STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:2457–2475
Published online 24 June 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2127

Group sequential clinical trials for longitudinal data with analyses using summary statistics

John M. Kittelson[1,*,†], Katrina Sharples[2] and Scott S. Emerson[3]

[1]*Department of Preventive Medicine and Biometrics, Box B-119, University of Colorado Health Sciences Center, Denver, Colorado 80262, U.S.A.*
[2]*Department of Preventive and Social Medicine, University of Otago, PO Box 913, Dunedin, New Zealand*
[3]*Department of Biostatistics, Box 357232, University of Washington Seattle, Washington 98195, U.S.A.*

SUMMARY

Longitudinal endpoints are used in clinical trials, and the analysis of the results is often conducted using within-individual summary statistics. When these trials are monitored, interim analyses that include subjects with incomplete follow-up can give incorrect decisions due to bias by non-linearity in the true time trajectory of the treatment effect. Linear mixed-effects models can be used to remove this bias, but there is a lack of software to support both the design and implementation of monitoring plans in this setting. This paper considers a clinical trial in which the measurement time schedule is fixed (at least for pre-trial design), and the scientific question is parameterized by a contrast across these measurement times. This setting assures generalizable inference in the presence of non-linear time trajectories. The distribution of the treatment effect estimate at the interim analyses using the longitudinal outcome measurements is given, and software to calculate the amount of information at each interim analysis is provided. The interim information specifies the analysis timing thereby allowing standard group sequential design software packages to be used for trials with longitudinal outcomes. The practical issues with implementation of these designs are described; in particular, methods are presented for consistent estimation of treatment effects at the interim analyses when outcomes are not measured according to the pre-trial schedule. *Splus*/*R* functions implementing this inference using appropriate linear mixed-effects models are provided. These designs are illustrated using a clinical trial of statin treatment for the symptoms of peripheral arterial disease. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: clinical trial; longitudinal endpoint; mixed-effects model; area under curve; software

1. INTRODUCTION

In clinical trials treatment effects are often evaluated with a continuous outcome that is measured repeatedly over time on each subject. Longitudinal endpoints are used for many reasons

*Correspondence to: John M. Kittelson, Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, 4200 E 9th Avenue, Box B-119, Denver, CO 80262, U.S.A.
†E-mail: john.kittelson@uchsc.edu

Contract/grant sponsor: NIH; contract/grant number: HL69719

Received February 2004
Accepted October 2004

[1] all of which stem from a basic scientific interest in the nature of the time trajectory of treatment effects. Within-subject summary statistics such as the minimum [2], maximum [3], rate of change [4, 5], or average [6] can be used to focus statistical inference on a clinically relevant aspect of the time trajectory [7].

In most situations the time trajectory of treatment effects is likely to be non-linear, and data-driven parameterizations of non-linearities are unlikely to extend beyond the observed time range. Estimation of treatment effects in this setting can lead to inference that changes with the length and distribution of follow-up measurement times; thus for example, the average outcome will change over time unless treatment effects happen to be constant. A treatment effect estimator will be unbiased for a fixed follow-up distribution, but will not necessarily generalize to any pattern or length of follow-up.

In clinical trials it is common to plan for follow-up at regular intervals, and as a consequence, trial results are interpretable as the effects measured with complete follow-up over the particular choice of measurement times. Thus, in a fixed-sample study (i.e. without interim analyses) inference is unbaised for the follow-up distribution and generalization is conditional on that distribution. The addition of interim analyses to a fixed-sample trial introduces the possibility that the distribution of measurement times at interim analyses will differ from that of the fixed-sample study. This will happen when at interim analyses there are subjects who have not yet completed all follow-up measurements, and as a result, the inference at interim analyses will be biased relative to that of the final analysis [8]. This is an issue with all methods for selecting interim decision rules including frequentist methods [9], error spending approaches [10], curtailment designs [11], and Bayesian methods [12].

The analysis of longitudinal endpoints is often conducted using either linear mixed-effects models [13, 14], generalized estimating equations [15, 16], or a 2-stage approach in which treatment effects are compared after first calculating a meaningful summary statistic on each subject [6, 7] (additional discussion and references are provided by Jennison and Turnbull [17, p. 233]). Unbiased inference is possible as long as the incomplete data are missing at random or completely at random [18] which is commonly the case at interim analyses. Linear mixed-effects models [19] or equivalent methods for imputation [20] can be applied to get unbiased inference. The problem can also be approached by parameterizing the time trajectory [21, 17 p. 68], however continuous-time models are also subject to bias at the interim analyses if the time-trajectory is misspecified.

To illustrate, we will consider a controlled trial of statin treatment for the symptoms of peripheral arterial disease [22]. Peripheral arterial disease (PAD) can lead to pain in the legs that prevents walking, which in turn contributes to progression of the disease. There is interest in determining if statin treatment may assist in the treatment of the symptoms of PAD. The primary outcome is the time that a patient can walk on a treadmill before they are stopped by pain (i.e. peak walk time or PWT), and this will be measured at regular intervals following randomization. When designing the PAD trial we might choose the average PWT over the follow-up times as the scientifically meaningful measure of treatment effect within each patient so that treatments are compared based on the average of this within-patient summary measure.

The objective of this paper is to describe how to design and implement interim analyses in clinical trials with longitudinal endpoints. Although previous work has shown that unbiased inference at interim analyses is possible, there is a lack of software to support the pre-trial evaluation of interim decision rules and the implementation of the design when the

measurement times differ from the pre-trial plan. We present software to extend standard group sequential design packages to include longitudinal endpoints and to address the practical issues faced when implementing such a design. Section 2 presents the methods and Section 3 illustrates their application to the PAD example described above.

2. SUMMARY STATISTICS AND GROUP SEQUENTIAL TRIALS

2.1. Within-subject summary statistics

Let $Y_{ik}(t)$ denote the outcome for the kth individual in treatment group i ($i=0,1$) at time t after study entry. Suppose that there will be a total of N_J individuals in each treatment group (i.e. $k=1,\ldots,N_J$)[‡]. Furthermore, suppose that the outcome will be measured at times $t=T_0<T_1<\cdots<T_L$ for each individual (with $T_0=0$ denoting the baseline measurement). As discussed above, we assume a fixed follow-up interval in order to assure generalizability. Let $\mathbf{Y}_{ik}$ denote the outcome vector at the L time points. Suppose that $Y_{ik}(t)$ has expectation $\mu_i(t)$ (and $E(\mathbf{Y}_{ik})=\mu_i$) and variance $\mathrm{var}(\mathbf{Y}_{ik})=\mathbf{\Sigma}_i$. In the PAD example, $Y_{ik}(T_\ell)$ denotes the exercise tolerance (in minutes) for the kth patient in the ith treatment group ($i=0$ for placebo, $i=1$ for active treatment) at measurement time T_ℓ ($T_0=$ baseline; $T_1=3$ months; $T_2=6$ months; $T_3=9$ months; $T_4=12$ months). Note that it is more efficient (powerful) to condition on baseline levels when analyzing treatment effects, and although we advocate such an approach, we do not specifically incorporate it into our notation. The practical effect of conditioning on baseline levels would be to reduce the magnitude of the covariance matrix $\mathbf{\Sigma}_i$, which could be made explicit during trial design and/or estimated at each interim analysis (see Section 3.3).

