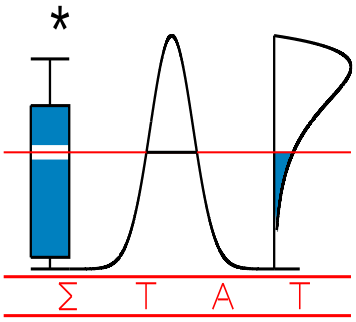
T E C H N I C A L

R E P O R T

06105

GENERALIZABILITY IN NON-GAUSSIAN LONGITUDINAL CLINICAL TRIAL DATA BASED ON GENERALIZED LINEAR MIXED MODELS

VANGENEUGDEN, T., MOLENBERGHS, G., LAENEN, A., ALONSO, A. and H. GEYS



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

Generalizability in Non-Gaussian Longitudinal Clinical Trial Data Based on Generalized Linear Mixed Models

Tony Vangeneugden[1,2] **Geert Molenberghs**[2]
Annouschka Laenen[2] **Ariel Alonso**[2]
Helena Geys[2,3]

[1] Tibotec, Johnson & Johnson, Mechelen, Belgium
Email: tvangene@tibbe.jnj.com
[2] Hasselt University, Center for Statistics, Diepenbeek, Belgium
[3] Janssen Pharmaceutica, Johnson & Johnson, Beerse, Belgium

Abstract

This work investigates how generalizability, an extension of reliability, can be defined and estimated based on longitudinal data sequences resulting from, for example, clinical studies. Useful and intuitive approximate expressions are derived based on generalized linear mixed models. Data from four double-blind randomized clinical trials in schizophrenia motivate the research and are used to estimate generalizability for a binary response parameter.

Keywords: *Binary data; Intraclass correlation; Generalizability; Random effects; Reliability; Variance components.*

1 Introduction

Many measurements in clinical research are based on clinicians' observations, are therefore prone to error, and hence call for assessment of observer reliability and agreement. The latter terms are often used interchangeably but,in principle,should be considered different concepts. *Reliability* coefficients express the ability to differentiate among subjects and are ratios of variances, in classical terms, the variance attributed to the difference among subjects divided by the total variance (Shrout and Fleiss 1979). *Agreement* refers to conformity, with corresponding parameters determining whether the same value is achieved if a measurement were performed twice, either

by the same or different observers. In homogeneous populations, one can imagine that reliability be low while agreement be high; in a heterogeneous population, reliability and agreement measures alledgedly will correspond well (Stratford 1989). The parameters for assessment of observer reliability and agreement differ according to the scale of measurement. For nominal and ordinal categorical measurements, the κ-coefficient and the weighted κ-coefficient, respectively, are measures of agreement. In case of continuous data, the intraclass correlation coefficient (ICC) is commonly used to measure observer reliability, although ICC-type quantities can be defined for binary and ordinal data as well (Fleiss 1981).

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 arms decreases. Fleiss describes as consequences of *unreliability*: attenuation of correlation in studies designed to estimate correlation between variables with poor reliability, biased sample selection in clinical studies where patients are selected based on attaining a minimum level of an unreliable measurement, and, last but not least, an increased sample size for trials with a primary, low-reliability parameter. 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 thus 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. Most often, the strategy is to use a scale that has been validated before and for which intra-rater (i.e., test-retest), inter-rater reliability, and internal consistency were 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, a new battery of reliability and validity testing might be warranted. Now, the classical framework may be deficient for the clinical trial setting since conventionally an observation is assumed to be a combination of an individual's *true* score and random measurement error. The assumption that all variance in scores can be divided into true and error variance may come across as a little simplistic. At the same time, once the trials finished and reported, it is astonishing how little attention is given to the observed reliability of a given scale. The focus usually is on estimating treatment effects and their significance. As a result, rarely is there any reflection on how reliable the scale was or how large the observed measurement error.

Vangeneugden *et al.* (2004) proposed a framework for studying *trial- or population-specific reliability* using clinical trial data. The appeal of this extension notwithstanding, next to the true score of an individual, multiple potential sources of error can exist. The goal is then to obtain the most precise estimate of the score a person should have if there were no sources of error contaminating our results. Each on among the variety of forms reliability can take, such as inter-rater reliability, test-retest reliability, and internal consistency, identifies and quantifies only one source of error variance at a time. *Generalizability theory* (GT, Cronbach 1963) enables considering all sources of variability simultaneously. The essence of the theory is the recognition that, in any measurement situation, there are multiple sources of error variance. The goal is to try and identify, measure, and thereby possibly find strategies to reduce the influence of these sources on the measurement under investigation (Shavelson, Webb, and Rowley 1989). Imagine that we could identify the most likely sources of error in a measurement of some characteristic of a person. We then have defined our "universe" of possible observations. If we subsequently proceed to average each person's score over all of these possible conditions, an unbiased estimate would result of that person's score over the universe as we have defined it. Note that there is no pretense that this is the "true" score; rather, it is conditional on the universe considered. Of course, more than one choice of universe can be considered.

In the context of clinical trials, by investigating sources of error, such as, for instance, country

or sub-category of diagnosis, the clinical trialist could learn a lot about performance of scales or other measurements in certain subgroups and what the impact of such factors is on reliability. Vangeneugden *et al.* (2005) applied such generalizability concepts to interval-scaled data from clinical trials, for which it is natural to assume a Gaussian distribution.

