Longitudinal Data Analysis of Symptom Score Trajectories Using Linear Mixed Models in a Clinical Trial

C. Engel[1,*], C. Meisner[2], A. Wittorf[3], W. Wölwer[4], G. Wiedemann[5], C. Ring[6], R. Muche[6] and S. Klingberg[3]

[1]*Center for Pediatric Clinical Studies, Biometry University Children's Hospital Tuebingen, Frondsbergstraße 23, 72070 Tübingen, Germany*

[2]*Department of Clinical Epidemiology and Applied Biometry, University of Tuebingen, Silcherstraße 5, 72076 Tübingen, Germany*

[3]*Department of Psychiatry and Psychotherapy, University of Tuebingen, Calwerstraße 14, 72076 Tübingen, Germany*

[4]*Department of Psychiatry and Psychotherapy, University of Düsseldorf, Bergische Landstraße 2, 40629 Düsseldorf, Germany*

[5]*Department of Psychiatry and Psychotherapy, Klinikum Fulda gAG, Pacelliallee 4, 36043 Fulda, Germany*

[6]*Institute of Epidemiology and Medical Biometry, University of Ulm, Schwabstraße 13, 89075 Ulm, Germany*

Abstract: In clinical trials, longitudinal data are often analyzed using T-tests, anovas or ancovas instead of the more powerful linear mixed models. The purpose of this paper is to demonstrate how the more sophisticated linear mixed models according to the approach of Singer and Willett, which allows special insight into the behaviour of the data, can be used in clinical trials. Individual trajectories of PANNS-MNS Scores from a controlled clinical trial were used to demonstrate all the steps needed for an analysis of longitudinal data. The model is built step by step, model assumptions are checked, time-variant and time-invariant factors are included and the results are interpreted. The unique needs of a clinical trial, such as the calculation of effect sizes or of an appropriate sample size, are taken into account. Finally, a flow chart is presented that would serve as an instruction tool for the analysis of longitudinal data in clinical trials.

Keywords: longitudinal studies, randomized controlled trial, linear models, sample size.

INTRODUCTION

Longitudinal data are often collected in clinical trials. For the analysis of these data, however, cross-sectional methods are used, which explains why, when data was collected, e.g., at 10 time-points, it is only the first and last values which are used in an analysis of covariance despite using appropriate methods. Failing that means power is lost for the analysis and money is also lost because unused data were collected in vain.

As far back as the 1920s the so called "slope-as-outcome" and "intercept-as-outcome" models were introduced. These models build up the intercept and the slope of a regression equation by means of a regression coefficient and an error term. This makes it possible to consider complex data and error structures and a hierarchical data ordering [1].

Primarily these methods were introduced for hierarchical data, e.g., data could be grouped according to siblings in families, or students in classes, and so on, as discussed [2]. Before the concept was created, however, the only correct method for analyzing longitudinal data was by means of estimations of individual regression equations. That means the individual measurements were grouped according to the individual involved, thus mirroring reality, as being described [3, 4]. Later on, ideas were proposed to evaluate the influence of specific factors on the course of the data [5]. These methods have been implemented for several years in standard statistical software programmes, e.g. as the procedure "proc mixed" in SAS since 1997, as described in SAS/STAT Software 1997 [6].

Nevertheless, these methods are still rarely used in the analysis of clinical trials. Recent literature includes analyses of longitudinal data using the t-test [7] or various kinds of analysis of variance as used [8-12], as an example. If linear mixed models are used, the model building is usually not described in a manner, that is easy to follow [13-18]. There even seems to be a trend to regard longitudinal data in a similar way as survival data and to define a certain event out of the measurements that may occur or not, as done [19-21],

*Address correspondence to this author at the Center for Pediatric Clinical Studies, Biometry University Children's Hospital Tuebingen, Frondsbergstraße 23, 72070 Tübingen, Germany; Tel: 0049-7071-2989111;
Fax: 0049-7071-294857; E-mail: corinna.engel@med.uni-tuebingen.de

E-ISSN: 1929-6029/13

as an example. In particular, the possibility of including time-variant factors in linear mixed models offers a wide field of applications, as in analysing the influence of stress on ward and whether doctors' workload impact supervision and ward activities of final-year students [22].

Well-established methods like analysis of variance, t-test or survival analysis offer the great advantage that all the necessary tools for planning and analyzing a study are implemented in standard software.

The problems that may arise when more sophisticated and, therefore, more appropriate methods are used in clinical studies, start with the calculation of an adequate sample size, continue with the build-up of an appropriate model and end with the calculation of effect sizes comparable to already known ones.

Based on the statistical analysis of a randomized clinical trial we will show how clinical studies in which quantitative longitudinal data is collected for measuring the outcome can be analyzed using appropriate methods. The approach of Singer and Willett (2003) [23] was chosen, because it facilitates insight in each step of the model building process in a very demonstrative manner applying the principle of multilevel analysis to longitudinal data. This is in contrast to the approach described for instance in SAS/STAT Software 1997 [6] regarding longitudinal data basically as repeated measurements.

We will demonstrate how model building can be done and effect sizes can be calculated. We conclude with a general flow chart which illustrates the analysis up to the final model. Finally, we propose a procedure for the sample size estimation by reviewing the literature.

MATERIAL AND METHODS

In the TONES study [24, 25], a multicenter clinical trial, 198 patients suffering from schizophrenia were recruited by three separate German centers. The primary endpoint of the study was the improvement of negative symptoms, which was quantified using a modified negative syndrome scale (PANNS-MNS) which is a subscale of the Positive and Negative Syndrome Scale (PANNS). The PANSS is an observer based rating scale for the assessment of symptoms associated with psychotic disorders and is based on a structured clinical interview. The PANSS includes three subscales: "Positive Symptoms", "Negative Symptoms", and "General Psychopathology". All the scales showed a sufficient fit with normal distribution for the use of parametric methods, as has been shown [26]. According to Klingberg *et al*. (2006) [27], a modified negative symptom scale has been chosen as primary endpoint. The TONES study compares the efficacy of two different psychological interventions: cognitive behavioural therapy (CBT), which is recommended in evidence based treatment guidelines for the reduction of positive symptoms, and cognitive remediation (CR). CR has been chosen as comparator as published evidence did not indicate treatment effects for negative symptoms at the time of designing this study. An effect size of 0.35 points superiority of CBT at the end of the study was expected. Sample size calculation was based on this assumption (α=0.05, ß=0.2) and done for classical ANOVA due to the fact, that software for the calculation of sample size for the analysis of longitudinal data using multilevel mixed model was not available. The design and sample size calculation of the study and the results of the primary and secondary endpoints were published [24, 25].