We consider a general parameterization of treatment effect obtained by a weighted sum of the response vector so that the within-subject summary statistic is the weighted sum of the individual's response vector. We refer to this statistic as the *weighted area under the response curve* (wAUC). Let $\mathbf{w}'=(w_0,\ldots,w_L)$ denote weights selected to express the scientific importance of effects at time points $T_0,\ldots,T_L$, and define the within-subject summary statistic $X_{ik}=\mathbf{w}'\mathbf{Y}_{ik}$. The average outcome with treatment i is given by $\hat{\theta}_i=\sum_k X_{ik}/N_J$. We let θ denote the effect of the new treatment relative to control, and estimate it by $\hat{\theta}_1-\hat{\theta}_0$. The following specific measures are included in the wAUC.

1. *Clinically relevant timeframe* (*last value*): $\mathbf{w}'=(0,..,0,1)$ when the scientific interest is on the response at T_L; that is, treatment effects are measured by

$$\theta=\mu_1(T_L)-\mu_0(T_L).$$

   It is also common to estimate $\mu_1(T_L)-\mu_0(T_L)$ by the *change* from baseline, which corresponds to weights $\mathbf{w}'=(-1,0,\ldots,0,1)$. This parameterization focuses entirely on the last time point even though outcome measurements are made at earlier times. In the PAD example, this outcome corresponds to measuring treatment effects by the difference in improvement in exercise tolerance after 12 months of treatment.

[‡] $j=1,\ldots,J$ will index interim analyses (Section 2.2), so N_J is the maximal sample size.

2. *Rate of change* (*slope*): Measure treatment effect by the difference in the linear time-trend for the treatment response:

$$\theta = \beta_1 - \beta_0$$

where β_i is the least squares approximation to the first order linear time trend (slope) in the response to treatment i. The slope statistic corresponds to a weight vector with ℓth element:

$$w_\ell = \frac{T_\ell - \bar{T}}{\sum_t (T_t - \bar{T})^2}$$

for $\ell = 0, \ldots, L$. In the PAD example $\mathbf{w}' = (-0.8, -0.4, 0, 0.4, 0.8)$, which corresponds to the difference in the annual rate of improvement in exercise tolerance between the two treatment groups.

3. *Area under the curve* (*auc*): Treatment effects are often parameterized as the area under the curve (usually standardized by the total measurement time T_L). Using a trapezoidal approximation corresponds to weights: $w_0 = (T_1 - T_0)/T_L$, $w_L = (T_L - T_{L-1})/T_L$, and for $\ell = 1, \ldots, L-1$:

$$w_\ell = (T_{\ell+1} - T_{\ell-1})/(2T_L)$$

The *auc* is often calculated using the difference from baseline which corresponds to using all of the above weights except $w_0 = (T_1 + T_0 - 3T_L + T_{L-1})/(2T_L)$. In the PAD example using $w_0 = -1$ and $w_\ell = 0.25$ corresponds to measuring treatment effects by the average change from baseline over all follow-up measurements.

Note that the wAUC is a function of the follow-up measurement times $T_1, \ldots, T_L$; thus, inference is referenced to the time range, and choosing a different time range will not necessarily give the same result.

The distribution of the estimated treatment effect $\hat{\theta}$ follows from the distribution of the within-subject summary statistic X_{ik}, which has expectation $E(X_{ik}) = \mathbf{w}'\mu_i$ and variance $\text{var}(X_{ik}) = \mathbf{w}'\boldsymbol{\Sigma}_i\mathbf{w}$. For large samples $\hat{\theta}$ is approximately normally distributed:

$$\hat{\theta} \sim \mathcal{N}(\theta, (V_1 + V_0)/N_J) \tag{1}$$

where $V_i = \text{var}(X_{ik})$. Thus, the fixed-sample trial with complete data on all participants can be designed using standard methods based on the above distribution for $\hat{\theta}$. Robustness to departures from normality in small sample sizes is discussed in Section 4.

Notice that in this mean-based inference with balanced data, the treatment effect is equivalently viewed as a contrast across the population mean outcomes; that is, treatment comparisons can be based on the weighted sum of the averages at each measurement time ($\hat{\theta}_i = \sum_t w(t)[\sum_k Y_{ik}(t)/N_J]$), or as the average of the within-individual summary measures ($\hat{\theta}_i = \sum_k [\sum_t w(t) Y_{ik}(t)]/N_J$). Without missing data these two approaches are identical, but with missing data naive application of the latter leads to bias. (Note that this equivalence will not hold in other settings (e.g. logistic regression) where there is a non-linear link between the mean and explanatory variables.) With large sample sizes the vector of the average outcome at each of the L time points is multivariate Normal with expectation μ_i

and variance $\boldsymbol{\Sigma}_i/N_J$, so that $\hat{\theta}$ is distributed as above (equation (1)). This alternative formulation only requires an appropriate covariance structure for the averages at each time point and does not require assumptions about the distribution of the outcome within an individual.

Notice also that because the treatment effect is defined according to the particular choice of weights $\mathbf{w}$, these weights also define the set of μ_i that will be considered as null and alternative hypotheses. For example, if statin treatment causes a transient increase in PWT that disappears by 12-months, then *auc* weights classify such effects as part of the alternative hypothesis whereas last-value weights classify the same effect as part of the null hypothesis. It is therefore essential that the weights reflect the clinical questions.

2.2. Interim analyses with summary statistics

Now suppose that the fixed-sample trial described above will be monitored in J interim analyses. Let $N_j^{(L)}$ denote the number of subjects who have completed follow-up at all L time points at the jth interim analysis in each treatment group. We assume that $N_1^{(L)} > 0$ so that at least one subject has completed follow-up at the time of the first interim analysis. To define the amount of information (complete and incomplete) at the jth interim analysis we let $N_j^{(\ell)}$ denote the number of subjects with outcome measured at times $T_1, \ldots, T_\ell$ (i.e. with exactly ℓ outcome measurements). For example, in the PAD trial $N_2^{(3)}$ would be 20 if at the second interim analysis there were 20 subjects with outcome measured at all but the final (12-month) time point (Section 3.2 provides a complete illustration). At the jth analysis the total number of subjects with one or more follow-up measurements is $N_j = \sum_{\ell=1}^{L} N_j^{(\ell)}$. We index the timing of interim analyses by the number of subjects with complete follow-up $N_j^{(L)}$.

It is possible to use standard group sequential designs if interim analyses are restricted to subjects who have completed follow-up. In this case the interim analyses occur after collecting $100 \times N_j^{(L)}/N_J$ per cent of the total information. Restricting attention to subjects with complete follow-up ignores the information that might be contained in the other subjects. It may be possible to increase efficiency by including all follow-up information, especially if the weights emphasize early effects over late effects.

To include all information we define $\mathbf{Y}_{ik}^{(\ell)}$ ($\ell = 1, \ldots, L$) as the outcomes in the $N_j^{(\ell)}$ subjects with exactly ℓ follow-up measurements; specifically, $\mathbf{Y}_{ik}^{(\ell)} = (Y_{ik}(T_0), \ldots, Y_{ik}(T_\ell))$. The expectation of $\mathbf{Y}_{ik}^{(\ell)}$ is the first ℓ elements of μ_i which we denote by $\mu_{i(\ell)}$ and its variance is given by the upper $\ell \times \ell$-element submatrix within $\boldsymbol{\Sigma}_i$ which we denote by $\boldsymbol{\Sigma}_{i(\ell)}$. Let $\bar{\mathbf{Y}}_{i\cdot j}^{(\ell)}$ denote the vector of averages among subjects with exactly ℓ outcome measurements at the jth interim analysis. For large sample sizes, $\bar{\mathbf{Y}}_{i\cdot j}^{(\ell)}$ will follow an ℓ-dimensional normal distribution with mean vector $\mu_i^{(\ell)}$ and variance matrix $\boldsymbol{\Sigma}_{i(\ell)}/N_j^{(\ell)}$. The likelihood is

