Sequential Testing of Model Fitting with Longitudinal Data

José C. Pinheiro
David L. DeMets, Ph.D.
Sequential Testing of Model Fitting with Longitudinal Data

José C. Pinheiro and David L. DeMets

Bell Laboratories
Lucent Technologies
Murray Hill, NJ

Department of Biostatistics
University of Wisconsin — Madison
Madison, WI

Longitudinal data are frequently collected in clinical trials to evaluate the evolution of treatment effects over time. A common statistical analysis for these data is to compare the rates of change over time for the different treatment groups in the study. Because the trial design and analysis depend on the assumed statistical model, the assumption of linearity for the response should be assessed at the time of the interim analyses. To control the rate of falsely rejecting the model, group sequential methods need to be used to account for the repeated testing. We propose a general method for the sequential testing of the linearity assumption of a longitudinal response, presenting a sequential model fitting test statistic with independent increments structure to which standard group sequential methods of analysis can be applied. We illustrate the use of the proposed sequential testing procedure using clinical trial data and Monte Carlo simulation.

key words: Group sequential methods; Mixed-effects models; repeated measures.

1 Introduction

Longitudinal data are frequently collected in clinical trials to evaluate the evolution of treatment effects over time. In many trials, the comparison of changes in a response variable among different treatment arms is the primary objective (e.g. the Intermittent Positive Pressure Breathing trial(IPPBT, 1983), the Lung Health Study(Anthonisen, 1989), and the ALS CNTF Treatment Study(CTS, 1996)). A common statistical analysis for these data is to compare the rates of change over time for the different treatment groups in the study.

Because of the longitudinal nature of these data, observations collected on the same patient will gen-
erally be correlated. A model that has been frequently used to account for this within-patient correlation is the linear mixed-effects models proposed by Laird and Ware (Laird and Ware, 1982). In this model, the observations collected on the $ith$ individual, denoted by $y_i$, are assumed to be a linear function of a set of fixed effects $\beta$ and patient-specific random effects $b_i$, as described below.

$$y_i = X_i \beta + Z_i b_i + \epsilon_i, \; i = 1, \ldots, M,$$

where $X_i$ and $Z_i$ denote respectively the fixed effects and the random effects regression matrices, $\epsilon_i$ denotes the within-patient error vector, and $M$ denotes the number of patients. The model assumes that the $b_i$ are independent for different patients and follow a $\mathcal{N}(0, D)$ distribution, with $D$ representing a general covariance matrix; the $\epsilon_i$ are independent for different patients, independent of the $b_i$, and follow a $\mathcal{N}(0, \Lambda_i)$ distribution, with $\Lambda_i$ representing a patient-specific covariance matrix, parametrized by a fixed, generally small, set of parameters. The correlation between observations collected on the same patient in model (1) follows from the fact that these observations share the same random effects and the possibly non-diagonal structure of the $\Lambda_i$ matrices.

Clearly, any analysis based on the comparison of rates of change over time will only be valid if, for the duration of the study, the response variable being sequentially measured varies with time, at least approximately, according to a straight line. This assumption may be violated either because the treatments alter the natural evolution of the response variable with time, resulting in a nonlinear relationship, or because the natural evolution of the response variable was not linear to begin with.

Interim analyses of accumulating data in clinical trials is compelling for both ethical and scientific reasons. Typical interim analyses evaluate primary, secondary and safety outcome measures. The validity of the interim analyses is just as critical as it is for scheduled final analysis, since decisions about the protocol continuation, modification or termination must be evaluated with each interim analyses. Interim analyses with an inappropriate model might easily lead to an incorrect decision, jeopardizing the trial or preventing the initial hypothesis to be properly addressed.

Given that the interim analyses methods need to reflect the final analysis, the appropriateness of the model must be tested repeatedly, but not falsely rejected. That is, the rate of falsely rejecting the model must be controlled since the model will be tested repeatedly, thus increasing the chance of a false rejection if not otherwise controlled. Rejection of the model has immediate consequence for the design of the trial. Most trial designs depend on a statistical model to evaluate the primary outcome, in this case a linear change in the primary outcome measure over time of follow up. Sample size, for example, is based on this linearity assumption. If the model is not correct, the design of the trial and sample size would have to be modified.
Thus, an incorrect rejection of the statistical model has serious design implications. However, continued assessment of the primary outcome with an inappropriate statistical model would also have serious consequences. In this paper, we propose a general method for the sequential testing of the linearity assumption of a longitudinal response, employing the linear mixed-effects model (1). This procedure uses standard group sequential methods techniques.

