Robustness of the linear mixed model to misspecied error distribution

Hélène Jacqmin-Gadda, Solenne Sibillot, Cécile Proust, Jean-Michel Molina, Rodolphe Thiébaut

Robustness of the linear mixed model to misspecified error distribution

Hélène Jacqmin-Gadda[1,2,4], Solenne Sibillot[1,2], Cécile Proust[1,2], Jean-Michel Molina[3], Rodolphe Thiébaut[1,2]

April 26, 2006

Authors' affiliations :

[1] Institut National de la Santé et de la Recherche Médicale, Equipe de biostatistique E0338, Bordeaux, France.

[2] Université Victor Segalen Bordeaux II, Bordeaux, France

[3] Département de maladies infectieuses, Assistance Publique Hôpitaux de Paris, Hôpital Saint Louis, Paris France

[4]**Corresponding author** : Hélène Jacqmin-Gadda, INSERM E0338, ISPED, case 11, 146 rue Léo Saignat, 33076 Bordeaux cedex, France. Tel: (33) 5 57 57 45 18; Fax (33) 5 56 24 00 81; e-mail: helene.jacqmin-gadda@bordeaux.inserm.fr

Abstract:

A simulation study is performed to investigate the robustness of the maximum likelihood estimator of fixed effects from a linear mixed model when the error distribution is misspecified. Inference for the fixed effects under the assumption of independent normally distributed errors with constant variance is shown to be robust when the errors are either non-gaussian or heteroscedastic, except when the error variance depends on a covariate included in the model with interaction with time. Inference is impaired when the errors are correlated. In the latter case, the model including a random slope in addition to the random intercept is more robust than the random intercept model. The use of Cholesky residuals and conditional residuals to evaluate the fit of a linear mixed model is also discussed.

Keywords:mixed model, robustness, random-effect, misspecification, maximum likelihood estimator

1 Introduction

The linear mixed model (Laird and Ware, 1982) is widely used for the analysis of longitudinal continuous data because it takes correlation between repeated measures into account and the maximum likelihood estimators are easily obtained using standard softwares. In longitudinal studies, the growth curve model including two random effects (intercept and slope) normally distributed and an independent gaussian error is probably the most routinely used to study change over time of a quantitative outcome. In most of these studies, the focus is on the fixed effects estimates which measure change over time and association of covariates with it; the random effects are only included to obtain reliable inference on the fixed effects by taking intra-subject correlation into account. Despite its wide use, consequences of misspecifying assumptions of the linear mixed model are not well known.

Apart from the specification of the expectation of the outcome, the key assumptions for the standard linear mixed model concern the specification of the covariance structure and of the error distribution :

(i) normality of the random effects distribution

(ii) independency of the response given the random effects i.e. independency of the errors

(iii) normality of the error

(iv) homoscedasticity of the error

Several studies have shown that maximum likelihood inference on fixed effects is robust to non-gaussian random effects distribution (Butler and Louis 1992; Verbeke and Lesaffre, 1997; Zhang and Davidian 2001). Some results also suggest robustness to misspecification of the covariance structure. First, Liang and Zeger (1986) have demonstrated convergence of fixed effects estimates obtained by Generalized Estimating Equations (GEE) whatever the working covariance matrix. Given that, for the linear model, estimating equations obtained by derivation of the maximum likelihood are identical (except for the covariance estimator) to GEE with appropriate covariance structure, this result demonstrates convergence of MLE for fixed effects in the linear mixed model when the covariance structure is not correct. On the other hand, Liang et Zeger (1986) demonstrated that variance estimates of fixed effects may be biased when the covariance structure is not correct and they recommend the use of the robust sandwich estimate (Royall, 1986). However this robust estimate may be instable for small sample sizes.

For the balanced and complete case, Lange and Laird (1989) have demonstrated that, in general, the variance of MLE of fixed effects depends strongly on the assumed covariance structure but the linear growth curve model with two random effects lead to unbiased variance estimates of fixed effects even if the true covariance structure implies more random effects. In a simulation study using the same linear growth curve model, Taylor, Cumberland and Sy (1994) have also found that the coverage rate of confidence intervals for fixed effects was not impaired when the true covariance was underlain by a quadratic random effects model, an integrated Ornstein-Uhlenbeck process or a Brownian motion. Thus the above results suggest that the linear growth curve model with random intercept and slope leads to robust inference for the fixed effects when the covariance structure is misspecified.

The two last assumptions, normality and homoscedasticity of the error, have been less studied. As maximum likelihood estimates of fixed effects are equivalent to weighted least square estimates with a weighting depending on the estimated covariance matrix, one can expect robustness of the LMM. However, in practice, it is frequent to analyse a transformation

of the outcome of interest, such as logarithm (Tsiatis, Degruttola and Wulfsohn, 1995), square root (McNeil and Gore, 1996; Jacqmin-Gadda et al, 1997) or fourth root (Taylor and Law, 1998), to move closer to normality or constant variance assumptions. Major drawbacks of these transformations are that results may depend on the transformation used and they are more difficult to understand particularly for non statisticians. Consequently, if the robustness of LMM may be assessed, it would generally be better to analyse data in their natural scale.

The aim of this paper is thus to study by simulation, sensitivity of inference for the fixed effects from a LMM when the error distribution is misspecified that is when the error is correlated, non gaussian or heteroscedastic. Results for the random intercept model and for the model with random intercept and slope are compared. The next section presents the linear mixed model and model diagnostic procedures. Analysis of data from the ALBI clinical trials (Molina et al, 1999) is described in section 3 to illustrate the problem and the simulation study is presented in section 4.

2 Linear mixed Model

2.1 Model

Assuming data come from a longitudinal study, we denote Y_{ij} the outcome for subject i, $i = 1, ..., N$, measured at occasion j, $j = 1, ..., n_i$. The general linear mixed model (Laird and Ware, 1982) has the following form :

$$Y_{ij} = X_{ij}^t \beta + Z_{ij}^t \alpha_i + e_{ij}, \tag{1}$$

where X_{ij} is a p-vector of covariates including time, Z_{ij} is a q-sub-vector from X_{ij}, β is the vector of fixed effects and α_i the vector of random effects assumed to be normally distributed with mean 0 and covariance matrix G. In the more general model, the measurement error, which is assumed to be gaussian, may be correlated with a covariance matrix denoted by Σ_i. Random effects α_i and errors e_{ij} are always assumed to be independent. Model (1) may be equivalently written using the marginal formulation for the n_i vector Y_i of reponses for subject i :

$$Y_i \sim N(X_i\beta, V_i = Z_i G Z_i^t + \Sigma_i), \tag{2}$$

with X_i (and Z_i) the design matrix $n_i \times p$ with row X_{ij}^t (and $n_i \times q$ with row Z_{ij}^t respectively). We focus on the case where the error is assumed independently distributed : $\Sigma_i = \sigma_e^2 I_{n_i}$.