The therapeutic alliance proved to be a common ingredient of all psychotherapeutic interventions and to be at least modestly correlated with outcome [28]. Thus, ratings of the therapeutic alliance were assessed at the end of each session of the study therapies using items from the short versions of the Bernese Post-Session Report for patients (BPSR-P) developed by Grawe and his co-workers at the University of Bern, Switzerland [29]. The following three items cover the factor of "therapeutic alliance" as perceived by the patient: (1) "The therapist and I understand each other", (2) "Today I felt at ease with the therapist", and (3) "I think the therapist is really interested in my well-being". The statements are rated on a 7-point Likert scale anchored by -3, "absolutely not true", 0, "neither nor", and +3, "absolutely true".

In this article we use the TONES study to exemplify how linear mixed models according to the approach of [23] can be used for designing and analyzing a randomised trial.

The factors taken into account were therapy (CBT versus CR) and center (1 versus 2 versus 3) as examples for level 2 (time-invariant) factors. At level 1 (time-variant factors) the following factors were used: time (12 monthly measurements during a period of 0-11 months), at least one session visited in the time interval between the last and the actual rating of

negative symptoms (yes versus no), and the factor "therapeutic alliance" from the BPSR-P.

Singer and Willet (2003) [23] proposed building up the model by estimating individual regression parameters for each subject followed by the estimation of mean parameters out of these individual ones. Intercept and slope are considered as random effects due to the fact that in the next sample other subjects are included, and an unstructured covariance matrix is assumed. This approach facilitates insight into the model building in terms of the variance components. Only after having built and checked the whole model according to this principle it is finalized using a repeated measurement approach, as demonstrated below.

First the null model is tested. This model has no independent factor and therefore calculates the individual mean and the individual and overall deviation from this mean.

Null model:

$$Y_{ij} = \gamma_{00} + (\zeta_{0i} + \varepsilon_{ij}) \quad (1)$$

Where i = patient 1 to i and j = measurement 0 to j

γ_{00} = overall mean

ζ_{0i} = difference between person specific mean and overall mean

ε_{ij} = difference between individual measurement and person specific mean

If there is enough variation between the subjects, and normal distribution is verified, the next step is to check the unconditional growth model. In this model, time is the only independent factor that is analyzed for its influence.

Unconditional growth model:

$$Y_{ij} = \left[\gamma_{00} + \gamma_{10} * time_{ij}\right] + \left[\zeta_{0i} + \zeta_{1i} * time_{ij} + \varepsilon_{ij}\right] \quad (2)$$

Where i and j as above.

γ_{00} = mean intercept

ζ_{0i} = difference between individual and mean intercept

γ_{10} = mean slope

ζ_{0i} = difference between individual and mean slope

ε_{ij} = difference between the individual measurement and the individual growth curve.

Also, heteroscedasticity and autocorrelation between time points are evaluated.

Heteroscedasticity:

$$\sigma^2_{\mathrm{Re}sidual_j} = \sigma_0^2 + \sigma_1^2 * time_j^2 + 2\sigma_{01} * time_j + \sigma_\varepsilon^2 \quad (3)$$

Autocorrelation:

$$\rho_{\mathrm{Re}sidual_j \mathrm{Re}siudual_{j'}} = \frac{\sigma_0^2 + \sigma_{01}(time_j + time_{j-1}) + \sigma_1^2 * time_j * time_{j-1}}{\sqrt{\sigma^2_{\mathrm{Re}sidual_j} * \sigma^2_{\mathrm{Re}sidual_{j-1}}}} \quad (4)$$

Where i and j as above

σ_0^2, σ_1^2 = residuals of level 2 (unestimated part of deviation of individual intercept and slope from mean intercept and slope)

σ_{01} = population's covariance

σ_ε^2 = residual of level 1 (unestimated part of individual deviation)

Only after having verified the existence of autocorrelation ($\rho \neq 0$) and heteroscedasticity ($\sigma^2_{\mathrm{Re}sidual_j} \neq \sigma^2_{\mathrm{Re}sidual_{j'}}$) in the unconditional growth model according to Gruber (1982) [30], the model analyzing the influence of fixed factors (in our case: the two therapies and adjustment for centers) is built.

Standard Model:

$$Y_{ij} = \left[\gamma_{00} + \gamma_{10} * time_{ij} + \gamma_{01} * X_i + \gamma_{11} * X_i * time_{ij}\right] + \left[\zeta_{0i} + \zeta_{1i} * time_{ij} + \varepsilon_{ij}\right] \quad (5)$$

Where i, j, γ_{00}, ζ_{0i}, γ_{00}, ζ_{0i} and ε_{ij} as above

X_i = fixed factor

Pseudo-R^2 statistics are calculated using the equations below and indicate the degree of explanation that an additional independent factor can contribute to the variability at this level.

Residual of level 1:

$$Pseudo\ R^2_\varepsilon = \frac{\hat{\sigma}^2_\varepsilon(\text{null model}) - \hat{\sigma}^2_\varepsilon(\text{unconditional growth model})}{\hat{\sigma}^2_\varepsilon(\text{null model})} \quad (6)$$

Residuals of level 2:

$$Pseudo\ R_{\zeta}^2 = \frac{\hat{\sigma}_{\zeta}^2(\text{unconditional growth model}) - \hat{\sigma}_{\zeta}^2(\text{following model})}{\hat{\sigma}_{\zeta}^2(\text{unconditional growth model})} \quad (7)$$

The final model built by means of this approach is then re-analyzed by a model that takes the repeated structure of the data into account and models the variance matrix properly. Models are compared by using the AIC (Akaike's information criterion) and BIC (Bayesian information criterion). This re-analysis can be regarded as a means of making final adjustments of the regression parameters.

