, we introduce individual patient data from four double-blind randomized clinical trials, comparing the effects of risperidone to conventional anti psychotic agents for the treatment of chronic schizophrenia. Schizophrenia has long been recognized as a heterogeneous disorder with patients suffering from both "negative" and "positive" symptoms. Negative symptoms are characterized by deficits in social functions such as poverty of speech, apathy and emotional withdrawal. Positive symptoms entail more florid symptoms such as delusions and hallucinations, which are superimposed

on the mental status. Several measures can be considered to assess a patient's global condition. The *Positive and Negative Syndrome Scale* (PANSS) consists of 30 items that provide an operationalized, drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia (Kay, Fiszbein, and Opler 1987). Classical reliability of the PANSS has been studied previously (Kay, Opler and Lindenmayer 1988; Bell *et al* 1992; Peralta and Cuesta 1994). The *Clinical Global Impression* (CGI) of overall change versus baseline is a 7-grade scale used by the treating physician to characterize how well a subject has improved since baseline. The levels are: "very much improved", "much improved", "minimally improved", "no change", "minimally worse", "much worse", "very much worse". Clinical response is often defined as a CGI score of "very much improved" or "much improved". Note that, while the psychometric characteristics of the PANSS scale are known to be very good, CGI has been in common use, for regulatory and other purposes. It is therefore important to study its merits, also relative to PANSS. This may guide researchers and regulators when choosing endpoints in future trials. Since the label in most countries recommend doses ranging from 4-6 mg/day, we include in our analysis only patients who received either these doses of risperidone or an active control (haloperidol, perphenazine, or zuclopenthixol). Depending on the trial, treatment was administered for a duration of 6-8 weeks. For example, in the international trials by Peuskens *et al* (1995), Marder and Meibach (1994), and Hoyberg *et al* (1993) patients received treatment for 8 weeks; while in the study by Huttunen *et al* (1995) patients were treated over a period of 6 weeks. The sample sizes were 453, 176, 74, and 71, respectively. Measurements were taken at Week 1, 2, 4, 6, and 8.

3 Methodology

First, Vangeneugden *et al* (2004) derived the intra-class correlation coefficient (ICC) of reliability for the classical linear mixed-effects model. Then we introduce the generalized linear mixed model and subsequently we derive an approximate formula for the variance-covariance matrix based on a GLMM. The latter will be the basis for general correlation coefficient computations.

3.1 ICC for a Linear Mixed-effects Model

A linear mixed-effects model with serial correlation can be written as (Verbeke and Molenberghs 2000): $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{W}_i + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{Y}_i$ is the n_i dimensional response vector for subject i, $1 \leq i \leq N$, N is the number of subjects, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $(n_i \times p)$ and $(n_i \times q)$ known design matrices, $\boldsymbol{\beta}$ is the p dimensional vector containing the fixed effects, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$ is the q dimensional vector containing the random effects, $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$ is a n_i dimensional vector of measurement error components, and $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N$ are assumed to be independent. Apart from the random-effects variances (the diagonal elements of D), there can be covariances as well, allowing for correlation between the random effects. For example, with growth curves, if subjects starting high also grow fast, the random intercept and random slope in time would be positively correlated. If in addition, there is serial correlation, it is captured by the realization of a Gaussian stochastic process, $\boldsymbol{W}_i$, which is assumed to follow a $N(\boldsymbol{0}, \tau^2\boldsymbol{H}_i)$ law. The serial correlation matrix $\boldsymbol{H}_i$ only depends on i through the number n_i of observations and through the time points t_{ij} at which measurements are taken. The structure of the matrix $\boldsymbol{H}_i$ is determined through the autocorrelation function $\rho(t_{ij} - t_{ik})$. A first simplifying assumption is that it depends only on the time interval between two measurements Y_{ij} and Y_{ik}, i.e., $\rho(t_{ij} - t_{ik}) = \rho(|t_{ij} - t_{ik}|)$, where $u = |t_{ij} - t_{ik}|$ denotes time lag. This function decreases such that $\rho(0) = 1$ and $\rho(+\infty) = 0$. Finally, $\boldsymbol{D}$ is a general $(q \times q)$ covariance matrix with (i, j) element $d_{ij} = d_{ji}$.

In this setting, it is easy to show that for subject i on time point j and k we have $\text{Var}(Y_{ij}) = \boldsymbol{z}_j\boldsymbol{D}\boldsymbol{z}_j' + \tau^2 + \sigma^2$, $\text{Var}(Y_{ik}) = \boldsymbol{z}_k\boldsymbol{D}\boldsymbol{z}_k' + \tau^2 + \sigma^2$, and $\text{Cov}(Y_{ij}, Y_{ik}) = \boldsymbol{z}_j\boldsymbol{D}\boldsymbol{z}_k' + \tau^2(\boldsymbol{H}_i)_{jk}$, and therefore, the correlation between time point j and k, the ICC of reliability can be written as:

$$\rho_{jk} = \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\boldsymbol{z}_j\boldsymbol{D}\boldsymbol{z}_k' + \tau^2(\boldsymbol{H}_i)_{jk}}{\sqrt{\boldsymbol{z}_j\boldsymbol{D}\boldsymbol{z}_k' + \tau^2 + \sigma^2}\sqrt{\boldsymbol{z}_k\boldsymbol{D}\boldsymbol{z}_k' + \tau^2 + \sigma^2}}. \tag{1}$$

In case of a simple random-intercept model, $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + b_i + \boldsymbol{\varepsilon}_i$, with no serial correlation, (1) simplifies to:

$$\rho_{st} = \text{Corr}(Y_{is}, Y_{it}) = \frac{d}{d + \sigma^2}, \tag{2}$$

where d is the variance of the random intercept b_i, i.e., the variance between patients, and σ^2 the measurement error. See Vangeneugden *et al* (2004) for the derivation of the ICC of reliability for more complex linear models.