$$L = \prod_{\ell=1}^{L} \sqrt{\frac{N_j^{(\ell)}}{(2\pi)^\ell} \det \boldsymbol{\Sigma}_{i(\ell)}^{-1}} \exp\left\{ \frac{-N_j^{(\ell)}}{2} (\bar{\mathbf{Y}}_{i\cdot j}^{(\ell)} - \mu_i^{(\ell)})' \boldsymbol{\Sigma}_{i(\ell)}^{-1} (\bar{\mathbf{Y}}_{i\cdot j}^{(\ell)} - \mu_i^{(\ell)}) \right\}$$

which is the same as that given by Jennison and Turnbull [17] or Galbraith and Marshner [19]. The maximum likelihood estimate of the mean outcome vector at the jth interim

analysis $\hat{\mu}_{ij}$ is

$$\hat{\mu}_{ij} = \mathbf{V}_{ij} \sum_{\ell=1}^{L} \boldsymbol{\Sigma}_{i(\ell)}^{-1} \bar{\mathbf{Y}}_{i\cdot j}^{(\ell)} N_j^{(\ell)}$$

where $\mathbf{V}_{ij}$ is the variance of $\hat{\mu}_{ij}$:

$$\mathbf{V}_{ij} = \left[\sum_{\ell=1}^{L} \boldsymbol{\Sigma}_{i(\ell+)}^{-1} N_j^{(\ell)} \right]^{-1}$$

(Notation: $\boldsymbol{\Sigma}_{i(\ell+)}^{-1}$ denotes $\boldsymbol{\Sigma}_{i(\ell)}^{-1}$ augmented with 0's to increase the dimension to $L \times L$ so that the above sum is properly defined.) At the jth interim analysis, the treatment effect $\hat{\theta}_j$ and its variance are given by:

$$\hat{\theta}_j = \mathbf{w}'(\hat{\mu}_{1j} - \hat{\mu}_{0j})\mathbf{w}$$

$$\text{var}(\hat{\theta}_j) = \mathbf{w}'(\mathbf{V}_{1j} + \mathbf{V}_{0j})\mathbf{w} \tag{2}$$

Using this variance incorporates the information on subjects who have not yet completed follow-up into the usual group sequential design framework.

Standard mixed-effects models can be used to obtain $\hat{\mu}_{ij}$ and its estimated variance $\hat{\mathbf{V}}_{ij}$. Such models would use time as a factor variable with an unstructured covariance matrix with possible adjustment for covariates (see Section 2.4) The average outcome in treatment i at the jth interim analysis is then given by $\hat{\theta}_{ij} = \mathbf{w}'\hat{\mu}_{ij}$ which has variance $\text{var}(\hat{\theta}_{ij}) = \mathbf{w}'\hat{\mathbf{V}}_{ij}\mathbf{w}$.

2.3. The design of group sequential trials with longitudinal endpoints

To design a group sequential trial requires the test statistic, its distribution, and the timing of the interim analyses. The test statistic and its distribution are given in the previous section, and analysis timing is determined by the amount of statistical information that has been accrued. With a longitudinal endpoint the information is a function of the covariance, the total number of subjects, and the amount of follow-up on each subject; thus, the timing of the jth interim analysis is described by $\boldsymbol{\Sigma}_i$ and $N_j^{(\ell)}$ (for $\ell = 1, \ldots, L$).

The appendix (Section A.1) describes an *Splus* (or *R* [23]) function `LMEinfo` that calculates the timing of interim analyses from $\boldsymbol{\Sigma}_i$ and $N_j^{(\ell)}$ for $\ell = 1, \ldots, L$ and $j = 1, \ldots, J$. The function returns the timing of the interim analyses in terms of the effective sample size; i.e. the product of the information and the maximal sample size. The effective sample size is between the total number of subjects enrolled and the number of subjects who have completed follow-up, and can be used in standard software packages such as *Splus SeqTrial* [24], *Pest* [25], or *EaST* [26] to describe interim analysis timing. The use of this function is illustrated using the PAD example in Section 3.

2.4. Flexibility during design implementation

For the purposes of pre-trial planning, the above designs have assumed that follow-up measurements are taken at the same time points in all subjects. However methods for implementation of the design must allow for deviations from this pre-trial plan. In fact it is possible that the distribution of follow-up measurements is nearly continuous even though the pre-trial plan calls for only a few follow-up measurement times.

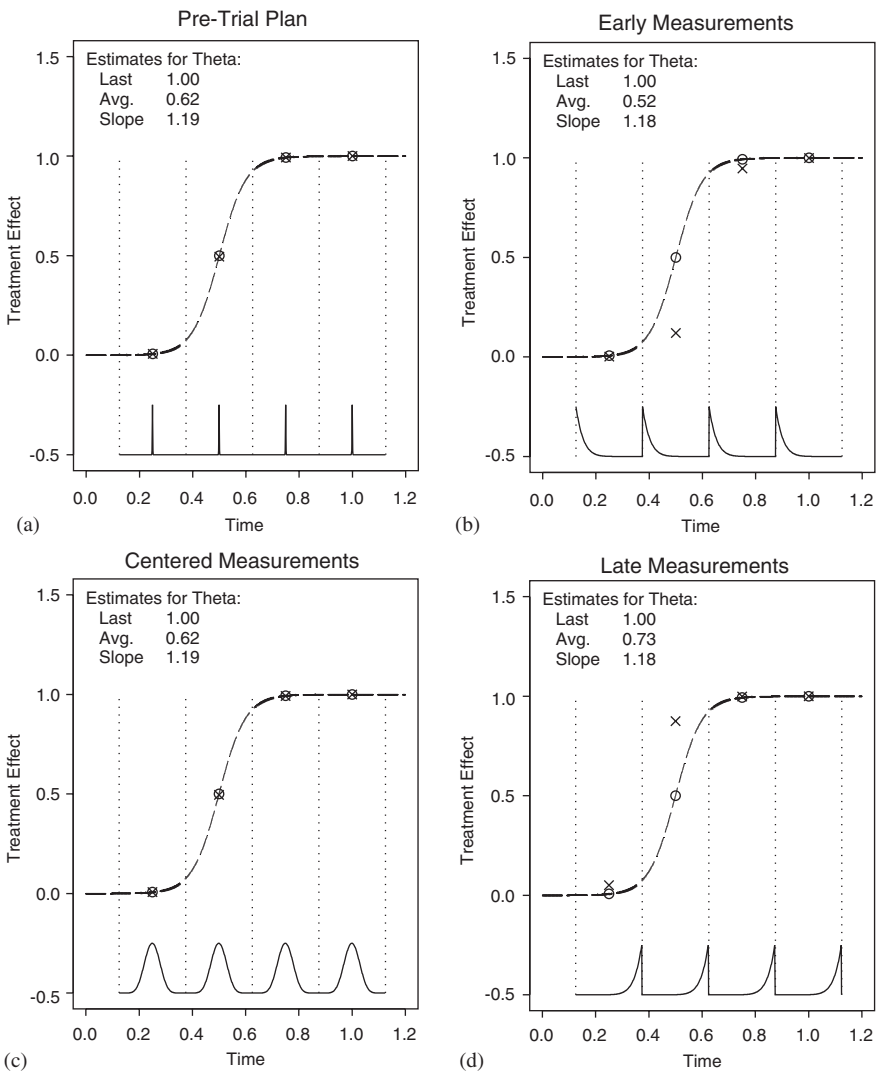

Figure 1. Estimation by mapping each measurement to its pre-trial time point when treatment effects are non-linear. Dotted vertical lines denote mapping window, solid lines at bottom of each panel denote the distribution of measurement times, circle denotes true effect, X denotes the estimate obtained by mapping.

