Available online at www.sciencedirect.com
SCIENCE DIRECT®
Controlled Clinical Trials 25 (2004) 13–30
Controlled Clinical Trials
www.elsevier.com/locate/conclintrial

Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements

Tony Vangeneugden[a,b,*], Annouschka Laenen[b], Helena Geys[b], Didier Renard[b,c], Geert Molenberghs[b]

[a] *Johnson & Johnson Pharmaceutical Research and Development, Beerse, Belgium*
[b] *Limburgs Universitair Centrum, tUL, Center for Statistics, Biostatistics, Diepenbeek, Belgium*
[c] *Eli Lilly, Mont-Saint-Guibert, Belgium*

Received 31 March 2003; accepted 8 August 2003

Abstract

Repeated measures are exploited to study reliability in the context of psychiatric health sciences. It is shown how test–retest reliability can be derived using linear mixed models when the scale is continuous or quasi-continuous. The advantage of this approach is that the full modeling power of mixed models can be used. Repeated measures with a different mean structure can be used to usefully study reliability, correction for covariate effects is possible, and a complicated variance–covariance structure between measurements is allowed. In case the variance structure reduces to a random intercept (compound symmetry), classical methods are recovered. With more complex variance structures (e.g., including random slopes of time and/or serial correlation), time-dependent reliability functions are obtained. The methodology is motivated by and applied to data from five double-blind randomized clinical trials comparing the effects of risperidone to conventional antipsychotic agents for the treatment of chronic schizophrenia. Model assumptions are investigated through residual plots and by investigating the effect of influential observations.

Keywords: Reliability; Linear mixed model; Repeated measurements; Psychiatry; Rating scale

1. Introduction

Many measurements in medical practice and research are based on observations made by clinicians. As measurements are prone to error, observer reliability and agreement are important issues in medicine.

* Corresponding author. Janssen Farmaceutica, Turnhoutseweg 30, 2340 Beerse, Belgium. Tel.: +32-14-603595; fax: +32-14-606342.
E-mail address: tvangene@prdbe.jnj.com (T. Vangeneugden).

0197-2456/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2003.08.009

The terms ''observer reliability'' and ''agreement'' are often used interchangeably, but in theory they are different concepts. Reliability coefficients express the ability to differentiate among subjects. They are ratios of variances: in general, the variance attributed to the difference among subjects divided by the total variance [1]. Agreement refers to conformity. Agreement parameters determine whether the same value is achieved if a measurement is performed twice, either by the same observer or by different observers. In homogeneous populations one can imagine that reliability might be low while agreement is high; in a heterogeneous population, reliability and agreement measures will correspond well [2]. The parameters for assessment of observer reliability and agreement differ according to the scale of measurement. For nominal and ordinal categorical measurements, respectively, the κ-coefficient and the weighted κ-coefficient (κ_W) are measures of agreement. In case of continuous data, the intraclass correlation coefficient (ICC) is used to measure observer reliability, although the ICC also can be used for ordinal categorical data.

As stated by Fleiss: ''The most elegant design of a clinical study will not overcome the damage by unreliable or imprecise measurement'' [3]. 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 that has been validated before and for which intrarater (test–retest) and interrater 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, 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 was the scale or how large was the observed measurement error. In this paper, we want to propose a framework to study trial or population specific reliability. Attention will be restricted to quantitative, interval-scaled measurements. The goal is to use clinical trial data at hand and to evaluate reliability of the measurement. 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 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 measurements in clinical trials are often ''unstable,'' in a psychometric sense, due to present treatment and time effect. In contrast, in the classical theory setting, reliability testing is always done on patients in a steady-state

condition, resulting in "parallel measurements" within the patients. Therefore, one of the biggest challenges is to find a way to extract these effects and to make the bridge to the classical reliability coefficient, a well-known and established concept in psychometrics.

The next section reviews the concept of reliability, introduces a new and flexible way to calculate the reliability of continuous measurement scales measured repeatedly over time, and describes possible approaches to investigate model assumptions. The following section introduces data from a meta-analysis of five clinical trials comparing antipsychotic agents for the treatment of chronic schizophrenia and then applies the methods just introduced on these data. The final section contains some concluding remarks.

2. Methodology

First, we give a general outline of the concept of reliability. Thereafter, we will introduce the model families that will further be used to approach this quantity in a longitudinal setting; subsequently we will discuss diagnostic tools to evaluate the fit of these models; and finally, we derive the general formula for reliability for this family of models.

2.1. Reliability

In the classical test theory, the outcome of an interval scaled test is modeled as

$$X = \tau + \epsilon,$$

where X represents an observation or measurement, τ is the true score, and ϵ the corresponding measurement error. It is assumed that the measurement errors are mutually uncorrelated as well as with the true scores. If this assumption is correct, we obtain

$$\mathrm{Var}(X) = \mathrm{Var}(\tau) + \mathrm{Var}(\epsilon).$$

The reliability of a measuring instrument is defined as the ratio of the true score variance to the observed score variance, i.e.,

$$R = \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(X)} = \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(\tau) + \mathrm{Var}(\epsilon)}. \tag{1}$$

One can easily show that the reliability coefficient is in fact an intraclass correlation coefficient. Suppose we have two measurements of the same patient, either from two raters or from the same rater, taken at two instances not too far apart, $X_1 = \tau + \epsilon_1$ and $X_2 = \tau + \epsilon_2$, with $\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \mathrm{Var}(X)$ and $\mathrm{Var}(\epsilon_1) = \mathrm{Var}(\epsilon_2) = \mathrm{Var}(\epsilon)$, i.e., parallel measurements. Further, the covariance of the two measurements equals

$$\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(\tau + \epsilon_1, \tau + \epsilon_2) = \mathrm{Var}(\tau),$$

and the correlation between the two measurements can be written as

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)}\sqrt{\mathrm{Var}(X_2)}} = \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(\tau) + \mathrm{Var}(\epsilon)} = R. \tag{2}$$

This shows that reliability is in fact an intraclass correlation coefficient with patient taken as the class. Stronger, as Bartko stated, a reliability coefficient defined as a ratio of variances that are estimated by a linear model, can only be correct when it can be interpreted as a "correlation coefficient" [4].

