MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.1787 Published online in Wiley Online Library

Estimating the reliability of repeatedly measured endpoints based on linear mixed-effects models. A tutorial

Wim Van der Elst,[a*] Geert Molenberghs,[a,b] Ralf-Dieter Hilgers,[c] Geert Verbeke,[a,b] and Nicole Heussen[c]

There are various settings in which researchers are interested in the assessment of the correlation between repeated measurements that are taken *within* the same subject (i.e., reliability). For example, the same rating scale may be used to assess the symptom severity of the same patients by multiple physicians, or the same outcome may be measured repeatedly over time in the same patients.

Reliability can be estimated in various ways, for example, using the classical Pearson correlation or the intra-class correlation in clustered data. However, contemporary data often have a complex structure that goes well beyond the restrictive assumptions that are needed with the more conventional methods to estimate reliability.

In the current paper, we propose a general and flexible modeling approach that allows for the derivation of reliability estimates, standard errors, and confidence intervals – appropriately taking hierarchies and covariates in the data into account. Our methodology is developed for continuous outcomes together with covariates of an arbitrary type.

The methodology is illustrated in a case study, and a Web Appendix is provided which details the computations using the R package *CorrMixed* and the SAS software.

Keywords: within-cluster correlation; test-retest reliability; intra-class correlation

1. INTRODUCTION

Reliability essentially refers to the reproducibility (or, predictability) of outcomes that are repeatedly measured *within* the same individuals. In particular, this metric quantifies the extent to which a repetition of a measurement under the same general conditions leads to the same result.

Conventional methods to estimate reliability. The concept of reliability is grounded in the so-called classical test theory [1]. In this paradigm, the outcome of a measurement procedure is modeled as $X = \tau + \varepsilon$, where X is the observed score of a subject, τ is the unobserved (latent) true score of this person, and ε is the measurement error. In classical test theory, it is assumed (i) that the measurement errors are mutually uncorrelated, and (ii) that the measurement errors are uncorrelated with the true scores. Under these assumptions, $\text{Var}(X) = \text{Var}(\tau) + \text{Var}(\varepsilon)$ and the reliability of the measurement (R) is defined as

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

Equation (1) is intuitively appealing because it defines reliability as the fraction of the observed test score variance that is attributable to the true score variance. If a test is perfectly reliable, the true score and observed score variances are equal, and thus $R = 1$. Unfortunately, reliability cannot be directly estimated based on Eq. (1) because τ cannot be observed. Instead, reliability will have to be estimated indirectly. A classical solution to the problem is to introduce the concept of *parallel tests* [2]. Parallel tests are tests that have the same true score for each subject and equal error variances. For example, suppose that we have two measurements X_1 and X_2 for the same subjects that are assessed at two instances of time with a short lag (such that τ does not change), or that are obtained from two raters at the same point in time. Then $X_1 = \tau + \varepsilon_1$ and $X_2 = \tau + \varepsilon_2$ with $\text{Var}(X_1) = \text{Var}(X_2) = \text{Var}(X)$ and $\text{Var}(\varepsilon_1) = \text{Var}(\varepsilon_2) = \text{Var}(\varepsilon)$, that is, X_1 and X_2 are parallel measurements. The covariance of the two measurements then equals

$$\begin{aligned}\text{Cov}\,(X_1,\, X_2) &= \text{Cov}(\tau + \varepsilon_1,\, \tau + \varepsilon_2)\\ &= \text{Var}(\tau) + \text{Cov}\,(\tau, \varepsilon_1) + \text{Cov}\,(\tau,\, \varepsilon_2) + \text{Cov}\,(\varepsilon_1, \varepsilon_2)\\ &= \text{Var}(\tau),\end{aligned}$$

[a] I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium

[b] I-BioStat, Katholieke Universiteit Leuven, Leuven, Belgium

[c] Department of Medical Statistics, RWTH Aachen University, Aachen, Germany

*Correspondence to: Wim Van der Elst, I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium.
E-mail: wim.vanderelst@gmail.com

and the correlation between X_1 and X_2 can be written as

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

Limitations of the conventional methods. Equation (2) provides a convenient and straightforward way to compute reliability, but it is important to stress that the assumption that the measurements are parallel is crucial. This assumption is often violated in practice [3]. For example, it seems implausible to assume that patients in a clinical trial or in medical practice do not exhibit a systematic change over time as a result of their treatment. Another limitation of Eq. (2) is that only two measurements can be considered, and these measurements should have the same test–retest interval for all subjects. In practice, data may be available for more than two measurement moments and/or with different test–retest intervals. Further, the use of Eq. (2) is less-than-ideal when data are missing, because subjects who have a missing observation for either X_1 or X_2 are discarded from the analysis. This approach does not only lead to a loss of information, but it also ignores the missing data generating mechanism. Basically, to obtain unbiased estimates for R using Eq. (2), the assumption that the data are missing completely at random should be valid. This means that the missingness should not depend on the observed or the unobserved outcomes [4,5]. This is a strong and often unrealistic assumption, for example, in a clinical trial setting, it is conceivable that subjects who have lower scores at the first measurement in time (poorer health) are more likely to drop out of the study at the second measurement in time (missing value for X_2).