The simplest approach to a non-discrete distribution of follow-up measurements is to map each measurement to one of the planned times. Mapping a measurement will produce unbiased estimates as long as the treatment effect is constant within the mapping window which is more likely if the window is small. Figure 1 illustrates the potential for bias when the time-trajectory for the treatment effect is non-linear and measurements do not occur according to the pre-trial schedule. Panel (a) shows the value of three summary measures (last value,

average, and slope) when all subjects are measured according to the pretrial plan. Panels (b)–(d) show these same summary measures when measurement times are shifted to early in the window, centred within the window, or shifted to late in the window. Bias is large if treatment effects are non-linear and measurements are not centred around the pre-trial time point (e.g. second and third measurement windows in panels b and d). However, the estimates have little bias if either treatment effects are constant (first and fourth windows in all panels) or if measurements are centred around the pre-planned time (all windows in panel c). Figure 1 also shows that although differences from the pre-trial measurement plan can lead to bias in the overall summary measure (e.g. the average measure in panels b and d), it is also possible that the bias at one measurement time will be offset by an opposite bias at another time so that the overall summary measure has little bias (e.g. the slope summary measure does not differ much across panels).

If it is not reasonable to map measurements to a discrete time point, then it is necessary to account for possible time trends in the treatment effect. With truly continuous time measurements it is possible to use growth curve models to estimate a weighted area under the treatment curve; i.e. $\theta_i = \int w(t)\mu_i(t)\,\mathrm{d}t$. However this approach either requires some knowledge of the form of the time trend or enough data to deduce an approximate form. In clinical trials we do not usually know how treatment effects are likely to evolve, and at interim analyses there is often insufficient data to deduce treatment effect time trends, so a more robust approach is required.

To minimize bias we consider using a piece-wise linear approximation to the treatment time-trend within estimation windows centred on the pre-specified measurement times. The treatment effects at the discrete times can be estimated from these linear approximations and then used to estimate θ. We define estimation windows which span the interval from the midpoints between the discrete time points; i.e. the ℓth window runs from $(T_{\ell-1}+T_\ell)/2$ to $(T_\ell+T_{\ell+1})/2$), with 0 as the lower limit of the first window and the maximum time measurement as the upper limit of the Lth window. Within each window a least-squares linear approximation to the treatment time trend is used to estimate the outcome at the time point of interest. The appendix (Section A.2) describes an *Splus*/*R* function (`ThetaEst`) that defines the estimation windows, fits a piece-wise linear mixed effects model, obtains estimates for $\mu_i(T_\ell)$ using this model, and uses these to estimate θ and its variance. Figure 2 illustrates estimation based on a piece-wise linear approximation to the non-linear function in Figure 1. The biases that were present in Figure 1 are reduced by the piecewise interpolation method. It is of course possible for other types of non-linear treatment effects or other follow-up measurement patterns to produce bias, and in such cases other interpolation methods (e.g. splines or lowess) might be useful.

Finally, recall that the use of a discrete follow-up distribution was motivated by the need for reproducible results if treatment effects are non-linear. The above approaches use a contrast across μ_i at the pre-defined times even though the follow-up measurements are not at these same times. This approach should also be reproducible as long as the piece-wise linear approximation is adequate.

In addition to deviations from the pre-trial measurement plan, group sequential trials for longitudinal endpoints must also allow flexibility in the number and timing of the interim analyses. Methods for non-longitudinal endpoints are readily applied to longitudinal endpoints to account for misspecification of the covariance matrix and for deviations in the pre-trial distribution of follow-up information (see Section 3.3). Furthermore, analysis of trial results should also be adjusted for the bias introduced by sequential testing. Standard approaches [27]

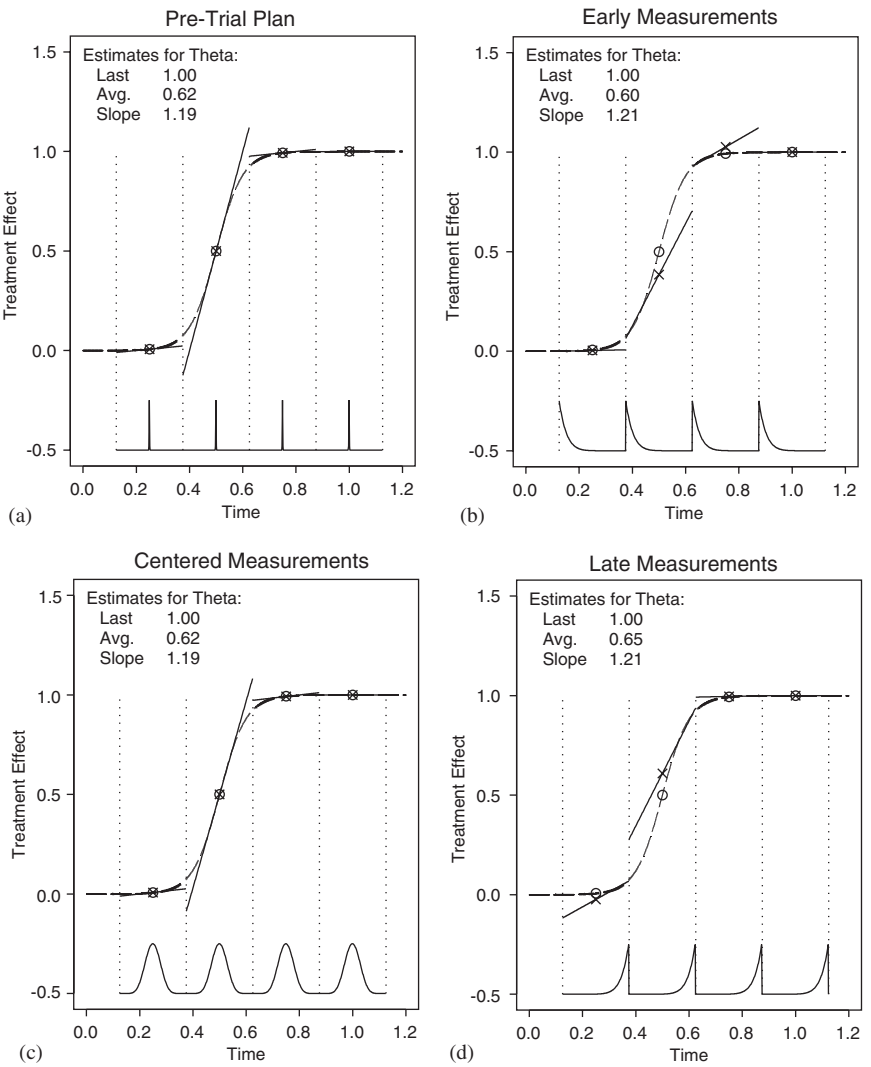

Figure 2. Estimation using a piece-wise linear approximation (lines within each mapping window) to the non-linear function. Plotting symbols are the same as described for Figure 1.

(and software) for bias adjustment also apply to the treatment effect estimate $\hat{\theta}_j$ derived from the longitudinal data.