The present work extends generalizability to non-Gaussian outcomes. While frequently encountered in repeated-measures clinical trials, especially when of a binary type, model formulation is less than straightforward. 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 likely the most prominent random-effects model (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, which limits its use when the correlation is of scientific interest, e.g., in view of the ICC. The GLMM, on the other hand, has a full likelihood basis, but fails to produce the marginal correlations in an easy fashion, owing to the presence of a non-linear link function, combined with a non-trivial mean-variance link, forcing the variance to change with the mean and hence with the regressors (Molenberghs and Verbeke 2005, Chapter 16). In spite of these considerations, we will show the GLMM provides a viable framework when correlations are of interest, with particular emphasis on the use of generalizability theory.

In Section 2, the motivating case study is introduced, while methodology is described in Section 3. In Section 4, we will estimate reliability and generalizability of a binary response variable, thereby underscoring the versatile use that can be given to the generalizability framework.

2 Motivating Studies

Consider 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 within 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 weeks 1, 2, 4, 6, and 8.

3 Methodology

After having given a general outline of the concepts of reliability and generalizability from a classical view-point, we will provide an introduction to the generalized linear mixed model paradigm, thus offering a framework within with reliability and generalizability can be derived based on longitudinal data from clinical trials or other studies, not specifically designed in view of reliability

or generalizability.

To fix ideas, let us give an example as to how the observed clinical trial data are typically decomposed:

$$Y_{pdt} = h(\mu + b_p + \mu_d + \mu_t + \mu_{dt}) + \varepsilon_{pdt}, \tag{1}$$

where $h(.)$ is a known link function. Further, b_p denotes the random effect for patient $p = 1, \ldots, N$, μ_d the fixed time effect at day $d = 1, \ldots, n_p$, μ_t the fixed effect of treatment $t = 1, \ldots, T$, μ_{dt} their interaction. Finally, ε_{pdt} refers to the residual error, the distribution of which is chosen in accordance with the outcome type. For example, when Y_{pdt} is a binary indicator, it is customary to adopt for $h(\cdot)$ the antilogit function and for ε_{pdt} the Bernoulli distribution with success probability $h(\mu + b_p + \mu_d + \mu_t + \mu_{dt})$. When other design levels are present, e.g., country or center, Model (1) can be extended in a straightforward fashion and various instances will be given in subsequent sections.

3.1 Reliability

In the classical test theory, reliability frequently materializes as the intraclass correlation coefficient (ICC). For instance, if one wishes to estimate test-retest reliability in case of Gaussian data, the outcome of a test can be modeled as

$$Y_{pd} = \mu + b_p + \mu_d + \varepsilon_{pd}, \tag{2}$$

where μ is an overall intercept, $b_p \sim N(0, \sigma_p^2)$ a random effect for patient, μ_d a fixed effect for day of measurement, and $\varepsilon_{pd} \sim N(0, \sigma_E^2)$ the corresponding measurement error. Then, the reliability is a function of two sources of variability, deriving from the patient and residual levels, respectively:

$$\widehat{R} = \frac{\widehat{\sigma}_p^2}{\widehat{\sigma}_p^2 + \widehat{\sigma}_E^2} = \frac{\mathsf{BMS} - \mathsf{WMS}}{\mathsf{BMS} + (n_D - 1)\mathsf{WMS}}, \tag{3}$$

where n_D is the number of measurements per patient, i.e., the number of days at which measurements are taken. The time effect should be zero because the classical theory assumes strictly

parallel measurements (Shavelson, Webb, and Rowley 1989). It is easy to show that R is the correlation between measurements of the same patient, on different but given days, i.e., conditioning on days and thereby keeping them fixed:

$$R = \mathsf{Corr}(Y_{pd}, Y_{pd'} \mid d, d'), \tag{4}$$

with notation as in (2). For parallel measurements, this correlation coefficient indeed coincides with the ICC of reliability.

3.2 Generalizability

Generalizability theory recognizes that, in virtually all measurement situation, there are multiple sources of error variance. The goal is to try and identify, measure, and thereby possibly find strategies to reduce the influence of these sources on the measurement in question. For instance, if one measures a patient, not only on different days but also by different raters, one could investigate both test-retest reliability and inter-rater reliability, assuming that rater and day of observation are the most important sources of error, in addition to residual measurement error. For Gaussian data, a linear version of (2) allowing for rater effects, is:

$$Y_{prd} = \mu + b_p + \mu_r + \mu_d + \varepsilon_{prd}, \tag{5}$$

where, in addition to effects already included, μ_r now represents the fixed effect pertaining to rater $r = 1, \ldots, R$. The associated sources of variability are denoted by σ_p^2 for the patient level, σ_r^2 for the rater level, etc. Model (5) enables estimation of the variances' magnitude stemming from the various sources of error, which are patient, rater, day, and residual in the example above. If the sources that we have identified are trivial, while we have missed any important source of error, then the residual variance will typically be large.

In GT terminology, 'person' is a so-called *facet of differentiation*, while 'rater' and 'day' are called *facets of generalization*. The levels of the facets of generalization are named *conditions*. It is common to use ANOVA for estimating the various variance components, which in turn lead

to a *generalizability coefficient*, analogous to a reliability coefficient, found as the ratio of the estimated person variance component and a so-called estimated observed score variance.

GT distinguishes between decisions based on the relative standing of individuals and decisions based on the absolute value of a score (Shavelson, Webb, and Rowley 1989). Let us explain these in turn.