Group sequential methods are frequently used in data monitoring of clinical trials to detect early therapeutic benefit or unexpected toxicity that might lead to early termination of the study (DeMets, 1987). In the context of lack-of-fit testing, the sequential monitoring is important to identify, as rapidly as possible, when a model is not adequate. Different methods have been proposed to take into account the effect of repeatedly testing the data on the overall significance level of the testing procedure (Pocock, 1977; O'Brien and Fleming, 1979; Lan and DeMets, 1983). These methods are based on the idea of adjusting the significance levels (or, equivalently, the critical values) of the individual analyses, so that the overall significance level of the testing procedure is kept at a pre-specified level $\alpha$. The most flexible of these methods, which includes the other methods as particular cases, is the alpha-spending function, proposed by Lan and DeMets (Lan and DeMets, 1983).

Even though group sequential methods are of general applicability, their most common use (for which computer programs for calculating the sequential boundaries are available) is when the sequence $\{\hat{\theta}_n\}_{n \geq 1}$ of test statistics for the interim analyses has an independent increments structure (DeMets and Lan, 1994), characterized by

$$\text{cov}(\hat{\theta}_i, \hat{\theta}_j) = \text{var}(\hat{\theta}_j), i \leq j.$$  \hspace{1cm} (2)

Sequential testing for comparison of changes in a response variable with repeated measures data has been considered by several authors. Geary (Geary, 1988) studied the case in which the number and spacing of the repeated measurements are the same for all patients and the interim analyses coincide with the observation times. Lee and DeMets (Lee and DeMets, 1991) studied the sequential comparison of changes in repeated measures under the linear mixed-effects model (1). The sequential comparison of repeated measures data using the area under the expected response change curves between two treatment groups was discussed by Wu and Lan (Wu and Lan, 1992) for a simplified version of model (1) with $X_i = Z_i$ and $\Lambda_i = I$ (with $I$ denoting the identity matrix), but allowing for possibly informative censoring. Reboussin, Lan, and DeMets (Reboussin, Lan and DeMets, 1992) consider yet another version of model (1) with $D = \sigma_b^2 I$ and $\Lambda_i = I$, and show that the sequence of estimators for the fixed effects commonly used with the Laid and Ware model has an independent increments structure. Sequential monitoring of repeated measures
data using generalized estimating equations estimation has been considered by Gange and DeMets (Gange and DeMets, 1996). Recently, Scharfstein, Tsiatis, and Robins (Scharfstein, Tsiatis and Robins, 1997) have shown that the independence increments structure holds, at least asymptotically, for a broad class of efficient test statistics used with semiparametric or parametric models, which, in particular, includes tests for the difference in rates of change over time for longitudinal data.

In Section 2, we show that a sequence of maximum likelihood estimators for the fixed effects in model (1) has an independent increments structure. This result is used to derive a sequential test of model fitting, using the alpha-spending function approach of Lan and DeMets (Lan and DeMets, 1983). In Section 3, we illustrate the use of the sequential model fitting test with some clinical trial data. The performance of the test under different nonlinearity patterns for the response variable and different alpha-spending functions, using Monte Carlo simulation is considered in Section 4. Our conclusions and suggestions for future research are included in Section 5.

2 A Sequential Test of Linearity

When $D$ and the $\Lambda_i$ are known in the linear mixed-effects model (1), the maximum likelihood estimator (MLE) of the fixed effects $\beta$ is

$$\hat{\beta} = \left[ \sum_{i=1}^{M} X_i^T \left( \Lambda_i + Z_iDZ_i^T \right)^{-1} X_i \right]^{-1} \sum_{i=1}^{M} X_i^T \left( \Lambda + Z_iDZ_i^T \right)^{-1} y_i.$$ 

Letting $\hat{\beta}_k$ denote the MLE of $\beta$ at the $k$th interim analysis we have that

$$\text{cov} \left( \hat{\beta}_k, \hat{\beta}_l \right) = \text{var} \left( \hat{\beta}_l \right), \quad k \leq l. \quad (3)$$

The proof is presented in the Appendix. It follows from (2) and (3) that the sequence $\{\hat{\beta}_k\}_{k \geq 1}$ has independent increments structure. This result generalizes the ones in Wu and Lan (Wu and Lan, 1992) and Reboussin et al. (Reboussin et al., 1992) in the sense it holds for general $D$ and $\Lambda_i$. Because $\hat{\beta}_k$ can be viewed as a semiparametric efficient estimator of $\beta$, (3) is also a consequence of the general result in Scharfstein et al. (Scharfstein et al., 1997).