Importance of reliability. It is important to carefully consider the reliability of a measurement procedure, for example in the context of designing a clinical trial. Obviously, in particular in explorative or experimental small population group studies, serial measurements are gathered to understand the nature of the disease. However, unreliable measurement methods might lead to serious misinterpretation of the disease process. Indeed, even the most elegant study design will not overcome the damage that is caused by the use of unreliable measurement procedures [6]. For example, biased sample selection may occur when patients are selected based on an unreliable measurement procedure, and the sample size that is required to detect an important treatment difference (δ) may increase substantially when the outcome of interest is quantified using an unreliable measurement procedure. As an illustration of the latter issue, consider a situation where a t-test is used to evaluate the treatment effect on the primary endpoint in a clinical trial with two treatment groups. When the measurement procedure that is used to quantify the primary endpoint has perfect reliability (i.e., $R = 1$), the required sample size to detect δ equals n^*. However, when this measurement procedure has a less-than-perfect reliability (i.e., $R < 1$), the required sample size becomes $n = \frac{n^*}{R}$ (for details, [6]). Thus, for example, when $R = 0.50$, the required sample size to detect δ *doubles* compared with what would have been needed when $R = 1$. Clearly, an increase in the required sample size is an issue in nearly all clinical studies (e.g., increased study duration and cost) – and it may even make the conduct of the study infeasible (e.g., clinical trials in rare diseases).

Aim and organization of the paper The main aim of the present paper is to illustrate how reliability can be estimated in a flexible way using linear mixed-effects models (LMMs). As will be detailed below, LMMs can separate the mean and the variance structures in the data – which allows for relaxing the strong assumptions that are needed to apply the conventional methods to estimate reliability. Further, LMMs can deal with data structures where different subjects have a different number of repeated measurements (2 or more) – which may or may not be regularly spaced. Finally, LMMs are so-called likelihood-based methods that provide valid results when the missingness mechanism is missing at random (MAR) [7]. MAR means that the missingness may depend on the observed outcomes (e.g., the first measurement X_1) but not on unobserved outcomes. MAR is a substantially less restrictive assumption than missing completely at random,and is thus more likely to hold in practice [4].

The remainder of the paper is organized in the following way. In Section 2, a case study is introduced that will be used throughout this paper to illustrate the methodology. In Section 3, an exploratory analysis of the case study is conducted. In Section 4, the LMM-based approach to estimate reliability is detailed. Section 5 discusses the results. A Web Appendix is also provided in which additional materials are presented. In particular, it details all the required computations using the newly developed *R* software package *CorrMixed* and SAS.

2. CASE STUDY

Pikkemaat *et al.* [8] performed an experiment where the cardiac output and stroke volume of $N = 14$ pigs was changed by increasing positive end-expiratory pressure (PEEP) levels (0, 5, 10, 15, 20, and 25 cm H_2O). The number of times that a particular PEEP level was used varied from animal to animal. For each PEEP level, stroke volume was measured by the continuous approximately normally distributed variable electrical impedance tomography (EIT)-based stroke volume (SV)-related signal. In each animal, four identical experiments were conducted (referred to as Cycles 1 to 4). The number of repeated ZSV measurements across PEEP levels and cycles in an animal ranged between 9 and 47. In the analyses in the succeeding text, it is assumed that all the measurements are equally spaced.

Pikkemaat *et al.* [8] were interested in estimating the levels of association between the repeatedly measured ZSV and transpulmonary thermodilution outcomes within an animal. As detailed in Section 1, it is also worthwhile to evaluate the reliability of these repeated measurements. Such analyses (not considered in [8]) will be the focus of the current paper. Given the complex design of the study, it is recommended to use a flexible LMM-based technique to estimate reliability (Section 4) – rather than the conventional techniques that were discussed in the Introduction.

As noted earlier, the study included a total of 14 pigs. However, the data of $n = 2$ animals could not be evaluated because of technical reasons, and these animals were thus excluded from the analyses. Further, there were $n = 2$ animals who appeared to have a 'clinically deviating' profile (as judged by the experimenters). These animals were kept in the current analyses, but a sensitivity analysis showed that the estimated reliabilities were not substantially affected by the inclusion or exclusion of these animals (see Web Appendix Part II). Note that the data for PEEP level 25 were included in the current analysis, as well as in the Pikkemaat *et al.* [8] study, although they were not explicitly mentioned in the latter.

3. EXPLORATORY DATA ANALYSIS

Figure 1 shows the individual profiles (gray lines) of ZSV as a function of measurement moment. As can be seen, there is substantial between – as well as within – animal variability. Further, drop-out is substantial, that is, there are less observations at later measurement moments compared with earlier measurement moments. This is more clearly depicted in Figure 2, where the number of available observations at each of the different measurement moments are shown.

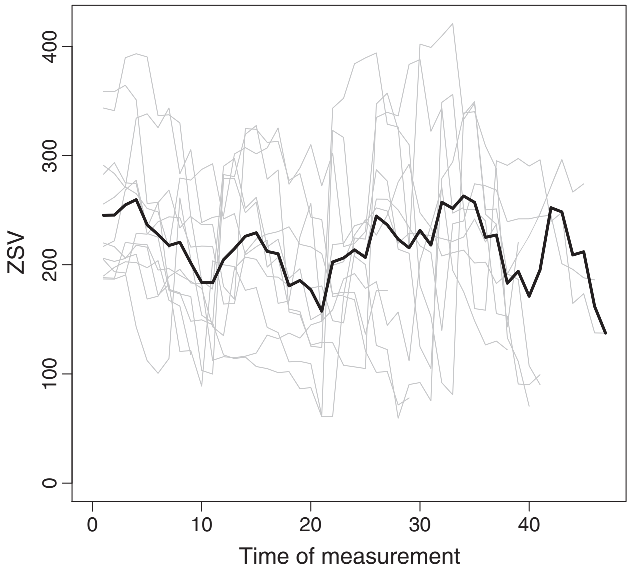

Figure 1. Individual profiles (gray lines) and mean values (black line) of the zero shear viscosity (ZSV) outcome as a function of time of measurement.