The outcomes X_1 and X_2 can, for example, be two subscores of a test, in which case we are also talking about split-halve reliability. If these outcomes represent two items of a scale, then this translates to internal consistency. If the scores are two measurements of the same instrument, measured at different moments in time, then we are dealing with test–retest reliability. When the scores are obtained by two different raters, at one moment in time, then the measure is called interrater reliability. Note that the assumption of steady state-behavior of the measurements, i.e., the assumption that measurements are parallel (same mean and same variance), is crucial. If, for instance, the patients are rated by the same investigator on two occasions that are too far apart, the patient's condition can have changed, translating into a low intraclass correlation coefficient, even in the case of highly reliable measures.

In the classical approach, reliability is estimated by the intraclass correlation coefficient [3–5]. For a simple replication study, this can be derived from a one way analysis of variance with patient as factor (Table 1).

The estimate for the intraclass correlation coefficient of reliability in Bartko [4] then is:

$$\hat{R}_c = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2} = \frac{\mathrm{BMS} - \mathrm{WMS}}{\mathrm{BMS} + (k-1)\mathrm{WMS}}.$$

Next, we will see how advantage can be made of linear mixed models in case repeated or longitudinal measures are taken, rather than a single measure or a pair of measures.

2.2. Linear mixed models

Methods for continuous data form the best developed and most advanced body of research, while the same is true for software implementation. This is natural, since the special status and the elegant properties of the multivariate normal distribution simplify model building and ease software development. It is in this area that the linear mixed model is situated [6,7]. Gaussian data can be modeled entirely in terms of their means, variances, and covariances. The parameters of the mean model are referred to as "fixed-effects" parameters, and the parameters of the variance–covariance model as "covariance parameters". The fixed-effects parameters capture the influence of explanatory variables on the mean structure, exactly as in the standard linear model. However, the occurrence of random effects and a structured covariance matrix distinguishes the linear mixed model from the standard linear model. The need for covariance modeling arises quite frequently in applications, such as when repeated

Table 1
ANOVA table to derive reliability coefficient from a simple replication study

Source of variation	*df*	Mean sum of sq.	Exp. sum of sq.
Between patient	$n-1$	BMS	$\sigma_e^2+k\sigma_p^2$
Within patients (error)	$n\ (k-1)$	WMS	σ_e^2
Total	$nk-1$		

k is the number of measurements per patient, n is the number of patients.

measurements are taken on the same experimental unit, with spatially correlated data, or when experimental units can be grouped into clusters and data from a cluster are correlated. One can distinguish between three components of variability. Part of the covariance structure arises from so-called "random effects" (i.e., additional covariate effects with random parameters). These are effects that arise from the characteristics of individual subjects. The variances of the random-effects parameters are commonly referred to as "variance components" [8]. Another component of the variability is the serial correlation, which captures that measurements taken close together in time are typically more strongly correlated than those taken further apart in time. On a sufficiently small time scale, this kind of structure is almost inevitable. The last component is the measurement error: when the measurement process involves difficult to quantify or "fuzzy" determinations, the results may show substantial variation even when two measurements are taken at the same time from the same subject.

A linear mixed-effects model with serial correlation can be written as

$$\boldsymbol{Y}_i = X_i\beta + Z_i\boldsymbol{b}_i + \boldsymbol{W}_i + \boldsymbol{\epsilon}_i, \tag{3}$$

where $\boldsymbol{Y}_i$ is the n_i dimensional response vector for subject i, $1 \le i \le N$, N is the number of subjects, X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ known design matrices, β is the p-dimensional vector containing the fixed effects, $\boldsymbol{b}_i \sim N(\boldsymbol{0},D)$ is the q-dimensional vector containing the random effects, $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma^2 I_{n_i})$ is an n_i-dimensional vector of measurement error components, and $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N$ 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 $N(\boldsymbol{0}, \tau^2 H_i)$ law. The serial covariance 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 $d_{ij} = d_{ji}$. Inference is based on the marginal distribution of the response $\boldsymbol{Y}_i$, which, after integrating over random effects, can be expressed as

$$\boldsymbol{Y}_i \sim N(X_i\beta, Z_i D Z_i' + \Sigma_i).$$

Here, $\Sigma_i = \sigma^2 I_{n_i} + \tau^2 H_i$ is a $(n_i \times n_i)$ covariance matrix grouping the measurement error and serial components.

2.3. Investigating influential observations via likelihood displacement and local influence

Fitting of mixed models just described is based upon maximum likelihood methods (maximum likelihood or restricted maximum likelihood). These methods can be sensitive to peculiar observations that can have an unusually large influence on the results of the analysis. Many diagnostic tools have been developed for linear regression models but the generalization of these methods is far from obvious. First, several kinds of residuals could be defined: the marginal residuals $\boldsymbol{Y}_i - X_i\hat{\beta}$, reflecting how a specific profile deviates from the overall population mean; the subject-specific residuals $\boldsymbol{Y}_i - X_i\hat{\beta} - Z_i\hat{\boldsymbol{b}}_i$, measuring how much the observed values deviate from the subject's own predicted regression line, and the estimated random effects $\hat{\boldsymbol{b}}_i$, reflecting how much specific profiles deviate

from the population average profile. Further, the linear mixed model involves two kinds of covariates. The matrix X_i represents the design matrix for the fixed effects, and Z_i is the design matrix for the random effects. Therefore, it is not clear how leverages should be defined, partially because the matrices X_i and Z_i are not necessarily of the same dimension. A classification of influential subjects can be based on Cook's distance, which measures how much parameter estimates change when a specific individual has been removed from the dataset [9]. In classical regression, closed-form expression exists, allowing easy calculation and also ascribing influence to the specific characteristics of the subjects (leverage, outlying). Unfortunately, this is no longer the case in linear mixed models. For exact Cook's distances, the iterative estimation procedure has to be used $N+1$ times, which can be extremely time-consuming. The local influence approach was first introduced by Cook [10]. The general idea is to give every individual its own weight in the calculation of the parameter estimates and to investigate how these estimates depend on the weights, locally around the equal-weight case, which is the ordinary maximum likelihood case. We restrict the discussion to models that assume conditional independence, hence no serial correlation and $\Sigma_i = \sigma^2 I_{n_i}$. Denote $\hat{\theta}$ as the maximum likelihood estimate for θ, obtained after maximizing $\ell(\theta)$ and $\hat{\theta}_\omega$ the estimate for θ after maximizing $\ell(\theta|\omega)$, any perturbed version of $\ell(\theta)$. The weight vector ω is N-dimensional and the original log-likelihood corresponds to $\omega = \omega_0 = (1, 1, \ldots, 1)'$. Cook proposed to measure the distance between $\hat{\theta}$ and θ_ω by the so-called "likelihood displacement", defined by