3.2 ICC Based on the Generalized Linear Mixed Model

The generalized linear mixed model (GLMM, Breslow and Clayton 1993) is the most frequently used random effects model for discrete outcomes. As before, Y_{ij} is the jth outcome measured for subjects $i, i = 1, \ldots, N, j = 1, \ldots, n_i$ and $\boldsymbol{Y}_i$ is the n_i-dimensional vector of all measurements available for cluster i. This model assumes that, conditionally on q-dimensional random effects $\boldsymbol{b}_i$, assumed to be drawn independent from the $N(\boldsymbol{0}, \boldsymbol{D})$, the outcomes Y_{ij} are independent with densities of the form $f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) = \exp\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij}, \phi)\}$, where the mean μ_{ij} is modeled through a linear predictor containing fixed regression parameters $\boldsymbol{\beta}$ as well as subject-specific parameters $\boldsymbol{b}_i$, i.e., $g(\mu_{ij}) = g(E(Y_{ij}|\boldsymbol{b}_i)) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i$ for a known link function $g(.)$, with $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ p-dimensional and q-dimensional vectors of known covariate values, with $\boldsymbol{\beta}$ a p-dimensional vector of unknown fixed regression coefficients, and with ϕ a scale parameter. With a natural link function this becomes $\theta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i$. The random effects $\boldsymbol{b}_i$ are assumed to be sampled from a (multivariate) normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{D}$.

In this GLMM setting, we can write the general model as follows:

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i, \tag{3}$$

where $\boldsymbol{\mu}_i$, the conditional mean, given the random effects, can be written as $\boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{\theta}_i) = \boldsymbol{\mu}_i(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i)$, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are known design matrices, $\boldsymbol{\beta}$ are fixed-effect parameters, $\boldsymbol{b}_i$ are random effects, and h is a known link function. Finally, $\boldsymbol{\varepsilon}_i$ is the residual error component. At first sight, decomposition (3) is somewhat unusual and may be perceived as restrictive. However, it is a device used by, for example, Nelder and Wedderburn (1972) and McCullagh and Nelder (1989), for univariate data, and by Wolfinger and O'Connell (1993) for vector-valued outcomes in the context of generalized linear mixed models, which is also our setting. It is merely a way of decomposing an observation into its systematic and stochastic components. For continuous outcomes, it is customary to assume the error is normally distributed. For binary data, the outcome can take two values only, and then some algebraic calculations show that the error structure automatically will be of a Bernoulli type, as it ought to. Therefore, (3) does not need to be seen as restrictive.

We will now derive a general formula for the variance-covariance matrix of $\boldsymbol{Y}_i$ without any restriction on the distribution of the outcome variable and allowing for serial correlation. This maximizes the

similarity with the case of continuous, normally distributed outcomes. However, a key distinction is that in the linear case there is no mean-variance link, whereas here the residual variance will follow from the mean. The variance covariance matrix can be derived as follows:

$$\boldsymbol{V}_i = \text{Var}(\boldsymbol{Y}_i) = \text{Var}(\boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i) = \text{Var}(\boldsymbol{\mu}_i) + \text{Var}(\boldsymbol{\varepsilon}_i) + 2\text{Cov}(\boldsymbol{\mu}_i, \boldsymbol{\varepsilon}_i). \tag{4}$$

It is easy to show that $\text{Cov}(\boldsymbol{\mu}_i, \boldsymbol{\varepsilon}_i) = \text{Cov}[\text{E}(\boldsymbol{\mu}_\text{i}|\boldsymbol{b}_\text{i}), \text{E}(\boldsymbol{\varepsilon}_\text{i}|\boldsymbol{b}_\text{i})] + \text{E}[\text{Cov}(\boldsymbol{\mu}_\text{i}, \boldsymbol{\varepsilon}_\text{i}|\boldsymbol{b}_\text{i})] = 0$ since the first term is 0 and the second term equals $\text{E}[\text{E}(\boldsymbol{\mu}_\text{i} - \text{E}(\boldsymbol{\mu}_\text{i}))(\boldsymbol{\varepsilon}_\text{i})|\boldsymbol{b_i}] = 0$ as $\boldsymbol{\mu}_i$ is a constant when conditioning on $\boldsymbol{b}_i$. For the first term in (4) we have, using a first-order Taylor series expansion around $\boldsymbol{b}_i = \boldsymbol{0}$:

$$\begin{aligned}
\text{Var}(\boldsymbol{\mu}_i) &= \text{Var}(\boldsymbol{\mu}_i(\boldsymbol{\eta_i})) = \text{Var}(\boldsymbol{\mu}_i(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i)) &(5)\\
&\cong \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{b}_i}\Big|_{\boldsymbol{b}_i=0}\right) \text{Var}(\boldsymbol{b}_i) \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{b}_i}\Big|_{\boldsymbol{b}_i=0}\right)' \\
&= \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta_i}}\frac{\partial \boldsymbol{\eta_i}}{\partial \boldsymbol{b}_i}\Big|_{\boldsymbol{b}_i=0}\right) \boldsymbol{D} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta_i}}\frac{\partial \boldsymbol{\eta_i}}{\partial \boldsymbol{b}_i}\Big|_{\boldsymbol{b}_i=0}\right)' = \boldsymbol{\Delta}_i\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{\Delta}_i', &(6)
\end{aligned}$$

where $\boldsymbol{\Delta}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta_i}}\Big|_{\boldsymbol{b}_i=0}$. For the second term in (4), we have:

$$\text{Var}(\boldsymbol{\varepsilon}_i) = \text{Var}[E(\boldsymbol{\varepsilon}_i|\boldsymbol{b}_i)] + E[\text{Var}(\boldsymbol{\varepsilon}_i|\boldsymbol{b}_i)] = E[\text{Var}(\boldsymbol{\varepsilon}_i|\boldsymbol{b}_i)] = \boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{\Sigma}_i\boldsymbol{\Phi}^{\frac{1}{2}}, \tag{7}$$