More specifically, the growth curve model with random intercept and slope may be written :

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \alpha_{0i} + \alpha_{1i} t_{ij} + e_{ij}, \tag{3}$$

where t_{ij} is the time of measurement for subject i at occasion j and $\alpha_{0i} \sim N(0, \sigma_0^2)$, $\alpha_{1i} \sim N(0, \sigma_1^2)$ and the error e_{ij} are independently distributed with $e_{ij} \sim N(0, \sigma_e^2)$. The uniform correlation model which includes a single random effect (the intercept) is defined by deleting $\alpha_{1i} t_{ij}$ from (3).

2.2 Estimation

We denote by θ the vector including all the parameters to be estimated, that is the fixed parameters β and the vector of covariance parameters denoted ϕ: $\phi^T = (Vec(G), \sigma_e^2)$. Under

the gaussian assumption (2), the vector θ is estimated by maximisation of the following log-likelihood :

$$l(\theta) = -\frac{1}{2}\sum_{i=1}^{N}\left\{n_i log(2\pi) + log|V_i| + (Y_i - X_i\beta)^T V_i^{-1}(Y_i - X_i\beta)\right\}. \quad (4)$$

Solving the score equations $\partial l(\theta)/\partial\theta = 0$, the maximum likelihood estimates for the fixed effects may be obtained by :

$$\hat{\beta}(\hat{\phi}) = (X^T V(\hat{\phi})^{-1} X)^{-1} X^T V(\hat{\phi})^{-1} Y, \quad (5)$$

which are identical to the weighted least square estimates with $V(\hat{\phi})^{-1}$ as weighting matrix. Thus unbiaseness of $\hat{\beta}$ requires only that $E(Y) = X^T\beta$ (Liang and Zeger, 1986). On the other hand, MLE of variance parameters ϕ have no closed form solution and must be estimated iteratively by maximizing $l(\phi, \hat{\beta})$ obtained by replacing β by $\hat{\beta}$ given by (5) in (4). Thus, unlike convergence of $\hat{\beta}$, convergence of $\hat{\phi}$ relies on correct specification of the covariance structure and on gaussian assumptions.

The asymptotic covariance matrix of the MLE is estimated by the inverse of the observed Hessian matrix at the optimum $-\partial l^2(\theta)/\partial\theta\partial\theta^t$. Given that $E(\partial l^2(\theta)/\partial\beta\partial\phi) = 0$, MLE of β and ϕ are asymptotically independent, and $var(\hat{\beta})$ may be estimated by :

$$Var(\hat{\beta}) = -(\partial l^2(\theta)/\partial\beta\partial\beta^T)^{-1} = (X^T V(\hat{\phi})^{-1} X)^{-1}, \quad (6)$$

which can also be directly derived from (5).

Strictly, convergence of the above estimates of Var($\hat{\beta}$) relies on correct specification of the model (2). When the covariance matrix or the random effects distribution may be misspecified, it is recommended to use sandwich-type robust variance estimator (Liang and Zeger, 1986; Verbeke and Lesaffre, 1997). However Verbeke and Lesaffre (1997) have shown in a simulation study that the difference between the model-based standard error of the fixed effects and the robust one is negligible even if the random effects distribution is misspecified and thus that model-based inference for the fixed effects remains valid. In the following, we investigate robustness of model-based inference for the fixed effects when the error distribution is misspecified.

2.3 Goodness of fit analysis

As residuals of LMM are correlated, they can not be directly used to check model assumptions. However, uncorrelated residuals may be obtained by using Cholesky decomposition of V_i^{-1}. Let us denote L_i the triangular matrix obtained by Cholesky decomposition of V_i^{-1} so that $V_i^{-1} = L_i L_i^t$, the Cholesky residuals are given by :

$$R_i = L_i^t(Y_i - X_i\hat{\beta}).$$

If the model is correct and neglecting variability of $\hat{\beta}$, R_i is approximately $N(0, I_{n_i})$. In our experience, based on a simulation study with moderate sample size (N=50 n_i=5, results not shown), the correction of the variance of the residuals to take variance of $\hat{\beta}$ into account was unecessary. Graphical evaluation of departure from the standard normal distribution using QQ-plot of R_{ij} is often sufficient. Normality may also be tested using, for instance, the Shapiro-Francia test (Royston, 1993). However, the goodness-of-fit tests are more powerful when the

sample size is large that is when the estimators are robust. Thus, it may be more useful to quantify departure from normality using a measure relatively independent from the sample size such as the V' statistic proposed by Royston (1993). As the Cholesky residuals are linear combinations of several observation-specific residuals $Y_{ij} - X_{ij}\hat{\beta}$, they cannot be plotted against $X_{ij}\beta$ or any element from X_{ij} (such as t_{ij}) to further explore potential misspecifications.

Park and Lee (2004) have proposed to summarize residuals for each subject using $(Y_i - X_i\hat{\beta})V_i(\hat{\phi})^{-1}(Y_i - X_i\hat{\beta})$ which, when model assumptions are correct, is χ_2 distributed with n_i degrees of freedom. These residuals are useful to detect influencial subjects or misspecification associated with a subject-specific covariate but they are little sensitive to departure from the model depending on time and, in our experience, less sensitive to departure from gaussian assumption than Cholesky residuals.

To explore constant variance assumption for the error, the conditional residuals Rc_i may be computed :

$$Rc_i = Y_i - X_i\hat{\beta} - Z_i\hat{\alpha}_i.$$

Using the empirical Bayes estimates of the random effects, we found that :

$$\begin{aligned} Rc_i &= Y_i - X_i\hat{\beta} - Z_iGZ_i^TV_i^{-1}(Y_i - X_i\hat{\beta}) \quad (7)\\ &= \sigma_e^2V_i^{-1}(Y_i - X_i\hat{\beta}), \quad (8)\end{aligned}$$

and thus, neglecting as above $var(\hat{\beta})$, the distribution of Rc_i when the model is correct is:

$$Rc_i \sim N(0, \sigma_e^4V_i^{-1}). \quad (9)$$

This result shows that the covariance matrix of the conditional residuals is not diagonal even if the residual error is assumed to be independent ($e_i = Y_i - X_i\beta - Z_i\alpha_i \sim N(0, \sigma_e^2I)$). This suggests also that $var(Rc_{ij})$ may depend on t_{ij} when V_i depends on t_i such as in the model including a random slope. For instance, in the growth curve model (3), the elements of V_i are $cov(Y_{ij}, Y_{ij'}) = \sigma_0^2 + \sigma_1^2t_{ij}t_{ij'} + \sigma_e^2I_{\{j=j'\}}$. However, in many applications when $\hat{\sigma}_1^2t_{ij}^2$ remains small compared to σ_0^2 and σ_e^2, one can roughly consider that $var(Rc_{ij})$ is time independent. In such a case, plotting the conditional residuals Rc_{ij} against $E(Y_{ij}|\hat{\alpha}_i)$ (or against X_{ij}) allows to explore heteroscedasticity of the error.

