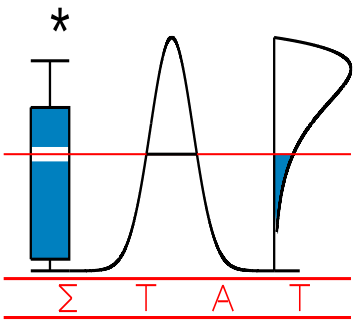
T E C H N I C A L

R E P O R T

0662

MARGINAL CORRELATION IN LONGITUDINAL BINARY DATA BASED ON GENERALIZED LINEAR MIXED MODELS

VANGENEUGDEN T., LAENEN A., GEYS H., RENARD D., and G. MOLENBERGHS



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

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

Tony Vangeneugden,[1,2,*] Annouschka Laenen,[2] Helena Geys,[2,3]

Didier Renard,[2,4] and Geert Molenberghs[2]

[1] Tibotec, Johnson & Johnson, Mechelen, Belgium

[2] Hasselt University, Center for Statistics, Diepenbeek, Belgium

[3] Janssen Pharmaceutica, Johnson & Johnson, Beerse, Belgium

[4] Eli Lilly, Mont-Saint-Guibert, Belgium

*email: tvangene@tibbe.jnj.com

SUMMARY. 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 parameter. 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.

Keywords: Binary data; Intraclass correlation; Random effects; Reliability; Variance Components.

1 Introduction

In applied sciences, one is often confronted with the collection of hierarchical data. For example, longitudinal data arise from measuring subjects repeatedly over time, while clustered data arise from measuring a characteristic in different but correlated subjects, such as siblings or foetuses stemming from the same dam. We will broadly refer to such settings as repeated measures. Methods for continuous longitudinal data form the best-developed and most advanced body of research, with a very prominent place taken by the linear mixed effects model (LMM, Laird and Ware, 1982; Verbeke and Molenberghs, 2000); the same is true for software implementation. This is natural, since the special status and the elegant properties of the normal distribution simplify model building and ease software development. In particular, 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. For example, deriving the intraclass correlation (ICC) from a random-intercept LMM is particularly straightforward and coincides with the correlation from a compound-symmetric structure. This is because the latter model is the marginalization of the former. For this reason, the LMM is a flexible tool to study psychometric reliability based on longitudinal data, as was shown by Vangeneugden *et al* (2004).

However, also non-Gaussian outcomes are very prominent in repeated measures studies. In this more general setting, model formulation is less straightforward than in the normal case. One distinguishes between marginal and random-effects model families and, unlike in the Gaussian situation, there is 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.

Detailed accounts can be found in Fahrmeir and Tutz (2001) and Molenberghs and Verbeke (2005). 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. This is due to the presence of a non-linear link function, as well as the mean-variance link (Molenberghs and Verbeke, 2005, Chapter 16).

Nevertheless, 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. As an application, we will derive the intra-class correlation coefficient of *reliability* of a derived binary response variable and we investigate correlation of this binary response variable with an established continuous, interval scaled variable to investigate the *criterion validity* of the derived response variable and the continuous response variable.

Indeed, *reliability* is an important aspect of any clinical parameter measured in a clinical trial. As stated by Fleiss (1986): '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 different subjects in the different treatment arm decreases. Fleiss describes a number of consequences of *unreliability*. He brings up attenuation of correlation in studies designed to estimate correlation between variables with poor reliability, biased sample selection in clinical studies where patients are selected with a minimum level of a certain measurement with low reliability, and last but not least, an increased sample size for trials with a primary parameter with low reliability. For the latter, one can 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 when responses are measured without error. It is very clear that a high reliability is important to the clinical trialist. 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 biostatistician and clinician are designing a new clinical study, they should have good information on the reliability of the measurements that are planned to be used in clinical studies. Most often, the strategy is to use a scale which has been validated before and for which intra-rater (test-retest), and inter-rater reliability and internal consistency are established. The validation is usually done on a selected small sample from the population for which the scale is intended. If the population of the trial is different (e.g., when a scale is used in adolescents while scale validation was done in adults), a new battery of reliability and validity testing might be warranted.