where $\boldsymbol{\Phi}$ is a diagonal matrix with the overdispersion parameters along the diagonal. In case there are no overdispersion parameters, $\boldsymbol{\Phi}$ is set equal to the identity matrix. We can expand the variance function $\boldsymbol{\Sigma}_i$ so that

$$\text{Var}(\boldsymbol{\varepsilon}_i) \quad = \boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{R}_i\boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{\Phi}^{\frac{1}{2}}, \tag{8}$$

where $\boldsymbol{R}_i$ is the correlation matrix, capturing serial correlation, similar in spirit to the linear mixed model where $\boldsymbol{H}_i$ is used to this end; $\boldsymbol{A}_i$ is a diagonal matrix containing the variances following from the generalized linear model specification of $\boldsymbol{Y}_{ij}$ given the random effects $\boldsymbol{b}_i = \boldsymbol{0}$, i.e., with diagonal elements $v(\mu_{ij}|\boldsymbol{b}_i = \boldsymbol{0})$. Using (6) and (8), we have the following expression for the variance-covariance matrix (4):

$$\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{\Delta}_i' + \boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{R}_i\boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{\Phi}^{\frac{1}{2}}. \tag{9}$$

If the canonical link is used, we have $\boldsymbol{A}_i = \boldsymbol{\Delta}_i$ and (9) can be written as: $\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{\Delta}_i' + \boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{\Delta}_i^{\frac{1}{2}}\boldsymbol{R}_i\boldsymbol{\Delta}_i^{\frac{1}{2}}\boldsymbol{\Phi}^{\frac{1}{2}}$. If in addition, conditional independence (no serial correlation) is assumed, then

(9) simplifies to: $\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' \boldsymbol{\Delta}_i' + \boldsymbol{\Phi}^{\frac{1}{2}} \boldsymbol{\Delta}_i \boldsymbol{\Phi}^{\frac{1}{2}}$. Further, if we reduce the random-effects part to a random-intercept model, i.e., $\boldsymbol{Z}_i = \boldsymbol{1}$ and $\boldsymbol{D} = d$, and (9) then reduces to $\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i (d\boldsymbol{J}) \boldsymbol{\Delta}_i' + \boldsymbol{\Phi}^{\frac{1}{2}} \boldsymbol{\Delta}_i \boldsymbol{\Phi}^{\frac{1}{2}}$. Note that, if we have a normal distribution with the canonical identity link, $\boldsymbol{\Delta}_i$ reduces to the identity matrix $\boldsymbol{I}$ and $\boldsymbol{\Phi} = \sigma^2 \boldsymbol{I}$, in which case it follows that $\boldsymbol{V}_i$ reduces to $d\boldsymbol{J} + \sigma^2 \boldsymbol{I}$, with $\boldsymbol{J}$ a square n_i dimensional matrix of ones, which is consistent with (2). Moreover, when we have a normal distribution with a general random-effects structure but without serial correlation, it is easy to show that $\boldsymbol{V}_i \cong \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + \sigma^2 \boldsymbol{I}$ and that subsequently ρ equals (1) when we leave out the serial correlation (τ). This shows that (9) can be seen as a generalization of (1). While the above derivation is referred to as a first-order Taylor series expansion, the exact same expression follows if a second-order expansion is considered, owing to terms vanishing. Therefore, we are authorized to refer to it as a second-order Taylor series expansion, too. In the following section we will derive the marginal correlation for the case of binary data when applying a random intercept model.

3.3 ICC for a Random-intercept Model for Binary Data

In this section, we will derive the formula for the ICC, being the marginal correlation function, in case of a random intercept model for binomial data with a logit link and assuming no overdispersion. In this case, $\boldsymbol{V}_i$ reduces to $\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i (d\boldsymbol{J}) \boldsymbol{\Delta}_i' + \boldsymbol{\Delta}_i = \boldsymbol{\Delta}_i (d\boldsymbol{J} + \boldsymbol{\Delta}_i^{-1}) \boldsymbol{\Delta}_i'$. Furthermore, $\boldsymbol{\Delta}_i$ is a diagonal matrix with $V_{ij}(0)$ as diagonal elements, where the variance function $V_{ij}(0) = \mu_{ij} \mid_{\boldsymbol{b}_i = 0} (1 - \mu_{ij} \mid_{\boldsymbol{b}_i = 0})$, and therefore $\boldsymbol{V}_i \cong \mathrm{diag}(V_{ij}(0))[d\boldsymbol{J} + \mathrm{diag}(V_{ij}(0))^{-1}]\mathrm{diag}(V_{ij}(0))$. In other words, the variance-covariance matrix for subject i is specified by the matrix with elements: $v_{ijj} = V_{ij}(0)[1 + V_{ij}(0)d]$, $v_{ijk} = dV_{ij}(0)V_{ik}(0), \quad (j \neq k)$. Based on these, we can determine a first-order approximation of the marginal correlation between time point j and k, which is the intra class correlation coefficient of reliability:

$$\rho_{ijk} = \mathrm{Corr}(Y_{ij}, Y_{ik}) = \frac{V_{ij}(0)V_{ik}(0)d}{\sqrt{\{V_{ij}(0)[1 + V_{ij}(0)d]\}\{V_{ik}(0)[1 + V_{ik}(0)d]\}}}. \tag{10}$$

It ought to be noted that this is a very special case of the general result (9). The latter is useful when the mean structure, the random-effects structure, and/or the serial structure takes a general form. The expression is useful though, because it allow to make a few simple but important observations. For any value of $V_{ij}(0)$ and $V_{ik}(0)$, $\rho_{ijk} = 0$ whenever $d = 0$, while ρ_{ijk} tends to 1 when d tends to