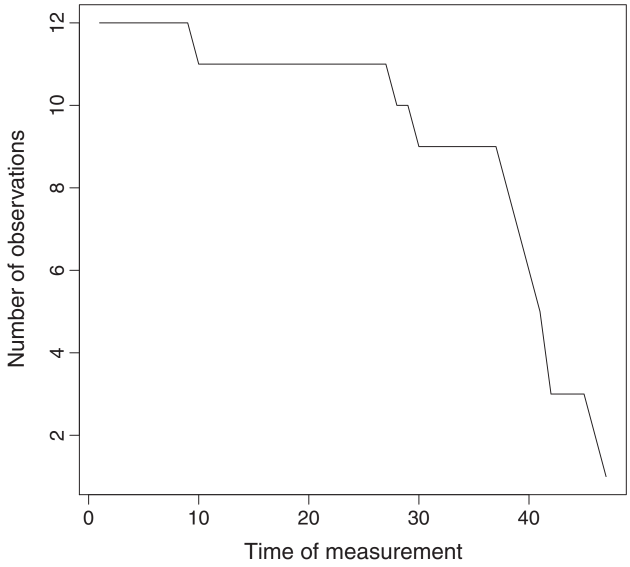

Figure 2. Number of observations for the zero shear viscosity outcome as a function of time of measurement.

Figure 1 also shows that the average evolution over time (solid black line) exhibits a rather complex shape that cannot be modeled in a straightforward way by using linear or quadratic polynomials. Therefore, it is useful to consider a more general family of parametric models that are based on so-called fractional polynomial functions [9].

Fractional polynomials. The idea is to fit regression models with m terms of the form t^p, where the exponents p are selected from a small predefined set S of both integer and non-integer values. The linear predictor for a fractional polynomial of order M for covariate t (here: measurement point in time) on the mean ZSV is then defined as

$$\beta_0 + \sum_{m=1}^{M} \beta_m t^{p_m}. \tag{3}$$

Each power p_m is chosen from a restricted set, typically $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Note that when $M = 2$ and $p_1 = p_2$, the linear predictor (3) becomes $\beta_0 + \beta_1 t^{p_1} + \beta_2 t^{p_1} \log(t)$. Also, when $p = 0$, this is taken to refer to $\log(t)$ [9]. In practice, all possible models of degree 1 to M are fitted. Thus, for $M = 1$, each of the eight values of S are used for the predictor t^{p_1}, for $M = 2$ each of the 36 combinations of powers are used for the predictors t^{p_1} and t^{p_2}, and so on. Subsequently, the 'best' fitting model is selected. This choice can be made in an informal way (i) based on Akaike's Information Criterion (AIC, where a lower value is indicative of a better model fit) and/or (ii) by graphically evaluating the fit of the model with the observed data. The AIC adds the number of model parameters as a penalty to the log likelihood of the model, which may help to avoid over-fitting (even though one still may want to be careful not to select an overly complex model, in particular when a large number of candidate powers is considered). The main advantage of using fractional polynomials (rather than regular polynomials) is that they allow for a much more flexible parametrization, that is, a large number of different shapes of curves can be captured by even a relatively small M. *Application to the case study* In the analysis of the case study, fractional polynomials of order $M = 1$ to $M = 5$ were considered using the standard set $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ for the powers p_m. Note that it is possible to use a more extensive set of values for S if the original set does not provide an adequate result, but the number of models that have to be fitted (and thus also the required computational time) increases sharply when the number of elements in S increases. For example, when the set S includes eight elements (the standard set), a total of 792 fractional polynomials of degree 5 can be made. However, when the set $S = \{-3, -2.75, \ldots, 3\}$ is used (25 elements), a total of 118,755 fractional polynomials of degree 5 can be made. Similarly, M can be increased, but this will again yield a sharp increase in the number of models to be evaluated.

Thus, regression models that included linear predictors for fractional polynomials of order $M = 1$ to $M = 5$ (Eq. (3)) were fitted to the data of the case study. Table I shows the powers p_m of the models of order 1 to 5 that had the lowest AIC values. As can be seen, the model with $M = 3$ had the lowest overall AIC value. Figure 3 shows the predicted mean ZSV as a function of measurement moment for this model.

Table I. Fractional polynomial results.

M	power p_m	AIC
1	−0.5	3788.703
2	0.5, 0.5	3786.096
3	2, 2, 3	3775.281
4	0.5, 1, 2, 2	3776.389
5	−2, −2, 0, 2, 3	3778.221

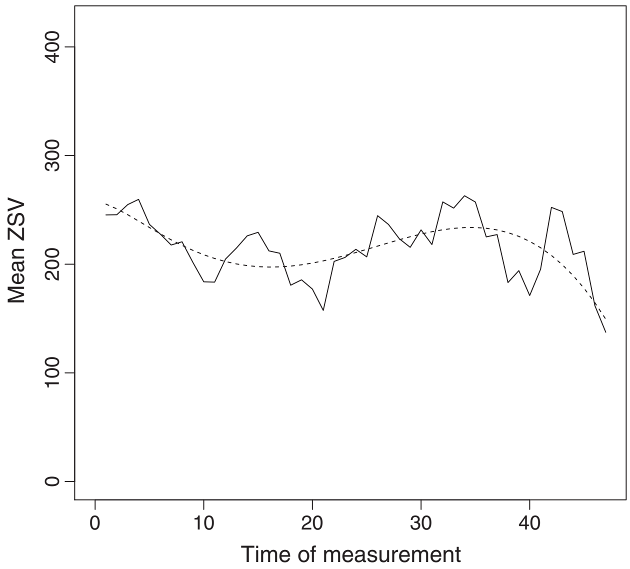

Figure 3. Observed means as a function of time of measurement (solid line) and fitted fractional polynomial of degree $m = 3$ (dashed line). ZSV, zero shear viscosity.

