

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

Peer-reviewed author version

VAN DER ELST, Wim; MOLENBERGHS, Geert; Hilgers, Ralf-Dieter; VERBEKE, Geert & Heussen, Nicole (2016) Estimating the reliability of repeatedly measured endpoints based on linear mixed-effects models. A tutorial. In: PHARMACEUTICAL STATISTICS, 15(6), p. 486-493.

DOI: 10.1002/pst.1787

Handle: <http://hdl.handle.net/1942/23064>

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

Wim Van der Elst¹, Geert Molenberghs^{1,2},
Ralf-Dieter Hilgers³, Geert Verbeke^{1,2} & Nicole Heussen³

Abstract

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, e.g., 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

¹I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium.

²I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium.

³Department of Medical Statistics, RWTH Aachen University, Aachen, Germany

1 Introduction

1 Reliability essentially refers to the reproducibility (or, predictability) of out-
 2 comes that are repeatedly measured *within* the same individuals. In particu-
 3 lar, this metric quantifies the extent to which a repetition of a measurement
 4 under the same general conditions leads to the same result.

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

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

14 Eq. (1) is intuitively appealing because it defines reliability as the fraction of
 15 the observed test score variance that is attributable to the true score variance.
 16 If a test is perfectly reliable, the true score and observed score variances are
 17 equal and thus $R = 1$. Unfortunately, reliability cannot be directly estimated
 18 based on Eq. (1) because τ cannot be observed. Instead, reliability will have
 19 to be estimated indirectly. A classical solution to the problem is to introduce
 20 the concept of *parallel tests* [2]. Parallel tests are tests that have the same
 21 true score for each subject and equal error variances. For example, suppose
 22 that we have two measurements X_1 and X_2 for the same subjects that are
 23 assessed at two instances of time with a short lag (such that τ does not
 24 change), or that are obtained from two raters at the same point in time.
 25 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
 26 $\text{Var}(\varepsilon_1) = \text{Var}(\varepsilon_2) = \text{Var}(\varepsilon)$, i.e., X_1 and X_2 are parallel measurements. The
 27 covariance of the two measurements then equals

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

28 and the correlation between X_1 and X_2 can be written as

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

29 **Limitations of the conventional methods** Eq. (2) provides a convenient and straightforward way to compute reliability, but it is important
30 to stress that the assumption that the measurements are parallel is crucial.
31 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
32 do not exhibit a systematic change over time as a result of their treatment.
33 Another limitation of Eq. (2) is that only two measurements can be considered, and these measurements should have the same test-retest interval for
34 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
35 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
36 approach does not only lead to a loss of information, but it also ignores the missing data generating mechanism. Basically, to obtain unbiased estimates
37 for R using Eq. (2), the assumption that the data are missing completely at random (MCAR) should be valid. This means that the missingness should
38 not depend on the observed or the unobserved outcomes [4, 5]. This is a strong and often unrealistic assumption, e.g., in a clinical trial setting it is
39 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
40 measurement in time (missing value for X_2).
41
42
43
44
45
46
47
48
49

50 **Importance of reliability** It is important to carefully consider the reliability of a measurement procedure, for example in the context of designing
51 a clinical trial. Obviously, in particular in explorative or experimental small population group studies, serial measurements are gathered to understand
52 the nature of the disease. However, unreliable measurement methods might
53 lead to serious misinterpretation of the disease process. Indeed, even the most elegant study design will not overcome the damage that is caused by
54 the use of unreliable measurement procedures [6]. For example, biased sample selection may occur when patients are selected based on an unreliable
55 measurement procedure, and the sample size that is required to detect an
56
57
58
59

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, see [6]). Thus, for example, when $R = 0.50$, the required sample size to detect δ *doubles* compared to 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 MCAR, 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.