As a consequence of (3), it follows that, for known $D$ and $\Lambda_i$, the sequence of MLEs of any linear combination of the fixed effects in model (1) has independent increments structure. To see this, let $\theta = c^T\beta$ denote an arbitrary linear combination of $\beta$. Then, the MLE of $\theta$ at the $k$th interim analysis is given by $\hat{\theta}_k = c^T\hat{\beta}_k$ and therefore

$$\text{cov} \left( \hat{\theta}_k, \hat{\theta}_l \right) = c^T \text{cov} \left( \hat{\beta}_k, \hat{\beta}_l \right) c = c^T \text{var} \left( \hat{\beta}_l \right) c = \text{var} \left( \hat{\beta}_l \right), \quad k \leq l. \quad (4)$$
In practical applications, \( D \) and \( \Lambda_i \) are usually unknown, but given the consistency of the MLE of these quantities (Pinheiro, 1994) the independent increments structure holds approximately when \( D \) and \( \Lambda_i \) are replaced by their MLEs.

Since linear (fixed-effects) models are particular cases of the linear mixed-effects model (1), the relationship (4) also holds for any linear combination of the usual least squares estimators (with general covariance structure for the error term). Equation (4) is also approximately true for the sequence of MLEs of any differentiable function of \( \beta \). Letting \( \theta = g(\beta) \), \( g \) differentiable, a standard application of the delta method (Efron and Tibshirani, 1993) and (4) gives

\[
\text{cov} \left( \hat{\theta}_k, \hat{\theta}_l \right) \simeq \frac{\partial g}{\partial \beta^T} \text{cov} \left( \hat{\beta}_k, \hat{\beta}_l \right) \frac{\partial g}{\partial \beta} = \frac{\partial g}{\partial \beta^T} \text{var} \left( \hat{\beta}_l \right) \frac{\partial g}{\partial \beta} \simeq \text{var} \left( \hat{\theta}_l \right)
\]

In applications, the straight line relationship between the response variable \( y \) and time is generally understood to be just an approximation to the true underlying functional relationship between these quantities. Let \( E(y_t) = f(\delta, x, t) \) represent the true, unknown functional relationship between the expected value of the response variable \( y \) at time \( t \) and a set of unknown parameters \( \delta \), a set of known covariates \( x \), and \( t \). The straight line model is generally based on the assumption that a first-order Taylor expansion of \( f \) provides a good representation of \( E(y_t) \) within the duration of the study. That is, it is assumed that, for some \( t_0 \) in the time span of the study,

\[
E(y_t) \simeq f(\delta, x, t_0) + f'(\delta, x, t_0)(t - t_0) = \alpha_0(\delta, x, t_0) + \alpha_1(\delta, x, t_0)t,
\]

where \( f' \) denotes the first-order partial derivative of \( f \) with respect to \( t \). For the straight line model to be adequate, \( f' \) should be approximately constant with respect to \( t \) for the duration of the study.

When the approximation in (5) fails to give a good representation of \( E(y_t) \), more terms are needed in the Taylor expansion of \( f \) to provide an adequate empirical model. The simplest case is when in the expansion is taken to the second-order, leading to the approximation

\[
E(y_t) \simeq f(\delta, x, t_0) + f'(\delta, x, t_0)(t - t_0) + f''(\delta, x, t_0)(t - t_0)^2/2
\]

\[
= \alpha_0(\delta, x, t_0) + \alpha_1(\delta, x, t_0)t + \alpha_2(\delta, x, t_0)t^2,
\]

where \( f'' \) denotes the second-order partial derivative of \( f \) with respect to \( t \). The extend to which \( \alpha_2(\delta, x, t_0) \) differs from zero gives an indication of the departure from linearity in the representation of \( E(y_t) \).