When the trials are finished and reported, it is astonishing how little attention is given to the observed reliability of a certain scale. The focus is on estimating treatment effects and their significance. Rarely is there any reflection on how reliable the scale was or how large the observed measurement error was. In this paper, we propose a framework to study *trial- or population-specific reliability*. As shown by Vangeneugden *et al* (2004, 2005), clinical trial data can be used to make progress when studying reliability as well as generalizability in case of interval scaled data, provided one is willing to make a number of modeling assumptions, enabling one to 'translate' biomedical data to a parallel measurements setting. Clearly, the softer an endpoint is, 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 more insight regarding 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, and an application will be provided for a binary response variable. 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. The advantage is that clinical trialists can learn before embarking on new trials in a similar population whether they feel comfortable using the same scale or response parameter again. 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. Also the term *face validity* is used to indicate whether the instrument appears to be assessing the desired qualities at face. This form of validity consists of a judgment by experts in the field. Construct validity refers to a wide range of approaches which are used when what we are trying to measure is a "hypothetical construct" (e.g., anxiety, irritable bowel syndrome, ...) rather than something that can be readily observed. The most commonly used methods to explore construct validity are: extreme groups (apply instrument for example to cases and non-cases), convergent and discriminant validity testing (correlate with other measures of this construct and not correlate with dissimilar or unrelated constructs) and the multitrait-multimethod matrix. 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. This also clearly links validity testing to surrogate marker validation as shown in Alonso *et al* (2002). 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 ICC and reliability functions. 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 parameter will be derived to investigate criterion validity between the derived binary response parameter and the more standard continuous response parameter.

In Section 2, the motivating case study is introduced, while methodology is described in Section 3. In Section 4, we will apply the derived formulae to the data introduced above to estimate reliability of a binary response variable. In Section 5, we will extend the methodology to calculate correlation between concurrently measured response parameters 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". 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, we derive 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 Models

A linear mixed-effects model with serial correlation can be written as (Verbeke and Molenberghs, 2000; Diggle *et al* 2002):

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{W}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

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, X_i and 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, $\boldsymbol{b}_i \sim N(\mathbf{0}, D)$ is the q dimensional vector containing the random effects, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 I_{n_i})$ 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. Serial correlation is captured by the realization of a Gaussian stochastic process, $\boldsymbol{W}_i$, which is assumed to follow a $N(\mathbf{0}, \tau^2 H_i)$ law. The serial correlation matrix 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 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, 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

$$\begin{aligned} \text{Var}(Y_{ij}) &= \boldsymbol{z}_j D \boldsymbol{z}_j' + \tau^2 + \sigma^2, \\ \text{Var}(Y_{ik}) &= \boldsymbol{z}_k D \boldsymbol{z}_k' + \tau^2 + \sigma^2, \\ \text{Cov}(Y_{ij}, Y_{ik}) &= \boldsymbol{z}_j D \boldsymbol{z}_k' + \tau^2 (H_i)_{jk}, \end{aligned} \tag{2}$$

and therefore, the correlation between time point j and k, the ICC of reliability can be written as:

$$\rho_{jk} = \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\boldsymbol{z}_j D \boldsymbol{z}_k' + \tau^2 (H_i)_{jk}}{\sqrt{\boldsymbol{z}_j D \boldsymbol{z}_k' + \tau^2 + \sigma^2}\sqrt{\boldsymbol{z}_k D \boldsymbol{z}_k' + \tau^2 + \sigma^2}}. \tag{3}$$

In case of a simple random-intercept model, $\boldsymbol{Y}_i = X_i \boldsymbol{\beta} + b_i + \boldsymbol{\varepsilon}_i$, with no serial correlation, (3) simplifies to:

$$\rho_{st} = \text{Corr}(Y_{is}, Y_{it}) = \frac{d}{d + \sigma^2}, \tag{4}$$

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 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 $i, i = 1, ..., N, j = 1, ..., n_i$ 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 $N(\boldsymbol{0}, D)$, the outcomes Y_{ij} are independent with densities of the form

$$f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) = \exp\left[\frac{y_{ij}\theta_{ij} - \psi(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right], \tag{5}$$

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., $g(\mu_{ij}) = g(E(Y_{ij}|\boldsymbol{b}_i)) = \boldsymbol{x\prime}_{ij}\boldsymbol{\beta}+\boldsymbol{z\prime}_{ij}\boldsymbol{b}_i$ 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 $\theta_{ij} = \boldsymbol{x\prime}_{ij}\boldsymbol{\beta} + \boldsymbol{z\prime}_{ij}\boldsymbol{b}_i$. The random effects $\boldsymbol{b}_i$ are assumed to be sampled from a (multivariate) normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{D}$.