3 Analysis of the ALBI data set

3.1 Analysis of crude data

The randomized clinical trial ALBI-ANRS 070 compared three anti-retroviral treatment regimens in HIV-1-infected adults naive of anti-retroviral therapy : stavudine (d4T) plus didanosine (ddI) for 24 weeks, zidovudine (AZT) plus lamivudine (3TC) for 24 weeks and a switching group with the d4T+ddI regimen for 12 weeks followed by the AZT+3TC regimen for 12 weeks (Molina et al, 1999). In this analysis, we compared the d4T+ddi group (50 patients) with the AZT+3TC group (50 patients). Measures of CD4+ count were collected at baseline and then approximately every 4 weeks til week 24. Because of missing data, the mean number of measurements by subject was 6.

To estimate change over time of CD4+, we fitted first the growth curve model defined by (5) with the addition of a treatment effect ($X_i = 1$ for the D4T+DDi group and 0 for AZT+3TC)

both on the baseline level and on the slope :

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 X_i + \beta_3 X_i \times t_{ij} + \alpha_{0i} + \alpha_{1i} t_{ij} + e_{ij}, \tag{10}$$

where Y_{ij} was the CD4+ measurement in their natural scale divided by 100, $\alpha_i^t = (\alpha_{0i}, \alpha_{1i}) \sim N(0, G)$ with G diagonal with elements σ_0^2 and σ_1^2 and $e_{ij} \sim N(0, \sigma_e^2)$. The time unit was 100 days so that the time variable range from 0 to 2. Results are presented in the first column of table 1.

Table 1 : Parameter estimates and standard-error of the linear mixed model (10) using the ALBI data set.

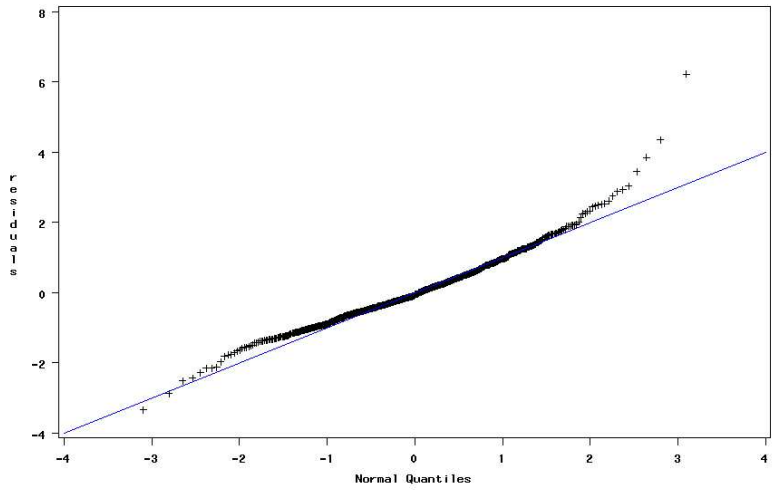
Parameter	Model for CD4/100		Model for $CD4^{1/4}$	
	estimate	SE	estimate	SE
β_0	4.75	0.194	4.62	0.045
β_1	0.27	0.085	0.07	0.021
β_2	-0.46	0.275	-0.10	0.063
β_3	0.27	0.122	0.07	0.029
σ_0^2	1.50	0.151	0.08	0.017
σ_1^2	0.02	0.084	0.002	0.008
σ_e^2	0.77	0.038	0.04	0.004

3.2 Goodness of fit analysis

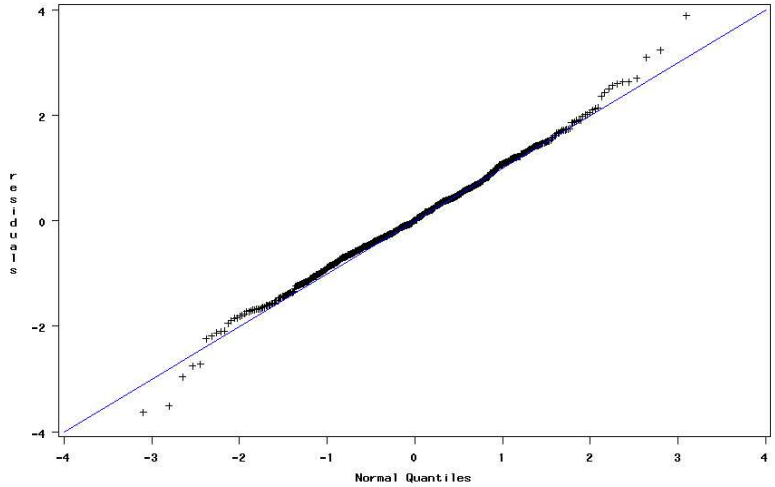
To check model assumptions, the QQ-plot of Cholesky residuals was drawn on figure 1a. It exhibits deviations from the normality assumption. Departure from the normal distribution was confirmed by the Shapiro-Francia test ($p<0.001$) and the high value of the Royston's V' statistic for normality (V'=16.7). All these results suggested model assumptions were not adequate for these data.

In this sample, the variance of the random slope σ_1^2 was small compared to the variance of the random intercept σ_0^2 and of the error σ_e^2 and the time variable was always less than 2. Thus, we could consider the variance of the estimated conditional residuals Rc_{ij} as approximately constant under correct specification of the model. However, the plot of the conditional residuals Rc_{ij} against the estimated conditional expectation exhibit a typical cone shape suggesting heteroscedasticity of the error (figure 2a).