$+\infty$. Even though this may seem obvious at first sight, especially because it is similar to the behavior of the intraclass correlation in the classical linear model for continuous data, one must give proper reflection to the impact of the binary nature of our outcomes, since certain correlation coefficients in certain models are highly constrained. For example, the correlation coefficients in the Bahadur (1961) model are highly constrained (Aerts *et al* 2002). These authors showed that in some realistic settings only a tiny interval around zero of allowable correlations remains. It is useful to realize that such constraints already apply to the Pearson correlation in a simple two by two contingency table. A mild form of the Bahadur constraints survives in generalized estimating equations, especially those of the second order. The multivariate probit model (Molenberghs and Verbeke 2005), on the other hand, is constrained only by the requirement that their correlations form a positive definite matrix. This advantage of the probit model is counterbalanced by its heavy computational burden. Also, the beta-binomial model (Molenberghs and Verbeke 2005) allows for all non-negative correlations as well as moderate negative values (Molenberghs and Verbeke 2005). The beta-binomial model suffers from its inability to accommodate within-cluster covariates, such as time in longitudinal studies. Thus, the proposed modeling framework is at the same time flexible, relatively easy from a numerical point of view, and does not face the strong constraints like in, for example, the Bahadur (1961) model.

One might wonder why no negative correlations are allowed. Also this aspect is similar to the linear mixed model, where the random-intercepts model, when its full hierarchical interpretation is adopted, does not allow for negative correlations. Once attention is restricted to the marginal model, some negative correlation can occur as well. Indeed, the compound-symmetry model can produce negative correlations, as long as the overall correlation matrix, of the form $\sigma^2\boldsymbol{I} + d\boldsymbol{J}$, remains positive-definite. Note that, while this article focuses on the correlation coefficient, also in line with classical reliability approaches, other measures of association between the outcomes, such as the odds ratio model (Molenberghs and Verbeke 2005) could be entertained. Arguably, this would require a fundamentally different approach, and is beyond the scope of this article.

4 Simulation Study

A reason for concern is the quality of approximation (10) since, unlike in the linear case, here a Taylor series expansion needs to be used. To provide a perspective on the impact of this issue, we conducted

a limited but insightful set of simulations. Precisely, we generated data from the Bahadur (1961) model, and then estimated the correlation coefficient using both generalized estimating equations (GEE, Liang and Zeger 1986) and our proposed approach. While it ought to be noted that a correlation coefficient for non-continuous data is a model-dependent concept, the relative agreement between the coefficients resulting from the various models still sheds some light on the quality of the approximation.

The Bahadur model is defined in terms of the marginal probability $\pi_{ij} = E(Y_{ij}) = P(Y_{ij} = 1)$ and standardized deviations $\varepsilon_{ij} = (Y_{ij} - \pi_{ij})/\sqrt{\pi_{ij}(1-\pi_{ij})}$ and $e_{ij} = (y_{ij} - \pi_{ij})/\sqrt{\pi_{ij}(1-\pi_{ij})}$, where y_{ij} is an actual value of the binary response variable Y_{ij}. Further, letting $\rho_{ij_1j_2} = E(\varepsilon_{ij_1}\varepsilon_{ij_2})$, $\rho_{ij_1j_2j_3} = E(\varepsilon_{ij_1}\varepsilon_{ij_2}\varepsilon_{ij_3})$, ..., $\rho_{i12\ldots n_i} = E(\varepsilon_{i1}\varepsilon_{i2}\ldots\varepsilon_{in_i})$, the general Bahadur model can be represented by the expression $f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i)c(\boldsymbol{y}_i)$, where $f_1(\boldsymbol{y}_i) = \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}$ and $c(\boldsymbol{y}_i) = 1 + \sum_{j_1<j_2} \rho_{ij_1j_2}e_{ij_1}e_{ij_2} + \sum_{j_1<j_2<j_3} \rho_{ij_1j_2j_3}e_{ij_1}e_{ij_2}e_{ij_3} + \ldots + \rho_{i12\ldots n_i}e_{i1}e_{i2}\ldots e_{in_i}$. For the purpose of our simulations, we will restrict this model to 2 and 3 measurements per subject, respectively. In the latter case, the three pairwise correlation will be set equal, while the third-order correlation will be set to zero. GEE can be viewed as a version of the Bahadur model where the higher-order correlations are left unspecified, and the pairwise correlation structure is considered a nuisance characteristic.

For the number of measurements equal to $n_i = n = 2$, the true correlations $\rho = 0.25$, 0.50, and 0.75 were considered, while for $n_i = n = 3$ we focused on $\rho = 0.20$, 0.40, and 0.60. For all six settings, 1000 datasets of size 200 patients were generated. For each such dataset, the pairwise correlation was estimated using both GEE and the proposed GLMM-based expression (10). Table 1 presents the results in terms of the simulation-averaged correlation together with its standard deviation. Note that, for the three-way model, there is a common GEE correlation, while for GLMM the correlation is specific to a pair of time points, indicated by ρ_{jk}, with $1 \leq j < k \leq 3$. Not surprising, the agreement between GEE and the generating Bahadur model is excellent, because GEE can be viewed as a restricted-moment version of the Bahadur model. We believe it is at the same time useful to study the performance of the random-effects models, because in many applications such will be the modeler's preference given that it allows for subject-specific inferences. At the same time, one might then want to derive marginal association parameters, underscoring the rationale for our work.