96 2 Case study

97 Pikkemaat *et al.* [8] performed an experiment where the cardiac output
 98 and stroke volume of $N = 14$ pigs was changed by increasing positive end-
 99 expiratory pressure (PEEP) levels (0, 5, 10, 15, 20, and 25 cm H₂O). The
 100 number of times that a particular PEEP level was used varied from animal
 101 to animal. For each PEEP level, stroke volume was measured by the contin-
 102 uous approximately normally distributed variable Electrical Impedance To-
 103 mography (ZSV). In each animal, four identical experiments were conducted
 104 (referred to as Cycles 1 to 4). The number of repeated ZSV measurements
 105 across PEEP levels and cycles in an animal ranged between 9 and 47. In the
 106 analyses below, it is assumed that all the measurements are equally spaced.

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

115 As noted above, the study included a total of 14 pigs. However, the data
 116 of $n = 2$ animals could not be evaluated due to technical reasons and these
 117 animals were thus excluded from the analyses. Further, there were $n = 2$
 118 animals who appeared to have a ‘clinically deviating’ profile (as judged by
 119 the experimenters). These animals were kept in the current analyses, but a
 120 sensitivity analysis showed that the estimated reliabilities were not substan-
 121 tially affected by the in- or exclusion of these animals (see Web Appendix
 122 Part II). Note that the data for PEEP level 25 were included in the current
 123 analysis, as well as in the Pikkemaat *et al.* [8] study, although they were not
 124 explicitly mentioned in the latter.

125 3 Exploratory data analysis

126 Figure 1 shows the individual profiles (grey lines) of ZSV as a function of
 127 measurement moment. As can be seen, there is substantial between- as well
 128 as within-animal variability. Further, drop-out is substantial, i.e., there are
 129 less observations at later measurement moments compared to earlier mea-

130 surement moments. This is more clearly depicted in Figure 2, where the
 131 number of available observations at each of the different measurement mo-
 132 ments are shown.

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

138

139 » Figures 1 and 2 about here «

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

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

145 Each power p_m is chosen from a restricted set, typically $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$.
 146 Note that when $M = 2$ and $p_1 = p_2$, the linear predictor (3) becomes
 147 $\beta_0 + \beta_1 t^{p_1} + \beta_2 t^{p_1} \log(t)$. Also, when $p = 0$, this is taken to refer to $\log(t)$ [9].
 148 In practice, all possible models of degree 1 to M are fitted. Thus for $M = 1$,
 149 each of the 8 values of S are used for the predictor t^{p_1} , for $M = 2$ each of the
 150 36 combinations of powers are used for the predictors t^{p_1} and t^{p_2} , and so on.
 151 Subsequently, the ‘best’ fitting model is selected. This choice can be made
 152 in an informal way (i) based on Akaike’s Information Criterion (AIC, where
 153 a lower value is indicative of a better model fit) and/or (ii) by graphically
 154 evaluating the fit of the model with the observed data. The AIC adds the
 155 number of model parameters as a penalty to the log likelihood of the model,
 156 which may help to avoid over-fitting (even though one still may want to be
 157 careful not to select an overly complex model, in particular when a large
 158 number of candidate powers is considered). The main advantage of using
 159 fractional polynomials (rather than regular polynomials) is that they allow
 160 for a much more flexible parametrization, i.e., a large number of different
 161 shapes of curves can be captured by even a relatively small M .

162 **Application to the case study** In the analysis of the case study, frac-
163 tional polynomials of order $M = 1$ to $M = 5$ were considered using the
164 standard set $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ for the powers p_m . Note
165 that it is possible to use a more extensive set of values for S if the original set
166 does not provide an adequate result, but the number of models that have to
167 be fitted (and thus also the required computational time) increases sharply
168 when the number of elements in S increases. For example, when the set S
169 includes 8 elements (the standard set), a total of 792 fractional polynomials
170 of degree 5 can be made. However, when the set $S = \{-3, -2.75, \dots, 3\}$
171 is used (25 elements), a total of 118,755 fractional polynomials of degree 5
172 can be made. Similarly, M can be increased but this will again yield a sharp
173 increase in the number of models to be evaluated.