Figure 1 : QQ-plot of the Cholesky residuals from the linear mixed model (11) estimated on the ALBI data set :
(a) CD4 in their natural scale
(b) $CD4^{1/4}$

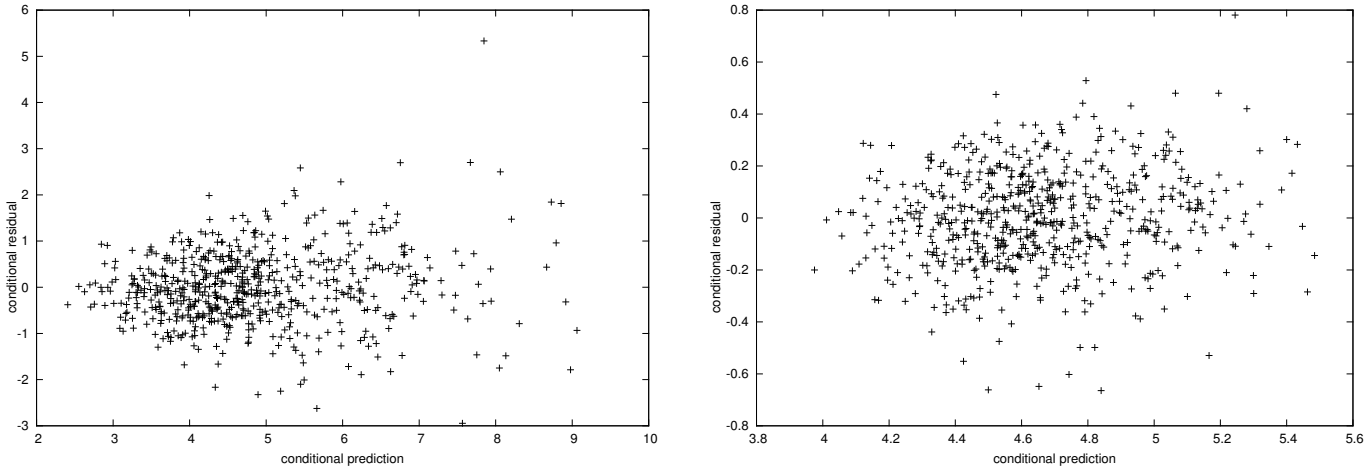


(a)



(b)

Figure 2 : Conditional residuals from the linear mixed model (11) estimated on the ALBI data set versus predictions including the random effects.
(a) CD4 in their natural scale
(b) $CD4^{1/4}$



(a) (b)

3.3 Analysis of transformed data

We tried several transformations of CD4 and found the fourth root led to the best results for adequacy of model (10) as it can be seen on the plot of conditional residuals (figure 2b) and the QQ-plot of Cholesky residuals (figure 1b). With this transformation the value of the Royston criteria was dramatically decreased to 2.5 (even if the Shapiro-Francia test was still significant with p=0.02). Parameter of model (10) estimated on the transformed data are displayed in table 1. Both models exhibit a treatment effect on CD4 counts change, a higher slope for the D4T+DDI group, but interpretation of parameter values is more difficult when the transformed variable is used. Futhermore, calculation of variance for the backtransformed prediction is not direct. In addition, adjustment to the normal distribution is not yet perfect and questions arise on the robustness of the model. Thus, it is interesting to investigate robustness of the linear mixed model to heteroscedasticity and other departures from the model to evaluate the need for such variable transformation.

4 Design of simulation

4.1 Overview

We carried out a simulation study to investigate robustness of parameter estimates from the LMM (1) with $\Sigma_i = \sigma_e^2 I_{n_i}$ when the errors e_{ij} are either correlated or non gaussian or of non constant variance. We compared robustness of the random intercept model (denoted RI) and of the model with random intercept and slope (denoted RIS). In the latter case, the estimated model was given by (10) with $\alpha_{0i} \sim N(0, \sigma_0^2)$, $\alpha_{1i} \sim N(0, \sigma_1^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. The random intercept model was defined by deleting $\alpha_{1i} t_{ij}$ from (10). This model allows to evaluate impact of misspecification on both coefficients for the time-dependent variables (β_1 and β_3) and for a time-independent covariate β_2. Optimisations were performed using a Newton-Raphson

like algorithm (Marquardt, 1963) and positivity constraints on the variance parameters were satisfied by using a square root transformation.

4.2 Sampling scheme

The sample size N was 50 or 200 and three intra-subject sampling schemes were used :

- sparse balanced : n_i=5 measures by suject at time 0,0.5,1,1.5,2
- intensive balanced : n_i=9 measures by suject at time 0,0.25,0.5,0.75,1,1.25,1.5,1.75,2
- unbalanced : n_i randomly selected between 3 and 7 and $t_{ij} = 0, 0.5, 1, ...0.5 * (n_i - 1)$

Thus, in the three cases, the mean follow-up time was 2 and the mean time of measurement was 1.

4.3 Models for data generation

Data were generated using either the RIS or RI model with different modifications detailed below. Parameter values were those obtained when model (10) was estimated on the ALBI data set using the CD4 in their natural scale (see table 1). However, two values of σ_e^2 were used (0.81 and 1.62) because the impact of a misspecifyed error distribution may depend on the variance of the error compared to the other sources of variation in the data. In fact, as in most simulated cases, results were very similar for both σ_e^2 values, we present only results for σ_e^2=1.62 except for the correlated error. Complete results are available upon request from the first author. The binary covariate X_i was generated using a Bernouilli distribution with probability p=0.5.

We used the random number generator Rcarry adapted fom James (1990) for U(0,1) variables and Norran adapted from Marsaglia and Tsang (1984) for the N(0,1) variables.

Standard linear mixed model

Data were generated by the correct model (RIS or RI) to validate the estimation method.

Heteroscedastic error

Three cases were studied.

- Variance depending on the conditional expectation : $Var(e_{ij}) = \sigma_e^2 + aE(Y_{ij}|\alpha_i)^4$. Parameters value were chosen to obtain Royston V' criterion for departure from normality around the value observed on the ALBI data set. We chose $\sigma_e^2 = 0.1$ and $a = 0.001$ so that V'=17.7 with $var(e_{ij}) = 1$ in the samples with N=200 and n_i=5.
- Increasing variance with time : $Var(e_{ij}) = \sigma_e^2 + at_{ij}$.
  with $\sigma_e^2 = 0.81$ and $a = 0.81$ ($E\{var(e_{ij})\} = \sigma_e^2 + aE(t_{ij}) = 1.62$).
- Variance depending on the covariate : $Var(e_{ij}) = \sigma_e^2 + aX_i$
  with $\sigma_e^2 = 0.82$ and $a = 1.60$ ($E\{var(e_{ij})\} = \sigma_e^2 + a/2 = 1.62$).