Therefore, by fitting a second-degree polynomial in time to the response variable at each interim analysis and testing if the second order coefficients are significantly different from zero, one obtains a sequential test of linearity. One has to be careful when choosing the model(s) to be used for this sequential lack-of-fit
testing, to ensure that the test statistics have the correct distribution under the null hypothesis of linearity. We propose fitting separate models for each treatment arm to prevent the lack of linearity in one arm from affecting the test statistic for the other treatment arms. The proposed linear mixed-effects model for testing the linearity assumption in treatment group $g$ is

$$y_{ij} = [\beta_{0g} + b_{0i} + [\beta_{1g} + b_{1i}] t_{ij} + \beta_{2g} t_{ij}^2] + \epsilon_{ij}, \quad i = 1, \ldots, n_{ig}, \quad j = 1, \ldots, m_g, \quad g = 1, \ldots, G \quad (6)$$

where $y_{ij}$ is the response for the $i$th patient at time $t_{ij}$, $\beta_{0g}$ is the intercept fixed effect, $\beta_{1g}$ is the slope fixed effect, $\beta_{2g}$ is the second order term fixed effect, $b_{0i} = (b_{00}, b_{01})$ represents the random effects for the $i$th patient, $\epsilon_{ij}$ represents the within-patient error term, $G$ is the number of treatment arms, $m_g$ is the number of patients in treatment arm $g$, and $n_{ig}$ is the number of visits for the $i$th patient in treatment group $g$. The distributional assumptions about the $b_{0i}$ and the $\epsilon_{ij}$ are the same as in model (1). Note that no random effects are included for the second order term, as this would affect the distribution of the lack-of-fit test statistics under the assumption of linearity.

Let $\tilde{\beta}^{(k)}_{2g}$ denote the MLE of $\beta_{2g}$ and $\tilde{\sigma}^{(k)}_{2g}$ its estimated standard deviation at the $k$th interim analysis, $k = 1, \ldots, K$. The model fitting test statistics for treatment group $g$ at interim analysis $k$ is $L_g^{(k)} = \tilde{\beta}^{(k)}_{2g} / \tilde{\sigma}^{(k)}_{2g}$. It follows from (4) that the sequence $\{\tilde{\beta}^{(k)}_{2g}\}_{k=1,...,K}$ has approximately independent increments structure and, therefore, standard group sequential methods can be used to derive critical boundaries for the sequential model fitting tests. In this paper we use the alpha-spending function method of Lan and DeMets (Lan and DeMets, 1983).

To obtain the critical boundaries for the sequential lack-of-fit tests, one needs to specify the overall significance level $\alpha$, the information fractions at each interim analysis (Lan, Reboussin and DeMets, 1994), and the alpha-spending function.

When choosing the significance level for the sequential lack-of-fit tests, one must take into account that the rejection of the statistical model will have immediate and direct consequences in the design and analysis of the trial. Thus, an incorrect rejection of the model has serious implications and should be carefully considered. We propose using the same $\alpha$ for testing the primary outcome and the model lack-of-fit. Because a different lack-of-fit test is used for each treatment arm, some adjustment for multiple testing should be used to obtain the individual significance levels. The sequences of test statistics used for the lack-of-fit tests will be independent for different treatment arms and, therefore, a Bonferroni adjustment, defined as $\alpha_g = \alpha / G$, gives an adequate adjustment.

Ideally, one should use $\tau_k = \text{var} \left( \tilde{\beta}^{(K)}_{2g} \right) / \text{var} \left( \tilde{\beta}^{(k)}_{2g} \right)$ as the information fraction at interim analysis $k$. However, because $\text{var} \left( \tilde{\beta}^{(k)}_{2g} \right)$ involves unknown parameters, usually a surrogate information fraction $\tilde{\tau}_k$ is
used in practice (Lan et al., 1994).

The choice of alpha-spending function will determine the ease with which the linearity hypothesis will be rejected at earlier interim analyses. The two most common alpha-spending functions used in practice have been proposed by Pocock (Pocock, 1977) and O'Brien and Fleming (O'Brien and Fleming, 1979). The Pocock alpha-spending function results in easier early rejection of the null hypothesis, while the O'Brien–Fleming alpha-spending function makes it harder to reject the null hypothesis at early stages into the trial. It should be noticed that the sequential boundaries used for the lack-of-fit tests do not have any relation to the boundaries used for testing treatment efficacy/harm.

In the next two sections we illustrate the use of the sequential test of linearity using clinical trial data and analyze its performance under different patterns of nonlinearity for the response, using Monte Carlo simulation.