$$LD(\omega) = 2(\ell(\hat{\theta}) - \ell(\hat{\theta}_\omega)).$$

$LD(\omega)$ will be large if $\ell(\theta)$ is strongly curved at $\hat{\theta}$, which means that θ is estimated with high precision and $LD(\omega)$ will be small if θ is estimated with high variability [10]. From this perspective, a graph of $LD(\omega)$ versus ω contains essential information. Ideally, we would like a complete influence graph to assess influence for a particular model and a particular dataset. However, this is very difficult in high-dimensional situations. One method to extract the most relevant information from an influence graph is local influence, which uses normal curvatures. See Verbeke and Molenberghs [7] for more detail. Denote C_h as the normal curvature at the surface of $(\omega, LD(\omega))$ at ω_0, in the direction h. Large values of C_h indicate sensitivity to the induced perturbations in the direction h. There are several choices for h. One evident choice corresponds to the perturbation of the ith weight only. This is obtained by taking h equal to the vector h_i, which contains zeros everywhere except on the ith position. One can prove that C_i can be approximated by

$$C_i = -2(\hat{\theta} - \hat{\theta}^1_{(i)})'\ddot{L}_{(i)}\ddot{L}^{-1}\ddot{L}_{(i)}(\hat{\theta} - \hat{\theta}^1_{(i)}).$$

where $\ddot{L}$ and $\ddot{L}_{(i)}$ are respectively the matrix of second-order derivatives of full log-likelihood and of the log-likelihood calculated after deleting the i case and where $\hat{\theta}^1_{(i)}$ is the one-step approximation of $\hat{\theta}_{(i)}$ from a single Newton–Rapson step in the maximization procedure of $\ell_{(i)}(\theta)$, using $\hat{\theta}$ as starting value. One can also show that for sufficiently large N, C_i can be interpreted as an approximation to the global case-deletion diagnostics. Lesaffre and Verbeke have shown that C_i can be decomposed into five interpretable components: the "length" of the standardized covariate in the mean structure, overall measure for how well the observed data for the ith subject are predicted by the mean structure $X_i\beta$, two similar components for the covariance structure, and finally the size of the variability of the ith subject [11].

2.4. Estimation of reliability in the linear mixed models framework

The general formula to calculate the intraclass correlation coefficient for model (3) can be derived via Eq. (2). Denote Y_{it} the observed measurement of subject i on time point t; s will also be used to denote (a second) time point. Then

$$\begin{aligned}
\mathrm{Var}(Y_{is}) &= \boldsymbol{z}_s D \boldsymbol{z}_s' + \tau^2 + \sigma^2, \\
\mathrm{Var}(Y_{it}) &= \boldsymbol{z}_t D \boldsymbol{z}_t' + \tau^2 + \sigma^2, \\
\mathrm{Cov}(Y_{is}, Y_{it}) &= \boldsymbol{z}_s D \boldsymbol{z}_t' + \tau^2 (H_i)_{st}.
\end{aligned} \tag{4}$$

Therefore, reliability in this general setting with multiple time points is time- or lag-dependent. Denote the test–retest reliability between time point s and t by $R(s, t)$. From Eq. (4) we have

$$R(s,t) = \mathrm{Corr}(Y_{is}, Y_{it}) = \frac{\boldsymbol{z}_s D \boldsymbol{z}_t' + \tau^2 (H_i)_{st}}{\sqrt{\boldsymbol{z}_s D \boldsymbol{z}_s' + \tau^2 + \sigma^2}\sqrt{\boldsymbol{z}_t D \boldsymbol{z}_t' + \tau^2 + \sigma^2}}. \tag{5}$$

In the next section, in different settings, we will apply Eq. (5) above and derive the reliability of psychiatric symptom scales from such models, thereby generalizing the classical developments as outlined in previously.

3. Case study

In this section we introduce and analyze individual patient data from five double-blind randomized clinical trials, comparing the effects of risperidone to conventional antipsychotic 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, hallucinations, and disorganized thinking, 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, drugsensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia [12]. Classical reliability of the PANSS has been studied previously [13–15].

Since the label in most countries recommends that risperidone is most effective in schizophrenia at doses ranging from 4 to 6 mg/day, we include in our analyses only patients who received either these doses of risperidone or an active control like haloperidol, levomepromazine, perphenazine, or zuclopenthixol. Depending on the trial, treatment was administered for a duration of 4–8 weeks. For example, in the international trials by Peuskens et al. [16], Chouinard et al. [17], Marder and Meibach

[18], and Hoyberg et al. [19] patients received treatment for 8 weeks; in the study by Blin et al. [20] patients received treatment for 4 weeks, while in the study by Huttunen et al. [21] patients were treated over a period of 6 weeks. The sample sizes were 453, 176, 74, 49, and 71, respectively. Measurements were taken at weeks 1, 2, 3, 4, 6, and 8.

Let us now apply the previously developed methodology on the pooled data. We will assess the reliability for the PANSS, using the SAS procedure MIXED. The SAS codes for fitting the subsequent models and their respective reliabilities can be obtained from the authors' website. As mentioned earlier, the PANSS scale is a continuous response with 30 items. For this response we consider in turn four different models and calculate the corresponding reliability measures.

3.1. Model 1

First, we assume a linear mixed model with a random intercept and with time, treatment, and their interaction as fixed effects. Time is modeled as a factor with seven levels such that we obtain a saturated cell means model for time and treatment. In that case Eq. (3) becomes $Y_i = X_i\beta + Z_i b_i + \epsilon_i$, with Z_i an n_i-dimensional vector of ones, $\epsilon_i \sim N(\mathbf{0}, \sigma^2 I)$, and $b_i \sim N(\mathbf{0}, d)$. This can be rewritten as:

$$Y_{ijk} = \mu_{jk} + b_i + \epsilon_{ijk},$$

where Y_{ijk} is the measure at time point j for subject i under treatment k; μ_{jk} groups the fixed-effects structure, b_i is still the random intercept; and ϵ_{ijk} is the measurement error. The fitted variance components are $\hat{d} = 311.00$ and $\hat{\sigma}^2 = 125.14$. From Eq. (5) we can easily derive the formula for the reliability for this simple model. Since τ equals 0 (no serial correlation) and since z_s is 1 and $D=d$, the variability of the random intercept model, we have:

$$R = \frac{d}{d + \sigma^2}. \tag{6}$$

The reliability expresses the ratio of the variance explained by the model to the total observed variance. The link of Eq. (6) with the intuitive definition of reliability as we have expressed in Eq. (1) is obvious. For data containing two measurements per subject, this value equals the test–retest reliability of the instrument. For any series of repeated measurements, this value gives an overall measure of the intraclass correlation between all the measurements within subjects. For the PANSS data this global reliability measure yields a value of $R = 0.713$ (SE 0.012). The standard error is calculated using the delta method. If we apply Fisher's variance stabilizing transformation and the delta method, the 95% confidence interval is [0.688, 0.736].