Error in *relative decisions* arises from all nonzero variance components associated with rank ordering of individuals, other than the component for the object of measurement (persons). Specifically, variance components associated with the interaction of person with each facet or combinations of facets define error. For Model (5), we distinguish between σ^2_{pr}, σ^2_{pd}, and $\sigma^2_{prd} = \sigma^2_e$. So, if one wishes to generalize from a rating by one rater on a particular day to a rating by a different rater at another point in time, the following generalizability coefficient can be constructed as the ratio of the universe-score variance to the expected rater-score variance:

$$E_{\rho^2\text{Rel}} = \text{Corr}(Y_{prd}, Y_{pr'd'} \mid r, r', d, d') = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{\text{Rel. Error}}} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pr} + \sigma^2_{pd} + \sigma^2_{prd}}, \qquad (6)$$

having the form of an ICC. Indeed, it is easy to show that (6) can be derived as a conditional correlation coefficient where we condition on rater and day, while at the same time allowing each one of them to take on different values. Alternatively, we can derive a test-retest or an inter-rater reliability coefficient, by either merely generalizing over day of observation and fixing rater or by merely generalizing over rater and fixing day of observation:

$$R_{\text{test-retest, Rel}} = \text{Corr}(Y_{prd}, Y_{prd'} \mid r, d, d') = \frac{\sigma^2_p + \sigma^2_{pr}}{\sigma^2_p + \sigma^2_{pr} + \sigma^2_{pd} + \sigma^2_{prd}}, \qquad (7)$$

$$R_{\text{inter-rater, Rel}} = \text{Corr}(Y_{prd}, Y_{pr'd} \mid r, r', d) = \frac{\sigma^2_p + \sigma^2_{pd}}{\sigma^2_p + \sigma^2_{pr} + \sigma^2_{pd} + \sigma^2_{prd}}. \qquad (8)$$

Decisions based on the level of observed score, disregarding the performance of others, are called *absolute decisions*. All variance components associated with such a score, except the component

for the object of measurement, are considered error. Then, (6) transforms to

$$\begin{aligned} E_{\rho^2\text{Abs}} = \text{Corr}(Y_{prd}, Y_{pr'd'}) &= \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2_{\text{Abs. Error}}} \\ &= \frac{\sigma_p^2}{\sigma_p^2 + \sigma_r^2 + \sigma_d^2 + \sigma_{pr}^2 + \sigma_{pd}^2 + \sigma_{rd}^2 + \sigma_{prd}^2}. \end{aligned} \tag{9}$$

Also here, (9) can be considered an ICC, this time conditioned neither on rater nor on day. Similar to the above, we can derive an *absolute* test-retest or inter-rater reliability coefficient:

$$\begin{aligned} R_{\text{test-retest, Abs}} &= \text{Corr}(Y_{prd}, Y_{prd'}) = \frac{\sigma_p^2 + \sigma_r^2 + \sigma_{pr}^2}{\sigma_p^2 + \sigma_r^2 + \sigma_d^2 + \sigma_{pr}^2 + \sigma_{pd}^2 + \sigma_{rd}^2 + \sigma_{prd}^2}, \\ R_{\text{inter-rater, Abs}} &= \text{Corr}(Y_{prd}, Y_{pr'd}) = \frac{\sigma_p^2 + \sigma_d^2 + \sigma_{pd}^2}{\sigma_p^2 + \sigma_r^2 + \sigma_d^2 + \sigma_{pr}^2 + \sigma_{pd}^2 + \sigma_{rd}^2 + \sigma_{prd}^2}. \end{aligned} \tag{10}$$

This example, aimed to enhance insight in the various uses of GT, was based on a simple so-called *crossed* design with two factors, each one occurring at all levels of the other. GT can be used with more complex designs as well, for example, including more factors, and even in *nested* designs, exhibiting a more complex factor structure. As discussed in Streiner and Norman (1995), the general principle remains untouched: one begins by isolating the various sources of variance in the scores, and then generating a family of coefficients that depend on the particular factors and that are allowed either to vary or to remain fixed.

A first type of study, designed to estimate variance components underlying a measurement process, is called a *G-study*. Second, having generated the variance estimates, one can then study the impact on generalizability of such decision as changing the number of observations or adding a further rater. Since this second type of study explores the impact of certain decisions, they are termed *Decision studies* or *D-studies*. Interestingly, D-studies can be undertaken solely using paper and pencil, or a computer. In planning a D-study, the decision maker defines the universe of generalization and specifies the proposed interpretation of the measurement. The goal is to identify important sources of variability in a particular measurement situation from the outset, and then one quantifies these sources.

Obviously, GT is broad and versatile. In the next section, we will show how this can be expanded

by embedding it in the flexible generalized linear mixed model framework. Apart from dealing with non-Gaussian outcomes, it will be possible to include further sources of variability, such as serial (temporal) correlation, commonly encountered in longitudinal studies, superimposed on the random-effects structure.

3.3 Generalized Linear Mixed Models

The generalized linear mixed model (GLMM, Breslow and Clayton 1993) has been the most frequently used random effects model for non-Gaussian outcomes, although alternative paradigms, such as laid out in Lee, Nelder, and Pawitan (2006), exist and are of interest, too.