3.3 ICC Based on GLMM

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

$$\boldsymbol{Y_i} = \boldsymbol{\mu_i} + \boldsymbol{\varepsilon_i}, \tag{6}$$

where the mean (systematic part) can be written as

$$\boldsymbol{\mu_i} = \boldsymbol{\mu_i}(\eta_i) = h(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i}), \tag{7}$$

X_i and 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 random component. 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 nor on the complexity of the model, e.g., allowing for serial correlation or not. The variance covariance matrix can be derived as follows:

$$\begin{aligned} \boldsymbol{V_i} &= \mathrm{Var}(\boldsymbol{Y_i}) = \mathrm{Var}(\boldsymbol{\mu_i} + \boldsymbol{\varepsilon_i}) \\ &= \mathrm{Var}(\boldsymbol{\mu_i}) + \mathrm{Var}(\boldsymbol{\varepsilon_i}) + 2\mathrm{Cov}(\boldsymbol{\mu_i}, \boldsymbol{\varepsilon_i}). \end{aligned} \tag{8}$$

It is easy to show that $\mathrm{Cov}(\boldsymbol{\mu_i}, \boldsymbol{\varepsilon_i}) = \mathrm{Cov}[\mathrm{E}(\boldsymbol{\mu_i}|\boldsymbol{b_i}), \mathrm{E}(\boldsymbol{\varepsilon_i}|\boldsymbol{b_i})) + \mathrm{E}(\mathrm{Cov}(\boldsymbol{\mu_i}, \boldsymbol{\varepsilon_i}|\boldsymbol{b_i})] = 0$ since the first term is 0 and the second term equals $\mathrm{E}[\mathrm{E}(\boldsymbol{\mu_i} - \mathrm{E}(\boldsymbol{\mu_i}))(\boldsymbol{\varepsilon_i})|\boldsymbol{b_i}] = 0$ as $\boldsymbol{\mu_i}$ is a constant when

conditioning on $\boldsymbol{b_i}$. For the first term in (8) we have:

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\mu_i}) &= \mathrm{Var}(\boldsymbol{\mu_i}(\boldsymbol{\eta_i})) = \mathrm{Var}(\boldsymbol{\mu_i}(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i)) \qquad (9)\\
&\cong \left(\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{b_i}}\,\Big|_{\boldsymbol{b_i}=0}\right) \mathrm{Var}(\boldsymbol{b_i}) \left(\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{b_i}}\,\Big|_{\boldsymbol{b_i}=0}\right)'\\
&\cong \left(\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\eta_i}}\frac{\partial \boldsymbol{\eta_i}}{\partial \boldsymbol{b_i}}\,\Big|_{\boldsymbol{b_i}=0}\right) \boldsymbol{D} \left(\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\eta_i}}\frac{\partial \boldsymbol{\eta_i}}{\partial \boldsymbol{b_i}}\,\Big|_{\boldsymbol{b_i}=0}\right)'\\
&\cong \Delta_i Z_i D Z_i' \Delta_i', \qquad (10)
\end{aligned}$$

where $\Delta_i = \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\eta_i}}\,\Big|_{\boldsymbol{b_i}=0}$.

For the second term in (8), we have:

$$\mathrm{Var}(\boldsymbol{\varepsilon_i}) = \mathrm{Var}[\mathrm{E}(\boldsymbol{\varepsilon_i}|\boldsymbol{b_i})] + \mathrm{E}[\mathrm{Var}(\boldsymbol{\varepsilon_i}|\boldsymbol{b_i})] = \mathrm{E}[\mathrm{Var}(\boldsymbol{\varepsilon_i}|\boldsymbol{b_i})] = \Phi^{\frac{1}{2}}\Sigma\Phi^{\frac{1}{2}}, \qquad (11)$$

where Φ is a diagonal matrix with the overdispersion parameters along the diagonal. In case there are no overdispersion parameters, Φ is set equal to the identity matrix. We can express the variance function Σ_i so that

$$\mathrm{Var}(\varepsilon_i) = \Phi^{\frac{1}{2}}\boldsymbol{A_i}^{\frac{1}{2}}\boldsymbol{R_i}\boldsymbol{A_i}^{\frac{1}{2}}\Phi^{\frac{1}{2}}, \qquad (12)$$

where $\boldsymbol{R}_i$ is the correlation matrix and $\boldsymbol{A}_i$ 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 (10) and (12), we have the following expression for the variance-covariance matrix (8):

$$\boldsymbol{V_i} \cong \Delta_i Z_i D Z_i' \Delta_i' + \Phi^{\frac{1}{2}}\boldsymbol{A_i}^{\frac{1}{2}}\boldsymbol{R_i}\boldsymbol{A_i}^{\frac{1}{2}}\Phi^{\frac{1}{2}}. \qquad (13)$$