A requirement of linear mixed models is the normal distribution of the composed residual

$$r_{ij} = \left[\varepsilon_{ij} + \zeta_{0j} + \zeta_{1j} time_j \right] \quad (8)$$

According to Hox (2002) [2], the residuals of the null model and the residuals of the final model were evaluated graphically for normal distribution. Calculating the same model with robust standard errors, and a second time with asymptotical ones, allows one to assess whether a possible deviation from normal distribution affects the results, as has been shown [31, 32].

Effect sizes for the primary endpoint were estimated using Hedges' g [33] adjusted for sample size.

$$d \cong \left(1 - \frac{3}{4N - 9}\right) * g \quad (9)$$

with

$$g = (\bar{Y}^E - \bar{Y}^C) / s$$

and

$$s = \sqrt{\frac{(n^E - 1) * (s^E)^2 + (n^C - 1) * (s^C)^2}{n^E + n^C - 2}}$$

$\bar{Y}^E - \bar{Y}^C$ = adjusted difference between experimental and control group, resulting from linear mixed model

s = pooled standard deviation

n^E, n^C = number in experimental/control group

s^E, s^C = standard deviation of experimental/control group

95% Confidence limits are given by

$$\delta_L = d - C_{\alpha/2}\hat{\sigma}(d) \qquad \delta_U = d + C_{\alpha/2}\hat{\sigma}(d) \quad (10)$$

with

$$\hat{\sigma}^2(d) = \frac{n^E + n^C}{n^E * n^C} + \frac{d^2}{2 * (n^E + n^C)} \quad (11)$$

δ_L / δ_U : lower/upper limit of confidence interval

$C_{\alpha/2}$: critical value of standard normal distribution

$\hat{\sigma}(d)$: standard deviation of d

Based on the argument posed [34], the standard deviation of the baseline values was used for s^E / s^C.

In addition to the fixed factors the aforementioned time-variant factors were exemplified. Their influence is on the individual level. Therefore changes in the level 1 variance component are considered.

Models were built up using SAS 9.1. Macros are available from the author.

RESULTS

The linearity of the dependent variable "MNS-score", i.e. the severity of negative symptoms, over the 10 assessments was checked by graphs of single individuals and the mean course. The observed course was compatible with the assumption of linearity.

The null model according to formula (1) revealed enough variation within as well as between patients and a sufficient fit to normal distribution. The unconditional growth model (formula 2) demonstrated a linear change over the time of the study, different from zero, a heteroscedasticity according to formula (3) between 0.64 and 0.86 and an autocorrelation (formula 4) between 0.79 and 0.9 at the different time points. The factor time explains about 35% of the individual variability at level 1, estimated as pseudo-R^2 statistic of level 1 residuals (formula 6). Building up the standard model according to formula (5) led to the following final model, and took into account the features required for randomized clinical trials. At any rate the patients of one center should have the same mean baseline values, regardless of the therapy they receive later. Therefore the following model was chosen to mirror this fact.

$$Y_{ij} = \begin{bmatrix} \gamma_{00} + \gamma_{10} * time_{ij} + \gamma_{01} * Centre_i \\ + \gamma_{11} * Centre_i * Therapy_i * time_{ij} \end{bmatrix} + \left[\zeta_{0i} + \zeta_{1i} * time_{ij} + \varepsilon_{ij} \right] \quad (12)$$

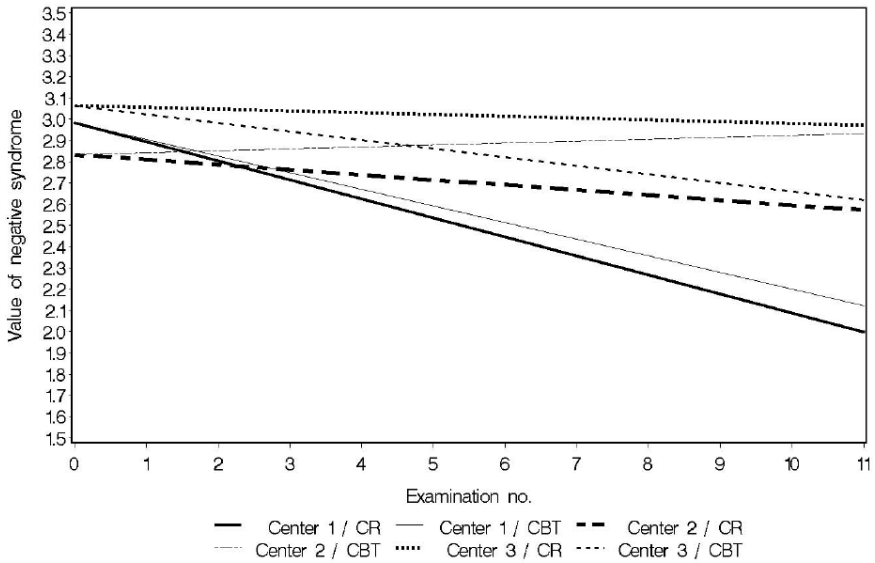

Figure 1: Mean course of negative syndromes in the two study groups within three centers, mirroring the fact that patients from one center should have the same mean baseline values.

No influence of the therapy on the intercept was assumed due to the fact that only the part modeling the slope $(y_{11} * Centre_i * Therapy_i * time_{ij})$ included the factor "therapy". This model explains 34% of the variability of the slopes but only 1.5% of the variability of the intercepts, i.e. of the variability of the groups at level 2, according to formula (7).

As shown in Figure **1**, a strong interaction was found between center and therapy.

After having built this final model, with intercept and slope as random effects, and assuming an unstructured covariance matrix, the model was re-analyzed as a model with repeated measurements for which a Toeplitz covariance structure revealed the best results. Table **1** shows the results and corresponding effect sizes (calculated according to formula 9, 10 and 11) for the primary endpoint "difference in negative symptoms between therapy groups".