3 A Clinical Trial Example

The data considered in this section are modified from an as yet unpublished randomized, double-blind clinical trial. The modification was done to maintain the confidentiality of the data. Two treatment arms were used in the study: a placebo and an active drug. A total of 211 patients, 102 in the active drug group and 109 in the placebo group, took part in the study. The primary outcome was the difference in the rates of change over time for a given response variable, between the two treatment groups. Measurements were taken at baseline, 3, 6, 12, 18, 24, and 30 months. Due to staggered entry in the study and missing visits, different number of observations were available on each patient at each interim analyses and at the end of the trial. Interim analyses took place at 6, 12, 18, and 24 months, and a final analysis occurred at 30 months. Figure 1 shows the mean response value per visit, for each treatment arm. It is clear from the plot that the drug altered the natural evolution of the response with time, leading to a nonlinear relationship. The placebo group mean response curve is approximately linear with respect to visit.

The linear mixed-effects model that the investigators believed valid before the beginning of the trial was

\[ y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + \delta_1 T_i + b_{1i}) t_{ij} + \epsilon_{ij} \]  

where \( y_{ij} \) is the response for the \( i \)th patient at time \( t_{ij} \), \( T_i \) is an indicator variable assuming the value 1 when the \( i \)th patient is in the treatment group and 0 otherwise, \( \beta_0 \) is the common intercept, \( \beta_1 \) is the slope for the placebo group, \( \delta_1 \) is the difference in slope between the treatment and the placebo groups, \( b_i = (b_{0i}, b_{1i}) \) represents the random effects for the \( i \)th patient, and \( \epsilon_{ij} \) represents the within-patient error term. The \( b_i \) are
Figure 1: Mean response curves per treatment.

assumed to be independent and to follow a $\mathcal{N}(0, D)$ distribution and the $\epsilon_{ij}$ are assumed to be independent, to follow a $\mathcal{N}(0, \sigma^2)$ distribution, and to be independent of the $b_i$.

Two separate linear mixed-effects models, as defined in (6), were fit to the placebo and active drug arms. The overall $\alpha$ chosen for this example was 5%, but, because two sequential tests are involved, we applied a Bonferroni adjustment to obtain a 2.5% significance level for each test.

The information fraction $\tau_k$ at interim analysis $k$ is determined by $\text{var}\left(\hat{\beta}_{2g}^{(k)}\right) / \text{var}\left(\hat{\beta}_{2g}^{(k)}\right)$, which is a function of the unknown relative dispersion matrix $\Delta = D / \sigma^2$ and the individual visit times. An estimate of $\Delta$ and a predicted visit pattern need to be available at the time the protocol is designed, so that sample size calculations can be done. These can also be used for estimating the information fractions. In this example, we use the $\Delta$ estimated from the complete data at the end of the trial and assume that all patients entered the study at the same time and did not miss any visits. Table 1 gives the estimated information fractions obtained using this approach and, for comparison, the true information fractions obtained using the $\text{var}\left(\hat{\beta}_{2g}^{(k)}\right)$ estimated at each interim analysis (the reported values correspond to the average of the observed information fractions for each treatment arm). There is a very close agreement between the estimated and observed information fractions.

We use the Pocock and the O'Brien–Fleming alpha-spending functions and the estimated information fractions in Table 1 to obtain the sequential boundaries for the tests.

Figure 2 displays the values of the test statistics $L_g^{(k)}$ for each interim analysis, together with the Pocock
Table 1: Estimated and observed information fractions in the clinical trial example.

<table>
<thead>
<tr>
<th>Type</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated</td>
<td>0.001</td>
<td>0.021</td>
<td>0.106</td>
<td>0.368</td>
<td>1.000</td>
</tr>
<tr>
<td>Observed</td>
<td>0.002</td>
<td>0.027</td>
<td>0.125</td>
<td>0.389</td>
<td>1.000</td>
</tr>
</tbody>
</table>

and the O'Brien–Fleming boundaries. The assumption of linearity would have been rejected for the active drug treatment group under both alpha-spending functions, but earlier for Pocock boundaries (first interim analysis) than for O'Brien–Fleming boundaries (second interim analysis). It would not have been rejected at all for the placebo group under both alpha-spending functions.

Figure 2: Sequential model fitting tests statistics per treatment.

4 Test performance under different non-linearity patterns

In this section, we investigate the performance of the sequential test of linearity introduced in Section 2 under different patterns of non-linearity, using Monte Carlo simulation. Each Monte Carlo replication consists of simulating data for a complete clinical trial, with a non-linear response, and using the sequential test of Section 2 to assess the linearity assumption at each planned analysis. The overall power, that is, the
percentage of times that non-linearity is detected, and the average time until detection of non-linearity (censored at the time of the last analysis) are used to evaluate the performance of the test.