3. EXAMPLE

3.1. Fixed-sample design

Consider a fixed-sample design for a randomized placebo-controlled clinical trial of statin treatment for PAD as described in the introduction. Suppose that peak walk time will be measured

after 12-months of treatment only (without repeated measurements); i.e. $\theta = \mu_1(T_L) - \mu_0(T_L)$ and $\hat{\theta} = \bar{Y}_{1\cdot}(T_L) - \bar{Y}_{0\cdot}(T_L)$. Based on previous trials in this setting [22] a reasonable between-subject variance for PWT is 160^2, and for design purposes we set the between-subject variance in subjects assigned to statin treatment at 180^2 to reflect a likely increase in variability due to the intervention. It follows that with 160 subjects per group $\text{var}(\hat{\theta}) = \frac{160^2}{160} + \frac{180^2}{160} = 19.04^2$, so if treatment effects are measured by the 12-month difference, the study will have 88 per cent power to detect a 60-second difference in PWT (a reasonable design point based on previous trials).

Addition of a baseline measurement can improve power, and it is common to measure outcome by the *change* in PWT $X_{ik} = Y_{ik}(T_L) - Y_{ik}(T_0)$ with treatment effect estimated by $\hat{\theta} = \bar{X}_1 - \bar{X}_0$. The variance of this estimate is a function of the correlation between measurements, which for placebo treatment is approximately $\rho = 0.6$. If statin treatment does not affect correlation, then $\text{var}(\hat{\theta}) \approx 17$, so that power is about 94 per cent. Power can be further increased by analyzing treatment effects conditional on baseline PWT.

Since PAD is a progressive disease it might be scientifically relevant to interpret the change from baseline as an annual rate of change. With this scientific interpretation and repeated measurements at times 0 (baseline) 3, 6, 9, and 12-months, the annual rate of change can be measured by the slope (i.e. $\mathbf{w}' = (-0.8, -0.4, 0, 0.4, 0.8)$). The power of a design for the general summary measure $\theta = \mathbf{w}'(\mu_1 - \mu_0)$ depends on the nature of the time trajectory and the covariance matrixes $\boldsymbol{\Sigma}_i$. As commonly happens at the design stage, $\boldsymbol{\Sigma}_i$ is not known, and we must evaluate operating characteristics under assumed covariance matrixes. Suppose that the placebo covariance matrix is exchangeable with correlation 0.6 and variance 160^2 but that active treatment induces a mean-variance relationship as described in the appendix (Section A.3 and Table V). Obviously, these covariance matrixes are only pre-trial guesses that would be revised using data at the interim analysis (see Section 3.3).

We explore the power of the design to detect the following alternative hypotheses:

Late effect:	$\mu_1' = (0, 0, 0, 0, 60)$
Immediate effect:	$\mu_1' = (0, 60, 60, 60, 60)$
Intermediate effect:	$\mu_1' = (0, 0, 0, 60, 60)$
Linear effect:	$\mu_1' = (0, 15, 30, 45, 60)$
Transient effect:	$\mu_1' = (0, 60, 60, 60, 0)$

We assume $\mu_1(0) = 0$ and $\mu_0' = (0, 0, 0, 0, 0)$ which does not affect study operating characteristics. All but the last of these alternatives show a 60-second improvement in PWT after 12-months of treatment, and represent non-null effects under the last-value, slope, or average summary measures. The transient effect is a null effect under either the last-value or slope summary measures, but is part of the alternative for the average summary measure. In a fixed sample design the value for θ and the standard error of its estimate are calculated according to equation (2). Table I shows these values and study power under the various design alternatives μ_1 and possible weights $\mathbf{w}$. The power is consistently high if 'last-value' weights are used (except for transient effects), but may be reduced under other weights.

3.2. Group sequential design

Now consider a group sequential design for the PAD trial. With non-longitudinal endpoints, the timing of interim analyses is usually described by the number of subjects enrolled at the

Table I. Values for θ, the standard error of $\hat{\theta}$, and power (β) in a fixed-sample trial under various weightings and values for the true treatment effect μ_1.

	Weighting (**w**)								
	Last value			*auc*			Slope		
μ_1	θ	SE($\hat{\theta}$)	β	θ	SE($\hat{\theta}$)	β	θ	SE($\hat{\theta}$)	β
(0, 0, 0, 0, 60)	60.0	17.3	0.93	15.0	12.8	0.22	48.0	15.3	0.88
(0, 60, 60, 60, 60)	60.0	17.3	0.93	60.0	14.3	0.99	48.0	15.3	0.88
(0, 0, 0, 60, 60)	60.0	17.3	0.93	30.0	13.1	0.63	72.0	16.3	0.99
(0, 15, 30, 45, 60)	60.0	17.3	0.93	37.5	13.3	0.81	60.0	15.8	0.97
(0, 60, 60, 60, 0)	0.0	16.0	0.025	45.0	13.6	0.91	0.0	14.3	0.025

time of the analysis. For the purposes of pre-trial evaluation of monitoring plans we assume that there will be 5 interim analyses after enrolling groups of 80 patients (40/treatment arm). The information at each interim analysis is also a function of the distribution of follow-up measurements (i.e. $N_j^{(\ell)}$ for $\ell = 1, \ldots, L$), and we choose the following as a reference for evaluation of design properties:

Interim	Number of subjects			
analysis (j)	$N_j^{(1)}$	$N_j^{(2)}$	$N_j^{(3)}$	$N_j^{(4)}$
1	10	10	10	10
2	10	10	10	50
3	10	10	10	90
4	10	10	10	130
5	0	0	0	160

For example at the second interim analysis 10 subjects (per arm) have completed only the first follow-up assessment, 10 subjects have completed the first and second follow-up assessments, 10 subjects have completed the first, second, and third assessments, and 50 subjects have completed all four assessments. We exclude any subjects who have been enrolled (and therefore have a baseline measurement), but who have not yet had the first response measurement.

The amount of information at each interim analysis for each treatment group is calculated using the function `LMEInfo`. Table II shows the `LMEInfo` output for average weights $\mathbf{w}' = c(-1, 0.25, 0.25, 0.25, 0.25)$ and for a treatment covariance matrix corresponding to the linear evolution of treatment effects $\mu_1 = c(0, 15, 30, 45, 60)$. Table III shows the effective sample size and variance under all of the combinations of treatment effects and weights described above. This effective sample size and variance can be used in standard group sequential design software packages to evaluate and select stopping boundaries.

Table IV shows stopping boundaries for group sequential designs with longitudinal outcomes using a Pocock or O'Brien–Fleming boundary shape [9]. It is apparent that using the longitudinal data allows smaller critical values than if only complete cases are used. Notice that the 'average' weights incorporate a much greater proportion of the total information at the interim analyses compared with either the 'last-value' or 'slope' weights (Table III). Although

Table II. Comparing the standard error of $\hat{\theta}$ and information growth for analyses based on all follow-up measurements versus analyses based only on complete cases. $\mu_1' = (0, 15, 30, 45, 60)$ and $\mathbf{w}' = (-1, 0.25, 0.25, 0.25, 0.25)$.

	All follow-up data		Complete cases	
Interim analysis	SE($\hat{\theta}$)	Effective sample size*	SE($\hat{\theta}$)	Actual sample size†
1	30.83	29.70	53.14	10
2	19.68	72.92	23.77	50
3	15.78	113.47	17.71	90
4	13.56	153.70	14.74	130
5	13.29	160.00	13.29	160

*Information at the interim analyses when all data are used at the interim analysis.
†Number of subjects who have completed 12-months of follow-up.

Table III. Effective sample size (per group) at each interim analysis under various weightings and several possible treatment effects.

