Published in final edited form as:

Biometrics. 2015 June ; 71(2): 469–477. doi:10.1111/biom.12287.

Longitudinal Clinical Trials with Adaptive Choice of Follow-up Time

Neal O. Jeffries* and Nancy L. Geller**

Office of Biostatistics Research, National Heart, Lung, and Blood Institute, Bethesda, MD. 20892

Summary

In longitudinal studies comparing two treatments with a maximum follow-up time there may be interest in examining treatment effects for intermediate follow-up times. One motivation may be to identify the time period with greatest treatment difference when there is a non-monotone treatment effect over time; another motivation may be to make the trial more efficient in terms of time to reach a decision on whether a new treatment is efficacious or not. Here we test the composite null hypothesis of no difference at any follow-up time versus the alternative that there is a difference at at least one follow-up time. The methods are applicable when a few measurements are taken over time, such as in early longitudinal trials or in ancillary studies. Suppose the test statistic Z_{t_k} will be used to test the hypothesis of no treatment effect at a fixed follow-up time t_k . In this context a common approach is to perform a pilot study on N_1 subjects, and evaluate the treatment effect at the fixed time points $t_1, \dots, t_K$ and choose t^* as the value of t_k for which Z_{t_k} is maximized. Having chosen t^* a second trial can be designed. In a setting with group sequential testing we consider several adaptive alternatives to this approach that treat the pilot and second trial as a seamless, combined entity and evaluate Type I error and power characteristics. The adaptive designs we consider typically have improved power over the common, separate trial approach.

Keywords

Adaptive design; Adaptive follow-up time; Adaptive longitudinal trial; Longitudinal study; Model-free longitudinal analysis

1. Introduction

Suppose a clinical trial centers on determining differences between randomized groups that are evaluated longitudinally, with non-monotone differences anticipated over time. The group differences may evolve as the effects of the intervention change over time – e.g. a drug may have an effect in the short term and then its effect dissipates. As an example consider quality of life data which may improve with treatment initially, but reflect a loss of

*nealjeff@nhlbi.nih.gov. **gellern@nhlbi.nih.gov.

7. Supplementary Materials

Web Appendix A (referenced in Section 3), R code for conducting simulations in Section 4, Web Table 1 (in Web Appendix B and referenced in Section 4), and Web Appendix C (referenced in Section 6) are available with this paper at the Biometrics website on Wiley Online Library.

effectiveness over time. In such circumstances it may be of interest to determine if the treatment has an effect at any time.

Assume a normally distributed response will be observed at K follow-up times $\{t_1, t_2, \dots, t_K\}$ and the i^{th} individual's change from baseline to the k^{th} evaluation is denoted by X_{it_k} in the active treatment group and by Y_{it_k} for the i^{th} individual in the comparison group. For a fixed set of individuals one may form K test statistics, Z_{it_k} , $\{k \in 1, \dots, K\}$ to test the null hypothesis that there is no difference at time t_k .

Further, one may consider a fixed follow-up time t_k and evaluate a sequential set of test statistics as groups of individuals reach the k^{th} follow-up period. Let $N_1 < \dots < N_M = N$ denote increasing sample sizes where N denotes the maximum sample size determined before the trial begins. The resulting sequence of test statistics for follow-up time t_k may be denoted $Z_{t_k}(1, N_1), \dots, Z_{t_k}(1, N_m), \dots, Z_{t_k}(1, N_M)$ where $(1, N_m)$ indicates the range of individuals contributing to the test statistic, e.g. the first N_m individuals randomized. Such sequences are frequently evaluated using group sequential methods pioneered by Pocock (1977), O'Brien and Fleming (1979), and Lan and DeMets (1983) and summarized by Jennison and Turnbull (2000).

This work investigates the properties of sets of test statistics that may include different information fractions (i.e. different N_m) as well as different time points for follow-up (i.e. different t_k). As a simple example consider a hypothetical drug trial in which the investigators are interested in evaluating differences between active drug and placebo after both 1 year and 2 years of follow-up. A testing strategy might first test for differences when one half of the maximum sample size, $N/2$, has reached one year of follow-up. The test statistic would be denoted as $Z_1(1, N/2)$. Should that difference not reach the statistically significant threshold, one could test again when that group of participants reaches the two year timepoint; that test statistic would be denoted as $Z_2(1, N/2)$. Should that result also not reach statistical significance, one could test the whole sample at the two year timepoint using $Z_2(1, N)$. Alternatively, one might have chosen to replace the second test, $Z_2(1, N/2)$, with one using the whole sample evaluated at one year follow up, $Z_1(1, N)$, or both tests could have been used yielding a set of 4 potential test statistics: $Z_1(1, N/2)$, $Z_2(1, N/2)$, $Z_1(1, N)$, $Z_2(1, N)$. A test statistic based upon the follow-up time t_k and data from individuals who enrolled as or after the N_1+1^{th} person enrolled and as or before the N_2^{th} person enrolled will be denoted as $Z_{t_k}(N_1+1, N_2)$.

We examine how to choose thresholds so that an overall error rate across the set of tests can be maintained, and how investigators can choose a sequence of tests adaptively, rather than following a predetermined, fixed sequence of test statistics. We present Type I error rates and power for several approaches and apply the methods in a heart failure clinical trial.

2. Choice of Threshold

Let $\{1 < N_1 < N_2 < \dots < N_M = N\}$ denote a set of sample sizes and $\{t_1 < t_2 < \dots < t_K\}$ a set of follow-up times for evaluation of group differences. Next let Ω denote a predetermined set of test statistics, $\{Z_{t_1}(1, N_1), \dots, Z_{t_k}(1, N_m), \dots, Z_{t_k}(1, N)\}$ that could be evaluated

during interim and final analyses for a clinical trial; Ω need not contain statistics based on every combination of sample size and follow up period.