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

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

184

185 » Table 1 about here «

186 » Figure 3 about here «

187 4 Estimating reliability using mixed-effects mod- 188 els

189 In this section the reliability of the ZSV will be estimated using a flexible
190 approach that is based on LMMs. The LMM is briefly introduced in Section
191 4.1 (for more details, see e.g., [7, 10, 11]), and the LMM-based approach to
192 estimate reliability is applied to the case study in Section 4.2. For conciseness,
193 in the latter section only a summary of the main results is given and no
194 reference to software tools that can be used to obtain the results is made.
195 However, full details can be found in the Web Appendix Parts I–V.

196 4.1 The linear mixed-effects model

197 A LMM can be written as

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

198 where \mathbf{Y}_i is the response vector for subject i (with $i = 1, 2, \dots, n$ subjects
199 in the study), \mathbf{X}_i and \mathbf{Z}_i are the known design matrices for the fixed and
200 random effects, $\boldsymbol{\beta}$ is the vector that contains the fixed effects, \mathbf{b}_i is the vector
201 that contains the random effects, and $\boldsymbol{\varepsilon}_i$ is the vector that contains the mea-
202 surement 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 $\boldsymbol{\Sigma}_i$
203 are general variance-covariance matrices). Model (4) thus assumes that the
204 vector of repeated measurements for each subject follows a linear regression
205 model where some of the parameters are population-specific (that is, param-
206 eters that are the same for all subjects in the population; the fixed effects)
207 and other parameters are subject-specific (that is, parameters that differ for
208 all subjects; the random effects).

209 The residual component $\boldsymbol{\varepsilon}_i$ is often further decomposed as $\boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}_{(1)i} +$
210 $\boldsymbol{\varepsilon}_{(2)i}$. Here, $\boldsymbol{\varepsilon}_{(2)i}$ is a component of serial correlation and $\boldsymbol{\varepsilon}_{(1)i}$ is a component
211 of measurement error. Serial correlation results from the fact that within
212 a subject, the (residuals of) observations that are closer in time are often
213 ‘more similar’ (i.e., more strongly correlated) than observations that are more
214 distant in time. It is assumed that $\boldsymbol{\varepsilon}_{(1)i} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i})$ (with \mathbf{I}_{n_i} an identity
215 matrix of dimension n_i = the number of repeated measurements in a subject)
216 and $\boldsymbol{\varepsilon}_{(2)i} \sim N(\mathbf{0}, \tau^2 \mathbf{H}_i)$ (with \mathbf{H}_i the serial correlation matrix that only
217 depends on i through the number of repeated measurements n_i and the time
218 points j and k at which the measurements are taken). The (j, k) element h_{ijk}
219 of \mathbf{H}_i can then be modeled as $h_{ijk} = g(|t_{ij} - t_{ik}|)$ for a decreasing function g .
220 Two frequently used functions are the exponential and Gaussian correlation
221 functions, defined as $g(u_{j,k}) = \exp(-\phi u_{j,k})$ and $g(u_{j,k}) = \exp(-\phi u_{j,k}^2)$,
222 respectively.

223 4.2 Case study analysis

224 **The mean structure of the model** The LMMs that will be fitted to
225 the case study dataset include an intercept, measurement moment, PEEP,
226 and Cycle as fixed effects. PEEP and Cycle are dummy-coded with 5 and
227 3 dummies, respectively. The relation between measurement point and the

228 ZSV outcome is modeled as $\beta_1 t^3 + \beta_2 t^2 + \beta_3 t^2 \log(t)$ (see the Fractional poly-
 229 nomial section).

230 **The covariance (correlation) structure of the model** In the analy-
 231 ses below, three LMMs with the same fixed-effect structure (see previous
 232 paragraph) but different variance structures will be considered.