With notation similar to the one used in previous sections, let Y_{pd} be the outcome recorded on day $d = 1, \ldots, n_p$ for subject $p = 1, \ldots, N$, and let $\boldsymbol{Y}_p$ be the n_p-dimensional vector of all measurements available for subject (cluster) p. The GLMM assumes that, conditionally on a q-dimensional random $\boldsymbol{b}_p$, alledgedly drawn independently from a $N(\boldsymbol{0}, D)$, the outcomes Y_{pd} are independent with densities of the form

$$f_p(y_{pd}|\boldsymbol{b}_p, \boldsymbol{\beta}, \phi) = \exp\left[\frac{y_{pd}\theta_{pd} - \psi(\theta_{pd})}{\phi} + c(y_{pd}, \phi)\right], \tag{11}$$

where the mean $\mu_{pd} = \partial\psi(\theta_{pd})/\partial\theta_{pd}$ is modeled through a linear predictor containing fixed regression parameters $\boldsymbol{\beta}$ as well as subject-specific parameters $\boldsymbol{b_p}$, i.e., $h^{-1}(\mu_{pd}) = h^{-1}(E(Y_{pd}|\boldsymbol{b}_p)) = \boldsymbol{x}'_{pd}\boldsymbol{\beta} + \boldsymbol{z}'_{pd}\boldsymbol{b}_p$ for a known link function $h(\cdot)$, with $\boldsymbol{x}_{pd}$ and $\boldsymbol{z}_{pd}$ r-dimensional and q-dimensional vectors of known covariate values, respectively, with $\boldsymbol{\beta}$ an r-dimensional vector of unknown fixed regression coefficients, and with ϕ a scale parameter. Employing a natural link function (McCullagh and Nelder 1989), this becomes $\theta_{pd} = \boldsymbol{x}'_{pd}\boldsymbol{\beta} + \boldsymbol{z}'_{pd}\boldsymbol{b}_p$. Estimation of model parameters is slightly cumbersome, since no explicit formula for a subject's likelihood contribution exists. Therefore, one has to resort to numerical integration or expansion methods. Molenberghs and Verbeke (2005) provide an overview. We also refer to Lee, Nelder, and Pawitan (2006) for details on the computation using so-called h-likelihood methods. Most inferential approaches are based on maximum likelihood, Bayesian methods, or a variation there upon.

3.4 Correlation Between Two Observations Using the GLMM Framework

We will now derive a general formula for the correlation between two observations, within the GLMM framework. In the spirit of (1), and with notation consistent with Section 3.3, we can write the general model as:

$$Y_{pdt} = \mu_{pdt} + \varepsilon_{pdt}, \tag{12}$$

where

$$\mu_{pdt} = \mu_{pdt}(\eta_{pdt}) = h(\boldsymbol{x}'_{pdt}\boldsymbol{\beta} + \boldsymbol{z}'_{pdt}\boldsymbol{b}_{pdt}). \tag{13}$$

Model (13) allows for a variety of distributions for the outcome variable and a wide range of link functions, while the modeler has the freedom to include or leave out serial correlation. To calculate correlation $\mathsf{Corr}(Y_{pdt}, Y_{p'd't'})$, we first derive a general expression for the variance:

$$\mathsf{Var}(Y_{pdt}) = \mathsf{Var}(\mu_{pdt} + \varepsilon_{pdt}) = \mathsf{Var}(\mu_{pdt}) + \mathsf{Var}(\varepsilon_{pdt}) + 2\mathsf{Cov}(\mu_{pdt}, \varepsilon_{pdt}). \tag{14}$$

It is easy to show that

$$\mathsf{Cov}(\mu_{pdt}, \varepsilon_{pdt}) = \mathsf{Cov}[\mathsf{E}(\mu_{pdt}|\boldsymbol{b}_{pdt}), \mathsf{E}(\varepsilon_{pdt}|\boldsymbol{b}_{pdt})] + \mathsf{E}(\mathsf{Cov}(\mu_{pdt}, \varepsilon_{pdt}|\boldsymbol{b}_{pdt})] = 0,$$

since the first term is zero and the second term equals $\mathsf{E}[\mathsf{E}(\mu_{pdt} - \mathsf{E}(\mu_{pdt}))(\varepsilon_{pdt})|\boldsymbol{b}_{pdt}] = 0$ as μ_{pdt} is constant when conditioning on $\boldsymbol{b}_{pdt}$. For the first term in (14) we have:

$$\begin{aligned}
\mathsf{Var}(\mu_{pdt}) &= \mathsf{Var}(\mu_{pdt}(\eta_{pdt})) = \mathsf{Var}[\mu_{pdt}(\boldsymbol{x}'_{pdt}\boldsymbol{\beta} + \boldsymbol{z}'_{pdt}\boldsymbol{b}_{pdt})] \\
&\cong \left(\frac{\partial \mu_{pdt}}{\partial \boldsymbol{b}_{pdt}}\bigg|_{\boldsymbol{b}_{pdt}=0}\right) \mathsf{Var}(\boldsymbol{b}_{pdt}) \left(\frac{\partial \mu_{pdt}}{\partial \boldsymbol{b}_{pdt}}\bigg|_{\boldsymbol{b}_{pdt}=0}\right)' \\
&\cong \left(\frac{\partial \mu_{pdt}}{\partial \eta_{pdt}}\frac{\partial \eta_{pdt}}{\partial \boldsymbol{b}_{pdt}}\bigg|_{\boldsymbol{b}_{pdt}=0}\right) \boldsymbol{D} \left(\frac{\partial \mu_{pdt}}{\partial \eta_{pdt}}\frac{\partial \eta_{pdt}}{\partial b_{pdt}}\bigg|_{b_{pdt}=0}\right)' \\
&\cong \Delta_{pdt}\boldsymbol{z}'_{pdt}D\boldsymbol{z}_{pdt}\Delta'_{pdt}.
\end{aligned} \tag{15}$$