The elements of Ω exhibit correlation that depends in part on the N_m and on t_k where $m \in \{1, 2, \dots, M\}$ and $k \in \{1, 2, \dots, K\}$. The test statistic, $Z_{t_k}(1, N_m)$, is of the following general form:

$$Z_{t_k}(1, N_m) = \frac{(\bar{X}_{t_k} - \bar{Y}_{t_k})}{\left(\hat{\sigma}_{X_{it_k}}^2 / N_X(1, N_m) + \hat{\sigma}_{Y_{it_k}}^2 / N_Y(1, N_m) \right)^{1/2}} \quad (1)$$

where $N_X(1, N_m)$ denotes how many of the N_m individuals were randomized to treatment X and $N_Y(1, N_m)$ is similarly defined for treatment Y . The $\bar{X}_{t_k}$ and $\bar{Y}_{t_k}$ denote the sample means for follow-up period t_k for X_{it_k} , $i = 1, \dots, N_X(1, N_m)$ and Y_{it_k} , $i = 1, \dots, N_Y(1, N_m)$ and

$\hat{\sigma}_{X_{it_k}}^2$ ($\hat{\sigma}_{Y_{it_k}}^2$) is the estimated variance of X_{it_k} (Y_{it_k})

The covariance between two elements in Ω , say $Z_{t_j}(1, N_1)$, and $Z_{t_k}(1, N_2)$ where $1 \leq N_1 \leq N_2 \leq N$ and $t_j, t_k \in \{t_1, \dots, t_K\}$ may be derived as

$$\begin{aligned} \text{Cov}(Z_{t_j}(1, N_1), Z_{t_k}(1, N_2)) &\approx \frac{1}{\tau_{t_j}(1, N_1)\tau_{t_k}(1, N_2)} \times \\ &\left[\frac{\#\{i: X_i \text{ contributes to both } Z_{t_j}(1, N_1) \text{ and } Z_{t_k}(1, N_2)\}}{N_X(1, N_1)N_X(1, N_2)} \text{Cov}(X_{it_j}, X_{it_k}) \right. \\ &\left. + \frac{\#\{i: Y_i \text{ contributes to both } Z_{t_j}(1, N_1) \text{ and } Z_{t_k}(1, N_2)\}}{N_Y(1, N_1)N_Y(1, N_2)} \text{Cov}(Y_{it_j}, Y_{it_k}) \right] \quad (2) \end{aligned}$$

where $\tau_{t_j}(1, N_1) = \left(\sigma_{X_{it_j}}^2 / N_X(1, N_1) + \sigma_{Y_{it_j}}^2 / N_Y(1, N_1) \right)^{1/2}$. The covariances $\text{Cov}(X_{it_j}, X_{it_k})$ and $\text{Cov}(Y_{it_j}, Y_{it_k})$ can be estimated from observed correlation among individuals who have data at both follow-up occasions and the $\sigma_{X_{it_j}}^2$ and $\sigma_{Y_{it_j}}^2$ terms can also be estimated in an unbiased manner. When $t_j = t_k$ and there are no missing data this expression simplifies to the usual expression in group sequential analysis: $\text{Cov}(Z_{t_j}(1, N_1), Z_{t_j}(1, N_2)) = \sqrt{\frac{N_1}{N_2}}$.

Given estimated covariances like those shown above, a prespecified set of W test statistics in Ω of the form $Z_{t_k}(1, N_m)$, and a prespecified set of W α values, $\alpha_1, \alpha_2, \dots, \alpha_W$ to spend on each test, one can use standard integration approaches and an assumption of multivariate normality to determine an associated set of thresholds b_w , $w = 1, \dots, W$ such that $P_{H_0}(Z_1 \leq b_1) = \alpha_1$ and $P_{H_0}(Z_w \leq b_w \text{ and } Z_v < b_v; 1 \leq v < w) = \alpha_w$ for $2 \leq w \leq W$ where for convenience we have ordered the set of W test statistics $\{Z_{t_k}(1, N_m)\}$ as $\{Z_1, Z_2, \dots, Z_W\}$. The ordering is done by the calendar time at which the indicated range of individuals (e.g. 1st through N_m^{th} person enrolled) has reached the indicated follow-up time, e.g. t_k . To compute the values of b_w first note $b_1 = \Phi^{-1}(1 - \alpha_1)$ where Φ denotes the cumulative distribution function of a standard normal variate. Given $b_1, b_2, \dots, b_{w-1}$ we compute b_w by using (2) to compute the joint multivariate normal distribution of $Z_w, Z_{w-1}, \dots, Z_1 = f(z_w, z_{w-1}, \dots, z_1; \Sigma)$ and find b_w such that

$$\int_{-\infty}^{b_w} \int_{-\infty}^{b_{w-1}} \cdots \int_{-\infty}^{b_1} f(z_w, z_{w-1}, \dots, z_1; \Sigma) dz_1 \cdots dz_{w-1} dz_w = 1 - \sum_{k=1}^w \alpha_k$$

where Σ is the estimated variance-covariance matrix for $Z_1, \dots, Z_w$. The **R** programming language (R Core Team, 2013) and the software package mvtnorm (Genz et al. 2012) were used to evaluate such rectangular integrals.

3. Adaptively choosing a set of tests

Consider a trial with follow-up measurements at six, nine, and twelve months and for each follow-up period statistics might be computed using 1/2, 3/4, or all of the individuals who could be enrolled. It may not be clear how to prospectively choose a set of tests for evaluation. Here we propose new adaptive methods for choosing a sequence of tests to conduct that avoids full prespecification.