	Interim analysis				
μ_1'	1	2	3	4	5
Complete cases only:					
(All μ_1)	10.0	50.0	90.0	130.0	160
Including all follow-up information:					
$\mathbf{w}' = (-1, 0, 0, 0, 1)$					
$(0, 0, 0, 0, 60)$	13.2	56.6	97.5	138.0	160
$(0, 60, 60, 60, 60)$	15.1	59.7	100.8	141.4	160
$(0, 0, 0, 60, 60)$	14.0	57.5	98.4	138.9	160
$(0, 15, 30, 45, 60)$	14.3	58.3	99.3	139.8	160
$(0, 60, 60, 60, 0)$	13.4	57.1	98.1	138.5	160
$\mathbf{w}' = (-1, 0.25, 0.25, 0.25, 0.25)$					
$(0, 0, 0, 0, 60)$	29.1	72.8	113.4	153.6	160
$(0, 60, 60, 60, 60)$	31.3	74.2	114.7	154.9	160
$(0, 0, 0, 60, 60)$	28.9	72.2	112.8	153.0	160
$(0, 15, 30, 45, 60)$	29.7	72.9	113.5	153.7	160
$(0, 60, 60, 60, 0)$	30.8	74.0	114.5	154.7	160
$\mathbf{w}' = (-0.8, -0.4, 0, 0.4, 0.8)$					
$(0, 0, 0, 0, 60)$	13.8	56.5	97.1	137.4	160
$(0, 60, 60, 60, 60)$	15.3	58.6	99.3	139.7	160
$(0, 0, 0, 60, 60)$	15.0	57.6	98.2	138.4	160
$(0, 15, 30, 45, 60)$	15.1	58.0	98.7	139.0	160
$(0, 60, 60, 60, 0)$	14.2	57.1	97.8	138.1	160

this might be used to motivate the use of average weights, caution is warranted since these weights can result in a substantial loss of power if treatment effects evolve slowly (Table I). In the absence of a strong scientific motivation for using one of the three weighting schemes, we would be inclined to use 'last-value' weights because they maintain power and still allow a reasonable increase in efficiency at the interim analyses.

Table IV. Group sequential stopping boundaries.

Interim analysis	O'Brien–Fleming boundary shape			Pocock boundary shape		
	Effective sample size	Lower boundary	Upper boundary	Effective sample size	Lower boundary	Upper boundary
O'Brien–Fleming boundary shape:						
Complete cases only						
1	10	−487.2	556.8	10	−83.8	167.5
2	50	−41.8	111.4	50	8.8	74.9
3	90	7.7	61.9	90	27.9	55.8
4	130	26.8	42.8	130	37.3	46.5
5	160	34.8	34.8	160	41.9	41.9
All data and $\mathbf{w}' = (-1, 0, 0, 0, 1)$						
1	14	−328.8	398.6	14	−57.3	140.4
2	58	−26.5	96.2	58	14.1	69.0
3	99	13.4	56.4	99	30.3	52.8
4	139	29.6	40.1	139	38.5	44.6
5	160	34.9	34.9	160	41.5	41.5
All data and $\mathbf{w}' = (-1, 0.25, 0.25, 0.25, 0.25)$						
1	29	−117.1	183.6	29	−13.6	91.4
2	73	−6.4	72.9	73	20.2	57.6
3	113	19.4	47.1	113	31.5	46.3
4	154	32.0	34.6	154	38.1	39.6
5	160	33.3	33.3	160	38.9	38.9

3.3. Implementation

Suppose that the PAD trial is designed using the average statistic ($\mathbf{w}' = (-1, 0.25, 0.25, 0.25, 0.25)$) and the corresponding O'Brien–Fleming boundary shape (Table IV). Suppose that the first interim analysis deviates from the timing of the pre-trial plan by occurring when 30 subjects have been followed for 12 months, 20 subjects for 9 months, 15 for 6 months, and 20 for 3 months. To illustrate issues in implementation, data were simulated using a covariance structure that differed from the pre-trial assumptions described above and with a non-discrete measurement distribution using piece-wise linear interpolation to estimate θ. The data frame containing measurements for these 85 subjects was then analyzed using `ThetaEst` giving $\hat{\theta} = 63.1$ with variance $= 329.9$.

The effect of the above deviations from the pre-trial plan is that the first interim analysis is occurring at a different point in information-time. The interim analysis could proceed by ignoring these differences and using the pre-trial stopping rule. The actual operating characteristics with this approach do not differ dramatically from the pre-trial characteristics as long as the deviations are not major [28]. Alternatively, the stopping rule can be recalculated using the observed proportion of total information. If the trial were to continue until completion, then the variance of $\hat{\theta}$ can be estimated by $\mathbf{w}'(\hat{\boldsymbol{\Sigma}}_0 + \hat{\boldsymbol{\Sigma}}_1)\mathbf{w}/160$ where $\hat{\boldsymbol{\Sigma}}_i$ are the observed covariance matrixes calculated by `ThetaEst`. In the simulated example $\mathbf{w}'(\hat{\boldsymbol{\Sigma}}_0 + \hat{\boldsymbol{\Sigma}}_1)\mathbf{w}/160 = 81.69$; thus, $100 \times 81.69/329.9 = 25$ per cent of the total information has been accrued. A recalculated stopping rule can be obtained by interpolating between the pre-trial rules, by using an error-spending function [10] that approximates the selected OBF design, or by recalculating the OBF stopping rule using the actual information at the first analysis [29]. Suppose we

choose to interpolate; 25 per cent of total information corresponds to an effective sample size of 40 per group which is 25 per cent of the distance between the effective sample sizes of 29 and 73 that represent the first and second analyses in the pre-trial plan (Table IV). Interpolating between the corresponding stopping rules implies that the trial should be stopped for lack of efficacy if $\hat{\theta} < -89.4$ and for efficacy if $\hat{\theta} > 155.9$. Since $\hat{\theta} = 63.1$ the trial would not be stopped. If a termination recommendation were warranted, then the estimated effect $\hat{\theta}$ should be adjusted for the bias due to sequential testing using the same methods that are used for non-longitudinal endpoints [27]. As in any group sequential design, the revised stopping rule and analysis timing would require revision of future stopping rules in order to maintain operating characteristics [29].

4. DISCUSSION

The methods and software described in this paper allow the design and implementation of group sequential monitoring plans with longitudinal endpoints. We have based these methods on a population-level contrast across treatment outcomes at discrete measurement times which we interpret as a general weighting of the area under the response curve. We use this framework in order to assure generalizability in trial results even when the time trajectory for treatment effects is non-linear. This approach will be less efficient than using continuous-time (growth curve) mixed-effects models if those models are structured around the true time trajectory, however it will be not be biased when the time trajectory is misspecified.

When outcomes are not measured on a regular follow-up schedule our approach is to use the data to estimate treatment effects at pre-defined reference time points which in turn estimate the *w*AUC. It is also possible to define a continuous weight function and use a non-parametric mixed-effects smoothing function [30] to estimate the weighted area under the continuous response curve. In such situations study generalizability must still be conditional on the length of the follow-up interval.

We have considered the problem of conducting interim analyses when at least one subject has completed the study (i.e. has a measurement at time T_L). It is possible that early interim analyses will be necessary before observing any outcome at time T_L. For example, in a surgery trial it may be necessary to weigh early morbidity against the potential for later benefit when there are no data to estimate that benefit. Although the statistical problems with extrapolating beyond the range of the data are well known, it may still be necessary to make interim decisions in such situations. Similarly, the use of mixed-effects models with very small sample sizes (e.g. when only a few subjects have complete follow-up) can give inaccurate inference, yet interim decisions are still required. Methods that allow interim decision making when there is little or no information at longer-term follow-up times are subjects of current research.

We note that if there are non-constant treatment effects, then an analysis using repeated measurements can in fact have *less* power than an analysis using just the final measurement (Table I). Although this might argue against using repeated measurements in a fixed-sample trial, these early measurements may be important in a group sequential trial because at interim analyses they can be used to predict later treatment effects.