Fig. 1 displays the standardized subject-specific residuals (Fig. 1A and B) for this model to assess the model fit and also investigates the distribution of the random intercept (Fig. 1C and D) and identifies influential observations (Fig. 1E and F).

The local influence method described above revealed five influential observations (Fig. 1E and F), two on the estimation of the fixed effects (81, 86) and three on the estimation of the variance components (240, 297, and 820). If we omit observations 81 and 86, this has little or no influence on estimation of variance components and the reliability coefficient remains $R = 0.71$. If we omit 240, 297, and 820, the

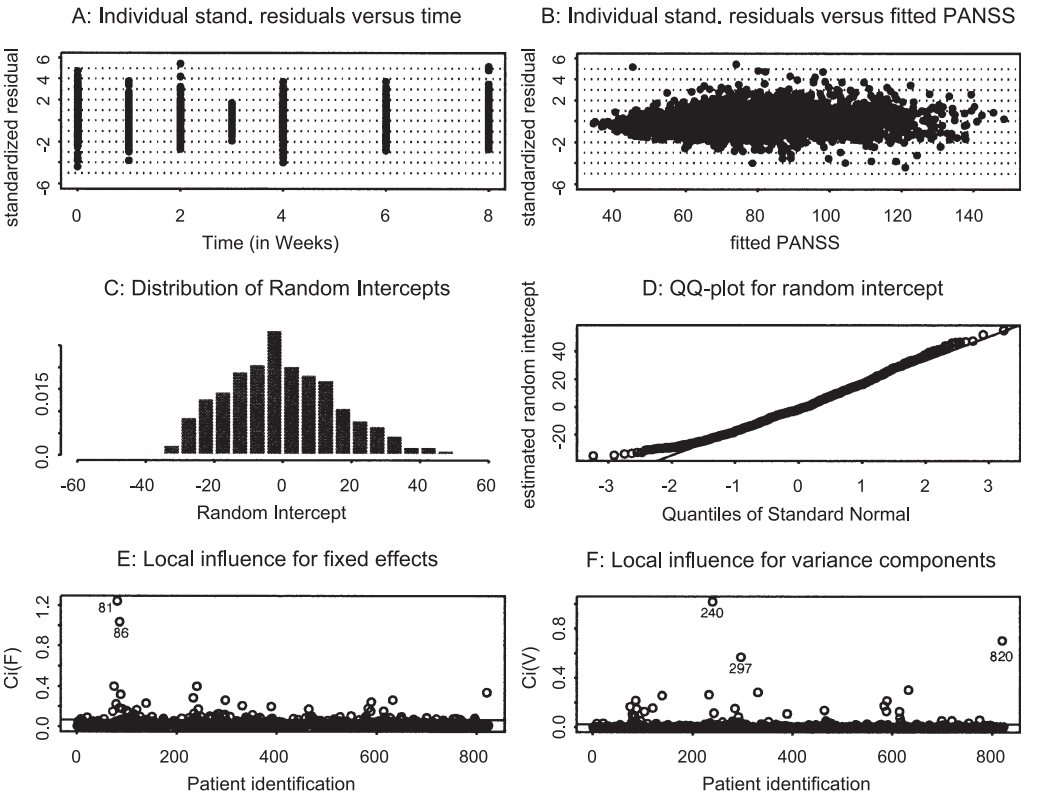

Fig. 1. Model 1. Diagnostic plots.

reliability increases to $R = 0.72$, which shows that the most influential measurements have little or no impact on the estimation of the overall reliability coefficient.

Note that the assumption of parallel measurements is not met. The mean PANSS decreases from 92.4 at baseline to 68.8 at endpoint. Even though classical reliability studies usually require the assumption of parallel measurements, our approach, due to the flexibility of the linear mixed model, obviates the need for this, since the mean and variability structures can be clearly separated. In particular, the linear mixed model will account for time and treatment effects by including them into the fixed effects component of the model. Although steady state is not taken care of by design as it would be in classical test–retest designs in psychometrics, steady state is provided through modeling at the analysis stage. A conceptually useful way to think about this is through the two-stage approach as the mixed effects model has been introduced, historically, by Laird and Ware [6]. If we derive the individual residuals for the model above and subsequently apply a random intercept model on these residuals without a fixed effect component ($\mu_{jk} = 1$), the same estimates for $\hat{d}$ and σ^2 would be obtained. Furthermore, there are three additional advantages: the mixed model approach can be applied when (1) there are more than two measurement occasions, (2) not all subjects have the same number of measurements (due to missingness or irregularly spaced measurement times), and (3) more complicated variance–covariance structures within subjects exist. To study these advantages further, we will consider more elaborate models in subsequent sections.

3.2. Model 2

The use of random effects in the assessment of reliability dates back to Bartko and has been described by Dunn [4,5]. Model 1 builds upon this work. In addition, we will introduce serial correlation and then generalize the calculation of reliability to this situation. Explicitly, the second model combines a random intercept with serial correlation. Typical choices for such serial correlation structures are based on exponentially or Gaussian decaying processes. These are standardly available in the SAS procedure MIXED [22]. In order to choose the covariance structure that best fits the data, an empirical variogram was created, which is shown in Fig. 2. For a formal introduction to the variogram in the context of longitudinal data, we refer to Diggle et al. [23] or Verbeke and Molenberghs [7]. The value of the variogram at time lag zero is an indication for the relative importance of the measurement error, the discrepancy between the variogram at the largest time lag, and the process variance (represented as a level straight line at the top of the plot) is an indication for the importance of the random intercept. The shape of the variogram describes the serial correlation process. The strength of the process is indicated by the amount of increase between zero and maximum time lags, while the shape of the curve is indicative of the shape of the process of serial decay.