We focus on testing whether there is a difference between an experimental and control intervention at any follow-up time. If a total of N individuals may be randomized one approach is to initially collect data from $N_1 < N$ participants and evaluate the associated test statistics $Z_{t_1}(1, N_1), \dots, Z_{t_K}(1, N_1)$ and base subsequent examinations of the data on these initial findings. We consider one-sided hypotheses of interest and determine t^* as that follow-up period showing best evidence of a difference, i.e. $t^* = \arg \max_{t_1 \dots t_K} Z_{t_k}(1, N_1)$.

To illustrate we consider a setting with 6, 9, and 12 month follow-up periods and a maximum sample size of $N = 400$. Suppose accrual occurs uniformly over a 24 month period and $N_1 = 133$ so that the test statistics for the three follow-up periods are evaluated after approximately 1/3 of the data are obtained for each. Figure 1 shows how accrual and evaluation are related to study time. Six month follow-up data begin accumulating after 6 months of study time and are complete at 30 months. Given uniform accrual the 133rd data point for the 6 month period is available after about 8 months of accumulation – around the 14th month of study time. At 14 months, only 5 months of the 9 month follow-up data are available – this corresponds to about 83 individuals providing 9 month follow-up period data ($83 \approx 400 \times 5/24$). Similarly, only about 33 individuals have provided 12 month follow-up period data after 14 months of study time.

The 133rd observation for the 9 month follow-up data is not available until 17 months of study time when $Z_9(1, 133)$ can be obtained. Then there are approximately 183 and 83 observations for the 6 and 12 month periods, respectively. The 133rd observation for the 12 month follow-up data is not available until after 20 months of study time.

After 20 months of study time the three test statistics $Z_6(1, 133)$, $Z_9(1, 133)$, and $Z_{12}(1, 133)$ are compared and t^* corresponds to either the 6, 9, or 12 month follow-up period such that $Z_{t^*}(1, 133) = \max\{Z_6(1, 133), Z_9(1, 133), Z_{12}(1, 133)\}$. Note that one could use all available information at study month 20 and choose t^* based on the values of $\{Z_6(1, 233), Z_9(1, 183), Z_{12}(1, 133)\}$. However, this approach may leave one with very little 6 or 9 month data to use in a second stage of the study and may affect the power for a second test based only on the

remaining observations. Notwithstanding, different amounts of information may be used for the follow-up period test statistics and the methods described below could be changed accordingly.

If $Z_t^*(1, N_1)$ is sufficiently large to provide evidence of a significant difference then the study would be stopped. Otherwise, enrollment continues and the data are reexamined after all N individuals have been assessed. A number of different approaches may be implemented in a second stage of the study that have consequences for type I error and power. We compare five different approaches.

3.1 Approach A1: Joint analysis of $Z_t^*(1, N_1)$ and $Z_t^*(1, N)$

Let α denote the nominal size for the joint procedure, e.g. 0.05. Choose $\alpha_1 < \alpha$ as a threshold for stopping early after $N_1 < N$ individuals provide follow-up period data. In practice α_1 may be determined from a spending function, e.g. $\alpha_1 = \alpha \times h(N_1/N)$ where $h(\cdot)$ is an increasing function with $0 \leq h(t) \leq 1$ for $t \in [0, 1]$. Given α_1 the threshold for the test statistic $Z_t^*(1, N)$ is b_1^* where $P_{H_0}(Z_t^*(1, N_1) > b_1^*) = \alpha_1$ and H_0 corresponds to the composite null hypothesis of no treatment difference at any follow-up period. Given estimated correlations among $\{Z_{t_1}(1, N_1), \dots, Z_{t_K}(1, N_1)\}$ one can use multivariate integration to find b_1^* .

If $Z_t^*(1, N_1)$ exceeds b_1^* then the recommendation would be to stop the study early with a conclusion of superiority. Alternatively, if the one-sided p value exceeds α_1 then the evidence is not sufficiently strong and the study continues until all N individuals are randomized and evaluated at the follow-up periods. When data are complete one compares $Z_t^*(1, N)$ to b_2^* where b_2^* is chosen to satisfy $P_{H_0}(Z_t^*(1, N_1) \leq b_1^*, Z_t^*(1, N) \leq b_2^*) = 1 - \alpha$. The threshold b_2^* can be found using similar integration ideas given estimates of the required correlations. Details of the computation of thresholds b_1^* and b_2^* (and thresholds for other approaches) are available in Web Appendix A. An approach using conceptually similar thresholds has been employed by Stallard and Todd (2003) in the context of choosing the best intervention among multiple treatment arms compared to a common control arm. The context here differs as there are only 2 treatment arms and K follow-up periods thereby inducing a more general correlation pattern.

3.2 Approach A2: Joint analysis of $Z_t^*(1, N_1)$ and $\max_k Z_{t_k}(1, N)$

Structurally, this is quite similar to the previous approach. Given α_1 , an early threshold for stopping, b_1^* can be determined to satisfy $P_{H_0}(Z_t^*(1, N_1) > b_1^*) = \alpha_1$.