233 Model 1 is a random intercept model, i.e., a LMM that only contains a
 234 random intercept in the random part of the model:

$$Y_{ij} = \mu_{ij} + b_{0i} + \varepsilon_{ij}, \quad (5)$$

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

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

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

246 Therefore, Model 2 extends Model 1 by adding a serial correlation com-
 247 ponent:

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

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

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

252 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
 253 no longer assumes that R remains constant for all pairs of measurements.

254 Instead, it models R as a function of the time lag u_{jk} between two measure-
 255 ments. As can be seen, a stronger serial effect (ρ^2) leads to a faster decreasing
 256 $R(u_{jk})$.

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

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

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

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

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

268 Table 2 summarizes the covariance structures that are used in the differ-
 269 ent models and their impact on the estimated R .

270

271 » Table 2 about here «

272 4.2.1 Model 1: random intercept model

273 When Model 1 was fitted to the case study dataset, it was obtained that $\hat{d} =$
 274 1901.611 and $\hat{\sigma}^2 = 2413.022$, yielding $\hat{R} = 0.441$ (see Eq. (6)). A CI around \hat{R}
 275 can be computed by using a (non-parametric) bootstrap or the Delta method
 276 (for details, see the Web Appendix Part VI). The bootstrap-based 95% CI
 277 (using 500 bootstrap samples) equaled [0.198; 0.618]. The Delta method-
 278 based CI was similar and largely overlapped, i.e., [0.189; 0.636]. Figure 4
 279 (top left) illustrates the results (the bootstrap-based CI is shown).

280 Overall, it can be concluded that \hat{R} is moderate and that there is sub-
 281 stantial uncertainty in \hat{R} (which is not surprising given the small number of

282 animals in the study).

283

284 » Figure 4 about here «

285 4.2.2 Model 2: random intercept and serial correlation

286 When Model 2 was fitted to the data of the case study, the estimated co-
287 variance parameters were $\hat{d} = 1349.650$, $\hat{\tau}^2 = 2489.351$, $\hat{\rho} = 3.581$, and
288 $\hat{\sigma}^2 = 382.795$. Thus, after correction for the fixed effects, the covariance
289 parameter estimates showed considerable remaining serial components.

290 Figure 4 (top right) shows the estimated $R(u_{jk})$ (see Eq. (8)) and their
291 95% CIs based on a bootstrap (the Delta method-based CIs were similar; data
292 are shown in the Web Appendix Part I). As can be seen, the estimated R were
293 high for small time lags (e.g., $\hat{R}(u_{jk} = 0) = 0.865$ and $\hat{R}(u_{jk} = 1) = 0.751$)
294 and subsequently decreased until they remained essentially constant at $\hat{R} \approx$
295 0.320 for measurements with time lags of about $u_{jk} = 10$ and higher. It can
296 also be observed that the CIs around $\hat{R}(u_{jk})$ were narrower for measurements
297 with smaller time lags (e.g., for time lags $u_{jk} = 0$ and $u_{jk} = 1$, the $CI_{95\%} =$
298 $[0.817, 0.906]$ and $CI_{95\%} = [0.654, 0.836]$, respectively) and subsequently
299 widened until they remained stable around time lag $u_{jk} = 10$ with $CI_{95\%} =$
300 $[0.045, 0.530]$.

301 4.2.3 Model 3: random intercept, slope, and serial correlation

302 When Model 3 was fitted to the data of the case study, the estimated covari-
303 ance parameters were $\hat{\tau}^2 = 1952.970$, $\hat{\rho} = 3.290$, $\hat{\sigma}^2 = 373.043$, and