CBT revealed the anticipated success in the treatment of negative symptoms only in one center, the overall effect size of CBT over all centers being lower by only 0.12 points for negative symptoms after CR. Therefore the explorative search for subgroups that might especially benefit from this therapy started. No fixed effects were found for characterizing a special subgroup (data not shown). Therefore time varying factors came into the focus. It was assumed that the actual attendance by a patient of the therapy session would influence the success of these therapies. At least one session had to be attended in the time interval between the last and the actual rating of negative symptoms, in order for it to be counted as "treated in this interval". Although it seems obvious that "no therapy" meant "no progression", the model mirroring this assumption did not prove to be the best one with regard to the negative symptoms. The best model according to AIC and BIC criterion was the following,

Table 1: Results and effect sizes for the primary endpoint "difference in negative symptoms between therapy groups" stratified for centers (N1: number in CBT group, N2: number in CR group, STD1: standard deviation in CBT group, STD2: standard deviation in CR group, raw difference between the two groups at end of study, effect size and 95%-confidence limits of the difference between the two groups at end of study according to formula 9, 10 and 11)

Difference	N1	N2	STD1	STD2	Raw Difference	Effect size	Lower limit	Upper limit
center 1, CBT-CR	33	33	0.6593	0.8176	0.1246	0.16580	-0.31755	0.64914
center 2, CBT-CR	33	33	0.8198	0.7611	0.3616	-0.42703	-0.91502	0.060957
center 3, CBT-CR	33	33	0.6877	0.8742	-0.3481	0.45435	-0.03435	0.94306

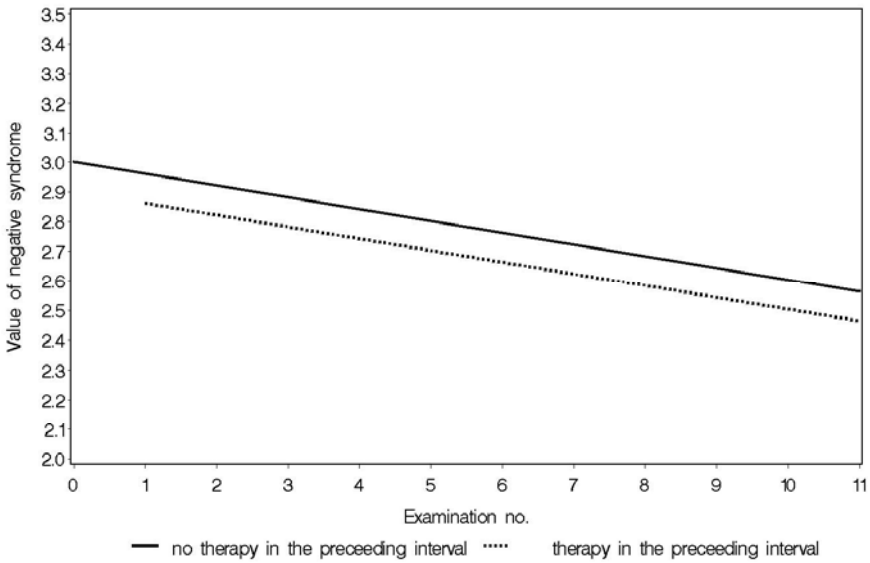

Figure 2: Mean course of negative syndromes of the time variant factor "therapy in the preceding time interval" – mean courses of real patients "switch" from one time interval to the next between these two hypothetical curves.

and included only an influence of the factor "therapy in the preceding time interval" on the intercept.

$$Y_{ij} = \left[\gamma_{00} + \gamma_{10} * time_{ij} + \gamma_{20} * time\_\text{var}iant\_faktor_{ij}\right] + \left[\zeta_{0i} + \zeta_{1i} * time_{ij} + \varepsilon_{ij}\right] \quad (13)$$

Figure **2** shows the resulting curves. Despite the interpretation of time invariant effects like therapy or sex, the curves for time varying effects do not represent real characteristics. There may be patients that visited all their therapy sessions, but there may be patients that visited none of them. The mean patient will switch more or less often between the two curves. Therefore the two curves are only the marginal conditions of the real individual curves. The difference between the two curves is small (0.11 points), and the fact that the curve for "no therapy in the preceding interval" also decreases, reveals a longer lasting effect of previous sessions. Nevertheless, there was no difference between therapies and no explanation for the interaction of centers with therapy groups.

Therefore the therapeutic alliance rating of each patient assessed at the end of each therapy session, was taken into account. The therapeutic alliance was elicited *via* different items ranging from -3 (not at all) to +3 (absolutely) and varied from one assessment to the next. Whether a time varying factor is dichotomous or ordinal does not change the analysis but complicates the interpretation of the results as Figure **3** shows (curves representing negative emotional engagement are missing because this did not occur). Again, the curves only mark the marginal conditions. Due to the fact that the therapeutic alliance may change from one session to the next, each patient has his own real curve, switching between these marginal conditions. Figure **3** reveals that the more successful the therapy, the less positive the patient's alliance rating. But, again, a strong interaction between therapy and center plays a role in the background. This result only holds true for two centers. In the third center, therapeutic alliance did not matter at all, as wide ranging success was observed.

No difference was observed with regard to the composition of the patient groups in a given center which explained the interaction between therapy and center. Due to this fact we assumed, that differences were linked to therapists, because most centers had only one or two therapists. To explore this, a patients' estimations of his or her therapist were assessed over the whole course of the study. A subgroup of patients could be found with at least two of the three following characteristics: emotional engagement always between 0.5 and 2 points, satisfaction with therapy between 0 and 1 point and possibility of co-management of therapy between 0 and 1. This subgroup revealed an overall effect of CBT compared to CB of 0.3 points (see Figure **4**), which is close to the expected 0.35 points. But, again, a strong interaction was found, when integrating the factor "center" into the analysis (see Table **2**). In this subgroup, however, CBT revealed an improvement of negative symptoms for two of the three centers.