Based on these results, the fractional polynomial of degree 3 was retained as the 'best' model for the subsequent analyses. Thus, in the LMM analyses detailed below, the relation between time of measurement t and the mean ZSV will be modeled as $\beta_1 t^2 + \beta_2 t^2 \log(t) + \beta_3 t^3$.

4. ESTIMATING RELIABILITY USING MIXED-EFFECTS MODELS

In this section, the reliability of the ZSV will be estimated using a flexible approach that is based on LMMs. The LMM is briefly introduced in Section 4.1 (for more details, e.g., [7,10,11]), and the LMM-based approach to estimate reliability is applied to the case study in Section 4.2. For conciseness, in the latter section, only a summary of the main results is given, and no reference to software tools that can be used to obtain the results is made. However, full details can be found in the Web Appendix Parts I–V.

4.1. The linear mixed-effects model

A LMM can be written as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \tag{4}$$

where $\mathbf{Y}_i$ is the response vector for subject i (with $i = 1, 2, \ldots, n$ subjects in the study), $\mathbf{X}_i$ and $\mathbf{Z}_i$ are the known design matrices for the fixed and random effects, $\boldsymbol{\beta}$ is the vector that contains the fixed effects, $\mathbf{b}_i$ is the vector that contains the random effects, and $\boldsymbol{\varepsilon}_i$ is the vector that contains the measurement error (with $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$ and $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\mathbf{D}$ and $\boldsymbol{\Sigma}_i$ are general variance–covariance matrices). Model (4) thus assumes that the vector of repeated measurements for each subject follows a linear regression model where some of the parameters are population-specific (i.e., parameters that are the same for all subjects in the population; the fixed effects) and other parameters are subject-specific (i.e., parameters that differ for all subjects; the random effects).

The residual component $\boldsymbol{\varepsilon}_i$ is often further decomposed as $\boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}_{(1)i} + \boldsymbol{\varepsilon}_{(2)i}$. Here, $\boldsymbol{\varepsilon}_{(2)i}$ is a component of serial correlation and $\boldsymbol{\varepsilon}_{(1)i}$ is a component of measurement error. Serial correlation results from the fact that within a subject, the (residuals of) observations that are closer in time are often 'more similar' (i.e., more strongly correlated) than observations that are more distant in time. It is assumed that $\boldsymbol{\varepsilon}_{(1)i} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}\right)$ (with $\mathbf{I}_{n_i}$ an identity matrix of dimension n_i = the number of repeated measurements in a subject) and $\boldsymbol{\varepsilon}_{(2)i} \sim N\left(\mathbf{0}, \tau^2\mathbf{H}_i\right)$ (with $\mathbf{H}_i$ the serial correlation matrix that only depends on i through the number of repeated measurements n_i and the time points j and k at which the measurements are taken). The (j, k) element h_{ijk} of $\mathbf{H}_i$ can then be modeled as $h_{ijk} = g(\mid t_{ij} - t_{ik} \mid)$ for a decreasing function g. Two frequently used functions are the exponential and Gaussian correlation functions, defined as $g(u_{j,k}) = \exp(-\phi u_{j,k})$ and $g(u_{j,k}) = \exp\left(-\phi u_{j,k}^2\right)$, respectively.

4.2. Case study analysis

The mean structure of the model. The LMMs that will be fitted to the case study dataset include an intercept, measurement moment, PEEP, and Cycle as fixed effects. PEEP and Cycle are dummy-coded with five and three dummies, respectively. The relation between measurement point and the ZSV outcome is modeled as $\beta_1 t^3 + \beta_2 t^2 + \beta_3 t^2 \log(t)$ (check the Fractional polynomial section).

The covariance (correlation) structure of the model. In the analyses in the succeeding text, three LMMs with the same fixed-effect structure (previous paragraph) but different variance structures will be considered.

Model 1 is a random intercept model, that is, a LMM that only contains a random intercept in the random part of the model:

$$Y_{ij} = \mu_{ij} + b_{0i} + \varepsilon_{ij}, \tag{5}$$

where Y_{ij} is the observed endpoint at measurement time j for subject i, μ_{ij} is the mean as a function of the fixed effects, b_{0i} is the random intercept, and ε_{ij} is the residual. Based on this model, the reliability of the repeated observations taken at measurement times t_k and t_j can be estimated as (for details, [12]):

$$R(t_j, t_k) = R = \frac{d}{d + \sigma^2}, \tag{6}$$

where d is the variance of the random intercept and σ^2 is the residual variance. As can be seen in Eq. (6), the random intercept model assumes that any two observations measured at different times have the same R. This assumption is often not realistic when repeated measures are considered, that is, measurements that are closer in time can be expected to be more strongly correlated than measurements that are more distant in time.

Table II. Summary of the covariance structures used in Models 1–3, and the impact on the estimated reliabilities.