Although $\max_k Z_{t_k}(1, N_1) = Z_t^*(1, N_1)$, by definition, it is not necessarily the case that $\max_k Z_{t_k}(1, N) = Z_t^*(1, N)$. t^* is determined after N_1 individuals provide data and does not change thereafter. In particular if $\arg \max_{t_1, \dots, t_K} Z_{t_k}(1, N_1) \neq \arg \max_{t_1, \dots, t_K} Z_{t_k}(1, N)$ (i.e. different follow-up periods are associated with the max statistic using N_1 and N individuals), then $\max_k Z_{t_k}(1, N) \neq Z_t^*(1, N)$. Thus, while one can use multivariate integration to find b_2 such that $P_{H_0}(Z_t^*(1, N_1) \leq b_1^*, \max_k Z_{t_k} \leq b_2) = 1 - \alpha$ it will not generally be the case that $b_2 = b_2^*$ described above. This approach is like the previous one in that the study should

conclude with a recommendation of superiority if either $Z_{t^*}(1, N_1) > b_1^*$ or $\max_k Z_{t_k}(1, N) > b_2$. Kelly, Stallard, and Todd (2005) described a similar model for evaluating the best treatment among multiple treatments though the differences in context mentioned before apply here.

3.3 Approach A3: Conditional rejection probability approach

In this approach one uses the same first stage test statistic and threshold but may implement a second stage test statistic $Z(N_1+1, N)$ that is computed using data from the individuals enrolled after the N_1^{th} participant and the statistic is adaptively chosen depending upon what is observed from the first N_1 individuals. Here, the two choices for the second stage test statistic are $Z(N_1+1, N) = Z_{t^*}(N_1+1, N)$ or $Z(N_1+1, N) = \max_k Z_{t_k}(N_1+1, N)$.

The choice of $Z(N_1+1, N) = Z_{t^*}(N_1+1, N)$ may make most sense if the first stage data suggest there exists one follow-up period in which the effect size is much larger than in the other periods. Under such circumstances it is reasonable to focus on a test statistic for only that follow-up period. If there does not appear to be a dominant effect size, then one can choose a test with power for a wide range of alternatives by using the second choice of test statistic, $Z(N_1+1, N) = \max_k Z_{t_k}(N_1+1, N)$. The conditional rejection probability (CRP) approach described by Müller and Schäfer (2004) is used to generate an appropriate level for the second stage test regardless of which statistic is used.

The CRP approach begins with an initial design – in this case we use the previous design that rejects the composite null hypothesis if $Z_{t^*}(1, N_1) > b_1^*$ or $\max_k Z_{t_k}(1, N) > b_2$ where $P_{H_0}(Z_{t^*}(1, N_1) > b_1^*) = \alpha_1$ and $P_{H_0}(Z_{t^*}(1, N_1) \leq b_1^*, \max_k Z_{t_k} \leq b_2) = 1 - \alpha$. After the first N_1 individuals have provided data and if the study is not terminated early one computes the conditional rejection probability as

$$crp = P_{H_0} \left(\max_k Z_{t_k}(1, N) > b_2 \mid Z_{t^*}(1, N_1) = z \right).$$

Next, one can use confidence intervals to determine whether to set $Z(N_1+1, N) = Z_{t^*}(N_1+1, N)$ or $Z(N_1+1, N) = \max_k Z_{t_k}(N_1+1, N)$. Let d_{t_k} denote the expected value of the difference between X_{t_k} and Y_{t_k} with estimate $\hat{d}_{t_k} = \bar{X}_{t_k}(1, N_1) - \bar{Y}_{t_k}(1, N_1)$. From (1) we have $\hat{d}_{t_k} = \tau_{t_k}(1, N_1) Z_{t_k}(1, N_1)$ where $\tau_{t_k}(1, N_1)$ is an estimate of $\tau_{t_k}(1, N_1)$ in (2). Given the variance-covariance matrix for $Z_{t_1}(1, N_1), \dots, Z_{t_K}(1, N_1)$ and the relationship between $\hat{d}_{t_k}$ and $Z_{t_k}(1, N_1)$ it is easy to produce confidence intervals for $d_{t_k} - d_{t_j}$. Next, let $t_{m_1}, t_{m_2}, \dots, t_{m_{K-1}}$ denote the $K-1$ follow-up periods that do not correspond to t^* . Then one chooses $Z'(N_1+1, N) = Z_{t^*}(N_1+1, N)$ if the 50% confidence intervals for $d_{t^*} - d_{t_{m_1}}, d_{t^*} - d_{t_{m_2}}, \dots, d_{t^*} - d_{t_{m_{K-1}}}$ all lie above 0, and otherwise choose $Z(N_1+1, N) = \max_k Z_{t_k}(N_1+1, N)$. As t^* is data dependent, the confidence intervals as constructed here may not have nominal coverage matching their confidence level. However, the approach is intended to be a simple heuristic means to determining which method to choose for the second stage and the confidence intervals are not used for hypothesis testing. Also, the choice of intervals with 50%

confidence is of course arbitrary and may depend in part upon K , the number of follow-up periods.

After the data are collected from all N participants the second stage test statistic $Z(N_1 + 1, N)$ is assessed with level = crp . It may happen that $crp > 0.05$ if $Z_t^*(1, N_1)$ is sufficiently large. The second stage test for this approach uses only data from the $N_1 + 1^{th}$ through N^{th} individuals and may entail some loss of efficiency relative to tests that use data from all N individuals. Two stage approaches that base each stage on disjoint sets of information are not uncommon for adaptive designs (e.g. Fisher's combination test in Bauer and Köhne, (1994)) and the loss of efficiency may be offset by the method's flexibility.

3.4 Approach A4: Pilot Study and Definitive Study Approach

This simple idea is common in practice although it does not protect the familywise Type I error. An investigator will conduct an initial test based on N_1 individuals - sometimes called a pilot study. If the p-value for $Z_t^*(1, N_1) < \alpha/K$, i.e. meets a Bonferroni threshold, then superiority is concluded and further testing need not be performed. The p-value for $Z_t^*(1, N_1)$ is computed using the standard normal distribution for the largest of the K test statistics rather than the multivariate distribution of the underlying statistics, i.e. the bound for declaring significance is $\Phi^{-1}(1 - \alpha/K)$.