Correlated error

Data were generated using RIS or RI with correlated error including an autoregressive structure :

$$e_{ij} = w_{ij} + \epsilon_{ij},$$

where w_{ij} is a gaussian random variable with mean 0, constant variance σ_w^2 and covariance $\sigma_w^2 exp(-\gamma|t_{ij} - t_{ik}|)$ and the independent error ϵ_{ij} is iid $N(0, \sigma_\epsilon^2)$. Variance parameters values

were either $\sigma^2_\epsilon = 0.25$ and $\sigma^2_w = 0.56$ ($var(e_{ij}) = 0.81$) or $\sigma^2_\epsilon = 0.5$ and $\sigma^2_w = 1.12$ ($var(e_{ij}) = 1.62$). In both cases, $\gamma = 0.64$ so that $corr(e_{ij}, e_{ij'}) = 0.5$ if $|t_{ij} - t_{ij'}| = 0.5$ (the minimum in the unbalanced case) and $corr(e_{ij}, e_{ij'}) = 0.1$ if $|t_{ij} - t_{ij'}| = 3$ (the maximum in the unbalanced case).

Non gaussian error

Data were generated by RIS or RI with various non gaussian errors (figure 3).

- Bimodal symmetric mixture of two gaussian distributions :

$$0.5N(-1.14, 0.57^2) + 0.5N(1.14, 0.57^2).$$

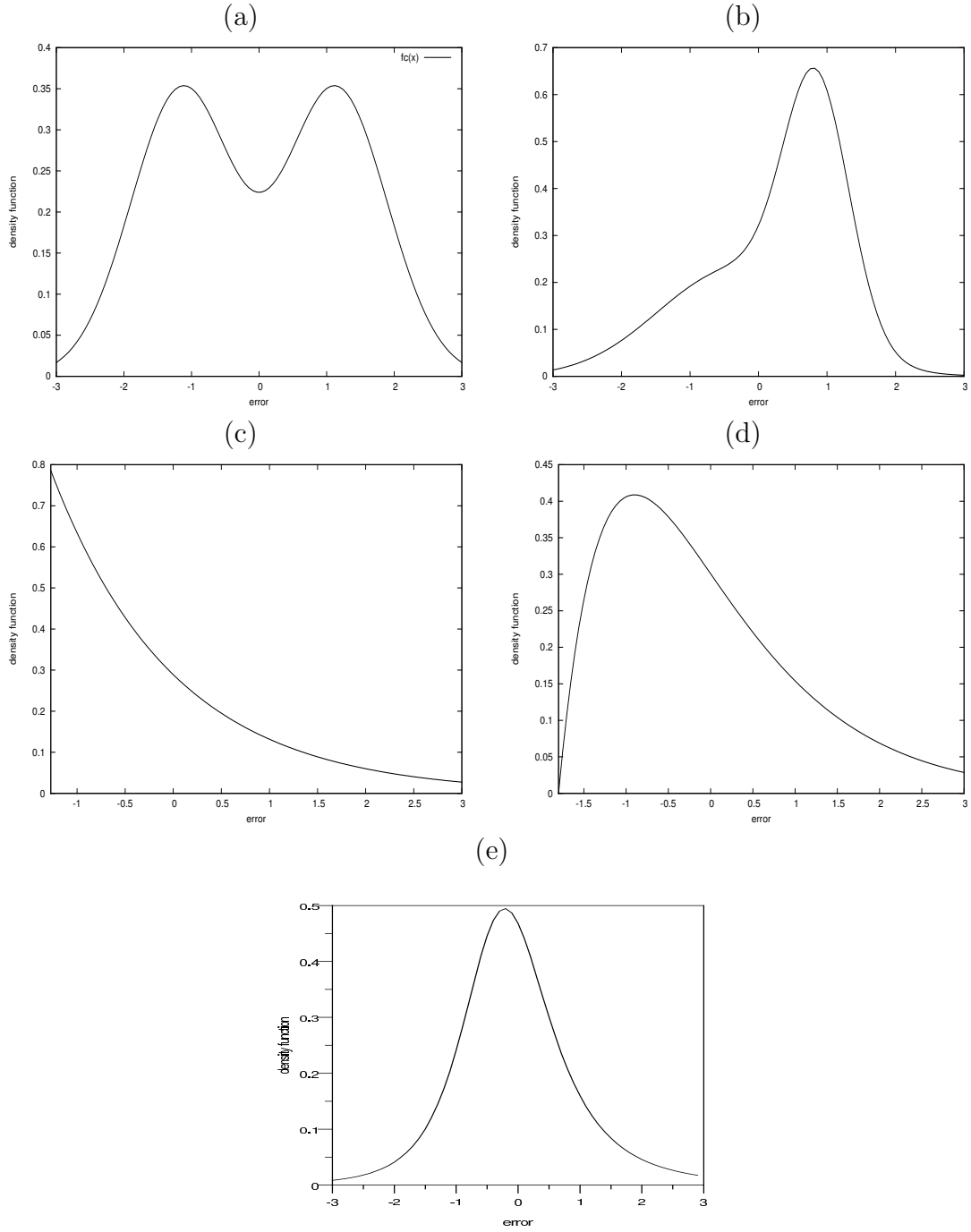
- Slightly asymmetric mixtures of gaussian distributions given by :

$$0.3N(0.848, 0.895^2) + 0.7N(-0.363, 1.237^2).$$

- A Gamma($\sqrt{\sigma^2_e/2}$,2) distribution which is franckly asymmetric.
- An Exponential distribution which is highly asymmetric
- A g-and-h distribution (Hoaglin, 1985)with g=0.2 and h=0.2 which is skewed and heavy-tailed.

The three last distributions were centred and the g-and-h was rescaled to obtain $E(e_{ij}) = 0$ and $var(e_{ij}) = 1.62$

Figure 3 : Densities of the non-gaussian distributions used in the simulation study
(a) Bimodal symmetric mixture of two gaussian distributions : $0.5N(-1.14, 0.57^2)+0.5N(1.14, 0.57^2)$
(b) Asymmetric mixture of gaussian distributions : $0.3N(0.848, 0.895^2)+0.7N(-0.363, 1.237^2)$.
(c) Gamma($\sqrt{1.62/2}$,2)
(d) Exponential with variance 1.62
(e) G-and-h distribution with g=h=0.2 and variance 1.6



5 Results

Coverage rates for the 95% confidence interval (CI) of the fixed parameter estimates are presented in tables 2,3 and 4.

Table 2 : Coverage rates of the 95% confidence intervals of the fixed effects from the random intercept model (RI) and from the model with random intercept and slope (RIS) computed using 1000 simulated data sets with correctly specified or heteroscedastic error distribution.