The variogram is essentially flat. This implies that the largest component of variability is attributable to a random intercept (i.e., the within-unit correlation comes from a subject-specific intercept rather than from a serial correlation). However, there is a hint that a perhaps small serial component may be present; we opt for a Gaussian serial process. Then Σ_i, the covariance matrix grouping the measurement error and serial components in Eq. (3), is defined by the matrix with elements

$$\Sigma_{ss} = \sigma_{ss} = \tau^2 + \sigma^2,$$

$$\Sigma_{ss} = \sigma_{ss} = \tau^2 \exp(-u_{st}^2/\rho^2), \quad s \neq t,$$

where σ^2 denotes the measurement error variance and the remaining part is the serial variance component with u_{st} the time lag between measurements Y_{isk} and Y_{itk} for subject i and treatment k.

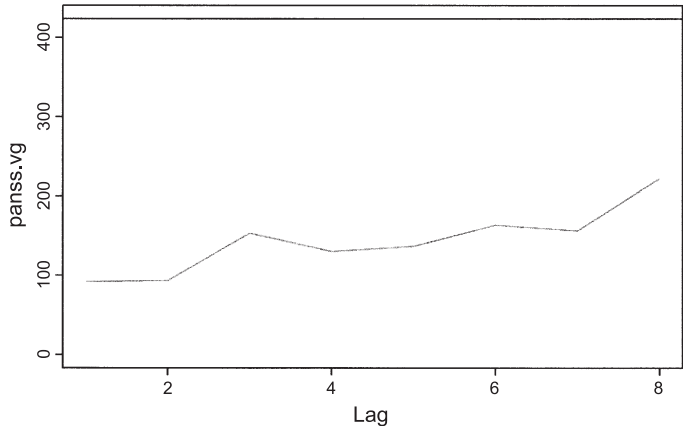

Fig. 2. Empirical variogram of the PANSS data.

The estimated covariance parameters of this model, applied to the PANSS data, are $\hat{d}=103.21$, $\hat{\tau}^2=274.97$, $\hat{\rho}=6.38$, and $\hat{\sigma}^2=65.21$.

The reliability can again be derived from Eq. (5) and is a function of time lag u_{st} between two observations measured at time point s and t

$$R(u_{st}) = \frac{d + \tau^2 \exp\left(\frac{-u_{st}^2}{\rho^2}\right)}{d + \tau^2 + \sigma^2}. \tag{7}$$

After correction for the fixed-time and treatment effects, the covariance parameter estimates show a considerable remaining serial component in the PANSS data. As can be seen from Eq. (7), a strong serial effect will lead to a fast decreasing reliability for increasing time lags. Fig. 3 shows that reliability is 0.80 or higher for measurements no further apart than 2 weeks but declines rapidly thereafter. This is consistent with the general consensus regarding the appropriate interval: generally speaking, a retest interval of 2 days to 2 weeks is appropriate [24]: if the interval is too short, the patients may remember their previous responses, if the interval is too long, things may have changed. A big advantage of model 2 is that this type of model allows to study the effect of lag time on the reliability.

The individual, subject-specific residuals of this model as well as the distribution of the random effects are displayed in Fig. 4. Although the standardized residuals are not as large as for model 1, Fig. 4B shows that the standardized residuals tend to increase with higher fitted PANSS values.

Influential observations were determined by means of likelihood displacement (Fig. 4E) instead of local influence due to the presence of serial correlation. Fig. 4 determines five influential observations: 79, 80, 240, 297, and 775. Removing these influential observations has little or no impact on estimation of reliability; the reliability does not differ more than 0.014 with or without the five influential observations.

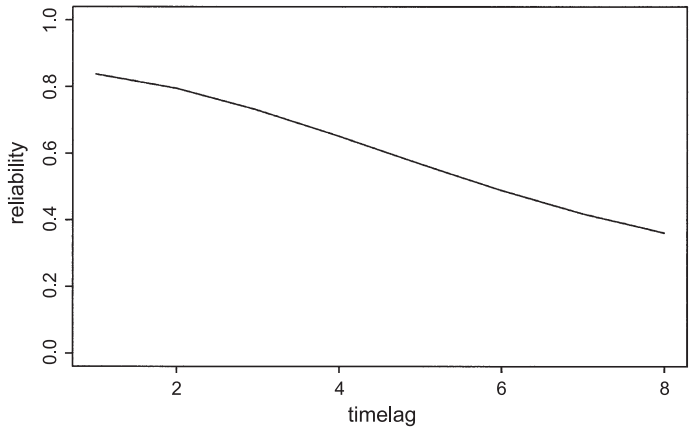

Fig. 3. PANSS. Reliability as a function of the time-lag u between any two measurements.

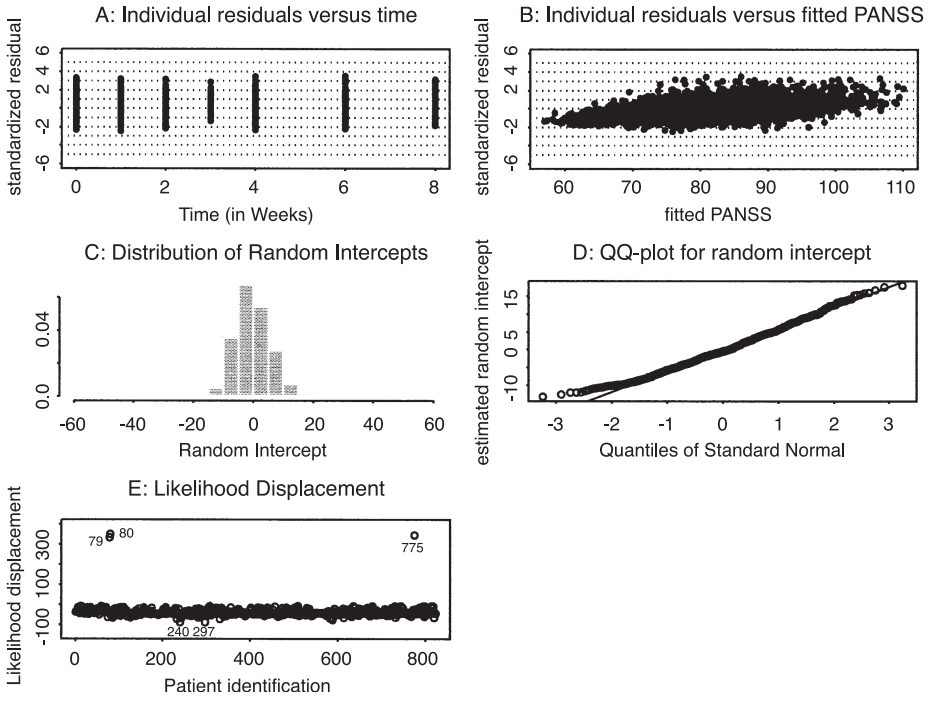

Fig. 4. Model 2. Diagnostic plots.