Table 1: *Results of the simulation study. n refers to the number of measurements per subject. 'True' is the correlation used in the generating Bahadur model. For both GEE and GLMM, the simulation-averaged correlation coefficients and their simulation standard deviations are reported.*

		GEE	GLMM	
$n = n_i$	True ρ	Est.(s.d.)	Coeff	Est.(s.d.)
2	0.25	0.248 (0.07)	ρ	0.270 (0.08)
2	0.50	0.499 (0.06)	ρ	0.554 (0.07)
2	0.75	0.753 (0.05)	ρ	0.568 (0.30)
3	0.20	0.199 (0.05)	ρ_{12}	0.225 (0.06)
			ρ_{13}	0.228 (0.06)
			ρ_{23}	0.237 (0.06)
3	0.40	0.398 (0.05)	ρ_{12}	0.467 (0.06)
			ρ_{13}	0.474 (0.06)
			ρ_{23}	0.497 (0.06)
3	0.60	0.598 (0.04)	ρ_{12}	0.666 (0.05)
			ρ_{13}	0.678 (0.05)
			ρ_{23}	0.723 (0.04)

Importantly for our purposes, the behavior of the GLMM-based expression (10) is quite acceptable. While, as stated earlier, the correlation is model-dependent, it falls everywhere within the same range as the one of the generating model. Note that, for our approach when $n = 3$, we have three coefficients, one for each pair of measurements. It would, in principle, be possible to replace the three estimates with a common one. Since this would come down in averaging the three correlations, it would further enhance stability. This is why we have chosen this somewhat more variable and therefore conservative presentation in terms of three separate coefficients.

Additionally, a simulation based on an actual GLMM was performed, using a simple random-effects model with $\boldsymbol{X}_i\boldsymbol{\beta} = \beta_0$. In this simulation, 10,000 datasets with 200 subjects were generated, each subject having 5 measurements as in the application of Section 5. Here, $\beta_0 = -1.61$ and the variance

of the random intercept, $d = 6.57$, was taken as observed in the application. Also here, the pairwise correlation was estimated using both GEE as well as the proposed GLMM-based expression (10). For the GEE, the mean correlation and its standard deviation was observed to be 0.465 (s.d. 0.04) and for GLMM the results were very similar, leading to a mean correlation of 0.473 (s.d. 0.05). Note that the GLMM based correlation of the real data was estimated to be 0.48.

Thus, we conclude that the correlation, based on GLMM, is a practically acceptable indication for association. In principle, it would be possible to further enhance performance using Monte-Carlo Markov Chain based methods, including the bootstrap. While such an approach would increase the computational burden somewhat, it certainly falls within the realm of practical feasibility.

5 Data Analysis

Let us now apply the concepts described above to the pooled data described in Section 2. We will calculate the ICC for response defined as obtaining either *very much improved* or *much improved* on the CGI of overall change versus baseline. The focus of this analysis is not to study treatment differences, but rather to investigate correlation between longitudinal binary data. To do so, we will calculate the ICC under different assumptions, with gradually increasing modeling complexity. For simplicity, we will focus on models with random intercepts and no serial correlation. Of course, as stated in Section 3, the extension to the more general case is straightforward but algebraically a bit more tedious.

5.1 Observed Response Rate and Correlation

The observed response rate increases over time from 0.15 at Week 1 to 0.47 at Week 8. Also note that only 490 from the 774 subjects who started treatment have an observed CGI score at Week 8 due to attrition. The correlation is high if we compare Week 1 and 2, but decreases slightly over time, when the lag time between observations is increased. On the other hand, the correlation between Week 6 and 8 is higher.

Table 2: *Summary of different subgroup analysis investigating time and treatment effect. Standard errors are calculated from the delta method.*

	Intraclass correlation ρ (s.e.)		
time points included	combined treatments	risperidone	active control
all time points	0.48 (0.026)	0.55 (0.038)	0.40 (0.035)
Week 1 and Week 8	0.11 (0.045)	0.11 (0.066)	0.10 (0.060)
Week 6 and Week 8	0.85 (0.026)	0.87 (0.032)	0.82 (0.043)

5.2 Initial Analysis

To exemplify computations, let us assume there are no covariates. Then, $\boldsymbol{X}_{is}\boldsymbol{\beta} = \beta$ is constant and (10) simplifies to: $V_{ij}(0) = V(0) = \exp(\beta)/(1+\exp(\beta))^2$ and $\rho_{ijk} = \rho = V(0)d/(1+V(0)d)$. When using this expression for a variety of subgroups and/or combination of times, a detailed picture can emerge but, as we will illustrate in what follows, it is possible and more elegant to incorporate the ICC into a fully specified model.

We can use the SAS procedure NLMIXED to fit this random-effects model, using adaptive Gaussian quadrature. Table 2 summarizes the results for a selection of subgroups. Before discussing these, let us note that subgroup analyses can rightfully be considered unsatisfactory by some. Therefore, we will revisit the concept of subgroups, but then in a more principled modeling approach, in Section 5.3. An added advantage of this approach is that the quality of the fit will be enhanced, owing to the high-quality approximation to the integration, required for likelihood evaluation. This is important, not only for the determination of the correlation coefficient, but also for other assessments, such as whether there is a significant treatment difference. Of course, one should be aware that reaching convergence with the NLMIXED procedure or related software for non-linear models is not straightforward. Tools exploiting linearity of the predictor are somewhat easier, but often based on poor approximations such as first-order PQL or MQL (Molenberghs and Verbeke 2005). Such alternative procedures may be used, however, to obtain good starting values, upon which the use of the non-linear procedures becomes easier.