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

304 As noted earlier, based on Model 3 the estimated $R(t_k, t_j)$ are different for
305 all pairs of measurements (see Eq. (10)). Figure 4 (bottom) shows the re-
306 sults graphically. In this figure, the utmost left line (marked with t_1) depicts
307 the estimated $R(t_1, t_j)$, i.e., the estimated reliabilities of ZSV taken at mea-
308 surement times 1 and 2–45. The line next to that one shows the estimated
309 $R(t_2, t_j)$, etc. The figure shows that $\hat{R}(t_k, t_j)$ is high when the time lag u is
310 small and flattens out for longer time lags. Further, depending on the partic-
311 ular pair of measurement moments (t_k, t_j) that is considered, the slope and
312 amount of decline in $\hat{R}(t_k, t_j)$ as a function of time lag differs. For example,

when considering $\widehat{R}(t_1, t_j)$, 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 $\widehat{R}(t_{20}, t_j)$ 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 $\widehat{R}(t_1, t_j)$ (left) and $\widehat{R}(t_{20}, t_j)$ (right). As can be seen, the CIs increase as a function of time and tend to be wider for $\widehat{R}(t_{20}, t_j)$ than for $\widehat{R}(t_1, t_j)$ (as expected).

» Figure 5 about here «

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, see [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 to 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_1^2 and χ_2^2 is needed (denoted by $\chi_{1:2}^2$), 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, see [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 (see 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 above used Restricted Maximum Likelihood (REML), but valid LR tests for comparing nested models with

349 different covariance structures can still be obtained under REML estimation
350 when the models that are compared have the same mean structure [7] – which
351 was the case here, see above.

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

361 Model 3 is the model with the largest likelihood. It would be preferred
362 if we would solely rely on statistical arguments. However, from an applied
363 perspective – i.e., also considering the practical usefulness of the results for
364 a clinician or researcher – Model 2 is arguably to be preferred over Model
365 3 because the former leads to reliability estimates that only depend on the
366 time lag between two measurements. In contrast, Model 3 yields different
367 reliability estimates for all possible pairs of measurements. Model 2 thus
368 provides a much more parsimonious result compared to Model 3 – whilst the
369 fit of both models is roughly comparable. Notice that the likelihood ratio
370 tests identify the best fitting model among the models that were under con-
371 sideration. However, when a model has been selected, the question remains
372 whether this model fits the data sufficiently well. Residuals and influence
373 diagnostics are useful in this respect. In Part VII of the Supplementary Ma-
374 terials, a residual analysis is conducted and the extent to which particular
375 animals exert a strong influence on the results (i.e., the REML distances of
376 the models, the estimated fixed-effects parameters, the estimated covariance
377 components, and the estimated reliability coefficients) is evaluated. Overall,
378 the impact of excluding an animal on the results was relatively small for
379 Models 2–3. For Model 1, the impact of deleting an animal on the results
380 was more substantial. Further, the residual analysis showed that there were
381 no major departures of normality.

382

383

384 » Table 3 about here «

385 5 Discussion

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

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

421 Second, in the present paper the focus was entirely on the random effect

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

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

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

449 Finally, in the present paper no time-varying covariates (other than mea-
450 surement occasion itself) were considered, but depending on the study at
451 hand it may be useful to include such covariates. For example, consider a
452 setting where one is interested in estimating the reliability of a psychiatric
453 rating scale that was scored by different physicians at the different mea-
454 surement moments. When only a limited number of raters are involved in
455 the study, the methodology that was proposed above can still be used in a
456 straightforward way. Indeed, one can then simply include rater as a (dummy-
457 coded) fixed-effect in the mean structure of the model. On the other hand,
458 when the number of raters is large, it is more sensible to include rater in the
459 random-effect part of the model. Such a model cannot be fitted in the cur-

460 rent version of the *CorrMixed* package, but it is straightforward to fit such
461 a model using SAS.

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

Acknowledgements

Financial support from the IAP research network #P7/06 of the Belgian Gov-
ernment (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.

Web Appendix

A Web Appendix is available that contains (i) the Delta method-based 95%
CIs of the estimated reliability coefficients for ZSV, (ii) a sensitivity analysis
where the impact of 2 clinically deviating animals on the results is examined,

(iii) a sensitivity analysis where the impact of using a different fixed-effect structure on the results is examined, (iv) details on how the newly developed R package *CorrMixed* can be used to estimate reliability, (v) details on how reliability can be estimated using SAS, (vi) details on the computation of the Delta method-based CIs for \hat{R} , and, (vii) the results of a residual analysis.