3.3. Model 3

After adding serial correlation in model 2 to the random-intercept model 1, we now add random slope in time as well. The random-effects variance then equals

$$D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}.$$

The estimated covariance parameters for the PANSS data are $\hat{d}_{11} = 47.24$, $\hat{d}_{12} = 13.65$, $\hat{d}_{22} = -0.10$, $\hat{\tau}^2 = 247.39$, $\hat{\rho} = 5.82$, and $\hat{\sigma}^2 = 63.96$. The residuals shown in Fig. 5 display a clear trend, variance of the residuals increase for increasing PANSS values and decrease in time, indicating an nonoptimal fit.

The model can now be written as follows:

$$Y_{ijk} = \mu_{jk} + (b_{i0} b_{i1}) \begin{pmatrix} 1 \\ j \end{pmatrix} + w_{ij} + \varepsilon_{ijk}. \tag{8}$$

From Eq. (5) we can derive the test–retest reliability for observations at time point s and time point t and lag time u_{st} between them:

$$R(s,t) = \frac{z_s D z_t' + \tau^2 \exp\left(\dfrac{-u_{st}^2}{\rho^2}\right)}{\sqrt{z_s D z_s' + \tau^2 + \sigma^2}\sqrt{z_t D z_t' + \tau^2 + \sigma^2}}. \tag{9}$$

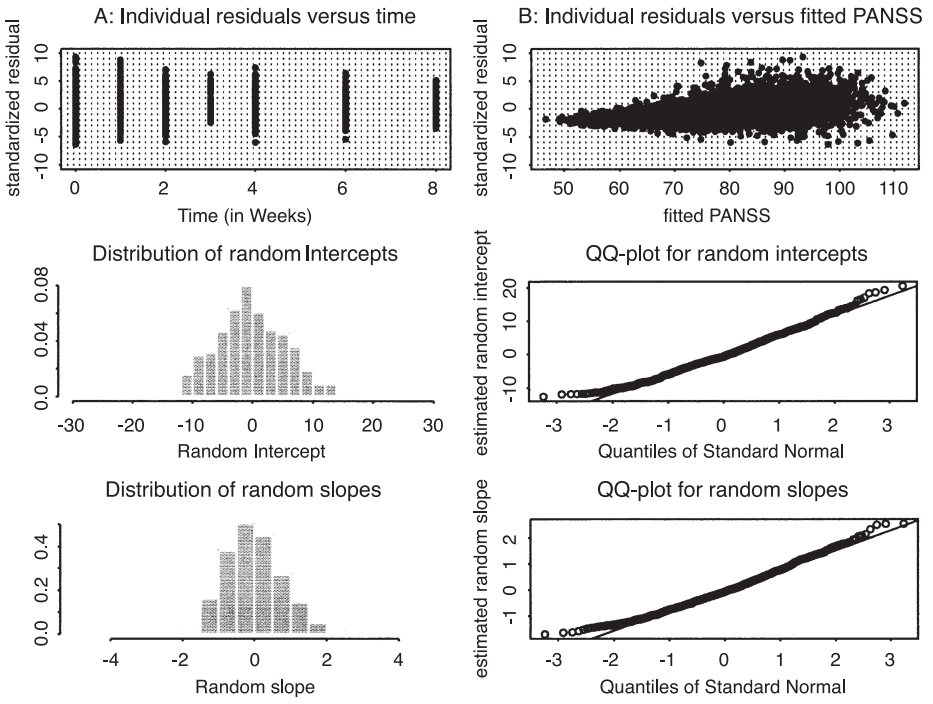

Fig. 5. Model 3. Diagnostic plots.

Here, z_s is the design row in Z corresponding to time s. Eq. (9) can be used to calculate the different reliabilities for any specific time point and for any given time lag. Due to the questionable fit, that will not be presented here. Instead, we will investigate a simpler model, by omitting the serial component.

3.4. Model 4

Only the random intercept and the random slope are retained in Eq. (8). The estimated covariance parameters for the PANSS data are $\hat{d}_{11} = 315.21$, $\hat{d}_{12} = -8.01$, $\hat{d}_{22} = 7.07$, $\hat{\sigma}^2 = 79.63$. Subsequently, the reliability of measurement observed on time s and time t:

$$R(s,t) = \frac{z_s D z_t'}{\sqrt{z_s D z_s' + \sigma^2}\sqrt{z_t D z_t' + \sigma^2}}.$$

Table 2 displays the reliability coefficients estimated from model 4; only the upper diagonal is shown for this symmetric test–retest reliability matrix. Again, we can observe that reliability is decreasing with increasing lag time. Another result that occurs is a slight increase in the reliability measure as time goes by, but for a fixed time lag.

Fig. 6 investigates the model diagnostics for this model and hints that the model fit has improved versus model 3. There are three influential observations for the variance components (240, 297, and 331) and three influential observations for the estimation of fixed effects (81, 86, and 88). After removing these influential measurements, the covariance parameters were estimated as $\hat{d}_{11} = 310.22$, $\hat{d}_{12} = -6.51$,

Table 2
Estimated test–retest reliabilities using model 4

Time point	Time point								
	0	1	2	3	4	5	6	7	8
0	0.80	0.79	0.76	0.72	0.68	0.62	0.57	0.52	0.47
1	·	0.79	0.79	0.76	0.73	0.69	0.65	0.61	0.57
2	·	·	0.80	0.79	0.78	0.75	0.72	0.69	0.66
3	·	·	·	0.81	0.81	0.80	0.78	0.75	0.73
4	·	·	·	·	0.82	0.82	0.82	0.80	0.79
5	·	·	·	·	·	0.84	0.84	0.84	0.83
6	·	·	·	·	·	·	0.86	0.86	0.86
7	·	·	·	·	·	·	·	0.87	0.88
8	·	·	·	·	·	·	·	·	0.89

$\hat{d}_{22} = 6.49$, $\hat{\sigma}^2 = 74.65$. The effect on estimation of the reliability coefficients is minimal. The largest difference is 0.03; e.g., test–retest reliability of observations on week 0 and week 8 increases from 0.47 to 0.5 after removal of the six influential observations.