	N=50						N=200					
n_i	3-7		5		9		3-7		5		9	
	RIS	RI	RIS	RI	RIS	RI	RIS	RI	RIS	RI	RIS	RI
Gaussian independent errors with constant variance												
β_0	93.7	96.0	94.6	95.0	94.5	94.4	94.9	94.7	96.0	95.8	94.2	95.5
β_1	95.2	95.8	95.1	95.5	95.1	95.5	95.4	96.1	96.2	94.6	93.8	95.7
β_2	94.7	95.2	93.7	95.2	94.3	93.4	94.1	94.4	94.5	**96.5***	95.9	95.8
β_3	95.1	95.6	93.6	95.0	94.9	95.3	95.4	96.1	94.7	95.2	95.2	95.7
Heteroscedastic errors: $Var(e_{ij}) = 0.1 + 0.001E(Y_{ij}\|\alpha_i)^4$												
β_0	95.0	94.5	93.9	93.6	94.5	95.7	94.9	94.7	95.2	96.2	95.4	95.7
β_1	95.5	93.9	95.4	94.3	94.4	**92.8**	95.2	94.1	93.7	93.8	95.3	94.6
β_2	94.5	94.2	94.5	94.0	95.0	94.5	94.5	94.7	95.0	96.0	95.6	95.1
β_3	94.8	94.4	94.7	94.8	95.4	**93.0**	95.2	94.6	94.2	93.8	95.4	**93.2**
Heteroscedastic errors : $Var(e_{ij}) = 0.81 + 0.81t_{ij}$												
β_0	96.3	95.1	95.1	95.7	94.9	94.1	94.9	95.8	95.3	**97.1**	95.6	96.0
β_1	94.6	**93.5**	94.7	94.9	**96.7**	**91.2**	95.5	**93.4**	94.9	94.0	95.4	**92.2**
β_2	96.2	95.0	95.2	**97.4**	94.5	95.4	96.1	96.4	95.1	**97.2**	96.0	94.8
β_3	95.6	93.7	95.7	95.4	**96.8**	**91.4**	95.8	94.0	96.2	94.9	**96.5**	94.4
Heteroscedastic errors: $Var(e_{ij}) = 0.82 + 1.60X_i$												
β_0	**97.2**	**97.0**	**96.7**	96.0	95.2	95.5	**96.7**	**97.2**	96.1	96.3	**96.7**	95.7
β_1	**99.4**	**99.5**	**99.2**	**99.4**	**99.4**	**99.1**	**99.2**	**99.3**	**99.4**	**99.5**	**99.5**	**99.4**
β_2	93.6	94.0	95.1	95.7	**93.0**	94.5	95.4	95.3	94.2	94.2	95.3	95.0
β_3	**93.4**	93.7	96.1	95.0	95.6	95.5	95.3	95.9	94.4	96.2	95.2	95.5

* Values significantly different from the nominal value 95% are in bold type

Table 2 shows that robustness to homoscedasticity assumption depends on the kind of heterogeneous variance simulated. In the first case, chosen to mimic heteroscedasticity in the ALBI data set, the variance of the error is a function of $E(Y_{ij}|\alpha_i)$ and the confidence interval are very robust : only tree coverage rates were significantly different from 95.0% , that is less than 93.6% or above 96.4%, for the RI model and none for the RIS model.

When the error variance highly depends on time ($Var(e_{ij}|t = 2) = 3 \times Var(e_{ij}|t = 0)$), the RIS model is robust : few coverage rates were significantly different from 95% and they remained close to the nominal value (between 94.5% and 96.8%). Once again, the RI model was less robust (coverage rate range : 91.2% -97.4%) than the RIS model. When simulating data with more moderate departure from homoscedasticity assumption ($Var(e_{ij}|t = 2) = 2 \times Var(e_{ij}|t = 0)$), every coverage rates were correct for the RIS model and only 6 values were significantly different from 95% while remaining in the range 92.3% - 96.4% (results not shown).

When the variance depends on the treatment group, the coverage rate of the CI for the slope in the reference group is biased. More precisely, when the variance is higher in the group

with X=1 ($Var(e_{ij}|X=1) = 3 \times Var(e_{ij}|X=0)$), the variance of $\hat{\beta}_1$ is overestimated and the confidence interval is too large (see table 2). On the reverse, when the variance is higher in the group with X=0, the variance of $\hat{\beta}_1$ is underestimated and the confidence interval is too narrow (results not shown, coverage rates between 88% and 92%). Additional simulations performed with modest heteroscedasticity ($Var(e_{ij}|X=1) = 1.5 \times Var(e_{ij}|X=0)$ also exhibited modest but significant bias for the variance of the estimator of the slope (coverage rates between 96.5% and 97.5%). Thus compared with the two previous heteroscadastic cases studied, the mixed model is less robust to heterogeneous variance associated with a covariate when the model includes an interaction between time and this covariate (when the interaction was not included, inference for the fixed effects was not impaired). However, in every simulations with $Var(e_{ij}) = f(X)$, the test for the treatment effect is robust since the variance of $\hat{\beta}_3$ (the difference between the slopes in the two groups) is correctly estimated (with nominal coverage rate). To summarize when the variance depends on the treatment group, the variance of the estimated slopes may be biased but the estimator of the variance of the difference between slopes is robust. Another important point is that sensitivity to this kind of heteroscedasticity is similar for the RIS and RI models.Lastly, in the various heteroscedastic data simulated, we observed neither bias on the fixed effect estimates nor increase of the mean square error.

Table 3 : Coverage rates of the 95% confidence intervals of the fixed effects from the random intercept model (RI) and from the model with random intercept and slope (RIS) computed using 1000 simulated data sets with correlated error.