Table 3: *Overall ICC (s.e.) matrix, marginal over treatment. Standard errors are calculated from the delta method.*

	Week			
Week	2	4	6	8
1	0.29 (0.029)	0.33 (0.030)	0.35 (0.029)	0.35 (0.029)
2	1	0.53 (0.032)	0.57 (0.030)	0.57 (0.029)
4		1	0.64 (0.027)	0.65 (0.026)
6			1	0.70 (0.024)

One observes that the ICC is somewhat larger in the risperidone treatment group. Additionally, we see that the ICC for observations measured at Week 1 and Week 8 is much smaller than the ICC measured from observations at Week 6 and Week 8. Here we should note that the ICC between Week 6 and 8 can truly be interpreted as an ICC of reliability in the psychometric sense. Indeed, the psychiatric condition of the patients was rather stable and did not change between Week 6 and 8: the mean total PANSS was 69.2 at Week 6 and 68.8 at Week 8. It is in such stable conditions that test-retest reliability of scale is evaluated, and often with a two-week time interval (Streiner and Norman 1995, Chapter 8). The same is not true when comparing Week 1 (mean PANSS of 80.8) and Week 8; that is, the ICC between Week 1 and 8 cannot be interpreted as an ICC of reliability but merely a correlation between two time points. As discussed in Vangeneugden *et al* (2004), appropriate models can be used to model and extract time and treatment effects, which avoids the need to assume that there is no change in a patient's situation over time. Thus, by using an appropriate model with well chosen covariate effects, a trial population is, in a broad sense, standardized towards a general population. By correcting for covariates, it is assumed that the correlation structure of the residuals can be approximated by an exchangeable structure, captured via a random intercept. While this may be perceived as somewhat more subjective than when a dedicated reliability study is undertaken, the important advantage is that data already collected can be used, which may have important practical, economic, and even ethical advantages. It is important to note that, in case a random intercept is deemed insufficient to capture the correlation structure, more versatile random-effects structures can be used, whilst maintaining the idea behind the calculations for the marginal correlation coefficients.

We will gradually take account of this, by first extracting time and then subsequently treatment effects. Of course, one ought not to forget that important but potentially complicated issues, such as dropout and non-compliance, may intervene. Since the method is likelihood-based, it is valid under the broad assumption of missingness at random, whereby missingness depends on observed outcomes and covariates but, given these, not further on unobserved outcomes. Likewise, when compliance issues intervene, it is important the covariates are chosen such that the causal interpretation of the resulting model be maintained. With good to perfect compliance, this is taken care of by virtue of randomization.

5.3 Accounting for Time and Treatment

If we adjust for time and ignore treatment, then ρ can be derived via (10) and it is easy to show that $V_{ij}(0) = \exp(\beta_j)/(1+\exp(\beta_j))^2$, where $\boldsymbol{\beta}_j$ is the estimated coefficient of the indicator variable representing time j, when we use a model without an intercept in the fixed effects. The variation of the random effect was estimated to be $\widehat{d} = 10.04$ and this time we had $\widehat{\beta_{W1}} = -3.79$, $\widehat{\beta_{W2}} = -2.25$, $\widehat{\beta_{W4}} = -1.50$, $\widehat{\beta_{W6}} = -3.79$ and $\widehat{\beta_{W8}} = -0.41$. Table 3 provides the estimated intra-class correlation coefficient matrix. This is in line with the well-known relationship between marginal and random-effects regression parameters (Molenberghs and Verbeke 2005), the correlations are determined by the random-intercept variance, together with the marginal probabilities factoring into the variance function: $\beta_j \cong \sqrt{1+0.346\,d}\cdot \text{logit}(p_j)$. Hence, these correlations are constant only in the simple case of a constant mean. Otherwise, they are functions of the covariates. Note that, in case a random-intercepts model is deemed too simple, a more elaborate random-effects structure can be assumed, whilst maintaining the essence of the proposed calculations.

When exploring Table 3, correlations clearly vary considerably. This indicates that pairs of measurements early in the sequence are less reliable for one another than pairs later in the sequence. Indeed, one can realistically assume that measurements earlier in the sequence are more prone to variability than later on, when subjects are more adapted to the study protocol and/or learning effects have taken place. If we repeat this for each treatment group separately, we consistently have a higher correlation coefficient in the risperidone treated subjects. Note that the ICC between observations from Week 6 and Week 8 ($\rho = 0.70$) is lower as estimated in the previous section ($\rho = 0.85$). In

the latter, however, only the subgroup of subjects with Week 6 and 8 was used, and if we apply the same model, accounting for time in this subgroup, then we have $\rho = 0.80$ instead of 0.70.

Jointly accounting for time and treatment produces a different ICC for each treatment group separately and also for each pair of time points. We allowed for interactions in the model. Table 4 summarizes the results. Apart from the estimated ICC, also the empirical Pearson (product-moment) correlation coefficients are added. The agreement between both is reasonable, especially when it is taken into account that the ICC does, but the Pearson correlation does not take the effect of covariates into account. After adjusting for time and treatment, the ICC between observations at Week 1 and 8 increased from 0.11 (2) to 0.40 in the risperidone group.

6 Extensions

So far, the application focused on correlation of repeated measures within a subject. As a specific application, the ICC was derived to estimate reliability of a binary response. Often, one is confronted with the situation that multiple response variables are measured over time, sometimes referred to as a family of responses. These different response variables can but do not have to be of the same type. Sometimes, the goal is to estimate treatment effects in a multivariate way, i.e., jointly estimate treatment effects on the binary and the continuous responses. In that case, one not only needs to take account of the correlation within a subject for a specific single response, but also take account of the correlation between the different responses for a specific subject. One application in the psychometric literature is the situation where one wants to estimate the correlation of a certain response variable with a gold standard to establish *criterion validity*. For instance, suppose we want to study the correlation between a continuous interval scaled parameter Y_{i1} and a binary response Y_{i2}, then a GLM can be extended too, as described in Molenberghs and Verbeke (2005):

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} = \begin{pmatrix} \mu_1 + \lambda b_i + \alpha_1 X_i \\ \frac{\exp[\mu_2 + b_i + \alpha_2 X_i]}{1 + \exp[\mu_2 + b_i + \alpha_2 X_i]} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix}.$$

Here, ε_{i1} and ε_{i2} are the error terms for the continuous and binary outcomes, respectively. Obviously, the first one will be normally distributed while the second one follows a Bernoulli distribution. We have included a scale parameter λ in the continuous component of an otherwise random-intercept

Table 4: *The first entries represent the overall ICC of reliability (s.e.) matrix, accounting for treatment, time and their interaction. Standard errors are calculated from the delta method. The second entries are the ordinary Pearson correlation coefficients between the pairs of measurements.*