For the second term in (14), we have:

$$\mathsf{Var}(\varepsilon_{pdt}) = \mathsf{Var}[\mathsf{E}(\varepsilon_{pdt}|\boldsymbol{b}_{pdt})] + \mathsf{E}[\mathsf{Var}(\varepsilon_{pdt}|\boldsymbol{b}_{pdt})] = \mathsf{E}[\mathsf{Var}(\varepsilon_{pdt}|\boldsymbol{b}_{pdt})] = \left(\Phi^{\frac{1}{2}}\Sigma\Phi^{\frac{1}{2}}\right)_{pdt}, \tag{16}$$

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_p so that

$$\mathsf{Var}(\boldsymbol{\varepsilon}_p) \quad = \Phi^{\frac{1}{2}} A_p^{\frac{1}{2}} R_p A_p^{\frac{1}{2}} \Phi^{\frac{1}{2}}, \tag{17}$$

where $\boldsymbol{\varepsilon}_p$ groups all error terms within subject p, R_p is the correlation matrix, and A_p is a diagonal matrix containing the variances following from the generalized linear model specification of Y_{pdt} given the random effects $\boldsymbol{b}_{pdt} = \mathbf{0}$, i.e., with diagonal elements $v(\mu_{pdt}|\boldsymbol{b}_{pdt} = 0)$. If the canonical link is used, we have $A_p = \Delta_p$ and then (14) becomes

$$\mathsf{Var}(\boldsymbol{Y}_p) \cong \Delta_p Z_p D Z_p' \Delta_p' + \Phi^{\frac{1}{2}} \Delta_p^{\frac{1}{2}} \boldsymbol{R_p} \Delta_p^{\frac{1}{2}} \Phi^{\frac{1}{2}}. \tag{18}$$

To determine $\mathsf{Corr}(Y_{pdt}, Y_{p'd't'})$, we still need to calculate $\mathsf{Cov}(Y_{pdt}, Y_{p'd't'})$. Similarly to the above, we have that $\mathsf{Cov}(\mu_{pdt}, \varepsilon_{p'd't'}) = \mathsf{Cov}(\varepsilon_{pdt}, \mu_{p'd't'}) = 0$. Therefore, we only need to derive $\mathsf{Cov}(\mu_{pdt}, \mu_{p'd't'})$:

$$\begin{aligned}
\mathsf{Cov}(Y_{pdt}, Y_{p'd't'}) &= \mathsf{Cov}(\mu_{pdt}, \mu_{p'd't'}) \\
&= \mathsf{Cov}[\mu_{pdt}(\boldsymbol{x}_{pdt}'\beta + \boldsymbol{z}_{pdt}'\boldsymbol{b}_{pdt}), \mu_{p'd't'}(\boldsymbol{x}_{p'd't'}'\boldsymbol{\beta} + \boldsymbol{z}_{p'd't'}'\boldsymbol{b}_{p'd't'})] \\
&\cong \left(\frac{\partial \mu_{pdt}}{\partial \boldsymbol{b}_{pdt}} \Big|_{\boldsymbol{b}_{pdt}=0} \right) \mathsf{Cov}(\boldsymbol{b}_{pdt}, \boldsymbol{b}_{p'd't'}) \left(\frac{\partial \mu_{p'd't'}}{\partial \boldsymbol{b}_{p'd't'}} \Big|_{\boldsymbol{b}_{p'd't'}=0} \right)' \\
&\cong \left(\frac{\partial \mu_{pdt}}{\partial \eta_{pdt}} \frac{\partial \eta_{pdt}}{\partial \boldsymbol{b}_{pdt}} \Big|_{\boldsymbol{b}_{pdt}=0} \right) \mathsf{Cov}(\boldsymbol{b}_{pdt}, \boldsymbol{b}_{p'd't'}) \left(\frac{\partial \mu_{p'd't'}}{\partial \eta_{p'd't'}} \frac{\partial \eta_{p'd't'}}{\partial \boldsymbol{b}_{p'd't'}} \Big|_{\boldsymbol{b}_{p'd't'}=0} \right)' \\
&\cong \Delta_{pdt} \boldsymbol{z}_{pdt}' \mathsf{Cov}(\boldsymbol{b}_{pdt}, \boldsymbol{b}_{p'd't'}) \boldsymbol{z}_{p'd't'} \Delta_{p'd't'}'.
\end{aligned} \tag{19}$$

The covariances $\mathsf{Cov}(b_{pdt}, b_{p'd't'})$ depend on which of the random effects are common when correlating Y_{pdt} and $Y_{p'd't'}$. Using (18) and (19), we can calculate the correlation for any given situation, any give GLMM. In the next section, we will specialize the correlation to the case of binary data with random effects and without serial correlation.

4 Data Analysis

Let us now apply the concepts of reliability and generalizability to the pooled data described in Section 2. We will investigate the impact of 'country' on measurement error. First, we will assess the overall reliability for CGI response, ignoring country effects. Subsequently, country effects will be extracted by including country as a fixed effect into the model. Next, we will investigate the impact of country on reliability through application of the same model to each country separately. We will also study the impact of a single country on overall reliability by leave-on-out ideas, i.e., by omitting one country at a time. Finally, we will assess the overall impact of country via generalizability theory.