	N=50						N=200					
n_i	3-7		5		9		3-7		5		9	
	RIS	RI	RIS	RI	RIS	RI	RIS	RI	RIS	RI	RIS	RI
Correlated errors: $\sigma^2_\epsilon = 0.25$, $\sigma^2_w = 0.56$ $\gamma = 0.64$												
β_0	94.0	95.2	**92.7**	93.9	**93.0**	**92.6**	94.5	94.2	**93.4**	**93.2**	94.3	**92.8**
β_1	**93.3**	**85.0**	**92.6**	**90.3**	**93.1**	**80.8**	**93.0**	**85.9**	93.7	**87.1**	**92.9**	**80.2**
β_2	**93.1**	**93.4**	**92.4**	93.8	**93.2**	**91.5**	94.6	95.3	93.8	95.1	94.5	**92.9**
β_3	**92.2**	**85.5**	**93.5**	**85.9**	93.6	**78.8**	93.8	**86.7**	**92.6**	**89.2**	94.0	**79.7**
Correlated errors: $\sigma^2_\epsilon = 0.5$, $\sigma^2_w = 1.12$ $\gamma = 0.64$												
β_0	**91.5**	**93.4**	**91.9**	**92.2**	**92.2**	95.7	94.3	93.9	94.1	94.8	93.9	95.5
β_1	**92.6**	**85.4**	**93.3**	**86.7**	**93.3**	**83.3**	**93.0**	**85.7**	93.8	**88.2**	94.6	**77.4**
β_2	**92.9**	93.6	**92.7**	**91.9**	**92.7**	95.1	94.3	94.3	**93.1**	94.3	94.1	94.6
β_2	**91.9**	**85.8**	**93.4**	**86.7**	**93.7**	**82.6**	**93.4**	**85.6**	**93.4**	**89.3**	94.4	**78.2**

When the error was correlated (with an autoregressive error structure), results displayed in table 3 confirm that the naive variance estimator of the estimated fixed effects may be biased and that the RIS model is more robust than the RI model. For the RIS model many coverage rates of CI were outside the range 93.6% - 96.4% but they remained close to the nominal value (above 91%). On the contrary, inference on the RI model was severely compromised when the correlation of the error was not taken into account : due to the underestimation of the variance of the estimators, some coverage rates dropped under 80%. Inference for fixed covariates was more robust than inference for time dependent covariates especially in large samples.

Finally, Table 4 shows robustness of both models (with or without random slope) to non gaussian error. While the simulated errors were far from the assumed normal ditribution and, in

some cases, severely skewed and heavy tailed, few coverage rates were significantly different from the nominal value 95% and they were always above 92.7%. As expected, parameter estimates were unbiased and simulations also showed that the mean square error of the fixed effects was not increased. The robustness was better for parameter associated with time-dependent variable ($\hat{\beta}_1$ and $\hat{\beta}_3$) .

Table 4 : Coverage rates of the 95% confidence intervals of the fixed effects from the random intercept model (RI) and from the model with random intercept and slope (RIS) computed using 1000 simulated data sets with non gaussian error distributions

	N=50						N=200					
n_i	3-7		5		9		3-7		5		9	
	RIS	RI	RIS	RI	RIS	RI	RIS	RI	RIS	RI	RIS	RI
Symmetric mixture: $0.5N(-1.14, 0.57^2) + 0.5N(1.14, 0.57^2)$												
β_0	94.1	**92.7**	94.0	94.4	**93.5**	**93.4**	95.3	94.6	94.0	94.1	93.8	95.1
β_1	94.9	95.1	96.2	96.0	95.8	94.1	95.8	**93.4**	95.3	95.8	**92.9**	95.7
β_2	94.1	94.2	**93.4**	**93.5**	95.2	93.9	95.8	93.7	**92.7**	95.2	**93.4**	95.7
β_3	96.1	93.9	95.3	95.1	95.7	95.7	94.1	94.7	95.8	95.1	93.7	95.7
Asymmetric mixture: $0.3N(0.848, 0.895^2) + 0.7N(-0.363, 1.237^2)$												
β_0	95.1	94.4	95.4	94.3	**93.2**	94.4	94.5	95.5	95.0	96.2	93.7	95.1
β_1	94.5	95.0	93.7	94.2	94.0	95.0	95.9	94.3	94.6	95.9	95.2	95.1
β_2	93.6	94.1	**93.3**	96.0	**93.0**	94.2	94.6	94.5	95.1	94.5	94.6	95.8
β_3	**93.3**	94.1	95.4	94.4	96.0	95.1	95.1	95.1	94.7	95.8	94.6	94.6
Gamma: $(\sqrt{1.62}/2,2)$												
β_0	94.0	**93.4**	**93.1**	94.7	94.7	94.2	95.0	95.3	94.8	94.3	**96.7**	94.5
β_1	94.4	94.5	94.6	94.6	95.4	94.6	95.6	96.1	96.4	94.6	95.8	95.5
β_2	94.0	94.6	94.2	95.2	94.5	94.6	95.3	94.2	96.1	94.7	95.2	95.0
β_3	94.9	94.7	95.2	94.0	96.1	95.0	95.2	95.0	95.2	95.8	94.5	94.8
Exponential: $\sigma_e^2 = 1.62$												
β_0	94.4	95.6	95.0	95.3	95.0	94.0	95.4	93.9	95.5	95.1	94.8	95.2
β_1	93.8	95.6	94.1	95.0	95.1	95.0	95.5	94.8	95.8	95.0	94.6	94.1
β_2	93.9	94.1	93.6	95.1	94.8	94.1	94.7	94.0	94.2	94.5	94.6	94.8
β_3	94.5	95.8	95.1	94.6	95.3	96.4	94.4	94.4	96.1	96.3	95.0	94.9
G-and-h distribution: $g = 0.2, h = 0.2, \sigma_e^2 = 1.62$												
β_0	94.4	95.5	95.1	94.0	**93.5**	94.3	**93.5**	94.4	94.9	95.4	94.4	95.1
β_1	94.1	94.7	96.0	94.8	94.8	95.0	95.0	95.7	95.1	95.1	94.6	94.9
β_2	94.0	95.3	95.1	94.5	94.3	93.7	94.7	94.8	94.3	96.0	94.3	94.5
β_3	94.1	95.7	95.5	95.0	95.0	94.9	95.1	96.1	94.9	95.6	95.0	95.4

6 Discussion

This study highlighted that, with moderate sample size (n=50 or n=200), inference on fixed effects from a linear mixed effects model assuming independent gaussian error with homogeneous variance was not impaired when the true error distribution was either non gaussian or heteroscedastic (except for covariate-dependent variance as discussed below). Knowing the robustness of the simple linear model, these results were not completely surprising. However, they can be of interest for applied statistician because they show that transformations of the

data, which complicate interpretation of model parameters, are not necessary. This is especially important for non-gaussian continuous responses because mixed models for such longitudinal data are not available in standard software while it is possible to estimate mixed models with heteroscedastic error using for instance SAS Proc Nlmixed.

Our results however underlined that inference on the slopes are sensitive to heterogeneous error variance when the variance depends on a covariate included in the model with interaction with time. Thus, when studying a treatment effect, it should be useful to compare variances of the residuals in the two groups or to check the homogeneous variance assumption by estimating a model with variance depending on the treatment. Neverthless, inference on the treatment effect (difference between the slopes in the two groups) is robust to this kind of heteroscedasticity.