If the first stage p-value for $Z_t^*(1, N_1)$ does not meet the Bonferroni threshold a second stage "definitive" study is conducted and the p-value for $Z_t^*(N_1 + 1, N)$ is compared to α . Such an approach does not preserve Type I error if the combined experiments are considered as a single process. The rationale for this design would be to use the first stage to define the follow-up time and once determined, α , e.g. 0.05, is used for testing significance in a second, independent sample. Such a rationale would be appropriate if there were not the opportunity to conclude superiority using only the first N_1 individuals.

3.5 Approach A5: Hybrid Bonferroni Approach

Like the previous approach this idea uses a Bonferroni correction but is similar in design to the second approach examining $Z_t^*(1, N_1)$ and $\max_k Z_{tk}(1, N)$. For some initial $\alpha_1 < \alpha$ chosen by a spending function, the first stage test compares the p-value for $Z_t^*(1, N_1)$ with α_1/K . Again, the p-value here is computed using the standard normal distribution and is correspondingly conservative as correlations are ignored.

If the first stage threshold for significance is not met then the second stage test statistic is based on comparing the p-value for $\max_k Z_{tk}(1, N)$ with b_2^H , a threshold determined by the remaining alpha available for testing subject to a Bonferroni adjustment, $(\alpha - \alpha_1)/K$. b_2^H is determined by satisfying the following equation

$$P(Z(1, N) > b_2^H, Z(1, N_1) < \Phi^{-1}(1 - \alpha_1/K)) = \frac{\alpha - \alpha_1}{K}$$

where $Z(1, N)$ and $Z(1, N_1)$ have a bivariate normal distribution with mean 0, variance 1, and covariance $\sqrt{N_1/N}$. This approach is conservative with respect to Type I error and power.

4. Simulations

Simulations were used to evaluate Type I error and power for these approaches and were generated in the following manner. Accrual occurred randomly in both arms at a uniform rate over an 18 month period. The outcome measure was assessed at baseline, 6 months, 9 months, and 1 year, with the post-baseline changes as the outcome of interest. X_{itk} , the change from baseline to follow-up time t_k for the i^{th} person in the active arm, was computed as follows: $\varepsilon_{it0}, \dots, \varepsilon_{itK}$ have a multivariate normal distribution with mean given by a $K + 1$ -length vector of zeros and pairwise covariance $Cov(\varepsilon_{itj}, \varepsilon_{itk}) = \exp(-|t_j - t_k|/5)$ where $t_0 = 0$ corresponds to baseline time 0. Then $X_{itk} = \mu_{Xt_k} + \varepsilon_{itk} - \varepsilon_{it0}$ for $k = 1, \dots, K$ where μ_{Xt_k} denotes the mean change between baseline and follow-up time t_k for the active arm. Y_{itk} is similarly defined for the control arm but $\mu_{Yt_k} = 0$ for all follow-up periods $t_1, \dots, t_K$.

The interim analysis was conducted when approximately N_1 of N individuals had data at all three follow-up periods. To reduce imbalances in randomization and ensure that the proportion of individuals in each arm with data at each timepoint was approximately N_1/N , not all available data were used – only enough so that approximately $100 \times N_1/N\%$ of each arm provide data at each timepoint. The one-sided level used for this initial test corresponds to $\alpha \times (N_1/N)^3$ for the first three methods, $\alpha/3$ for method A4, and $\alpha \times (N_1/N)^3/K$ for method A5. α was chosen as 0.05 for all of the simulations.

Table 1 provides Type I error and power for the simulated designs. **R** code for simulations is available in the supplementary material. The first two columns of Table 1 show the total number that can be randomized, N , and the fraction of participants that contribute to the first stage analysis. The next 3 columns show μ_{Xt_k} for 6, 9, and 12 month followup, i.e. the differences between the active and control arms. The last 5 columns show the study design options; the numbers in these columns show the proportion of times the composite null hypothesis of no difference at any follow-up period was rejected in favor of the one-sided alternative. For Type I error and power 10,000 simulations were run. For the conditional rejection probability approach (CRP in the table) the second number, $\hat{\pi}$, indicates the proportion of times $Z'(N_1 + 1, N) = Z_t^*(N_1 + 1, N)$ instead of $Z'(N_1 + 1, N) = \max_k Z_{t_k}(N_1 + 1, N)$. For each method the first stage and second stage test statistics are shown, e.g. $Z_t^*(1, N)$, though the first stage test statistic is the same for each method.

The simulations show the first three adaptive approaches have reasonably accurate Type I error rates. The last two approaches using a Bonferroni approach are suboptimal. Approach A4 using the pilot and definitive two study approach lacks control of Type I error because the α spent on the first stage is ignored when using $\alpha = 0.05$ in the second stage. Approach A5 is conservative; although the Bonferroni approach controls Type I error it does not take into account the correlation among test statistics and consequently sets too stringent a threshold. Further, in terms of power both of these perform worse than the three new approaches.

One might expect Approach A1 – the joint approach involving $Z_t^*(1, N_1)$ and $Z_t^*(1, N)$ – to perform best relative to the others when there is one follow-up period with the strongest effect size. Indeed, this method performs well when the 12 month difference is 0.40 and the other periods have no difference.