References

- [1] Lord FM, Novick MR. *Statistical theories of mental test scores*. Addison-Welsley Publishing Company, Reading, MA; 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*; 2008. Unpublished PhD thesis. Retrieved from <http://ibiostat.be/publications/phd/annouschkalaenen.pdf>
- [4] Molenberghs G, Kenward M. *Missing data in clinical studies*. New York: John Wiley & Sons; 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*. New York: Springer-Verlag; 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. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 1994; **43**:429-467.
- [10] Glaser D, Hastings RH. (2011). 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.)*. New York: CRC Press, Taylor & Francis Group; 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*. New York: Springer-Verlag; 2005.
- [17] Roy A. Estimating correlation coefficient between two variables with repeated observations using mixed effects model. *Biometrical Journal* 2006; **48**:286-301.

Tables

Table 1: 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

Table 2: 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)
<i>Note.</i> t_j = measurement at time j .	

Table 3: Fit indices of the different models for the ZSV outcome.

	# Pars.		logL	G^2	Test	p
	Rand.	Ser.				
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.

Figures

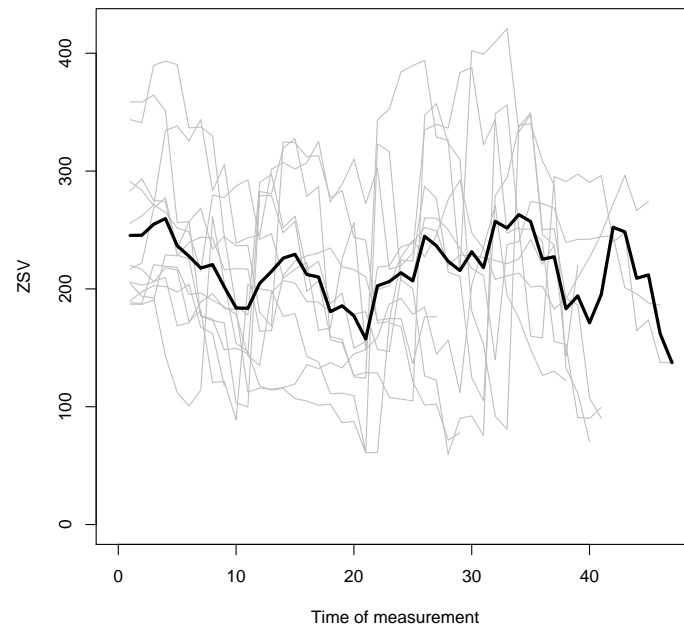


Figure 1: Individual profiles (grey lines) and mean values (black line) of the ZSV outcome as a function of time of measurement.

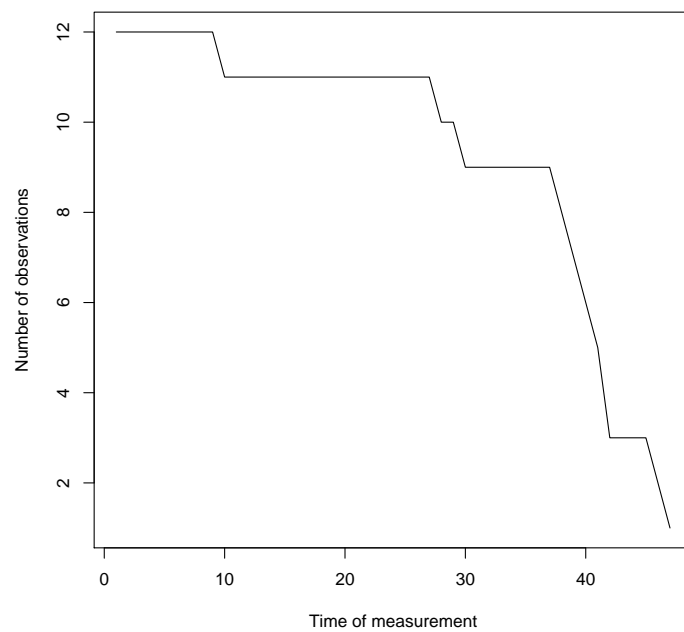


Figure 2: Number of observations for the ZSV outcome as a function of time of measurement.