If the canonical link is used, we have $\boldsymbol{A}_i = \Delta_i$ and (13) can be written as:

$$\boldsymbol{V_i} \cong \Delta_i Z_i D Z_i' \Delta_i' + \Phi^{\frac{1}{2}}\Delta_i^{\frac{1}{2}}\boldsymbol{R_i}\Delta_i^{\frac{1}{2}}\Phi^{\frac{1}{2}}. \qquad (14)$$

If in addition, conditional independence (no serial correlation) is assumed, then (13) simplifies to:

$$\boldsymbol{V_i} \cong \Delta_i Z_i D Z_i' \Delta_i' + \Phi^{\frac{1}{2}} \Delta_i \Phi^{\frac{1}{2}}. \tag{15}$$

Further, if we reduce the random-effects part to a random-intercept model, i.e., $\boldsymbol{Z}_i = \mathbf{1}$ and $\boldsymbol{D} = d$, and (13) then reduces to

$$\boldsymbol{V_i} \cong \Delta_i (d\boldsymbol{J}) \Delta_i' + \Phi^{\frac{1}{2}} \Delta_i \Phi^{\frac{1}{2}}. \tag{16}$$

Note that if we have a normal distribution with the canonical identity link, Δ_i reduces to the identity matrix $\boldsymbol{I}$ and $\Phi = \sigma^2 \boldsymbol{I}$, in which case it follows that (16) reduces to $d\boldsymbol{J} + \sigma^2 \boldsymbol{I}$ which is consistent with (4). Moreover, when we have a normal distribution with a general random-effects structure but without serial correlation, it is easy to show that $\boldsymbol{V_i} \cong Z_i D Z_i' + \sigma^2 I$ and that subsequently ρ equals (3) when we leave out the serial correlation (τ). This shows that (13) can be seen as a generalization of (3). In the following section we will derive the marginal correlation in case of binary data when applying a random intercept model.

3.4 ICC for a Random-intercept Model for Binary Data

In this section, we will derive the formula for the ICC in case of a random intercept model for binomial data with a logit link and assuming no overdispersion. In this case (16) reduces to $\boldsymbol{V_i} \cong \Delta_i (d\boldsymbol{J}) \Delta_i' + \Delta_i = \Delta_i (d\boldsymbol{J} + \Delta_i^{-1}) \Delta_i'$. Furthermore, Δ_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

$$\boldsymbol{V_i} \cong diag(V_{ij}(0))[d\boldsymbol{J} + diag(V_{ij}(0))^{-1}] diag(V_{ij}(0)). \tag{17}$$

In other words, the variance-covariance matrix for subject i is specified by the matrix with elements:

$$v_{ijj} = V_{ij}(0)[1 + V_{ij}(0)d] \quad (18)$$

$$v_{ijk} = dV_{ij}(0)V_{ik}(0), \quad (j \neq k). \quad (19)$$

Based on (18) and (19), 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:

$$\rho_{ijk} = \text{Corr}(Y_{ij}, Y_{ik}) = \frac{V_{ij}(0)V_{ik}(0)d}{\sqrt{\{V_{ij}(0)[1 + V_{ij}(0)d]\}\{V_{ik}(0)[1 + V_{ik}(0)d]\}}}. \quad (20)$$

This expression allows us 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, being of the Pearson type, are highly constrained, a phenomenon studied in detail by Declerck, Aerts, and Molenberghs (1998) and reported in 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 such constraints already apply to the Pearson correlation in a simple two by two contingency table (Diggle *et al*, 2002). A mild form of the Bahadur constraints survives in generalized estimating equations, especially those of the second order (Liang, Zeger, and Qaqish, 1992). The multivariate probit model (Ashford and Sowden, 1970; Molenberghs and Verbeke, 2005), on the other hand, is constrained only by the requirement that the tetrachoric correlations form a positive definite matrix. This advantage of the probit model is counterbalanced by its heavy computational burden. Also, the beta-binomial model (Skellam, 1948; Molenberghs and Verbeke, 2005) allows for all non-negative correlations

as well as moderate negative values (see also Molenberghs and Verbeke, 2005). The problem 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 leaves the intraclass correlation unconstrained within the unit interval.