Week	2	4	6	8
			risperidone	
1	0.36 (.045)	0.39 (.044)	0.40 (.042)	0.40 (.042)
	0.51	0.41	0.33	0.27
2	1	0.62 (.036)	0.64 (.033)	0.64 (.032)
		0.65	0.52	0.53
4		1	0.69 (.026)	0.69 (.026)
			0.70	0.61
6			1	0.71 (.023)
				0.75
			active control	
1	0.22 (.036)	0.27 (.038)	0.31 (.038)	0.31 (.038)
	0.52	0.34	0.33	0.27
2	1	0.42 (.046)	0.48 (.043)	0.49 (.041)
		0.59	0.49	0.43
4		1	0.57 (.039)	0.59 (.037)
			0.66	0.57
6			1	0.67 (.029)
				0.70

model, because the continuous and binary outcome are measured on a different scale. In this case,

we have

$$\boldsymbol{Z}_i = (\,\lambda\,)\,1, \boldsymbol{\Delta}_i = \begin{pmatrix} 1 & 0 \\ 0 & v_{i2}(0) \end{pmatrix}, \boldsymbol{\phi} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix},$$

with $v_{i2}(0) = \mu_i\,|_{\boldsymbol{b}_i=0}\,(1-\mu_i\,|_{\boldsymbol{b}_i=0})$. Note that $\boldsymbol{Z}_i$ is not a design matrix in the strict sense, since it contains an unknown parameter. Nevertheless, it is useful to consider this decomposition, implying that (9) becomes

$$\boldsymbol{V}_i = \begin{pmatrix} \lambda^2 & v_{i2}(0)\lambda \\ v_{i2}(0)\lambda & v_{i2}(0)^2 \end{pmatrix}\tau^2 + \begin{pmatrix} \sigma^2 & 0 \\ 0 & v_{i2}(0) \end{pmatrix} = \begin{pmatrix} \lambda^2\tau^2+\sigma^2 & v_{i2}(0)\lambda\tau^2 \\ v_{i2}(0)\lambda\tau^2 & v_{i2}(0)^2\tau^2 + v_{i2}(0) \end{pmatrix}.$$

Here, τ^2 is the random-intercept variance. As a result, we have the following approximation for the marginal correlation: $\rho(\beta) = v_{i2}\lambda\tau^2/\sqrt{\lambda^2\tau^2+\sigma^2}\sqrt{v_{i2}^2\tau^2+v_{i2}}$, which we can now apply to the same data set to estimate the correlation at Week 8 between the binary response variable defined above and the continuous response defined as the total PANSS, the sum of all 30 items of the PANSS. Table 5 summarizes the results. We can conclude that there was a high correlation between the response variable defined by the CGI and the total PANSS indicating criterion validity of the derived CGI response and the total PANSS. This correlation was similar in both treatment groups. Note that the correlation (-0.75 in the risperidone group and -0.74 in the control group) is negative because higher PANSS values indicate a more psychotic condition and response was coded 1 if the CGI was equal to "very much improved" or "much improved". In the classical approach, often either the Pearson, or the Spearman's rank correlation, or both, are calculated, based on subjects observed at Week 8, between the binary response and the continuous PANSS score, resulting in -0.59 and -0.61, for Pearson's and Spearman's correlation, respectively.

While in this section we have considered two outcomes of a different type, hence restricting attention to a cross-sectional setting, it is perfectly possible to combine the longitudinal ideas of previous sections with the multivariate setting considered here, thus producing a flexible method that can handle multivariate longitudinal data. One can then distinguish between various types of correlations, e.g., within-sequence (referring to the reliability concept), between two different measurements taken at the same time (of relevance in marker evaluation), and even between different measurements at different times. Details on how such models can be built and fitted are given in Molenberghs and Verbeke (Molenberghs and Verbeke 2005, Ch. 24).

Table 5: *Parameter estimates (standard errors) for a bivariate joint GLMM analysis to estimate criterion validity between response and total PANSS at Week 8. The SAS procedure NLMIXED has been used. Standard errors are calculated using the delta method.*

Endpoint	Effect	Parameter	Estimate	(s.e.)
Total PANSS	Intercept	μ_1	68.98	(1.59)
	Treatment	α_1	-0.41	(2.06)
	Standard deviation	σ_1	13.83	(0.43)
	Variation	σ_1^2	191.37	(11.90)
	Inflation	λ	-0.97	(0.61)
Response (CGI)	Intercept	μ_2	-2.56	(3.25)
	Treatment	α_2	0.96	(2.44)
Common parameters	R.I. st.dev.	τ	16.84	(10.73)
	R.I. var.	τ^2	283.74	(361.40)
	Corr. (control)	ρ_{cont}	-0.74	(0.026)
	Corr. (risperidone)	ρ_{ris}	-0.75	(0.022)

7 Discussion

We proposed an approximation to calculate correlations from longitudinal data from generalized linear mixed models. Whilst for continuous, interval scaled data, derivation of correlations, such as the ICC of reliability is rather straightforward, it is more complex for other types of data. A general formula was derived using the GLMM. This formula could be used for interval, binary or other types of data, such as counts. For our case study, the reliability coefficient was derived for a binary response, using a random-intercepts model. We observed that the correlation was higher between Week 6 and 8 as compared to Week 1 and Week 8. The slightly decreasing correlation, however, from Week 1 and Week 2 to Week 1 and Week 8 was not observed in the estimates. It should be noted that the random-effects model does also properly account for missing values due to attrition, provided the missing data are missing at random, which is not the case for the conventional

ad hoc analyses. In contrast, classical methods such as the kappa statistics, can only include paired observations. Another important advantage of the present method is that it becomes possible to estimate trial-specific or population-specific reliability. This is especially true because, even in studies designed to assess reliability, it is difficult to exclude fluctuations in the true scores and furthermore these studies are often conducted with different populations and in different circumstances. After extracting time and treatment effect and their interaction, clinical trial data can be used to make progress when studying test-retest reliability as a function of time. Indeed, reliability should not be perceived as a fixed quantity but changes with circumstances. Other covariates can bxincorporated into the model to study their effect on error variance and on reliability. Modeling other sources of variation, like for example country or rater, is therefore an interesting topic for further research. In psychometric theory, this is referred to as generalizability theory.