Conversely, when each follow-up period has the same effect size (e.g. 0.30 in the table) one might expect Approach A2 to perform better than Approach A1. Under Approach A1 the choice of t^* at the first stage may differ from the period with greatest effect at the end of the study after all N participants have been evaluated. This is more likely to happen when effect sizes are all the same. Consequently, Approach A2 may be expected to outperform Approach A1 and this is what is shown. For those scenarios with the same effect size of 0.30 for each follow-up period Approach A2 has power about 2–3% higher than Approach A1 – a significant difference given the standard error for power in these tables is approximately 0.004. For those intermediate scenarios (with effect sizes of 0.35, 0.20, and 0) neither Approach A1 nor Approach A2 appears to dominate.

Approach A3 was an attempt to blend Approaches A1 and A2 and pick the better approach depending upon the first stage. If it appeared that one effect size dominated then $Z_t^*(N_1 + 1, N)$ would be used to analyze the second stage of the data and if there were insufficient evidence of a dominant effect size then $\max_k Z_{t_k}(N_1 + 1, N)$ would be chosen. The desired effect is seen as this blending approach performs relatively well for the extreme cases of one dominant effect size or all three effect sizes the same. In the case with effect sizes of 0.35, 0.20, and 0 all three approaches performed similarly given the standard error due to simulation. While Approach A3 seems promising it should be noted that this good performance may not be generally guaranteed. Alternative CRP approaches using 75% confidence intervals to determine which approach to use and/or using a conditional error function approach (Proschan and Hunsberger, 1999) resulted in performance that was not typically superior to Approaches A1 and A2 (data not shown).

An additional set of simulations was conducted using identical parameters but included 20% missing data following a missing completely at random mechanism. Web Table 1 in Web Appendix B shows the same general patterns though power is reduced.

5. Application: Quality of Life in a Clinical Trial

The Studies of Left Ventricular Dysfunction (SOLVD) treatment trial was a double-blind, randomized, placebo-controlled trial to assess the effect of enalapril (an ACE inhibitor vasodilator) on mortality in a heart failure population (The SOLVD Investigators, 1991).

A quality of life survey was administered at baseline, 6 weeks, 1 year, and 2 years after baseline. The survey contained a question assessing the frequency of dyspnea, that is, labored and difficult breathing, a common symptom in heart failure. Frequency of dyspnea was recorded on a 6 point scale with higher scores indicating higher frequency. Table 2 shows the mean and standard deviation of the self-assessment of dyspnea at the various time points. Attention is restricted to those who completed the survey at all three follow-up time points.

The authors compared the treatment groups with respect to their differences in dyspnea scores at 6 weeks, 1 year and 2 years. They reported an improvement in dyspnea frequency at six months and one year “not accounting for multiple testing”. Here we are interested in determining if there is a difference at any of the follow-up times, accounting for the multiple testing. We take $N = N_{\text{Plac}} + N_{\text{Enal}} = 1055$ and N_1 such that $N_1/N \approx 0.40$. Other choices of N_1 may induce different outcomes.

The initial examination of the test statistics is performed when approximately 40% of the 538 enalapril (215) and 40% of the 517 placebo (207) participants have provided follow-up data at all three timepoints. For $N_1 = 215 + 207 = 422$ the three corresponding first stage test statistics are $Z_{6w}(1, 422)$, $Z_{1y}(1, 422)$, and $Z_{2y}(1, 422) = 1.139, 0.518$, and 0.255 , respectively, as shown in Table 3.

We note that the follow-up period corresponding to the maximum for the first stage data is 6 weeks (i.e. $t^* = 6$ weeks); however the follow-up period associated with the maximum for the second stage data and all the data is 1 year. This means that those approaches that use $Z_t^*(1, N)$ or $Z_t^*(N_1 + 1, N)$ for the second test statistic will likely not perform as well as those statistics based upon $\max_k Z_{tk}(1, N)$ or $\max_k Z_{tk}(N_1 + 1, N)$.

For a simple spending function of $0.05 \times t^3$ the one-sided threshold p-value for early termination at the first look is 0.0032. One can compute the corresponding threshold test statistic values for the 5 approaches discussed above. The test statistic threshold, 3.046, is the same for the first three approaches and is computed using multivariate integration and estimated correlations based on the first 40% of the participants in each arm. The threshold for Approach A4 is $\Phi^{-1}(1 - 0.05/3) = 2.128$ and the threshold for the hybrid Bonferroni approach (Approach A5) is $\Phi^{-1}(1 - 0.05 \times .4^3/3) = 3.071$. The boundaries for all these approaches exceed the maximum of the three first stage test statistics, 1.139, so under each approach one would proceed to the second stage test statistics. The observed statistics and the boundary thresholds for the various approaches using first stage data only are summarized in the first two rows of Table 4.

The last two rows of the table show the observed second stage test statistics for each approach and the relevant threshold for that statistic. Except for the CRP approach the second stage statistic fails to reach the threshold. For the CRP approach the second stage test statistic is determined by examining 50% confidence intervals for differences in the observed effect sizes and in this case these differences were not found sufficiently strong to suggest a single effect size dominates the others (data not shown). Subsequently, for this approach the second stage test statistic $Z'(N_1 + 1, N)$ is set to $\max_k Z_{tk}(N_1 + 1, N)$. Integration was used to determine 1.998 as the threshold. This threshold value was exceeded by the test statistic $\max_k Z_{tk}(N_1 + 1, N) = 2.088$ and thus for this approach a significant difference is obtained. The second approach using $\max_k Z_{tk}(1, N) = 1.957$ just failed to reach its threshold value of 2.067.