Model	Estimated reliabilities R
Model 1: Random Intercept	$\hat{R}$ is identical for all pairs $(t_j,\ t_k)$
Model 2: Random intercept and serial component	$\hat{R}$ only depends on the time lag $u_{jk} = t_k - t_j$
Model 3: Random intercept, slope, and serial component	$\hat{R}$ is different for all pairs $(t_j,\ t_k)$

Note. t_j = measurement at time j.

Therefore, Model 2 extends Model 1 by adding a serial correlation component:

$$Y_{ij} = \mu_{ij} + b_{0i} + \varepsilon_{(1)ij} + \varepsilon_{(2)ij}, \tag{7}$$

where μ_{ij}, b_{0i} are the same as in Model 1 and $\varepsilon_{(1)ij}, \varepsilon_{(2)ij}$ are measurement error and serial correlation components, respectively. Based on Model 2, the reliability of the repeated observations taken at measurement times t_k and t_j can be estimated as (for details, [12]):

$$R(t_j,\ t_k) = R(u_{jk}) = \frac{d + \tau^2 \exp\left(\frac{-u_{jk}^2}{\rho^2}\right)}{d + \tau^2 + \sigma^2}, \tag{8}$$

where $u_{jk} = t_k - t_j$, $\sigma^2 = \text{Var}(\varepsilon_{(1)i})$ and $\tau^2 = \text{Var}(\varepsilon_{(2)i})$. Model 2 thus no longer assumes that R remains constant for all pairs of measurements. Instead, it models R as a function of the time lag u_{jk} between two measurements. As can be seen, a stronger serial effect (ρ^2) leads to a faster decreasing $R(u_{jk})$.

Finally, Model 3 further extends Model 2 by including a random slope for measurement moment:

$$Y_{ij} = \mu_{ij} + b_{0i} + b_{1i}t_j + \varepsilon_{(1)ij} + \varepsilon_{(2)ij}, \tag{9}$$

where $\mu_{ij}, b_{0i}, \varepsilon_{(1)ij}, \varepsilon_{(2)ij}$ are the same as in Models 1 and 2, and b_{1i} is the random slope for measurement moment. Based on Model 3, the reliability of the repeated observations measured at times t_k and t_j can be estimated as (for details, [12]):

$$R(t_j,\ t_k) = \frac{\mathbf{z}_j\mathbf{D}\mathbf{z}_k' + \tau^2 \exp\left(\frac{-u_{jk}^2}{\rho^2}\right)}{\sqrt{\mathbf{z}_j\mathbf{D}\mathbf{z}_j' + \tau^2 + \sigma^2}\sqrt{\mathbf{z}_k\mathbf{D}\mathbf{z}_k' + \tau^2 + \sigma^2}}, \tag{10}$$

where $u_{jk} = t_k - t_j$, and $\mathbf{z}_j$, $\mathbf{z}_k$ are the design rows in $\mathbf{Z}$ corresponding to time j and k, respectively. As can be seen in Eq. (10), Model 3 no longer assumes that measurements taken at different time points, but with the same time lag have the same R. Instead, it provides estimates of reliability for all pairs of measurements.

Table II summarizes the covariance structures that are used in the different models and their impact on the estimated R.

4.2.1. Model 1: random intercept model. When Model 1 was fitted to the case study dataset, it was obtained that $\hat{d} = 1901.611$ and $\hat{\sigma}^2 = 2413.022$, yielding $\hat{R} = 0.441$ (Eq. (6)). A CI around $\hat{R}$ can be computed by using a (non-parametric) bootstrap or the Delta method (for details, see the Web Appendix Part VI). The bootstrap-based 95% CI (using 500 bootstrap samples) equaled [0.198; 0.618]. The Delta method-based CI was similar and largely overlapped, that is, [0.189; 0.636]. Figure 4 (top left) illustrates the results (the bootstrap-based CI is shown).

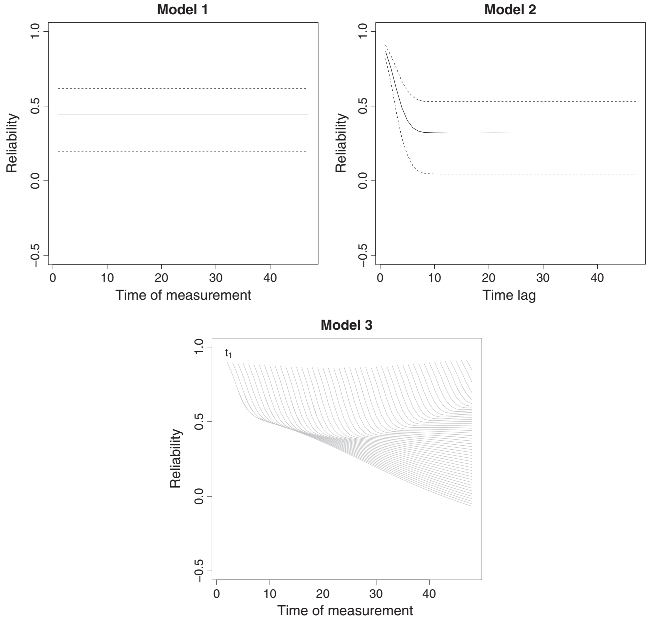

Figure 4. Estimated reliabilities (solid lines) and 95% Confidence Intervals (dashed lines) for zero shear viscosity based on Model 1 (upper left), Model 2 (upper right) and Model 3 (bottom). For Model 3, no Confidence Intervals are provided to avoid a cluttered figure. The utmost left line marked with t_1 depicts the estimated correlations between t_1 and all other measurements, the line next to that one depicts the correlations between t_2 and measurements 2–45, and so on.