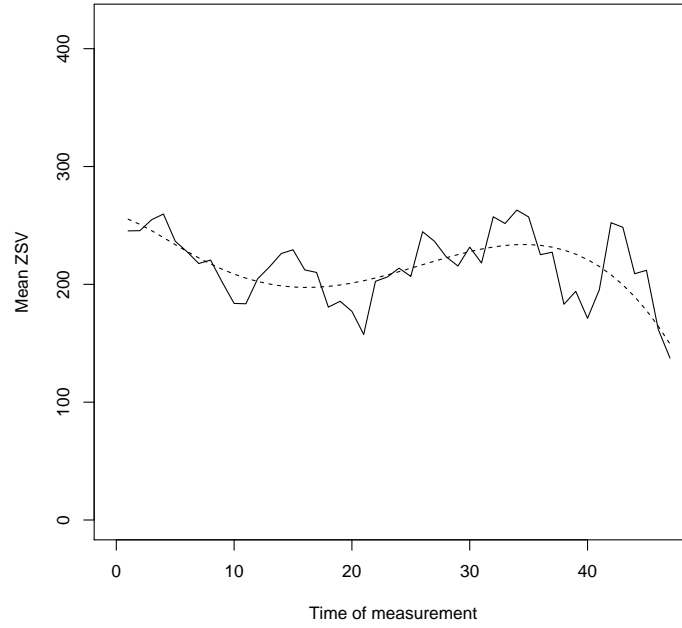
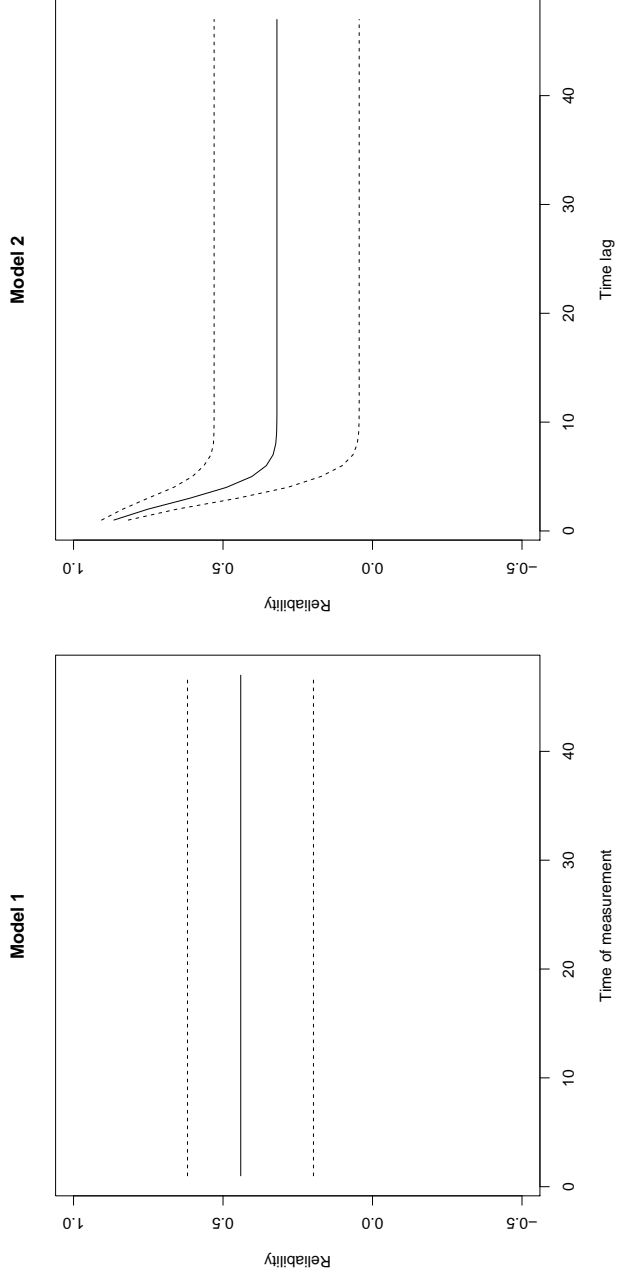