The design used for the simulated trials is similar to the one in the example of Section 3. Two treatment arms are present. For simplicity, the patients are assumed to enter the study at the same time and have visits at baseline, 3, 6, 12, 18, 24, and 30 months. Interim analyses are planned for 6, 12, 18, and 24 months and a final analysis is planned for the end of the study, at 30 months. To investigate the effect of number of patients on the test performance, two sample sizes are used in the simulation: 30 and 80 patients per treatment arm. Because the sequential tests are performed independently for each treatment arm, we only report here the results for one treatment arm. The overall size for the sequential tests of linearity is 5%, with a Bonferroni adjustment giving a 2.5% significance level per treatment arm. Pocock and O’Brien–Fleming alpha-spending functions are used to determine the sequential boundaries. The true information fractions (i.e. $\tau_k = \frac{\text{var} \left( \hat{\beta}_{2g}^{(k)} \right)}{\text{var} \left( \hat{\beta}_{2g}^{(k)} \right)}$, $k = 1, \ldots, K$) under the linear mixed-effects model (10) are used in the alpha-spending function calculations. A total of 1,000 Monte Carlo replications are used for each combination of model, curvature, and sample size.

Two non-linear functions are used to simulate a non-linear response.

- **asymptotic regression**: $E(y_i) = \beta_0 + \beta_1 \exp(-\beta_2 t)$, which gives an example of curvature with no inflection points;

- **logistic**: $E(y_i) = \beta_0 + \beta_1 / \{1 + \exp [-\beta_2(t - \beta_3)]\}$, which gives an example of symmetric curvature around a single inflexion point $\beta_3$.

In both cases, the amount of curvature is determined by the parameter $\beta_2$. Figure 3 shows examples of asymptotic regression and logistic curves with mild, moderate, and severe degrees of non-linearity. These curves are chosen so that $E(y_{30}) = 73$ and $E(y_{30}) = 78$, which gives approximately the same range for the average response as in the clinical trial example of Section 3 (cf. Figure 1).

To take into account the longitudinal nature of the response, the following nonlinear mixed-effects models (Lindstrom and Bates, 1990) are used to generate the data for the $j$th observation on the $i$th patient.

- **Asymptotic regression**

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) \exp \left[-(\beta_2 + b_{2i}) t_j \right] + \varepsilon_{ij} \tag{8}$$

- **Logistic**

$$y_{ij} = (\beta_0 + b_{0i}) + \frac{\beta_1 + b_{1i}}{1 + \exp \left[-(\beta_2 + b_{2i}) \left[t_j - (\beta_3 + b_{3i}) \right] \right]} + \varepsilon_{ij} \tag{9}$$
Figure 3: Asymptotic regression and logistic curves with mild, moderate, and severe non-linearity.

In both models \( b_{0i} \sim \mathcal{N}\left[0, (0.05\beta_0)^2\right] \), \( b_{1i} \sim \mathcal{N}\left[0, (0.05\beta_1)^2\right] \), \( b_{2i} \sim \mathcal{N}\left[0, (0.25\beta_2)^2\right] \), and \( \varepsilon_{ij} \sim \mathcal{N}(0, 1) \), with the random effects independent of each other and independent of the within-subject error \( \varepsilon_{ij} \). In the logistic model, \( b_{3i} \sim \mathcal{N}\left[0, (0.25\beta_3)^2\right] \), independent of the other random effects and of the within-subject error.

Three degrees of non-linearity are used with each model. For the asymptotic regression model (8) we use \( \beta_2 = 0.025, 0.075, \) and 0.2, corresponding, respectively, to the mild, moderate, and severe non-linear patterns on the left hand side plot of Figure 3. The remaining parameters are determined so that the average response at baseline is 73 and the average response at thirty months is 78. For the logistic model (9), we use \( \beta_2 = 0.15, 0.25, 0.40 \), corresponding, respectively, to the mild, moderate, and severe non-linear patterns on the right hand side plot of Figure 3. \( \beta_0 \) and \( \beta_1 \) are determined so that the average response at baseline was 73 and the average response at thirty months is 78 and the inflection point \( \beta_3 \) is set at 15, half of the planned duration for trial.