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 $\sigma^2 I + dJ$, remains positive-definite.

4 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 first describe the observed proportion of response and the correlation between response in the data at hand via graphical displays. Subsequently, we will calculate the ICC under different assumptions, with gradually increasing modeling complexity.

4.1 Observed Response Rate and Correlation

Figure 1 displays the response rate of both treatment groups combined across time. We can see that 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.

Figure 2 illustrates the correlation of the different responses over time. Circles are drawn at different response level combinations (no, yes) in this matrix plot, exhibiting the correlation between the observed responses at different pairs of time points. The diameter of the circle is proportional to the number of subjects at each response combination, e.g., the large circle for response=No at Week 1 and response=No at Week 2 indicates that many subjects who did not respond at Week 1 also did not respond at Week 2. The larger the diameter of the circles are at the bisector line ($y = x$), the larger the correlation is between the same level of response at time point s versus time point t. From the first 2 rows of the plot, we can see that correlation is high if we compare Week 1 and 2, but that this correlation decreases slightly in time, when the lag time between observations is increased. On the other hand, the correlation between Week 6 and 8 is noticeably higher, as the diameter of the circles on the bisector line (same response at Week 6 and 8) are large and almost zero for circles not on the bisector line (different response on Week 6 and 8). Alternatively, the area of the circles could be made proportional to the sample size, rather than the diameter. Arguably, this is a matter of choice.

4.2 Analysis Without Covariates

In this first analysis we ignore time and treatment and include all observations. This will generate an estimate of the overall ICC, which can be interpreted as an average ICC across all

time points. In this simplification, we have that $\boldsymbol{X}_{is}\boldsymbol{\beta} = \beta$ is constant across all subjects i and time points j. Then (20) simplifies to:

$$\begin{aligned} V_{ij}(0) &= V(0) = \frac{\exp(\beta)}{(1+\exp(\beta))^2}, \\ \rho_{ijk} &= \rho = \frac{V(0)d}{1+V(0)d}. \end{aligned} \tag{21}$$

If we use the SAS procedure NLMIXED to fit this random-effects model, using adaptive Gaussian quadrature, we have that $\widehat{\beta} = -1.61$, $\widehat{d} = 6.57$, $\widehat{V(0)} = 0.14$ and subsequently $\widehat{\rho} = 0.48$. We also applied this simple model to investigate time and treatment effect by looking at subgroups. Table 1 summarizes the results.

One observes that the ICC is somewhat lager 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, as we could expect from Figure 2. 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, 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, unlike in the continuous case, there are a wide variety of options to capture association between binary measures (Agresti, 2002; Molenberghs and Verbeke, 2005). Thus, while the correlation is certainly not a universally accepted measure of association, it is and remains a useful one, even though there is an intricate relationship between marginal means and association. This is unavoidable as soon as outcomes are non-Gaussian since, unlike in the Gaussian case, means and variances are related through the so-called mean-variance relationship. A commonly encountered measure of agreement to assess reliability in case of binary outcome is the kappa coefficient. Bloch and Kraemer (1989) showed that the ICC can also be interpreted as an ICC in case of parallel measurements, i.e., when the two measurements on each subject are interchangeable. When we calculate the kappa statistic for agreement between Week 6 and 8, we obtained 0.72, which is somewhat lower than the ICC from Table 1, while the kappa statistic for agreement between Weeks 1 and 8 was somewhat higher (0.20) than the ICCs of the table. An advantage of our approach is that it can easily deal with sequences of repeated measures. To date, there is no kappa-based alternative.

4.3 Accounting for Time

If we adjust for time and ignore treatment, then ρ can be derived via (20) and it is easy to show that

$$V_{ij}(0) = \frac{\exp(\beta_j)}{(1+\exp(\beta_j))^2},$$

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 2 provides the estimated intra-class correlation coefficient matrix. This is in line with the well-known relationship between marginal and random-effects regression parameters (Diggle *et al* 2002), the correlations are determined by the random-intercept variance, together with the marginal probabilities factoring into the variance function:

$$\beta_j \cong \sqrt{1 + 0.346\ d} \cdot \text{logit}(p_j). \tag{22}$$

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 2, 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. Note that the ICC ($\rho = 0.70$) between observations at Week 6 and 8 is close to the kappa coefficient (0.72).