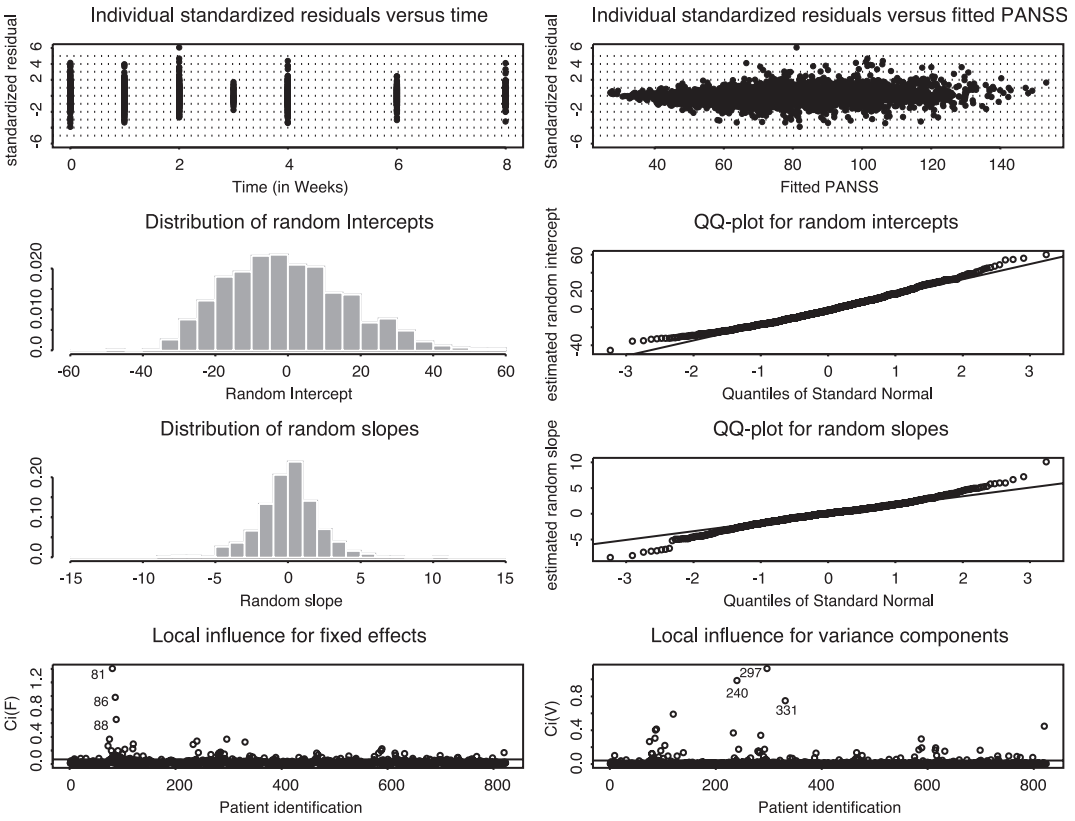

Fig. 6. Model 4. Diagnostic plots.

Table 3
Estimated variance component for models 1–4

Component	Par.	Estimates for various models			
		1	2	3	4
Var. rand. int.	d_{11}	311.00	103.21	47.24	315.21
Cov. (rand. int., rand. slope)	d_{12}			13.65	−8.01
Var. rand. slope	d_{22}			−0.10	7.07
Serial process variance	τ^2		274.97	247.39	
Serial process corr. par.	ρ		6.38	5.82	
Measurement error var.	σ^2	125.14	65.21	63.96	79.63
−2 log likelihood		33,870.7	33,232.4	33,192.2	33,331.4

3.5. Summary of different models

Table 3 summarizes the parameter estimates and the log likelihood of the different models described above. Model 3 is the model with the largest likelihood and would be the one of preference if one would rely purely on likelihood ratio testing. However, the diagnostic plots indicate that model 4 fits the data better, which is in line with the variogram where the random effect rather than the serial correlation dominates the within-subject correlation.

Note that our research is ancillary to the assessment of treatment effect. Indeed, by first considering an appropriate mean structure, one can concentrate on the variability structure, thus enabling the use of clinical trial data to study reliability.

4. Discussion

A body of research exists on reliability, especially in psychology and educational sciences. In the past decades the topic is also entered the field of health sciences and especially the psychiatric health sciences because of the inherent subjectivity of the measures employed in this field. Test–retest reliability as one of the classical approaches typically deals with the problem of time: how to disentangle the measurement error from real fluctuations in what you are measuring.

Wiley and Wiley were among the first authors to deal with this problem by assuming a linear relationship between two adjacent measurements [25]. In this way also reliability will have different values at both moments of measurement. Tisak and Tisak [26] also stressed the fact that reliability is not a fixed property of an instrument but changes with time. They proposed a method to calculate a time function of reliability. Dunn [5] describes a method that uses components of variance in the calculation of reliability. He further extends this method to a mixed model to deal with rater effects by taking the rater into the model as a random effect. The mixed model methodology indeed allows a study of variance components and fixed effects simultaneously. The variance–covariance structure is typically decomposed further into three components: (1) measurement error (process with memory 0), (2) serial correlation (process with finite memory), and (3) random effects (accommodating hierarchies, infinite memory process). Such hierarchies arise due to repeated measurements over time. Other hierarchies could be accommodated as well. Indeed, even in our current work, hierarchy arises due to the fact that data come from five trials. A proper account of this calls for the incorporation of (meta-analytic and

other) hierarchies into our modeling strategies. Some work exists to this effect and is known as generalizibility theory [27]. The combination of this work with ours is the subject of ongoing research.

While, for this reliability study, we are primarily interested in the variance components, mixed-model methodology provides an interesting opportunity to model the fixed effects as well. We do not have to make the unrealistic assumption that there is no change in a patient's situation over time or with treatment. Instead, such changes can be incorporated into the model.

When using repeated measurements a third source of variation can be taken into account when calculating reliability, the so-called "serial correlation". In this work, a method has been proposed that allows for serial correlation in the calculation of test–retest reliability, as well as random effects and measurement error.