The focus of this work is on trials with continuous outcome measurements. Similar issues exist in trials with survival endpoints in the presence of non-proportional hazards [31]. These issues also affect interim analyses in trials with Poisson outcomes or recurrent binary endpoints. Regardless of outcome type, the objective should be to maintain trial reproducibility and generalizability when the nature of the time trajectory of treatment effects is unknown. The approach of this paper is to select the distribution of outcome measurement times and direct estimation at a scientifically meaningful contrast across those pre-defined times. This basic approach can be applied to any type of outcome.

APPENDIX A: SOFTWARE[§]

A.1. Information at interim analyses (LMEInfo)

Description: Calculates the effective sample size at the interim analyses when including subjects with incomplete follow-up in a trial with longitudinal outcome measurements.

```
LMEInfo <- function(Wt,V0,V1,N) {
    L <- nrow(V0)
    J <- nrow(N)
    rslt <- NULL
    for (j in 1:J) {
        Nj <- matrix(NA,nc=L,nr=L)
        V0inv.j <- matrix(0,nc=L,nr=L)
        V1inv.j <- matrix(0,nc=L,nr=L)
        for (d in 1:L) {
          V0inv.j[1:d,1:d] <- V0inv.j[1:d,1:d] + solve(V0[1:d,1:d])*N[j,d]
          V1inv.j[1:d,1:d] <- V1inv.j[1:d,1:d] + solve(V1[1:d,1:d])*N[j,d]
        }
    V0j <- solve(V0inv.j)
    V1j <- solve(V1inv.j)
    tmp <- t(Wt) %*% (V0j + V1j) %*% Wt
    rslt <- c(rslt,sqrt(tmp))
    }
    tmp <- sqrt(t(Wt) %*% (V0 + V1) %*% Wt)/sqrt(N[,L])
    rslt <- cbind(SElme=rslt,
        Info.lme=(1/rslt)^2,
        SampleSize.lme=max(N)*(rslt[J]/rslt)^2,
        SEcompl=tmp, Info.compl = 1/tmp^2,
        SampleSize.compl=N[,L])
    rslt
}
```

[§]Functions can be downloaded from *http://www.uchsc.edu/pmb/biom/kittelson/jk.htm*

Arguments:

`Wt`: Weights used to define *w*AUC.

`V0`: Covariance matrix for control group.

`V1`: Covariance matrix for treatment group.

`N`: Matrix of dimension $J \times (L+1)$ giving $N_j^{(\ell)}$ where rows represent interim analyses ($j = 1, \ldots, J$) and columns denote measurement times ($\ell = 0, \ldots, L$).

Value: Matrix with J rows (one for each interim analysis) and 6 columns:

Column 1: Standard error of $\hat{\theta}$ at each interim analysis when all data are used.
Column 2: Statistical information at each interim analysis when all data are used.
Column 3: Effective sample size at each interim analysis when all data are used.
Column 4: Standard error of $\hat{\theta}$ when interim analyses are based on complete cases only.
Column 5: Effective sample size when interim analyses are based on complete cases only.
Column 6: Statistical information when interim analyses are based on complete cases only.

A.2. Estimation of treatment effects (`ThetaEst`)

Description: Estimation of θ, its standard error, and the covariance matrixes $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ using data at an interim analysis.

```
ThetaEst <- function(dta,T.Infer,Wt) {
    L <- length(T.Infer)
    breakpts <- (T.Infer[-1] + T.Infer[-L])/2
    breakpts <- c(0,breakpts,max(dta$T.meas)*1.00001)
    rslt <- NULL
    indx <- 0
    for (g in unique(dta$group)) {
    indx <- indx + 1
    grp <- dta$group == g
    grpdta <- dta[grp,]
    PieceLin.B0 <- matrix(0,nr=nrow(grpdta),nc=L)
    PieceLin.B1 <- matrix(0,nr=nrow(grpdta),nc=L)
    LinInterp <- rep(T,L)
    for (i in 1:L) {
        sub <- grpdta$T.meas >= breakpts[i] & grpdta$T.meas < breakpts[i + 1]
        PieceLin.B0[sub,i] <- 1
        PieceLin.B1[sub,i] <- grpdta$T.meas[sub]
        if (length(unique(grpdta$T.meas[sub])) == 1) LinInterp[i] <- F }
    X <- cbind(PieceLin.B0,PieceLin.B1[,LinInterp],grpdta$CoVars)
    y <- grpdta$y
    id <- grpdta$id
    REmat <- X[,1:L]
    zz <- lme(fixed = y ~ -1 + X, random= ~ -1 + REmat | id,method="ML")
    TimeCntrst <- cbind(diag(rep(1,L)),diag(T.Infer))
    TimeCntrst <- TimeCntrst[,c(rep(T,L),LinInterp)]
```



```
    dims <- length(TimeCntrst)
    Theta <- t(Wt )%*% (TimeCntrst %*% zz$coef$fix[1:dims])
    Var.Theta <- zz$varFix[1:dims,1:dims]
    Var.Theta <- TimeCntrst %*% Var.Theta %*% t(TimeCntrst)
    Var.Theta <- t(Wt) %*% Var.Theta %*% Wt
    Sigma <- (as.matrix(zz$model$re$id) + diag(rep(1,L)))*zz$sigma^2
    tmp <- list(Theta=c(Theta, Var.Theta),Sigma=Sigma)
    rslt[[indx]] <- tmp
    }
    names(rslt) <- paste("Group",unique(dta$group),sep="")
    rslt
}
```

Arguments:

`dta` Data frame of results at the interim analysis with elements:
- `$id` Factor variable with individual subject id number.
- `$group` 0-1 indicator for treatment group.
- `$y` Outcome measurement.
- `$T.meas` Time at which outcome was measured.
- `$CoVar` Optional matrix of covariates that are included in the mixed-effects model. In particular, baseline measurements can be included for increased precision.

`T.Infer` Vector of discrete measurement times used to calculate θ (in text: $T_0, \ldots, T_L$)

`Wt` Vector of weights (length L) used to calculate θ.

Value: List of lists (one for each treatment group). Each of the treatment group lists has two elements:

`$Theta`: Vector with first element $\hat{\theta}$ and second element var($\hat{\theta}$).

`$Sigma`: Estimated covariance matrix for the treatment group.

At an interim analysis, the variance is used to estimate the current information in the trial. The magnitude of $\hat{\theta}$ is compared with the design stopping rules to determine if termination is warranted. The above mixed-effects models will provide unbiased estimates of θ using all available data as long as any missing follow-up measurements are missing at random or missing completely at random.

A.3. Multiplicative mean-variance relationship

Many treatments inflate the variance. The following *Splus/R* simulation uses a random multiplicative offset (following a lognormal distribution) to get a covariance matrix for use as the alternative variance `V1` in the function `LMEInfo`. The argument `sd = 0.82` controls the magnitude of the variance, `mu` denotes the vector of means corresponding to μ_1, and `Sigma0` denotes the control-group covariance matrix $\mathbf{\Sigma}_0$.

Table A1. Covariance matrices for the active treatment group ($\boldsymbol{\Sigma}_1 = \boldsymbol{\sigma}_1 \boldsymbol{\sigma}_1' \times \mathbf{C}_1$).