Overall, it can be concluded that $\hat{R}$ is moderate and that there is substantial uncertainty in $\hat{R}$ (which is not surprising given the small number of animals in the study).

4.2.2. Model 2: random intercept and serial correlation. When Model 2 was fitted to the data of the case study, the estimated covariance parameters were $\hat{d} = 1349.650$, $\hat{\tau}^2 = 2489.351$, $\hat{\rho} = 3.581$, and $\hat{\sigma}^2 = 382.795$. Thus, after correction for the fixed effects, the covariance parameter estimates showed considerable remaining serial components.

Figure 4 (top right) shows the estimated $R(u_{jk})$ (Eq. (8)) and their 95% CIs based on a bootstrap (the Delta method-based CIs were similar; data are shown in the Web Appendix Part I). As can be seen, the estimated R were high for small time lags (e.g., $\hat{R}(u_{jk} = 0) = 0.865$ and $\hat{R}(u_{jk} = 1) = 0.751$) and subsequently decreased until they remained essentially constant at $\hat{R} \approx 0.320$ for measurements with time lags of about $u_{jk} = 10$ and higher. It can also be observed that the CIs around $\hat{R}(u_{jk})$ were narrower for measurements with smaller time lags (e.g., for time lags $u_{jk} = 0$ and $u_{jk} = 1$, the $CI_{95\%} = [0.817,\ 0.906]$ and $CI_{95\%} = [0.654,\ 0.836]$, respectively) and subsequently widened until they remained stable around time lag $u_{jk} = 10$ with $CI_{95\%} = [0.045,\ 0.530]$.

4.2.3. Model 3: random intercept, slope, and serial correlation. When Model 3 was fitted to the data of the case study, the estimated covariance parameters were $\hat{\tau}^2 = 1952.970, \hat{\rho} =$

3.290, $\hat{\sigma}^2 = 373.043$, and

$$\hat{\mathbf{D}} = \begin{pmatrix} 3219.869 & -77.377 \\ -77.377 & 3.686 \end{pmatrix}.$$

As noted earlier, based on Model 3 the estimated $R(t_k,\ t_j)$ are different for all pairs of measurements (Eq. (10)). Figure 4 (bottom) shows the results graphically. In this figure, the utmost left line (marked with t_1) depicts the estimated $R(t_1,\ t_j)$, that is, the estimated reliabilities of ZSV taken at measurement times 1 and 2–45. The line next to that one shows the estimated $R(t_2,\ t_j)$, etc. The figure shows that $\hat{R}(t_k,\ t_j)$ is high when the time lag u is small and flattens out for longer time lags. Further, depending on the particular pair of measurement moments $(t_k,\ t_j)$ that is considered, the slope and amount of decline in $\hat{R}(t_k,\ t_j)$ as a function of time lag differs. For example, when considering $\hat{R}(t_1,\ t_j)$, it can be seen that the estimated reliabilities decline particularly strong for the first few subsequent measurements (say, until about t_8) and continue to decline for all t_j afterwards at a slower pace. Instead, for $\hat{R}(t_{20},\ t_j)$, there is only a substantial decline in the estimated reliabilities for the first few subsequent measurements (say, until about t_{25}) after which the estimated reliabilities remain essentially constant.

Based on Model 3, estimates of reliability are provided for each pair of measurements, and the same obviously holds for the CIs. To avoid cluttered figures, no CIs were added to Figure 4 (bottom). By means of illustration, Figure 5 provides 95% bootstrap-based CIs for $\hat{R}(t_1,\ t_j)$ (left) and $\hat{R}(t_{20},\ t_j)$ (right). As can be seen, the CIs increase as a function of time and tend to be wider for $\hat{R}(t_{20},\ t_j)$ than for $\hat{R}(t_1,\ t_j)$ (as expected).

4.2.4. Selecting the most appropriate model. Based on the likelihood ratio (LR) test statistic G^2, the fit of Models 1–3 can be formally compared (for details, [7]). G^2 is equal to -2 times the difference of the log likelihoods of the models being compared. Before discussing the results for the case study, some general remarks are useful. First, when interest is in testing the need for including random effects in the model, the usual procedure where the test statistic G^2 is compared with a χ^2 distribution with the number of degrees of freedom equal to the difference in the model parameters to be estimated is no longer valid. For example, consider the situation where interest is in testing whether one or two random effects are needed (Model 2 versus Model 3). This corresponds to testing that $d_{12} = d_{21} = d_{22} = 0$. To test this hypothesis, a *mixture* with equal weights 0.5 for χ^2_1 and χ^2_2 is needed (denoted by $\chi^2_{1:2}$), because the variance d_{22} cannot be negative and thus the hypothesis test of interest is on the boundary of the parameter space (for details, [7]). Second, the results of the LR tests should be interpreted with caution because of the small sample size in the case study. Alternative testing procedures that are based on permutation tests (e.g., [13]) could provide a more viable alternative, but these methods are beyond the scope of the present paper. Third, the valid use of LR tests typically requires that the models are fitted using Maximum Likelihood estimation. The results provided earlier used Restricted Maximum Likelihood (REML), but valid LR tests for comparing nested models with different covariance structures can still be obtained under REML estimation when the models that are compared have the same mean structure [7] – which was the case here, as discussed earlier.