Although Approaches A2 and A5 have the same test statistic, the threshold for the hybrid Bonferroni approach is higher. Approaches A1 and A4 fare worse because their second stage test statistic is based on the follow-up period of 6 weeks – whereas the totality of the

data suggest the one year period may have the strongest effect. Note the second stage statistics for Approaches A3 and A4 are based only on the second part of the sample while the second stage statistics for the other approaches are based on the entire sample. The example illustrates that results depend on which approach is selected for the second stage. A predetermined choice of the CRP approach would indicate there is an improvement in dyspnea frequency with treatment, accounting for multiple testing.

6. Discussion

We presented several flexible two-stage adaptive approaches to analyzing longitudinal data which may have non-monotone treatment differences over time. The procedures are especially relevant when investigators do not know at what time point the maximum treatment difference will occur.

Our approaches maintain Type I error whereas the two Bonferroni approaches have poor power and Type I error characteristics. The power of our approaches depends on the configuration of effects at the different time points. If there is a time point with a substantially larger treatment difference than other time points then Approach A1 typically has highest power. In addition, Approach A1 would allow for discontinuation of follow up on some subjects if t^* is not the last of the time points. Also, if the trial is positive, one can conclude there is a treatment difference at time t^* .

If the treatment differences at two different time points are similar, a test that allows different time points for the first and second analysis may improve power, e.g. Approach A2. If the first stage yields a statistically significant difference, one would conclude that there is a treatment difference at time t^* . If the trial proceeds to the second stage and a statistically significant difference is found, one can conclude that there is a treatment difference at $\arg \max_{t_1, \dots, t_K} Z_{t_k}(1, N)$.

Identifying other time points for which there is a treatment difference could in principle be determined via a stepdown closed procedure based upon $\max_{t_k \in \Omega} Z_{t_k}(1, N_s)$ where $\Omega \subseteq \{t_1, \dots, t_K\}$ and $N_s = N_1$ (for the interim analysis) or $N_s = N$ (for the final analysis). More work is required for this question.

If the CRP Approach A3 is chosen and the first analysis yields a statistically significant difference, one would conclude that there is a treatment difference at time t^* . If the trial proceeds to the second stage and a statistically significant difference is found, one can only conclude that there is a treatment difference, but not when that difference occurs (unless $t^* = \arg \max_{t_1, \dots, t_K} Z_{t_k}(1, N)$). That is because the test statistic in the second stage is disjoint from that in the first stage. This is a disadvantage of the CRP approach. Also, this approach only uses data from individuals enrolled after the N_1^{th} person, in its second stage analysis. This is a feature of some Fisher combination and CRP approaches in the adaptive design literature (Bauer and Köhne, 1994 and Müller and Schäfer, 2004). In principle more powerful results might be obtained if all the data could be used in a way that maintains Type I error control. This is an avenue for further research.

Other work on repeated testing of longitudinal data makes assumptions about the trajectory of treatment differences such as linearity (Armitage, Stratton and Worthington, 1985; Geary, 1988), a common slope difference between groups over time (Lee and DeMets, 1991) or considering other summary measures (Wu and Lan, 1992; Kettelson, Sharples and Emerson, 2005). When results differ from the assumed models, our approaches typically have greater power (results not shown). Some of these alternative approaches may be more appropriate if a particular summary measure is of interest, e.g. area under a curve.

The approaches we present are specific for longitudinal settings. A related question concerns a more general approach toward selecting one endpoint among several alternatives. The CRP approach described by Müller and Schäfer, (2004) works in this more general setting. Kieser, Bauer, and Lehmacher (1999) describe approaches involving closed testing procedures, *a priori* prioritization of hypotheses, and a Holm-Bonferroni approach for rejecting hypotheses in a setting of testing multiple hypotheses. Kosorok, Shi, and DeMets (2004) develop a method for determining which among several coprimary outcomes have significant differences in a group sequential setting. Their approach further generalizes to determining if outcomes show benefit, non-detriment, or detriment. Such an approach could be extended to the longitudinal outcomes here and may be of interest if one wants to determine that the intervention has at least one follow-up period with positive effect and no periods of detrimental effect.

As with any group sequential or adaptive method, different proportions of the totality of data for the first stage can yield different results. This implies the need for precise design up front. Also, adaptive methods that may involve a change in primary endpoint may be most appropriate for early phase or exploratory studies as opposed to pivotal or confirmatory studies. Further, designs that allow for early stopping may provide little information for secondary endpoints or relationships between endpoints and the decision to terminate a study or stop collecting data on an endpoint should account for these considerations.

Parameter estimates from these adaptive designs will be subject to selection bias arising from choosing outcomes with the strongest observed effect sizes. This is an area for future research.

An alternative to basing the interim analysis on the first N_1 individuals is to use all available data at the interim analysis, e.g. an alternative statistic of

$$\max_k Z_{t_k}(\text{all available data}) = \max\{Z_{t_1}(1, 233), Z_{t_2}(1, 183), Z_{t_3}(1, 133)\} \quad (3)$$

for the example in Figure 1. This does have the advantage of using more data and hence increasing the likelihood of stopping early. However, there are some drawbacks. For a conditional likelihood approach employed in Approach A3, the second stage test statistic needs to be based on a set of observations that are independent of information used in determining the maximum in (3) for the first stage test. For Figure 1, this means if, say, follow-up t_2 had the maximum test statistic based on $Z_{t_2}(1, 183)$ then any second stage test would only be based on the 234th through 400th observations - the data from the 184th through 233rd individual could not be used. Simulations (not included) show that including

these omitted data lead to biased results that does not maintain nominal Type I error. Appropriate Type I error can be obtained by omitting these observations but at a substantial loss of power - a more detailed examination is available in Web Appendix C. A similar bias arises when trying to adapt Approach A1 for use with (3). Simulations show Approach A2 may have good characteristics when using (3) but further work is needed to draw general conclusions. A second disadvantage with using (3) is that this max statistic is based on different amounts of information and thus may miss important differences in effect sizes. In summary, there are important difficulties in using the max statistic with all available information at the interim analysis.