4.4 Accounting for Time and Treatment

Next, we will account for both time and treatment. This will yield a different ICC for each treatment group separately and also for each pair of time points. We allowed for interactions in the model. Table 3 summarizes the results.

After adjusting for time and treatment, the ICC between observations at Week 1 and 8 increased from 0.11 (1) to 0.40 in the risperidone group.

5 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 parameter. 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 (6) can be extended too, as described in Molenberghs and Verbeke (2005):

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

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

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

with $v_{i2}(0) = \mu_i \mid_{\boldsymbol{b_i}=0} (1 - \mu_i \mid_{\boldsymbol{b_i}=0})$. Note that 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 (13) becomes

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

Here, τ^2 is the random-intercept variance. As a result, we have the following approximation for the marginal correlation:

$$\rho(\beta) = \frac{v_{i2}\lambda\tau^2}{\sqrt{\lambda^2\tau^2+\sigma^2}\sqrt{v_{i2}^2\tau^2+v_{i2}}}. \tag{23}$$

We can now apply (23) to the same data set to estimate the correlation at Week 8 between the binary response variable defined above and the continuous response parameter defined as the total PANSS, the sum of all 30 items of the PANSS. Table 4 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 the Pearson or the Spearman correlation is calculated, including only data observed at Week 8 for both the binary response and the continuous PANSS score. Here, this resulted 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 (2005, Ch. 24).

6 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 parameter, 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 (Little and Rubin 2002), 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 be incorporated 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 parameter 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 (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.

Acknowledgments

The authors are thankful to J&J PRD for kind permission to use their data. We gratefully acknowledge support from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

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

Agresti, A. (2002). *Categorical Data Analysis (2nd ed.*. New York: John Wiley & Sons.

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2002). Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. *Journal of Biopharmaceutical Statistics* **12**, 161–179.

Ashford, J.R. and Sowden, R.R. (1970). Multivariate probit analysis. *Biometrics*, **26**, 535–546.

Bahadur, R.R. (1961). A representation of the joint distribution of responses to n dichotomous items. In: *Studies in item analysis and prediction,* H. Solomon (Ed.). Stanford mathematical studies in the social sciences VI. Stanford, CA: Stanford University Press.

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.

Bloch, D.A. and Kraemer, H.C. (1989). 2x2 Kappa coefficients: Measures of agreement or association. *Biometrics* **45**, 269–287.

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

Declerck, L., Aerts, M., and Molenberghs, G. (1998). Behaviour of the likelihood ratio test statistic under a Bahadur model for exchangeable binary data. *Journal of Statistical Computations and Simulations*, **61**, 15–38.

Diggle, P.J., Heagerty, P., Liang, K.Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data, 2nd edition.* Oxford: Clarendon Press.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modeling Based on Generalized Linear Models, 2nd edition.* 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* **88**, 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 Schizophrenics. *Psychiatric Research* **23**, 99–110.

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–974.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **73**, 13–22.

Liang, K.Y., Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, **54**, 3–40.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

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

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data.* New York: Springer.

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.

Streiner D.L and Norman G.R. (1995). *Health measurement scales.* Oxford University Press.

Skellam, J.G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials, *Journal of the Royal Statistical Society, Series B*, **10**, 257–261.

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.

Table 1: *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)

Table 2: *Overall ICC (s.e.) matrix, marginal over treatment. Standard errors are calculated from the delta method.*

			time		
Week	1	2	4	6	8
1	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)
8					1

Table 3: *Overall ICC of reliability (s.e.) matrix, accounting for treatment, time and their interaction. Standard errors are calculated from the delta method.*

Week	1	2	4	6	8
			risperidone		
1	1	0.36 (.045)	0.39 (.044)	0.40 (.042)	0.40 (.042)
2		1	0.62 (.036)	0.64 (.033)	0.64 (.032)
4			1	0.69 (.026)	0.69 (.026)
6				1	0.71 (.023)
8					1
			active control		
1	1	0.22 (.036)	0.27 (.038)	0.31 (.038)	0.31 (.038)
2		1	0.42 (.046)	0.48 (.043)	0.49 (.041)
4			1	0.57 (.039)	0.59 (.037)
6				1	0.67 (.029)
8					1