Subgroup analyses using a simple model and more versatile models accounting for time and treatment and their interaction suggested a higher ICC among subjects in the risperidone group than in subjects in the active control group, indicating that responses over time within the same subject were more consistent within the risperidone treatment group than in the active control group. The methodology can be used to derive population or trial-specific ICC of reliability in case of binary data. In particular, it extends the random intercepts model proposed in Vangeneugden *et al* (2004) to binary data. This general framework cannot only be used to derive the intraclass correlation coefficient or in general to study correlation of a single response variable of any type, but was also extended to investigate correlation between concurrently measured longitudinal data. Also here, a general framework was provided to deal with various possible situations. The correlation between the binary response derived from the CGI and the total PANSS was calculated and a high correlation was found between these two clinical endpoints at Week 8. A large number of general models, that can be fit using standard software, can be found in Molenberghs and Verbeke (Molenberghs and Verbeke 2005, Part V). Clearly, the quality of the proposed method hinges upon the accuracy of the Taylor series approximations employed. This not dissimilar to the well-known accuracy issues with Breslow and Clayton's (1993) PQL method. For a discussion, see Molenberghs and Verbeke (2005). Arguably, our method will perform reasonably well for two reasons. First, the parameter estimates plugged in are based on the accurate adaptive Gaussian quadrature, obtained with the NLMIXED procedure, rather than with the

expansion methods. Further, our method is a second-order rather than a first-order approximation, since the second-order terms vanish. This explains why the approximation is rather well, as confirmed by the simulation study. Therefore, even though, in principle, one could construct an approximation about the conditional estimates of the random effects, rather than around zero, this does not appear to be necessary for the correlation purposes of this work; this in itself is a nice feature, since such an approximation would involve lengthy and, implementation-wise, time-consuming computations. Furthermore, Rodríguez and Goldman (1995) documented that a Taylor series expansion might produce relatively accurate estimates for the variances, even though the parameter estimates, for binary data, might be biased.

The SAS developed used for this manuscript is available at the authors' web site, as well as upon simple request.

Acknowledgment

Financial support from the IAP research network #P6/03 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged.

References

Aerts, M., Geys, H., Molenberghs, G., Ryan, L. (2002). *Topics in Modelling of Clustered Data*. London: Chapman & Hall.

Alonso, A., Geys, H., Molenberghs, G., Vangeneugden, T. (2002). Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. *Journal of Biopharmaceutical Statistics*, **12**, 161–17.

Bahadur, R.R. (1961). A representation of the joint distribution of responses of p dichotomous items. In H. Solomon, editor, *Studies in item analysis and prediction*. Stanford, California, Stanford University Press.

Baker, S.G. and Kramer, B.S. (2003). A perfect correlate does not make a surrogate. *BioMed Central Medical Reseach Methodology*, **3**, 16.

Bell, M., Milstein, R., Beam-Goulet, J., Lysaker, P., and Cicchetti, D. (1992). The positive and negative syndrome scale and the brief psychiatric rating scale: Reliability, comparability, and predictive validity. *Journal of Nervous & Mental Disease*, **180**, 723–728.

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Burzykowksi, T., Molenberghs, G., and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints*. New York: Springer.

Fleiss, J.L. (1986). *Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons.

Hoyberg, O.J., Fensbo, C., Remvig, J., Lingjaerde, O., Sloth-Nielsen, M., and Salvesen, I. (1993). Risperidone versus perphenazine in the treatment of chronic schizophrenic patients with acute exacerbations. *Acta Psychiatrica Scandinavica*, **8**, 395–402.

Huttunen, M.O., Piepponen, T., Rantanen, H., Larmo, L., Nyholm, R., and Raitasuo, V. (1995). Risperidone versus zuclopenthixol in the treatment of acute schizophrenic episodes: a double-blind parallel-group trial. *Acta Psychiatrica Scandinavica*, **91**, 271–277.

Kay, S.R., Fiszbein, A., and Opler, L.A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, **13**, 261–276.

Kay, S.R., Opler, L.A., and Lindenmayer, J.-P. (1988) Reliability and validity of the Positive and Negative Syndrome Scale for schizophrenia. *Psychiatric Research*, **23**, 99–110.

Liang, K.-Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Marder, S.R. and Meibach, R.C. (1994). Risperidone in the treatment of schizophrenia. *American Journal of Psychiatry*, **151**, 825–835.

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

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *Journal of the Royal Statistical Society, Series B*, **135**, 370–384.

Peralta, V. and Cuesta, M.J. (1994). Psychometric properties of the positive and negative syndrome scale (PANSS) in schizophrenia. *Psychiatric Research*, **53**, 31–40.

Peuskens, J. and the Risperidone Study Group. (1995). Risperidone in the treatment of chronic schizophrenic patients: a multinational, multicentre, double-blind, parallel-group study versus haloperidol. *British Journal of Psychiatry*, **166**, 712–726.

Rodríguez, G. and Goldman, N. (1995) An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, **158**, 73–89.

Shrout, P.E. and Fleiss, J.L. (1979). Intraclass correlations: uses in assessing interrater reliability. *Psychological Bulletin*, **86**, 420–428.

Streiner, D.L. and Norman, G.R. (1995). *Health Measurement Scales*. Oxford University Press.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., and Molenberghs, G. (2004). Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Controlled Clinical Trials*, **25**, 13–30.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., and Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics*, **61**, 295–304.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Wolfinger, R. and O'Connell, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**, 233–243.