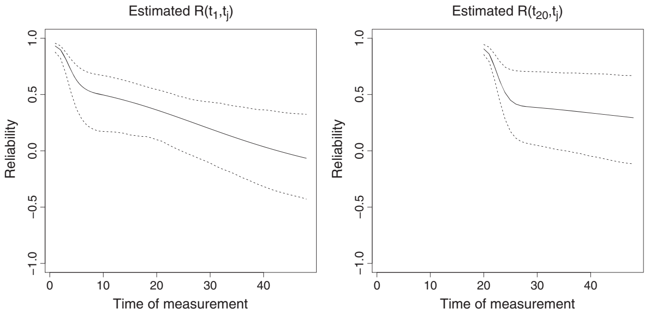

Figure 5. $\hat{R}(t_1,\ t_j)$ (left) and $\hat{R}(t_{20},\ t_j)$ (right) based on Model 3 and their 95% Confidence Intervals for the zero shear viscosity outcome.

The log likelihood values for Models 1–3 are shown in Table III. As can be seen, the random intercept model with serial correlation (Model 2) fitted the data significantly better than the random intercept model with no serial correlation (Model 1), $p < 0.001$. This test thus rejects the null hypothesis that there is no serial correlation process, that is, it can be concluded that observations that are closer in time are stronger correlated than observations that are more distant in time. Further, adding a random slope to the random intercept model with serial correlation (Model 3 versus Model 2) significantly improves the model fit, $p = 0.015$ – though the gain was quite modest.

Model 3 is the model with the largest likelihood. It would be preferred if we would solely rely on statistical arguments. However, from an applied perspective – that is, also considering the practical usefulness of the results for a clinician or researcher – Model 2 is arguably to be preferred over Model 3 because the former leads to reliability estimates that only depend on the time lag between two measurements. In contrast, Model 3 yields different reliability estimates for all possible pairs of measurements. Model 2 thus provides a much more parsimonious result compared with Model 3 – while the fit of both models is roughly comparable. Notice that the likelihood ratio tests identify the best fitting model among the models that were under consideration. However, when a model has been selected, the question remains

Table III. Fit indices of the different models for the ZSV outcome.

	# Pars.					
	Rand.	Ser.	logL	G^2	Test	p
Model 1	1	0	−2328.910			
Model 2	1	2	−2125.135	407.551	Model 2 vs. 1: χ^2_2	< 0.001
Model 3	3	2	−2121.399	7.472	Model 3 vs. 2: $\chi^2_{1:2}$	0.015

Note. logL = log likelihood, $G^2 = -2$ the difference of two log likelihood values. Rand., random effect parameters; ser., serial components; ZSV, zero shear viscosity.

whether this model fits the data sufficiently well. Residuals and influence diagnostics are useful in this respect. In Part VII of the Supporting Information, a residual analysis is conducted and the extent to which particular animals exert a strong influence on the results (i.e., the REML distances of the models, the estimated fixed-effects parameters, the estimated covariance components, and the estimated reliability coefficients) is evaluated. Overall, the impact of excluding an animal on the results was relatively small for Models 2–3. For Model 1, the impact of deleting an animal on the results was more substantial. Further, the residual analysis showed that there were no major departures of normality.

5. DISCUSSION

The conventional methods to estimate reliability (e.g., the well-known Pearson correlation coefficient) require assumptions that are often not met in real-life studies (e.g., parallel measurements, equally spaced test-retest intervals, etc.). The main aim of the current paper was to present a general and flexible approach to estimate reliability that is based on LMMs. It was shown that this approach can be successfully applied even in a 'challenging' dataset like in the presented case study – where the number of independent subjects is low, different subjects have a different number of repeated observations, and several covariates have to be taken into account. Overall, the analysis of the case study suggested that the reliability of ZSV was high (and its CIs narrow) when the time lag was small. For larger time lags, the reliability estimates decreased and their CIs widened.

Some critical remarks are in place. First, despite the major differences between the conventional and the LMM-based methods to estimate reliability, there are also some obvious similarities. For example, the expressions to estimate reliability based on Model 1 (Eq. (6)) and the conventional approach (Eq. (2)) are very similar (i.e., both are ratios of variances). However, a fundamental difference between both methods is that the LMM-based approach does not require the parallel measurement assumption. The reason for this is that the mean and variance structures can be clearly separated in LMMs (check discussion earlier). For example, when the means at different time points are different (as was observed in the case study, Figure 1), systematic effects of time and other covariates can be taken into account by including them into the fixed-effect part of the model (as was done here). In essence, the main difference between the conventional and LMM-based approaches to estimate reliability is that the former requires a set of assumptions that are taken care of in the study design, whereas the latter takes care of these assumptions through modeling at the analysis stage [3]. There is however a price to pay for the increased flexibility of the LMM-based approach, that is, it requires substantially more complex statistical analyses compared with the conventional methods to estimate reliability. We tried to circumvent this issue by developing an R package (*CorrMixed*) that allows for obtaining reliability estimates based on Models 1–3 in a relatively straightforward way. The Web Appendix (Parts IV and V) provides full details on how the analyses can be conducted in practice.

Second, in the present paper, the focus was entirely on the random effect structure of the models because we were interested in estimating the reliability of the outcomes. Apart from estimating reliability, medical practitioners are also often interested in obtaining so-called normative data. Normative data are used to convert a patient's 'raw' outcomes into relative measures that reflect the proportion of demographically-matched healthy controls in the population who have a lower outcome value compared with this patient. A well-known example are growth curves of young children. Such normative data (nomograms) for repeated measurements can be obtained without any substantial additional effort using the same type of models that were fitted in the present paper. The only difference is that the focus will then be on the fixed-effect part of the model – rather than on the random effect structure (for details, [14]).