4.1 Overall Reliability of CGI

First, we apply a simple random-intercept model, combined with fixed effects for treatment, time and their intercation. With the logit link, (12) becomes:

$$Y_{pdt} = \frac{\exp(\mu + b_p + \mu_d + \mu_t + \mu_{dt})}{1 + \exp(\mu + b_p + \mu_d + \mu_t + \mu_{dt})} + \varepsilon_{pdt}, \tag{20}$$

where μ_d, μ_t, and μ_{dt} denote the fixed effects for day, treatment, and their interaction, respectively, and b_p represents the random patient effect.

The overall correlation of observations within the same subject, on the same treatment, but on different time points, and conditioning on treatment and time points, can be expressed as $\text{Corr}(Y_{pdt}, Y_{pd't} \mid t, d, d')$. In this model, we have $\boldsymbol{Z} = \mathbf{1}$ and $\boldsymbol{D} = \sigma_p^2$, a scalar representing the variance of the random intercept, and since (20) does not include serial correlation we have that $\boldsymbol{R_p} = \boldsymbol{I}$. It is therefore easy to show that the variance covariance matrix (18) reduces to

$$\text{Var}(Y_p) \cong \Delta_p(\sigma_p^2 \boldsymbol{J})\Delta_p' + \Phi\Delta_p = \Delta_p(d\boldsymbol{J} + \Phi\Delta_p^{-1})\Delta_p',$$

where $\boldsymbol{J}$ is a rectangular matrix of ones. Furthermore, Δ_p is a diagonal matrix with $V_{pdt}(0)$ as diagonal elements, where the variance function $V_{pdt}(0) = \mu_{pdt}\,\big|_{b_{pdt}=0}\,(1 - \mu_{pdt}\,\big|_{b_{pdt}=0})$, and

therefore we have

$$\begin{aligned} \mathsf{Var}(Y_{pdt}) &\cong \mathrm{diag}(\mathrm{V_{pdt}}(0))[\sigma^2_{\mathrm{p}}\boldsymbol{J} + \Phi\mathrm{diag}(\mathrm{V_{pdt}}(0))^{-1}]\mathrm{diag}(\mathrm{V_{pdt}}(0)), \qquad (21) \\ \mathsf{Cov}(Y_{pdt}, Y_{pd't}) &\cong \mathrm{diag}(\mathrm{V_{pdt}}(0))[\sigma^2_{\mathrm{p}}\boldsymbol{J}]\mathrm{diag}(\mathrm{V_{pd't}}(0)). \qquad (22) \end{aligned}$$

Based on (21) and (22), we can determine a first-order approximation of the marginal correlation between time point d and d', which is the intraclass correlation coefficient of reliability:

$$\rho = \mathsf{Corr}(Y_{pdt}, Y_{pd't}) = \frac{\sigma^2_T\sqrt{V_{pdt}(0)V_{pd't}(0)}}{\sqrt{[\Phi_{pdt} + V_{pdt}(0)\sigma^2_N]\cdot[\Phi_{pd't} + V_{pd't}(0)\sigma^2_N]}}, \qquad (23)$$

where σ^2_T represents the covariance between the random effects and σ^2_N is the variance resulting from the random effects. In this model, $\sigma^2_T = \sigma^2_N = \sigma^2_p$ since all other covariates are fixed effects. The delta method can be usefully applied to estimate the standard error:

$$\begin{aligned} \frac{\partial\rho}{\partial(\boldsymbol{\beta},\boldsymbol{\lambda})} &= \left(\frac{\partial(\boldsymbol{\eta},\boldsymbol{\sigma}^2)}{\partial(\boldsymbol{\beta},\boldsymbol{\lambda})}\right)\left(\frac{\partial(V_{pdt}(0), V_{pd't}(0), \sigma^2_T, \sigma^2_N, \phi)}{\partial(\boldsymbol{\eta},\boldsymbol{\sigma}^2)}\right), \\ &\quad \times \left(\frac{\partial\rho}{\partial(V_{pdt}(0), V_{pd't}(0), \sigma^2_T, \sigma^2_N, \phi)}\right). \end{aligned}$$

Explicit expressions for the various components follow from straightforward linear algebra. The SAS V9.1 procedure GLIMMIX was used to estimate Φ, σ^2_p, and V_{pdt}. Table 1(a) summarizes the results.

In case of continuous data, a single-measure overall intraclass correlation coefficient reliability would have been obtained (Vangeneugden *et al.* 2005). Here, for the binary data case, a separate intraclass coefficient of reliability is produced for each treatment group and each time point. From Table 1(a), we observe that the correlation is somewhat higher in the risperidone arm and that the correlation between week 1 and other time points is lower than the correlation between any two other time points that do not involve week 1.

4.2 Overall Reliability of CGI Response Adjusting for Country

In Section 4.1, only treatment, time, and their interaction were included. Now, we will include countries as fixed effects, which will result in intraclass coefficients of reliability per treatment,

time point, and country combination. We will not present all coefficients but merely present the coefficients for one country, the U.S.A., in Table 1(b). Additionally, we list the ICC of reliability between weeks 6 and 8 in the risperidone group for all countries in Table 2. The results for the U.S.A. are consistent with the overall results, and when we investigate the correlation between weeks 6 and 8 in the risperidone group, we observe from column 3 in Table 2 that the ICC is rather stable across countries, the lowest correlation begin for Austria (0.65, s.e. 0.09) and the highest for the U.S.A., Sweden, and Spain (0.78, s.e. 0.02).