The method was applied to the PANSS, a psychiatric rating scale for schizophrenia. Several models were applied: model 1 resulted in an overall test–retest reliability coefficient, averaging reliability across the 8 weeks, models 2–4 allowed us to study the test–retest reliability as a function of time. We observed a gradual decrease of reliability with increasing time lag between measurements. As mentioned earlier, there are different possible scenarios to explain such effects, such as memory effect of the raters or other covariates that are not taken into account in the model. For the PANSS scale we obtained reliability estimates from almost 0.90 to 0.50. Up to a time interval of 5 weeks, the reliability does not go below 0.60, which is considerable. Another result that occurs quite consistently is a slight increase in the reliability measure as time goes by but for a fixed time lag. The reason for this is most likely a learning effect in the raters. In a different setting, one might also encounter learning effects in the study subjects. Of course, other perhaps complementary explanations cannot be excluded.

The present method stresses once again the fact that 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, such as country or rater, is therefore an interesting topic for further research on the present data. In psychometric theory, this is referred to as "generalizability theory" [27].

A further important advantage of the present method is that it becomes possible to estimate trial-specific or population-specific reliability in clinical studies. 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. Finally, when measurement sequences on a subset of respondents are incomplete, these data can still be used for analysis, unlike in the classical approaches. In our case, we have focused on population-level reliability. Should we calculate them trial-specific via model 1, we would obtain the values 0.72, 0.69, 0.72, 0.71, and 0.59.

While it seems variability in reliability over time could be ascribed to variability in study duration, we are protected against such spurious effects by the use of a likelihood framework, where shorter and longer sequences contribute to estimates at any time point, as in the missing data literature [28]. Further, the strength of our methodology is that a proper time variable can be included into the mean and variance model, allowing us to combine studies of variable length.

Of course, some of the fluctuation observed in reliability estimates may be due purely to random noise, due to limited sample sizes. A clear perspective on this can be obtained by calculating interval estimates, which can also be used to assess appropriate sample sizes.

When clinical trials are designed, only validated scales should be used. Therefore, validation should always happen before clinical trials are started. This should not prevent the statistician, however, from

studying how well the scale actually performed during the trial: was the test–retest reliability indeed as predicted? Most often, the only focus is estimating treatment effect, taking into account the observed variance, without investigating the latter. This is also a missed opportunity to increase knowledge about the scale: often the number of subjects and observations in studies designed to validate scales are rather low while the information coming from clinical trials can be very rich.

Acknowledgements

The second author was supported in part by the Minimal Psychiatric Data Registry of the Belgian Federal Ministry of Public Health and Social Affairs. The third author was supported by the "Fonds voor Wetenschappelijk Onderzoek (FWO) Vlaanderen," Belgium. In addition, the authors are grateful to Johnson & Johnson Pharmaceutical Research and Development 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

[1] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing interater reliability. Psychol Bull 1979;86:420–8.
[2] Stratford P. Consistency or differentiating among subjects? Phys Ther 1989;69:299–300.
[3] Fleiss JL. Design and analysis of clinical experiments. Wiley: New York; 1986.
[4] Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966;19:3–11.
[5] Dunn G. Design and analysis of reliability studies: the statistical evaluation of measurement errors. Edward Arnold: London; 1989.
[6] Laird NM, Ware JH. Random effects models for longitudinal data. Biometrics 1982;38:963–74.
[7] Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. Springer: New York; 2000.
[8] Searle SR, Casella G, McCulloch CE. Variance components. Wiley: New York; 1992.
[9] Cook RD. Detection of influential observations in linear regression. Technometrics 1977;19:15–8.
[10] Cook RD. Assessment of local influence. J R Stat Soc, Ser B 1986;48:133–69.
[11] Lesaffre E, Verbeke G. Local influence in linear mixed models. Biometrics 1998;54:570–82.
[12] Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. Schizophr Bull 1987;13:261–76.
[13] Kay SR, Opler LA, Lindenmayer JP. Reliability and validity of the Positive and Negative Syndrome Scale for schizophrenics. Psychiatr Res 1988;23:99–110.
[14] Bell M, Milstein R, Beam-Goulet J, Lysaker P, Cicchetti D. The Positive and Negative Syndrome Scale and the Brief Psychiatric Rating Scale: reliability, comparability, and predictive validity. J Nerv Ment Dis 1992;180:723–8.
[15] Peralta V, Cuesta MJ. Psychometric properties of the Positive and Negative Syndrome Scale (PANSS) in schizophrenia. Psychiatr Res 1994;53:31–40.
[16] Peuskens J, the Risperidone Study Group. Risperidone in the treatment of chronic schizophrenic patients: a multinational, multicentre, double-blind, parallel-group study versus haloperidol. Br J Psychiatry 1995;166:712–26.
[17] Chouinard G, Jones B, Remington G. A Canadian multicenter placebo-controlled study of fixed doses of risperidone and haloperidol in the treatment of chronic schizophrenic patients. J Clin Psychopharmacol 1993;13:25–40.
[18] Marder SR, Meibach RC. Risperidone in the treatment of schizophrenia. Am J Psychiatry 1994;151:825–35.
[19] Hoyberg OJ, Fensbo C, Remvig J, et al. Risperidone versus perphenazine in the treatment of chronic schizophrenic patients with acute exacerbations. Acta Psychiatr Scand 1993;88:395–402.
[20] Blin O, Azorin JM, Bouhours P. Antipsychotic and anxiolytic properties of risperidone, haloperidol and methotrimeprazine in schizophrenic patients. J Clin Psychopharmacol 1996;16:38–44.

[21] Huttunen MO, Piepponen T, Rantanen H, et al. Risperidone versus zuclopenthixol in the treatment of acute schizophrenic episodes: a double-blind parallel-group trial. Acta Psychiatr Scand 1995;91:271–7.
[22] Littell RC, Milliken GA, Stroup WW, Wolfinger RD. SAS system for mixed models. SAS Institute: Cary, SP; 1996.
[23] Diggle PJ, Liang K-Y, Zeger SL. Analysis of longitudinal data. Clarendon Press: Oxford; 1994.
[24] Streiner DL, Norman GR. Health measurement scales. Oxford Univ Press: Oxford; 1995.
[25] Wiley DE, Wiley JA. The estimation of measurement error in panel data. Am Sociol Rev 1970;35:112–7.
[26] Tisak J, Tisak MS. Longitudinal models of reliability and validity: a latent curve approach. Appl Psychol Meas 1996;20:275–88.
[27] Cronbach LJ, Rajaratnam N, Gleser GC. Theory of generalizability: a liberalization of reliability theory. Br J Stat Psychol 1963;16:137–63.
[28] Little RJA, Rubin DB. Statistical analysis with missing data. Wiley: New York; 1987.