Third, the outcome that was considered in the case study was a normally distributed (Gaussian) variable. One may also be interested in estimating the reliability of repeated measurements of outcomes of a different distributional nature, for example, binary (yes/no, health/sick) or categorical ordered outcomes. Such extensions are possible, but not trivial. The interested reader is referred to Vangeneugden *et al.* [15].

Fourth, in the analysis of the case study, the fixed-effect structures were kept constant for Models 1 to 3 (because we were primarily interested in evaluating the impact of different random-effect structures on the estimated reliabilities). In the Web Appendix (Part III), a sensitivity analysis is conducted where the impact of using different plausible fixed-effect structures on the estimated reliabilities is evaluated. Overall, the analyses indicated that the estimated reliabilities are not sensitive to the fixed-effect part of the model (provided that the mean structure of the model is supported by the data).

Finally, in the present paper, no time-varying covariates (other than measurement occasion itself) were considered, but depending on the study at hand it may be useful to include such covariates. For example, consider a setting where one is interested in estimating the reliability of a psychiatric rating scale that was scored by different physicians at the different measurement moments. When only a limited number of raters are involved in the study, the methodology that was proposed earlier can still be used in a straightforward way. Indeed, one can then simply include rater as a (dummy-coded) fixed-effect in the mean structure of the model. On the other hand, when the number of raters is large, it is more sensible to include rater in the random-effect part of the model. Such a model cannot be fitted in the current version of the *CorrMixed* package, but it is straightforward to fit such a model using SAS.

On a related note, in the present paper, interest was primarily in the estimation of the reliability of a single outcome that was repeatedly measured within the same subject. It might also be of interest to estimate how strongly the vectors of two outcomes are correlated *with each other*. For example, consider a setting where two raters assess all patients at all measurement moments. Here, it would be natural to study the correlation between the vectors of scores to evaluate the level of agreement between the two raters. Or, as another example, consider a setting where there are two alternative measurement procedures for the same latent variable. When one of the two measurement procedures is more 'difficult' to conduct (e.g., is more expensive, more painful for the patient, requires more time to obtain the test results, etc), it may be of interest to estimate the correlation between the measurements obtained by both procedures. Indeed, when it can be shown that there is a high correlation between the vectors of outcomes, the 'easier' measurement procedure may replace the more difficult one – in the same spirit as is done when a surrogate endpoint is used to replace the true endpoint in a clinical trial (individual-level surrogacy; for details, [16]). The quantification of the correlation between two vectors of outcomes is however beyond the scope of the present paper, as different statistical techniques are needed to estimate this quantity (e.g., [17]).

Acknowedgements

Financial support from the IAP research network #P7/06 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged. This project has received funding from the European Union's 7th Framework Programme for research, technological development and demonstration under the IDEAL Grant Agreement no 602552.

REFERENCES

[1] Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Addison-Welsley Publishing Company, Reading, Massachusetts, 1968.

[2] Spearman C. The proof and measurement of association between two things. *The American Journal of Psychology* 1904; **15**:72–101.

[3] Laenen A. Psychometric validation of continuous rating scales from complex data. Unpublished PhD thesis, KULeuven, 2008. Available at: http://ibiostat.be/publications/phd/annouschkalaenen.pdf (accessed 18.09.2016).

[4] Molenberghs G, Kenward M. *Missing Data in Clinical Studies*. John Wiley & Sons: New York, 2007.

[5] Rubin DB. Inference and missing data. *Biometrika* 1976; **63**:581–592.

[6] Fleiss JL. *Design and Analysis of Clinical Experiments*. Wiley: New York, 1986.

[7] Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Springer-Verlag: New York, 2000.

[8] Pikkemaat R, Lundin S, Stenqvist O, Hilgers RD, Leonhardt S. Recent advances in and limitations of cardiac output monitoring by means of electrical impedance tomography. *Anesthesia & Analgesia* 2014; **119**:76–83.

[9] Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (withdiscussion). *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 1994; **43**:429–467.

[10] Glaser D, Hastings RH. An introduction to multilevel modeling for anesthesiologists. *Anaesthesia & Analgesia* 2011; **113**:877–887.

[11] West BT, Welch KB, Galecki AT. *Linear Mixed Models. A Practical Guide Using Statistical Software* (2nd Ed.) CRC Press Taylor & Francis Group: New York, 2015.

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

[13] Lee OE, Braun T. Permutation tests for random effects in linear mixed models. *Biometrics* 2012; **68**:486–493.

[14] Van der Elst W, Molenberghs G, Van Boxtel MPJ, Jolles J. Establishing normative data for repeated cognitive assessment: a comparison of different statistical methods. *Behavior Research Methods* 2013; **45**:1073–1086.

[15] Vangeneugden T, Molenberghs G, Laenen A, Geys H, Beunckens C, Sotto C. Marginal correlation in longitudinal binary data based on generalized linear mixed models. *Communications in Statistics. Theory and Methods* 2010; **39**:3540–3557.

[16] Burzykowski T, Molenberghs G, Buyse M. *The Evaluation of Surrogate Endpoints*. Springer-Verlag: New York, 2005.

[17] Roy A. Estimating correlation coefficient between two variables with repeated observations using mixed effects model. *Biometrical Journal* 2006; **48**:286–301.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.