4.3 Overall Reliability of CGI by Country and Impact on Overall Reliability by Leaving Out a Country

When we apply the model to each country separately, we observe that the model did not always converge and estimates were less stable, especially and not surprisingly, in countries with few patients. Patients included in Finland had data up to week 6 only (Hoyberg *et al.* 1993). The results are summarized in the third column of Table 2. A different way to investigate impact of country on reliability is by leaving out one country at a time. If the overall reliability increases, this would provide evidence for a poor reliability in the specific country. The results are summarized in the fifth column of Table 2. Note that the impact was low for all countries, again suggesting that reliability is relatively consistent across countries.

4.4 Estimating Impact of Country From Generalizability Theory

Subgroup analysis by country as shown in the previous two sections can be enlightening. Now, we want to quantify their effect on measurement error and calculate a generalizability coefficient, thereby generalizing results across countries. We will add a random effect for country into the previous model, so that we have a model with time, treatment, and their interaction as fixed effects, and further country, indexed by c, and patient as random effects:

$$Y_{pdtc} = \frac{\exp(\mu + b_p + \mu_d + \mu_t + \mu_{dt} + b_c)}{1 + \exp(\mu + b_p + \mu_d + \mu_t + \mu_{dt} + b_c)} + \varepsilon_{pdtc}. \tag{24}$$

From (24) we can calculate the overall test-retest reliability coefficient as in Section 4.1, but this time accounting for country as a random effect instead of extracting it as a random effect. Then, $\sigma_T^2 = \sigma_N^2 = \sigma_p^2 + \sigma_c^2$ in (23). Table 1(c) shows that the results are consistent with the overall reliability coefficients.

This test-retest reliability coefficient for any given country and time point follows directly from analyzing the clinical trial, similar to generalizability coefficients that are computed after design and analysis of a G-study. In the spirit of D-studies, we can also generalize across countries. Indeed, although patients are nested within country in a clinical trial setting, we assume, by way of a thought experiment, that patients could switch from one country to another, with the aim to evaluate the impact of country. We then have that $\sigma_T^2 = \sigma_p^2$ and $\sigma_N^2 = \sigma_p^2 + \sigma_c^2$, needed to calculate Corr$(Y_{pdtc}, Y_{pd'tc'})$ as in (23). Table 1(d) provides the ensuing ICC coefficients.

Thus, generalizing across time points and countries, or taking account of impact of variance of country, reduces the overall test-retest reliability approximately by 5%: for risperidone the decrease in reliability amounted to between 4–7% and for active control this was between 3–6%. In this situation, the price for setting up an international trial instead of a single country is rather small. This insight is relevant and underscores the usefulness of the thought experiment.

Evidently, the methodology can easily be extended to more complex situations including, for example, serial correlation or random time effects but also additional variables, such as, for example, age and sex of the patient.

4.5 Estimating Impact of Baseline PANSS Negative Subtotal on Reliability of CGI Response

In the computations reported above, a relatively high generalizability coefficient suggested that country does not have an important impact on the test-retest reliability and on measurement error. We now investigate the impact of baseline PANSS Negative subtotal on measurement error. We included a random intercept for baseline PANSS Negative subtotal instead of country in

model (24). Subsequently, we derived the variance components and calculated the generalizability coefficient for baseline PANSS Negative subtotal, similar to how it was done for country. In this analysis, the reduction in generalizability coefficient was more substantial: in the risperidone group between week 6 and 8, we have that the ICC reduces from 0.55 (s.e. 0.13) to 0.39 (s.e. 0.13) when generalizing across baseline negative subtotal. Full details are given in Table 1(e). This indicates that baseline PANSS Negative subtotal reduces the test-retest reliability. A clinical explanation for this phenomenon could be that patients with a higher deficit in negative symptoms at baseline, such as poverty of speech, apathy, or emotional withdrawal, are more difficult to evaluate, resulting in higher measurement error and lower test-retest reliability. A practical conclusion would be that additional training is needed for professionals having to rate patients with a high baseline negative subtotal or, even more invasive, in the recommendation to use a different scale in this type of patients.

5 Concluding Remarks

In this paper, we have extended classical reliability measures and associated estimation procedures in four important ways. First, fully longitudinal data can be used, rather than paired measurements. Second, clinical trial data can be employed or, more generally data from other studies not expressly designed for the investigation of reliability, through adopting a modeling framework, obviating the need for parallel measurements. Third, the broad generalizability theory framework is invoked, encompassing the various classical reliability versions, such as inter-rater and test-retest reliability, and allowing for the study of such important factors' impact as day of measurement, rater, country, investigator, etc. Fourth, all calculations are conducted within the generalized linear mixed model paradigm, allowing one not only to accommodate all aforementioned aspects, but also to deal with Gaussian and non-Gaussian data alike. Specific emphasis was put on binary outcomes, but analogous computations for nominal, ordinal, or count data can be done as well. Unlike in the Gaussian case, the reliability and generalizability coefficients depend on the days, raters, countries, or whatever levels studied. This is due to the mean-variance link and the

nonlinear nature of the model.