To evaluate the performance of the sequential test under the null hypothesis of linearity, the following linear mixed-effects also is used in the simulation.

\[
y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) t_j + \varepsilon_{ij} \tag{10}
\]

where \( b_{0i} \sim \mathcal{N}\left[0, (0.05\beta_0)^2\right] \), \( b_{1i} \sim \mathcal{N}\left[0, (0.25\beta_1)^2\right] \), and \( \varepsilon_{ij} \sim \mathcal{N}(0, 1) \), with all three terms independent of each other. The fixed effects were chosen so that the expected value of \( y_{ij} \) at baseline was 73 and at the
end of the study was 78, that is, $\beta_0 = 73$ and $\beta_1 = 1/6$. Table 2 summarizes the simulation results.

Table 2: Power of sequential test of linearity and average trial duration (months) for simulated trials with 30 and 80 patients per treatment arm with O'Brien–Fleming(O–F) and Pocock (Po.) critical boundaries, for different patterns of non-linearity.

<table>
<thead>
<tr>
<th>Model</th>
<th>Curvature</th>
<th>30 patients/arm</th>
<th>80 patients/arm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power</td>
<td>Avg. Duration</td>
<td>Power</td>
</tr>
<tr>
<td>Linear</td>
<td>None</td>
<td>0.027 0.031 30.0 29.6</td>
<td>0.029 0.024 30.0 29.7</td>
</tr>
<tr>
<td>Asym. Reg.</td>
<td>Mild</td>
<td>0.614 0.445 29.9 28.3</td>
<td>0.967 0.916 29.3 25.7</td>
</tr>
<tr>
<td>Asym. Reg.</td>
<td>Moderate</td>
<td>1.000 1.000 24.5 18.3</td>
<td>1.000 1.000 19.8 14.4</td>
</tr>
<tr>
<td>Asym. Reg.</td>
<td>Severe</td>
<td>1.000 1.000 12.9 9.9</td>
<td>1.000 1.000 11.3 7.0</td>
</tr>
<tr>
<td>Logistic</td>
<td>Mild</td>
<td>0.048 0.300 29.9 26.3</td>
<td>0.134 0.742 29.1 20.5</td>
</tr>
<tr>
<td>Logistic</td>
<td>Moderate</td>
<td>0.366 0.929 26.5 17.8</td>
<td>0.985 1.000 18.1 14.3</td>
</tr>
<tr>
<td>Logistic</td>
<td>Severe</td>
<td>0.981 1.000 18.3 16.2</td>
<td>1.000 1.000 17.6 13.6</td>
</tr>
</tbody>
</table>

The main conclusions from the simulation study are:

- Under Pocock boundaries, the sequential tests of linearity have high overall power for moderate and severe curvatures with both the asymptotic regression and the logistic models, even for the relatively small sample size of 30 patients per treatment arm.

- The O'Brien–Fleming boundaries give tests with high power for moderate and severe curvature with the asymptotic regression model, but only performs well with the logistic model for severe curvature.

- The Pocock boundaries lead to significant earlier detection of non-linearity for both the asymptotic regression and the logistic models.

- The significance levels of the sequential linearity tests are close to the nominal value of 2.5%.

- There is a substantial increase in power for mild degrees of curvature and considerable reductions in average detection time when the sample size increases from 30 to 80 patients per treatment arm.

- The symmetric non-linearity pattern corresponding to the logistic model seems to be harder to detect than the concave type of curvature associated with the asymptotic regression model. This behavior is more evident for mild degrees of non-linearity.
5 Conclusions

We describe a sequential procedure for testing the assumption of linearity with longitudinal data. The corresponding test statistics can be calculated using available software for fitting linear mixed-effects models. We showed that the sequence of test statistics for the linearity assumption has approximately an independent increments structure, making it easy to implement the testing procedure with methods and software available for group sequential designs.

When Pocock sequential boundaries are used, the proposed test has high power under different non-linearity patterns and different degrees of curvature. In general, Pocock boundaries also lead to significantly earlier detection of non-linearity than O'Brien–Fleming boundaries. The significance level of the sequential tests is about the same for the two types of boundaries. Therefore, Pocock boundaries seem preferable to O'Brien–Fleming boundaries in the context of sequential lack-of-fit testing.