μ_1	$\boldsymbol{\sigma}_1'$	$\mathbf{C}_1$
(0, 0, 0, 0, 60)	(160, 160, 160, 160, 180)	$\begin{bmatrix} 1.00 & 0.53 & 0.53 & 0.53 & 0.60 \\ 0.53 & 1.00 & 0.68 & 0.68 & 0.53 \\ 0.53 & 0.68 & 1.00 & 0.68 & 0.53 \\ 0.53 & 0.68 & 0.68 & 1.00 & 0.53 \\ 0.60 & 0.53 & 0.53 & 0.53 & 1.00 \end{bmatrix}$
(0, 60, 60, 60, 60)	(160, 180, 180, 180, 180)	$\begin{bmatrix} 1.00 & 0.53 & 0.53 & 0.53 & 0.60 \\ 0.53 & 1.00 & 0.68 & 0.68 & 0.53 \\ 0.53 & 0.68 & 1.00 & 0.68 & 0.53 \\ 0.53 & 0.68 & 0.68 & 1.00 & 0.53 \\ 0.60 & 0.53 & 0.53 & 0.53 & 1.00 \end{bmatrix}$
(0, 0, 0, 60, 60)	(160, 160, 160, 180, 180)	$\begin{bmatrix} 1.00 & 0.53 & 0.53 & 0.53 & 0.60 \\ 0.53 & 1.00 & 0.68 & 0.68 & 0.53 \\ 0.53 & 0.68 & 1.00 & 0.68 & 0.53 \\ 0.53 & 0.68 & 0.68 & 1.00 & 0.53 \\ 0.60 & 0.53 & 0.53 & 0.53 & 1.00 \end{bmatrix}$
(0, 15, 30, 45, 60)	(160, 161, 165, 172, 180)	$\begin{bmatrix} 1.00 & 0.53 & 0.53 & 0.53 & 0.60 \\ 0.53 & 1.00 & 0.68 & 0.68 & 0.53 \\ 0.53 & 0.68 & 1.00 & 0.68 & 0.53 \\ 0.53 & 0.68 & 0.68 & 1.00 & 0.53 \\ 0.60 & 0.53 & 0.53 & 0.53 & 1.00 \end{bmatrix}$
(0, 60, 60, 60, 0)	(160, 180, 180, 180, 160)	$\begin{bmatrix} 1.00 & 0.53 & 0.53 & 0.53 & 0.60 \\ 0.53 & 1.00 & 0.68 & 0.68 & 0.53 \\ 0.53 & 0.68 & 1.00 & 0.68 & 0.53 \\ 0.53 & 0.68 & 0.68 & 1.00 & 0.53 \\ 0.60 & 0.53 & 0.53 & 0.53 & 1.00 \end{bmatrix}$

```
Ran.offset <- rlnorm(100000,mean=0,sd=0.82)
Ran.offset <- matrix(rep(Ran.offset,length(mu)),ncol=length(mu))
Ran.means <- Ran.offset*mu
tmp <- rmvnorm(100000,Ran.means,cov = Sigma0)
rslt <- cov(tmp)
```

The PAD designs of section 3 use covariance matrixes from the above simulations with

$$\Sigma_0 = 160 \times \begin{bmatrix} 1.0 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0.6 & 1.0 & 0.6 & 0.6 & 0.6 \\ 0.6 & 0.6 & 1.0 & 0.6 & 0.6 \\ 0.6 & 0.6 & 0.6 & 1.0 & 0.6 \\ 0.6 & 0.6 & 0.6 & 0.6 & 1.0 \end{bmatrix}$$

and under means corresponding to late, immediate, intermediate, linear, and transient effects as shown in Table A1.

ACKNOWLEDGEMENT

This research was supported by NIH grant HL69719.

REFERENCES

1. Shih WJ, Quan H. Planning and analysis of repeated measures at key time points in clinical trials sponsored by pharmaceutical companies. *Statistics in Medicine* 1999; **18**:961–973.
2. Ben-Josef E, Shamsa F, Forman JD. Predicting the outcome of radiotherapy for prostate cancer. *Cancer* 1998; **82**:1334–1342.
3. Molina L, Elosua R, Marrugat J, Pons S. Relation of maximum blood pressure during exercise and regular physical activity in normotensive men with left ventricular mass and hypertrophy. *American Journal of Cardiology* 1999; **84**:890–893.
4. Laird NM, Want F. Estimating rates of change in randomized clinical trials. *Controlled Clinical Trials* 1990; **11**:405–419.
5. Frison LJ, Pocock SJ. Linearly divergent treatment effects in clinical trials with repeated measures: efficient analysis using summary measures. *Statistics in Medicine* 1997; **16**:2855–2872.
6. Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine* 1992; **11**:1685–1704.
7. Matthews JN, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. *BMJ* 1990; **300**:230–235.
8. Zackin R, Marschner IC, Anderson J, Cowles MK, DeGruttola V, Hammer S, Fischl M, Cotton D. HIV-1 RNA endpoints in HIV clinical trials: issues in interim monitoring and early stopping. *Journal of Infectious Diseases* 1998; **177**:761–765.
9. Kittelson JM, Emerson SS. A unifying family of group sequential test designs. *Biometrics* 1999; **55**:874–882.
10. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
11. Lan KKG, Wittes J. The *B*-value: a tool for monitoring data. *Biometrics* 1988; **44**:579–585.
12. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society*, *Series A*, *General* 1994; **157**:357–387.
13. Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
14. Lee JW, DeMets DL. Sequential comparison of changes with repeated measurements data. *Journal of American Statistical Association* 1991; **86**:757–762.
15. Gange SJ, DeMets DL. Sequential monitoring of clinical trials with correlated responses. *Biometrika* 1996; **83**:157–167.
16. Lee SJ, Kim K, Tsiatis AA. Repeated significance testing in longitudinal clinical trials. *Biometrika* 1996; **83**: 779–789.
17. Jennison C, Turnbull BW. *Group Sequential Methods with Application to Clinical Trials*. Chapman & Hall/CRC: Boca Raton, 2000.
18. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.
19. Galbraith S, Marschner IC. Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. *Statistics in Medicine* 2003; **22**:1787–1805.
20. Fairclough DL. *Design and Analysis of Quality of Life Studies in Clinical Trials*. Chapman & Hall/CRC: Boca Raton, 2002; 161–167.
21. Spiessens B, Lesaffre E, Verbeke G, Kim K, DeMets D. An overview of groups sequential methods in longitudinal clinical trials. *Statistical Methods in Medical Research* 2000; **9**:497–515.
22. Simons M, Annes BH, Laham RJ, Kleiman N, Henry T, Dauerman H, Udelson JE, Gervina EV, Pike M, Whitehouse MJ, Moon T, Chronos NA. Pharmacological treatment of coronary artery disease with recombinant fibroblast growth factor-2 double-blind, randomized, controlled clinical trial. *Circulation* 2002; **105**:788–793.
23. *R* project for statistical computing. *http://www.r-project.org/*
24. *S+Seqtrial*. Insightful Corporation. Seattle Washington. *http://www.insightful.com/products/seqtrial/default.asp*
25. *East*. Cytel Statistical Software. Cambridge, Massachusetts. *http://www.cytel.com/East/Default.asp*
26. *Pest*. Medical and Pharmaceutical Statistics Research Unit. University of Reading: Reading, U.K. *http://www.rdg.ac.uk/mps/mps_home/software/software.htm*
27. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika* 1990; **77**:875–892.
28. Emerson SS, Fleming TR. Symmetric group sequential test designs. *Biometrics* 1989; **45**:905–923.
29. Burington BE, Emerson SS. Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics* 2003; **59**:770–777.
30. Diggle PJ, Heagerty P, Liang KY, Zeger SL. *Analysis of Longitudinal Data* (2nd edn). Oxford University Press: New York, 2002; 319–326.
31. Gillen DL, Emerson SS. Information growth in a family of weighted logrank statistics under repeated analyses. *Sequential Analysis* 2005; **24**:1-22.