The work was motivated by and applied to data from multi-country trial data collected in patients with chronic schizophrenia. Using the generalizability framework, we were able to establish that the reliability measures are rather stable across countries, and no single country as an undue effect on the overall reliability. Country-specific reliabilities varied in a usefully narrow range.

An important conclusion, never reached before, is that the price to pay for a multi-country study, rather than a single-country one, is a mere 5% in test-retest reliability. The ability to conduct multi-country studies is important in view of the availability of a larger pool of available patients, thereby reducing the length of the accrual period and/or increasing the sample size, and hence power.

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

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 and Mental Disease* **180**, 723–728.

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

Fleiss J.L. (1981). *Statistical Methods for Rates and Proportions*. New York: John Wiley.

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

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.

Lee, Y., Nelder, J.A., and Pawitan, Y. (2006). *Generalized Linear Models with Ranom Effects*. Boca Raton: Chapman & Hall/CRC.

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.

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.

Shavelson, R.J., Webb, N.M., and Rowley, G.L.(1989). Generalizability theory. *American Psychologist* **44**, 922–932.

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

Stratford P. (1989). Consistency or differentiating among subjects? *Physical Therapy* **69**, 299–300.

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.

Table 1: *ICC matrices (standard error), accounting for treatment, time and their interaction. Standard errors are calculated from the delta method. Five different situations are reported.*

	risperidone				active control			
Week	2	3	6	8	2	4	6	8
(a) Overall								
1	.52(.04)	.55(.04)	.55(.04)	.55(.04)	.42(.04)	.47(.04)	.50(.04)	.50(.04)
2	1	.74(.02)	.74(.02)	.74(.02)	1	.61(.04)	.65(.03)	.66(.03)
4		1	.78(.02)	.78(.02)		1	.72(.03)	.73(.02)
6			1	.79(.01)			1	.78(.02)
(b) By country: U.S.A.								
1	.52(.06)	.54(.06)	.54(.05)	.54(.05)	.38(.07)	.42(.07)	.46(.06)	.46(.06)
2	1	.73(.03)	.74(.03)	.74(.02)	1	.57(.06)	.62(.05)	.63(.05)
4		1	.77(.02)	.77(.02)		1	.69(.04)	.70(.04)
6			1	.78(.02)			1	.76(.02)
(c) Country as random effect: U.S.A.								
1	.53(.05)	.55(.05)	.56(.05)	.56(.05)	.40(.06)	.45(.06)	.48(.05)	.48(.05)
2	1	.74(.03)	.75(.02)	.75(.02)	1	.59(.05)	.64(.04)	.65(.04)
4		1	.78(.02)	.78(.02)		1	.71(.03)	.72(.03)
6			1	.79(.02)			1	.77(.02)
(d) Generalized across countries: U.S.A.								
1	.49(.05)	.51(.05)	.51(.05)	.51(.04)	.37(.05)	.41(.05)	.44(.05)	.45(.05)
2	1	.68(.03)	.69(.03)	.69(.03)	1	.55(.05)	.59(.04)	.60(.04)
4		1	.72(.03)	.72(.03)		1	.65(.04)	.66(.03)
6			1	.72(.03)			1	.71(.03)
(e) Generalized across baseline negative symptoms								
1	.37(.13)	.38(.13)	.39(.13)	.39(.13)	.29(.10)	.32(.11)	.35(.12)	.35(.12)
2	1	.51(.18)	.52(.18)	.52(.18)	1	.43(.15)	.46(.16)	.46(.16)
4		1	.54(.18)	.54(.18)		1	.50(.17)	.51(.17)
6			1	.55(.19)			1	.54(.18)

Table 2: *Reliability by country and impact of country on overall reliability table. ICC ρ (standard error) between Week 6 and 8 in risperidone, with (1) country as fixed effect, (2) country-specific analyzes, and (3) a given country omitted. (NA: not available by lack of data.)*

Country	Number of patients	Country as fixed effect	By country	Omitting a given country
Argentina	31	0.76 (0.04)	NA	0.78 (0.02)
Austria	29	0.65 (0.09)	0.02 (0.04)	0.78 (0.01)
Belgium	26	0.76 (0.04)	NA	0.78 (0.01)
Brazil	44	0.73 (0.05)	0.54 (0.14)	0.79 (0.01)
Canada	44	0.77 (0.02)	0.76 (0.10)	0.79 (0.01)
Denmark	47	0.77 (0.02)	0.65 (0.09)	0.80 (0.01)
Spain	32	0.78 (0.02)	0.88 (0.07)	0.79 (0.01)
Finland	71	0.66 (0.07)	NA	0.79 (0.01)
France	92	0.77 (0.02)	0.40 (0.11)	0.81 (0.01)
Great Britain	21	0.77 (0.03)	0.91 (0.05)	0.78 (0.01)
Germany	25	0.73 (0.06)	NA	0.78 (0.01)
Italy	39	0.70 (0.07)	NA	0.77 (0.02)
Mexico	36	0.76 (0.03)	0.92 (0.06)	0.78 (0.02)
Netherlands	17	0.74 (0.06)	0.71 (0.37)	0.78 (0.01)
Norway	37	0.71 (0.06)	0.91 (0.04)	0.78 (0.01)
South Africa	79	0.71 (0.05)	0.80 (0.09)	0.78 (0.02)
Sweden	30	0.78 (0.02)	0.94 (0.03)	0.78 (0.01)
U.S.A.	122	0.78 (0.02)	0.75 (0.04)	0.79 (0.02)