Once non-linearity is detected for a longitudinal response, the usual statistical analysis, based on the comparison of rates of change across treatment arms, will no longer be valid. If the pattern of non-linearity is the same for the different treatment groups, a linear polynomial mixed-effects model with higher order terms in time, or a nonlinear mixed-effects model could be fit to the data and some of its parameters used to test treatment differences. If a linear polynomial mixed-effects model is used, the independence increments structure for the estimated coefficients will still hold and standard group sequential methods of analysis can be used. In general, the covariance matrix for the estimators in a nonlinear mixed-effects model will be estimated using some local approximation to the objective function, which will depend on the current estimates of the parameters(Lindstrom and Bates, 1990). Because these approximations change with interim analysis, it is difficult to determine whether a sequence of test statistics obtained from a nonlinear mixed-effects model has an independent increments structure. Further research is needed on appropriate group sequential methods for nonlinear mixed-effects models.

When non-linearity occurs only in some of the treatment arms, both linear and nonlinear mixed-effects models may be needed to explain the evolution of the response with time. In this case, one will not be able to test for treatment efficacy using a test statistics based on a parameter that is present in all models. More investigation is needed on methods for comparing treatments with different response patterns, which take into account both the longitudinal characteristic of the measurements and the sequential nature of the tests.
References


Appendix

We include here the proof of result (3) in Section 2. In what follows, $X_{ik}$ and $Z_{ik}$ will denote respectively the fixed effects and the random effects regression matrices for the $i$th individual at the $k$th interim analysis.

Letting

$$A_k = \left[ \sum_{i=1}^{M_k} X_{ik}^T \left( \Lambda_{ik} + Z_{ik}DZ_{ik}^T \right)^{-1} X_{ik} \right]^{-1},$$

we have that

$$\text{cov} \left( \hat{\beta}_k, \hat{\beta}_l \right)$$

$$= A_k \sum_{i=1}^{M_k} X_{ik}^T \left( \Lambda_{ik} + Z_{ik}DZ_{ik}^T \right)^{-1} \text{cov} \left( y_{ik}, y_{il} \right) \left( \Lambda_{il} + Z_{il}DZ_{il}^T \right)^{-1} X_{il} A_l$$

$$= A_k \sum_{i=1}^{M_k} X_{ik}^T \left( \Lambda_{ik} + Z_{ik}DZ_{ik}^T \right)^{-1} \left[ \Lambda_{ik} + Z_{ik}DZ_{ik}^T : Z_{ik}DZ_{ik:il}^T \right] \left( \Lambda_{il} + Z_{il}DZ_{il}^T \right)^{-1} X_{il} A_l$$

$$= A_k \sum_{i=1}^{M_k} X_{ik}^T \left( \Lambda_{ik} + Z_{ik}DZ_{ik}^T \right)^{-1} [I : 0] X_{il} A_l$$

$$= A_k \sum_{i=1}^{M_k} X_{ik}^T \left( \Lambda_{ik} + Z_{ik}DZ_{ik}^T \right)^{-1} X_{ik} A_l$$

$$= A_k A_k^{-1} A_l = A_l = \text{var} \left( \hat{\beta}_l \right)$$

where $Z_{ik:il}$ denotes the rows of $Z_{il}$ corresponding to the new observations made on the $i$th individual between interim analyses $k$ and $l$ and $[A : B]$ denotes the matrix obtained by stacking the columns of the $A$ and $B$ matrices side by side.

In the proof above, we used the fact that

$$\Lambda_{il} + Z_{il}DZ_{il}^T = \begin{bmatrix} \Lambda_{ik} + Z_{ik}DZ_{ik}^T & Z_{ik}DZ_{ik:il}^T \\ Z_{ik:il}DZ_{ik}^T & \Lambda_{ik:il,ik:il} + Z_{ik:il}DZ_{ik:il}^T \end{bmatrix},$$

where $\Lambda_{ik:il,ik:il}$ denotes the covariance matrix of the new observations made on the $i$th individual between interim analyses $k$ and $l$. It follows that $[\Lambda_{ik} + Z_{ik}DZ_{ik}^T : Z_{ik}DZ_{ik:il}^T]$ represents the first $n_{ik}$ rows of $\Lambda_{il} + Z_{il}DZ_{il}^T$ and therefore

$$\left[ \Lambda_{ik} + Z_{ik}DZ_{ik}^T : Z_{ik}DZ_{ik:il}^T \right] \left( \Lambda_{il} + Z_{il}DZ_{il}^T \right)^{-1} = [I : 0].$$