The flow chart presented in Figure **5** indicates, how a model can be built, when to check for model

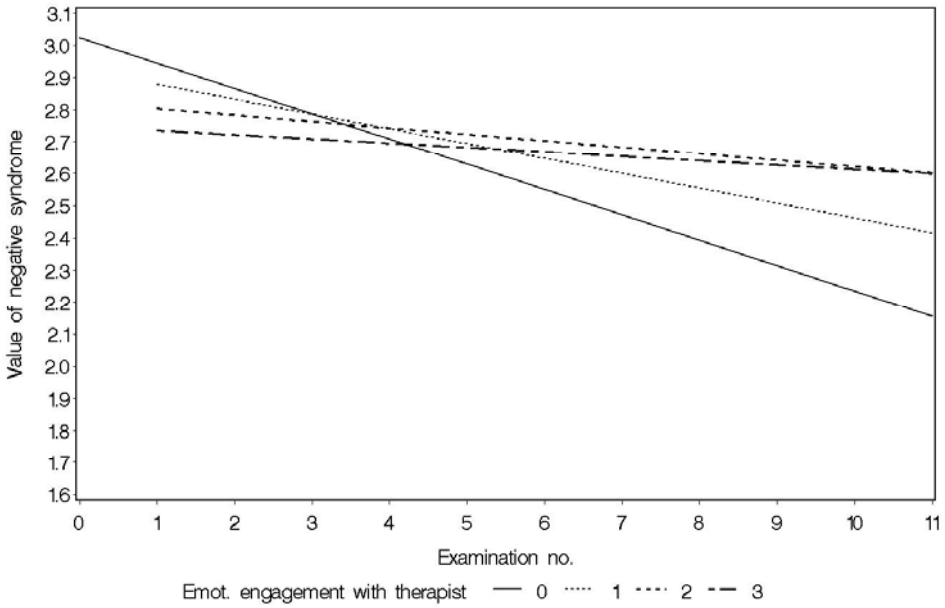

Figure 3: Example for a time variant factor with four possible values. These mean curves also only represent the hypothetical marginal states between which the real patients' mean course switches from one time point to the next.

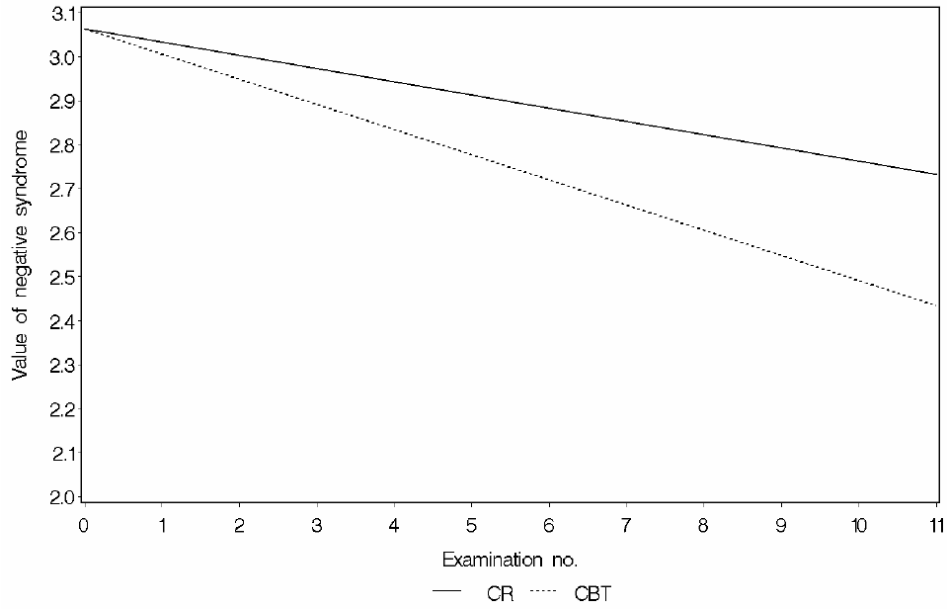

Figure 4: Mean course of negative syndrome for the two study groups within a subgroup of patients, revealing at least two of the three following characteristics: emotional engagement rather steady between 0.5 and 2 points, satisfaction with therapy between 0 and 1 point and opportunity of co-management of therapy between 0 and 1.

Table 2: Results for a subgroup of patients with at least two of the following characteristics regarding "therapeutic alliance": moderate emotional engagement, moderate satisfaction and moderate co-management

Difference	Estimated raw difference at end of study	p value
center 1, CBT-CR	-0.1194	0.5573
center 2, CBT-CR	0.3405	0.1896
center 3, CBT-CR	-0.4985	0.0181