As it was previously shown (Lange and Laird, 1989; Taylor *et al*, 1994), our results confirmed that the mixed model with random intercept and slope is more robust to misspecification of the covariance structure than the random intercept model. Poor coverage rates for the random intercept model are due to biased estimates of the variances of the fixed effects and could be corrected using the robust variance estimator (Liang and Zeger, 1986). In this field, we have only investigated the case of an autoregressive error structure because other structures had been previously studied (Lange and Laird, 1989; Taylor *et al*, 1994). Moreover, it is easy in standard software to include correlated error in a linear mixed model without blurring interpretation of fixed effects.

Some authors (Richardson and Welsh, 1995; Copt and Victoria-Feser, 2006) have investigated the robustness of maximum likelihood estimators to outlying observations by simulating the error from a comtaminated normal distribution $(0.9N(0,1) + 0.1N(0,11))$. We conducted an additional simulation study using this contaminated distribution and found, as these authors, that fixed effects estimators were robust while variance components estimators were impaired (results not shown). As expected, Copt and Victoria-Feser (2006) reported biases for the fixed effects only when the mean of the contaminated error distribution was not null (which is equivalent to misspecify the mean structure of the model).

More generally, this study only focused on inference for fixed effects which is the unique objective of the longitudinal analysis in most cases. It is obvious and it was checked in our simulation study (results not shown), that variance parameter estimates and random effects may be biased when the covariance structure is misspecified. For instance, some authors have investigated lack of robustness of variance components estimators Taylor and Law (1998) have shown that individual predictions are affected by misspecified covariance structure. Thus, when random effects estimates or individual predictions are of interest, the distribution assumption for the error must be carefully checked. Indeed, random effects or individual predictions may be used as covariates in survival model either in a two step approach (Thiebaut et al, 2003) or in a joint model (Tsiatis and Davidian, 2004) to study association between evolution of the repeated measures and an event. In such a study, graphical evaluation of the assumed N(0,1) distribution of Cholesky residuals may be useful but they do not allow to distinguish between misspecification of the error structure and of the fixed part of the model. It would be better but not always feasible with standard softwares, to evaluate sensitivity of the results to the various assumptions using model including heteroscedastic, correlated or non gaussian error.

Acknowledgments :

We thank the Scientific Committee of the ALBI trial for authorizing us to use their data, the management centre at INSERM Unit 330 for providing us with the data.This trial was supported by a grant from the Agence National de Recherche sur le SIDA.

References :

Butler, S.M. and Louis, T.A. 1992. Random effects models with non parametric priors. Statist. Med. 11,1981-2000.

Copt, S. and Victoria-Feser, M.P. 2006. High-breakdown inference for mixed linears models. J. Amer. Stat. Assoc. 101, 292-300.

Hoaglin, D.C. 1985. Summarizing shape numerically : the g-and-h distribution. In: Hoaglin,D.C, Mosteller, F., Tuckey, J.W. (Eds.), Exploring Data Tables,Trends and Shapes p 461-513, New-York, Wiley.

Jacqmin-Gadda, H., Fabrigoule, C., Commenges, D. and Dartigues, J.F. 1997. A five-year longitudinal study of Mini-Mental State Examination in normal aging. Amer. J. Epidemiol. 145, 498-506.

James, F. 1990. A Review of Random Number Generators. Comp. Phys. Comm. 60, 329-344.

Laird, N. and Ware J. 1982. Random-Effects Models for Longitudinal Data. Biometrics 38, 963-74.

Lange, N. and Laird, N. 1989 The effect of Covariance Structure on Variance Estimation in Balanced Growth-Curve Models With Random Parameters. J. Amer. Statist. Assoc. 84, 241-7

Liang, K. and Zeger, S. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13-22.

Mcneil, A.J. and Gore, S.M. 1996. Statistical analysis of zidovudine (AZT) effect on CD4 cell counts in HIV disease. Statist. Med. 15(1), 75-92.

Marsaglia, G. and Tsang, W.W. 1984. A Fast, Easily Implemented Method for Sampling from Decreasing or Symmetric Unimodal Density Functions. SIAM J. Sci. Stat. Comput. 5, 349-359.

Molina, J.M., Chne, G., Ferchal, F., Journot, V., Pellegrin, I., Sombardier, M.N. et al. 1999. The ALBI Trial:A Randomized Controlled Trail Comparing Stavudine Plus Didanosine with Zidovudine Plus Lamivudine and a Regimen Alternating Both Combination in Previously Untreated Patients Infected with Human Immonodeffiency Virus. J. Infect. Dis. 180, 351-8.

Park, T. and Lee, S.Y. 2004. Model Diagnostic Plots for Repeated Measures Data. Biometrical J. 46, 441-52.

Richardson, A.M. and Welsh, A.H. 1995. Robust restricted maximum likelihood in mixed linear models. Biometrics 51, 1429-1439.

Royall., R.M. 1986. Model robust confidence intervals using maximum likelihood estimators. Int. Statist. Rev. 2,221-226.

Royston, P. 1993. A pocket-calculator algorithme for the Shapiro-Francia test for non-normality: an application to medicine. Statist. Med. 12, 181-4.

Taylor J.M.G., Cumberland W.G., and Sy J.P. A stochastic model for analysis of longitudinal AIDS data. J. Amer. Statist. Assoc. 1994. 89,: 727-736.

Taylor J.M.G and Law N. Does the covariance structure matter in longitudinal modelling for the prediction of futur CD4 counts ? Statist. Med. 17, 2381-2394.

Thiebaut, R., Chene, G., Jacqmin-Gadda, H., Morlat, P., Mercie, P., Dupon, M., Neau, D., Ramaroson, H., Dabis, F., Salamon, R. and Groupe d'Epidemiologie Clinique du SIDA en Aquitaine 2003. Time-updated CD4+ T lymphocyte count and HIV RNA as major markers of

disease progression in naive HIV-1-infected patients treated with a highly active antiretroviral therapy: the Aquitaine cohort, 1996-2001.J. Acquir. Immune Defic. Syndr. 33(3),380-6

Tsiatis, A.A. and Davidian, M. 2004. Joint modeling of longitudinal and time-to-event data : an Overview. Statist. Sinica 14, 809-834.

Tsiatis, A.A., DeGruttola, V. and Wulfsohn, M.S. 1995. Modelling the relationship of survival to longitudinal data measured with error. Application to survival and CD4 counts in patients with AIDS. J. Amer. Statist. Assoc. 90, 27-37.

Verbeke, G. and Lesaffre, E. 1997. The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. Comput. Statist. Data Anal. 23, 541-556.

Zhang, D. and Davidian, M. 2001. Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data. Biometrics 57, 795-802.