Figure 3: Observed means as a function of time of measurement (solid line) and fitted fractional polynomial of degree $m = 3$ (dashed line).



Model 3

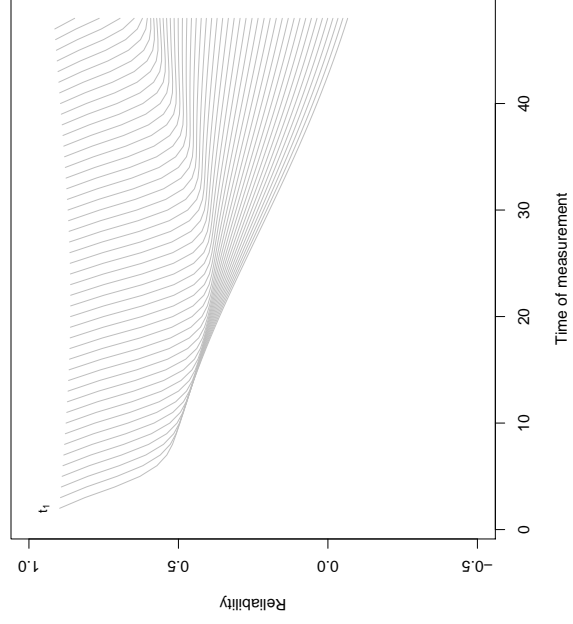


Figure 4: Estimated reliabilities (solid lines) and 95% Confidence Intervals (dashed lines) for ZSV based on Model 1 (upper left), Model 2 (upper right) and Model 3 (bottom). For Model 3, no Confidence Intervals are provided to avoid a cluttered figure. The utmost left line marked with t_1 depicts the 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.

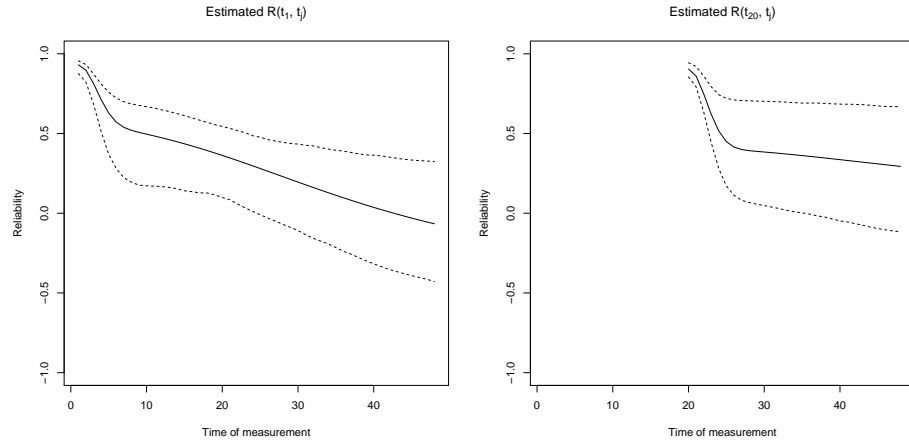


Figure 5: $\hat{R}(t_1, t_j)$ (left) and $\hat{R}(t_{20}, t_j)$ (right) based on Model 3 and their 95% Confidence Intervals for the ZSV outcome.