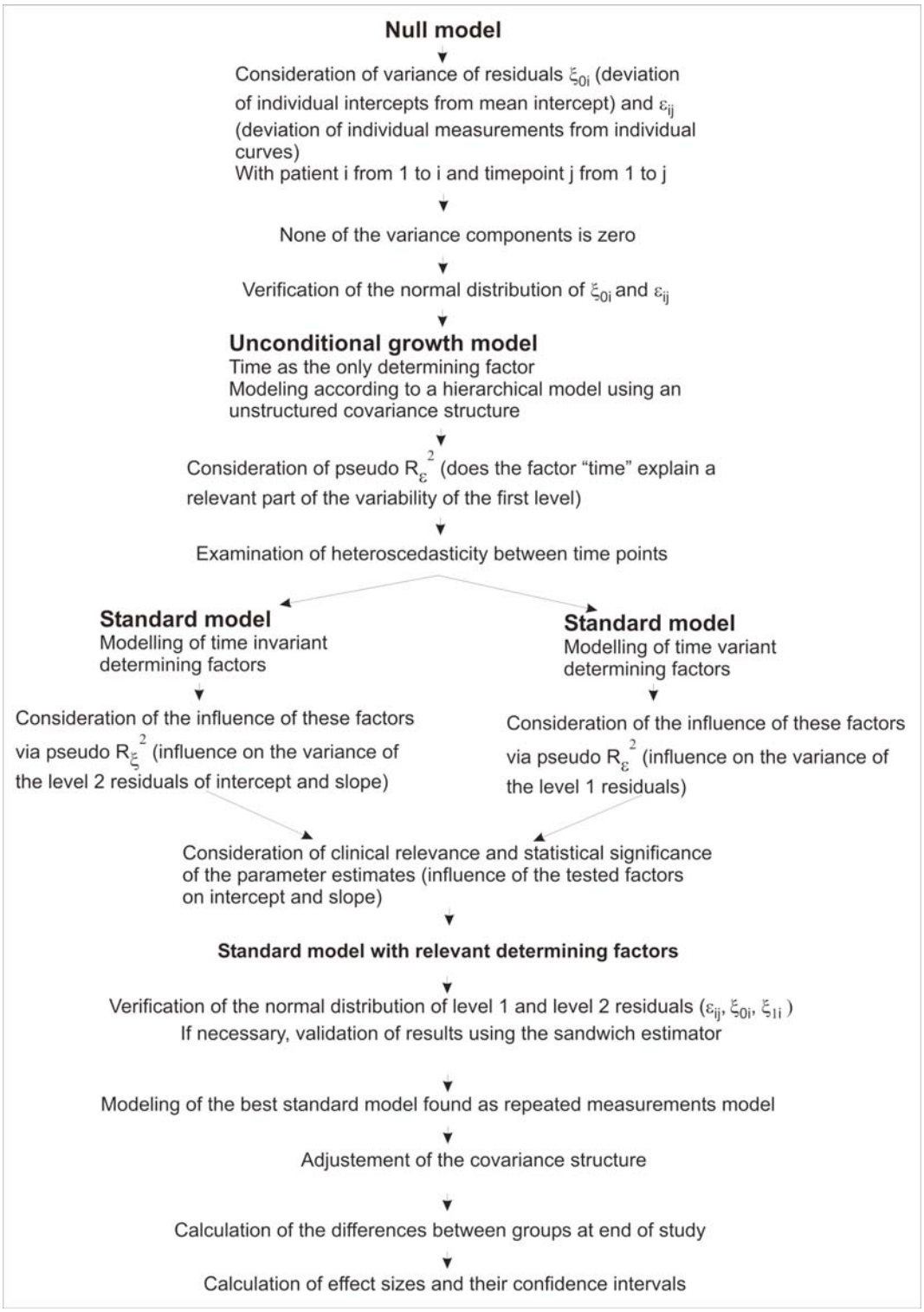

Figure 5: Recommendations for building a linear mixed model.

assumptions, when to include time-invariant and time-variant effects and how to finalize the model.

DISCUSSION

Through the TONES study, we show how linear mixed models can be used in clinical trials. This method offers great advantages as it can be used to analyze time-variant factors and to model specific assumptions like identical intercepts - or even slopes - in subgroups. By adopting the approach of Singer and Willett [23], better insight into model building can be achieved, as compared to starting with a repeated measurements model.

A major advantage of this method is the transparent and statistically specified handling of missing values, being estimated automatically by the method itself and results obtained by incomplete datasets are only slightly different from those obtained with complete datasets [37].

Sophisticated imputation methods, which were again proposed recently [35, 36] are superfluous. By comparing different methods for dealing with missing data, [38] also concluded that the best method involved mixed models because imputations can be dispensed with. Only when the sample size is small, and when there are many time points and a great proportion of missing data, did the multiple imputation method appear to be better than mixed models [38].

When using linear mixed models there is no need for identical intervals between the assessments for all patients, as the method estimates an individual curve for each subject. In the case of individual measurement intervals, only the re-analysis as repeated measurements model and the adjustment of the covariance structure are omitted.

Through linear mixed models, the sample size is minimized as compared to an analysis of covariance, due to the fact that the correlation between the data is taken into account. Although sample size calculation for linear mixed models is not yet implemented in standard software, the equation proposed [2], according to Kish (1987) [39], may be used for a calculation of sufficient precision:

$$n_{eff} = n / \left[1 + (n_{clus} - 1)\rho\right]$$

With n_{clus} = number of measurements per patient, n = number of data collected and ρ= autocorrelation between measurements, the effective sample size can be calculated. According to Willett and Singer (1998) [40] n_{clus} of between 8 to 10 is a good choice for modeling linear courses with an acceptable standard error.

Otherwise the sample size needed for analyzing a specific difference can be calculated by means of covariance analysis in a standard sample size calculation. By converting the equation above, the sample size needed for an analysis with linear mixed models can be calculated.

The use of mixed models in small sample N-of-1-designs was discussed recently [41].

Further fields of application for mixed models were evaluated recently, such as for reliability studies, as [42] demonstrated, according to the proposals of [43, 44] or for their use in discriminant analysis, as proposed [45]. Their application in genome-wide association studies was also been discussed [46, 47] and linear mixed models were also used for the analysis of non-linear data, modeling them for example as exponential growth or fractional polynomial curves, as done [48, 49].

These recent applications still lack user-friendly methods that fullfill all the requirements relevant to the context of planning and analyzing clinical studies.

Through the linear mixed models described here, we demonstrate their applicability for analyzing longitudinal data from clinical trials and show that important requirements in this context, such as sample size considerations and effect sizes, can be met. Thus, it is no longer necessary to apply inappropriate methods, such as covariance analysis or t-tests, at least in terms of linear longitudinal data.

However the GCP-compliance of linear mixed models for the confirmatory analysis of longitudinal data in clinical trials is under discussion. As shown the longitudinal data analysis using linear mixed modelling is a step-by-step procedure where every next step depends on the results of the analysis of the previous step. Therefore an analysis using linear mixed models is more or less data driven. This is in contradiction to the statistical principles of analysing clinical trials in the context of GCP [50] which prefers a detailed statistical analysis plan which must be fixed without any impression from real study data.

ACKNOWLEDGEMENTS

This study was funded publicly by the German Research Foundation (Deutsche Forschungsgemeinschaft, grants Kl 1179/2-1 and Kl 1179/3-1).

We thank Johannes Harbort, Michael Ruch, Hanna Smoltczyk, Maria Weickert for their contribution as study raters, Klaus Hesse, Sabine Kossow, Maria Memmou for their contribution as study therapists, Birgit Conradt for her contribution as psychotherapy supervisor, and Nicole Frommann for her contribution to the treatment manuals.

REFERENCES

[1] Langer W. Mehrebenenanalyse – Eine Einführung in Forschung und Praxis. 2. Aufl. Wiesbaden: Verlag für Sozialwissenschaften 2009.