Table 4: *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.*

End point	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)

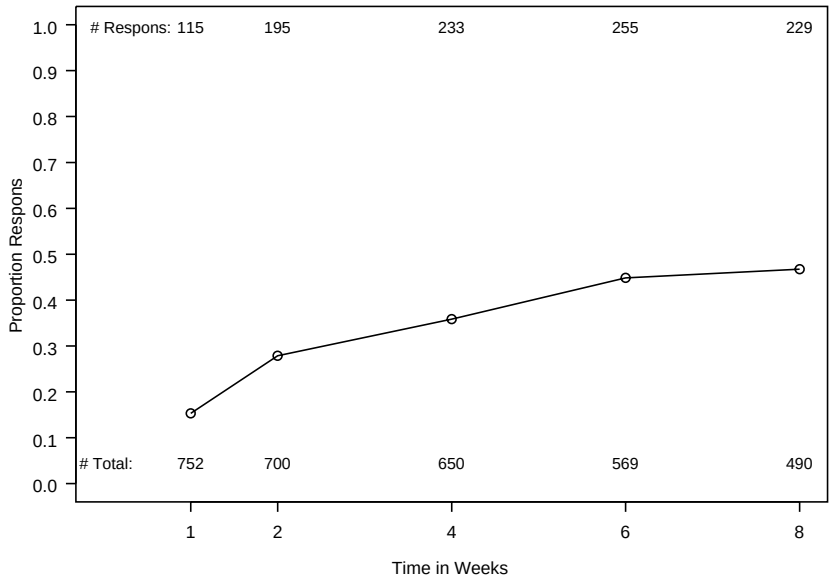


Figure 1: *Graphical representation of observed response over time.*

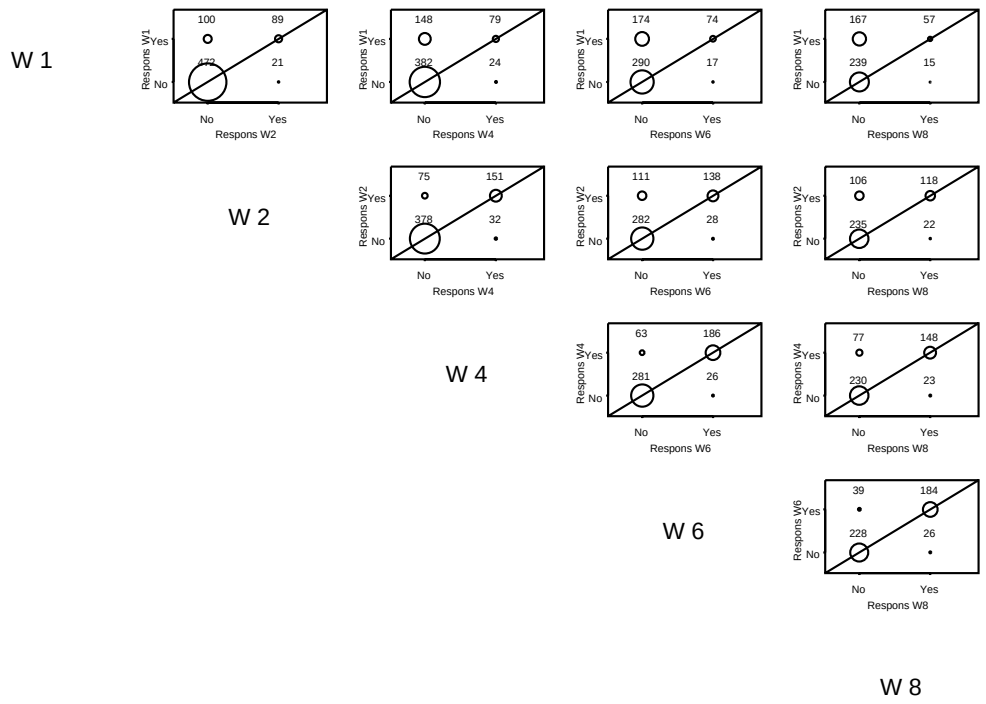


Figure 2: *Graphical representation of the correlation of observed response over time.*