[2] Hox J. Multilevel Analysis. Techniques and Applications. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc. 2002.

[3] Rogosa D, Brandt D, Zimowski M. A growth curve approach to the measurement of change. Psychol Bull 1982; 90: 726-48. http://dx.doi.org/10.1037/0033-2909.92.3.726

[4] Willett JB. Questions and answers in the measurement of change. In: Rothkopf E, editor. Review of research in education. Washington DC: AERA 1988; pp. 345-422.

[5] Rogosa DR, Willett JB. Understanding correlates of change by modelling individual differences in growth. Psychometrika 1985; 50(2): 203-28. http://dx.doi.org/10.1007/BF02294247

[6] SAS/STAT Software: Changes and enhancements through release 6.12. Cary, NC, USA: SAS Institute Inc. 1997.

[7] Bender S, Olbrich HM, Fischer W, Hornstein C, Schoene W, Falkai P, *et al*. Antipsychotic efficacy of the antidepressant trimipramine: a randomized, double-blind comparison with the phenothiazine perazine. Pharmacopsychiatry 2003; 36: 61-69. http://dx.doi.org/10.1055/s-2003-39043

[8] Schooler N, Rabinowitz J, Davidson M, Emsley R, Harvey PD, Kopala L, *et al*. Risperidone and haloperidol in first-episode psychosis: a long-term randomized trial. Am J Psychiatry 2005; 162(5): 947-53. http://dx.doi.org/10.1176/appi.ajp.162.5.947

[9] Lerner V, Miodownik C, Kaptsan A, Cohen H, Loewenthal U, Kotler M. Vitamin B6 as add-on treatment in chronic schizophrenic and schizoaffective patients: a double-blind, placebo-controlled study. J Clin Psychiatry 2002; 63(1): 54-58. http://dx.doi.org/10.4088/JCP.v63n0111

[10] Amiri A, Noorbala AA, Nejatisafa AA, Ghoreishi A, Derakhshan MK, Khodaie-Ardakani MR, *et al*. Efficacy of selegiline add on therapy to risperidone in the treatment of the negative symptoms of schizophrenia: a double-blind randomized placebo-controlled study. Hum Psychopharmacol Clin Exp 2007; 23: 79-86. http://dx.doi.org/10.1002/hup.902

[11] Wong M, Oakley SP, Young L, Jiang BY, Wierzbicki A, Panayi G, *et al*. Infliximab improves vascular stiffness in patients with rheumatoid arthritis. Ann Rheum Dis 2009; 68: 1277-84. http://dx.doi.org/10.1136/ard.2007.086157

[12] Rogers BA, David LA, Briggs TWR. Sequential outcome following autologous chondrocyte implantation of the knee: A six-year follow-up. Int Orthop (SICOT) 2010; 34: 959-64. http://dx.doi.org/10.1007/s00264-009-0842-x

[13] Nakonezny PA, Byerly MJ, Rush AJ. The relationship between serum prolactin level and sexual functioning among male outpatients with schizophrenia or schizoaffective disorder: a randomized double-blind trial of risperidone vs. quetiapine. J Sex Marital Therapy 2007; 33: 203-16. http://dx.doi.org/10.1080/00926230701267829

[14] Centorrino F, Meyers AL, Ahl J, Cincotta SL, Zun L, Gulliver AH, *et al*. An observational study of the effectiveness and safety of intramuscular olanzapine in the treatment of acute agitation in patients with bipolar mania or schizophrenia/ schizoaffective disorder. Hum Psychopharmacol Clin Exp 2007; 22: 455-62. http://dx.doi.org/10.1002/hup.870

[15] Kane JM, Osuntokun O, Kryzhanovskaya LA, Wen Xu, Stauffer VL, Watson SB, Breier A. A 28-week, ranodmized, double-blind study of olanzapine versus aripiprazole in the treatment of schizophrenia. JCP 2009; 70(4): 572-81.

[16] Pienkowski D, Resig JA, Talwalkar V, Tylkowski C. Novel three-dimensional MRI technique for study of cartilaginous hip surfaces in legg-calve-perthes disease. J Orthop Res 2009; 27: 981-88. http://dx.doi.org/10.1002/jor.20909

[17] Litt MD, Shafer DM, Kreutzer DL. Brief cognitive-behavioral treatment for TMD pain: Long-term outcomes and moderators of treatment. PAIN 2010; 151: 110-16. http://dx.doi.org/10.1016/j.pain.2010.06.030

[18] DuHamel KN, Mosher CE, Winkel G, Labay LE, Rini C, Meschian YM, *et al*. Randomized clinical trial of telephone-administered cognitive-behavioral therapy to reduce post-traumatic stress disorder and distress symptoms after hematopoietic stem-cell transplantation. J Clin Oncol 2010; 28: 3754-61. http://dx.doi.org/10.1200/JCO.2009.26.8722

[19] Csernansky JG, Mahmoud R, Brenner R. A comparison of risperidone and haloperidol for the prevention of relapse in patients with schizophrenia. N Engl J Med 2002; 346(1): 16-22. http://dx.doi.org/10.1056/NEJMoa002028

[20] Arato M, O'Connor R, Meltzer HY. A 1-year, double-blind, placebo-controlled trial of ziprasidone 40, 80 and 160 mg/day in chronic schizophrenia: the ziprasidone extended use in schizophrenia (ZEUS) study. Int Clin Psychopharm 2002; 17(5): 207-15. http://dx.doi.org/10.1097/00004850-200209000-00001

[21] Lipkovich I, Deberdt W, Csernansky JG, Buckley P, Peuskens J, Kollak-Walker S, *et al*. Predictors of risk for relapse in patients with schizophrenia or schizoaffecitve disorder during olanzapine drug therapy. J Psychiatr Res 2007; 41: 305-10. http://dx.doi.org/10.1016/j.jpsychires.2006.07.016

[22] Celebi N, Tsouraki R, Engel C, Lammerding-Köppel M, Baur P, Riessen R, Weyrich P. Does doctors' workload impact the supervision and ward activities of final-year medical students? A prospective study. BMC Med Educ 2012; 12: 24. http://dx.doi.org/10.1186/1472-6920-12-24

[23] Singer JD, Willet JB. Applied longitudinal data analysis. Modelling change and event occurrence. New York: Oxford University Press 2003. http://dx.doi.org/10.1093/acprof:oso/9780195152968.001.0001

[24] Klingberg S, Wittorf A, Herrlich J, Wiedemann G, Meisner C, Buchkremer G, *et al*. Cognitive behavioural treatment of negative symptoms in schizophrenia patients: study design of the TONES study, feasibility and safety of treatment. Eur Arch Psych Clin N 2009; 259(Suppl 2): 149-54. http://dx.doi.org/10.1007/s00406-009-0047-8

[25] Klingberg S, Wölwer W, Engel C, Wittorf A, Herrlich J, Wölwer W, *et al*. Negative symptoms of schizophrenia as primary target of cognitive behavioral therapy: results of the randomized clinical TONES study. Schizophrenia Bull 2011; 37(Suppl 2): S98-10. http://dx.doi.org/10.1093/schbul/sbr073

[26] Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophrenia Bull 1987; 13(2): 262-76. http://dx.doi.org/10.1093/schbul/13.2.261

[27] Klingberg S, Wittorf A, Wiedemann G. Disorganization and cognitive impairment in schizophrenia: independent symptom dimensions? Eur Arch Psych Clin N 2006; 256: 532-40. http://dx.doi.org/10.1007/s00406-006-0704-0

[28] Martin DJ, Garske JP, Davis MK. Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. JCCP 2000; 68: 438-50.

[29] Flückinger C, Regli D, Zwahlen D, Hostettler S, Caspar F. Der Berner Patienten- und Therapeutenstundenbogen 2000. Ein Instrument zur Erfassung von Therapieprozessen. Z Klin Psychol Psychother 2010; 39 (2): 71-79. http://dx.doi.org/10.1026/1616-3443/a000015

[30] Gruber J. Econometrics part 1, Introduction in multiple regression and econometrics. Hagen: University 1982.

[31] Verbeke G, Lesaffre E. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. Comput Stat Data An 1997; 23: 541-56. http://dx.doi.org/10.1016/S0167-9473(96)00047-3

[32] Maas CJM, Hox JJ. Robustness issues in multilevel regression analysis. Stat Neerl 2004; 58(2): 127-37. http://dx.doi.org/10.1046/j.0039-0402.2003.00252.x

[33] Hedges LV, Olkin I. Statistical methods for meta-analysis. London: Academic Press 1985.

[34] Feingold A. Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. Psychol Methods 2009; 14(1): 43-53. http://dx.doi.org/10.1037/a0014699

[35] Gunnes N, Seierstad TG, Aamdal S, Brunsvig PF, Jacobsen AB, Sundstrøm S, Aalen OO. Assessing quality of life in a randomized clinical trial: Correcting for missing data. BMC Med Res Methodol 2009; 9: 28. http://dx.doi.org/10.1186/1471-2288-9-28

[36] Aalen OO, Gunnes N. A dynamic approach for reconstructing missing longitudinal data using the linear increments model. Biostatistics 2010; 11(3): 53-72. http://dx.doi.org/10.1093/biostatistics/kxq014

[37] Twisk JWR. Applied longitudinal data analysis for epidemiology. Cambridge: University Press 2003.

[38] Elobeid MA, Padilla MA, McVie Th, Thomas O, Brock DW, Musser B, *et al*. Missing data in randomized clinical trials for weight loss: Scope of the problem, state of the field, and performance of statistical methods. PLoS ONE 2009; 4(8): e6624. http://dx.doi.org/10.1371/journal.pone.0006624

[39] Kish L. Statistical design for research. New York: Wiley 1987. http://dx.doi.org/10.1002/0471725196

[40] Willett JB, Singer JD. The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. Dev Psychopathol 1998; 10: 395-26. http://dx.doi.org/10.1017/S0954579498001667

[41] Ridenour TA, Hall DL, Bost JE. A small sample randomized clinical trial methodology using N-of-1 designs and mixed model analysis. Am J Drug Alcohol Abuse 2009; 35: 260-66. http://dx.doi.org/10.1080/00952990903005916

[42] Laenen A, Alonso A, Molenberghs G, Vangeneugden T. Reliability of a longitudinal sequence of scale ratings, Psychometrika 2009; 74: 49-64. http://dx.doi.org/10.1007/s11336-008-9079-7

[43] Laenen A, Alonso A, Molenberghs G. A measure for the reliability of a rating scale based on longitudinal clinical trial data. Psychometrika 2007; 72: 443-48. http://dx.doi.org/10.1007/s11336-007-9002-7

[44] Laenen A, Alonso A. The functional living index-cancer: estimating its reliability based on clinical trial data. Qual Life Res 2010; 19: 103-109. http://dx.doi.org/10.1007/s11136-009-9568-x

[45] Komárek A, Hansen BE, Kuiper EMM, van Buuren HR, Lesaffre E. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. Stat Med 2010; 29: 3267-83. http://dx.doi.org/10.1002/sim.3849

[46] Zhang Z, Buckler ES, Casstevens TM, Bradbury PJ. Software engineering the mixed model for genome-wide association studies on large samples. Brief Bioinform 2009; 10(6): 664-75. http://dx.doi.org/10.1093/bib/bbp050

[47] Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, *et al*. A unified mixed model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 2006; 38: 203-208. http://dx.doi.org/10.1038/ng1702

[48] Subtil F, Rabilloud M. Robust non-linear mixed modelling of longitudinal PSA levels after prostate cancer treatment. Stat Med 2010; 29: 573-87.

[49] Long J, Ryoo J. Using fractional polynomials to model non-linear trends in longitudinal data. Br J Math Stat Psychol 2010; 63: 177-203. http://dx.doi.org/10.1348/000711009X431509

[50] ICH Harmonised tripartied guideline: Statistical principles for clinical trials, E9. Washington: Federal Register, 16 September 1998; 63(179) 49583-49598.

Received on 06-08-2013 Accepted on 10-11-2013 Published on 11-11-2013

http://dx.doi.org/10.6000/1929-6029.2013.02.04.7