A reviewer has raised the problem of additional data being available when the first test is significant at level α_1 , often called “overrun”. One can introduce a second hurdle, retesting at α_1 using a max test based all of the available data. If the second test is not significant at α_1 , one could continue to stage 2. This is a closed testing procedure in the spirit of Tang and Geller (1999) and Glimm, Maurer, and Bretz (2010) that would strongly protect α , but affect power, the likelihood of stopping early, and any assessment of the time at which a difference occurs. This use of the max statistic in (3) is only as a second hurdle without changing the second stage statistic and does not replace the interim test statistic based on N_1 individuals; consequently the bias problem mentioned above is avoided. Further work on such procedures including overrun is in progress.

Other extensions to these proposals are possible. Approaches A1 and A2 readily extend to more than two stages and any spending function although the calculations and decisions become more complex. Although we assumed our test statistics to be normally distributed, the methodology can be extended to other asymptotically normal test statistics (such as the Mann-Whitney-Wilcoxon test). As well, further work on missing data in this problem would be of interest.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank an anonymous referee for suggesting the hybrid Bonferroni approach.

References

- Armitage P, Stratton IM, Worthington HV. Repeated significance tests for clinical trials with a fixed number of patients and variable follow-up. *Biometrics*. 1985; 41:353–359. [PubMed: 4027318]
- Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994; 50:1029–1041. [PubMed: 7786985]
- Geary DN. Sequential testing in clinical trials with repeated measurements. *Biometrics*. 1998; 75:311–318.
- Genz, A.; Fretz, F.; Miwa, T.; Mi, X.; Leisch, F.; Scheipl, F.; Hothorn, T. mvtnorm: Multivariate Normal and t Distributions. 2012. R package version 0.9-9992. <http://CRAN.R-project.org/package=mvtnorm>
- Glimm E, Maurer W, Bretz F. Hierarchical testing of multiple endpoints in group sequential trials. *Statistics in Medicine*. 2010; 29:219–228. [PubMed: 19827011]

- Jennison, C.; Turnbull, BW. Group Sequential Methods with Applications to Clinical Trials. Boca Raton: Chapman and Hall/CRC; 2000.
- Kelly PJ, Stallard N, Todd S. An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics*. 2005; 15:641–658. [PubMed: 16022169]
- Kieser M, Buaer P, Lehmacher W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal*. 1999; 41:261–277.
- Kittelson JM, Sharples K, Emerson SS. Group sequential clinical trials for longitudinal data with analyses using summary statistics. *Statistics in Medicine*. 2005; 24:2457–2475. [PubMed: 15977295]
- Kosorok MR, Shi Y, DeMets DL. Design and analysis of group sequential clinical trials with multiple primary endpoints. *Biometrics*. 2004; 60:134–145. [PubMed: 15032783]
- Lan K, DeMets D. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983; 70:659–663.
- Lee JW, DeMets DL. Sequential comparison of changes with repeated measures data. *Journal of the American Statistical Association*. 1991; 86:757–762.
- Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine*. 2004; 23:2397–2508.
- O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 35:549–556. [PubMed: 497341]
- Pocock SJ. Groups sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977; 64:191–199.
- Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics*. 1995; 51:1315–1324. [PubMed: 8589224]
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2013. <http://www.R-project.org/>
- Rodgers WJ, Johnstone DE, Yusuf S, Weiner DH, Gallagher P, Bittner VA, Ahn S, Schron E, Shumaker SA, Sheffield LT. Quality of life among 5,025 patients with left ventricular dysfunction randomized between placebo and enalapril: the studies of left ventricular dysfunction. *Journal of the American College of Cardiology*. 1994; 23:393–400. [PubMed: 8294693]
- Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine*. 2003; 22:689–703. [PubMed: 12587100]
- Tang DI, Geller NL. Closed testing procedures for group sequential clinical trial with multiple endpoints. *Biometrics*. 1999; 55:1188–1192. [PubMed: 11315066]
- The SOLVD Investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *The New England Journal of Medicine*. 1991; 325:293–302. [PubMed: 2057034]
- Wu MC, Lan KKG. Sequential monitoring for comparison of changes in a response variable in clinical studies. *Biometrics*. 1992; 48:765–80. [PubMed: 1420840]

Example: $N = 400$, 24 Month Accrual with Uniform Enrollment,
 $t_1 = 6$ months, $t_2 = 9$ months, and $t_3 = 12$ months

Figure 1.
 Temporal availability of test statistics

Table 1

Type I error and power (based on 10,000 simulations). Bolded entries denote the approach with highest power.

$N, N_1/N$	6		9		12		Approach A1 Joint		Approach A2 Joint		Approach A3 CRP		Approach A4 Simple 2 Study		Approach A5 Hybrid Bonferroni	
	Mos	Diff.	Mos	Diff.	Mos	Diff.	$Z_t^*(1, N_1)$	$Z_t^*(1, N)$	$Z_t^*(1, N_1)$	$\max Z_{t_k}(1, N)$	$Z_t^*(1, N_1)$	$Z'(N_1 + 1, N), \pi^*$	$Z_t^*(1, N_1)$	$Z_t^*(N_1 + 1, N)$	$Z_t^*(1, N_1)$	$\max Z_{t_k}(1, N)$
200, 1/3	0		0		0		0.0488		0.0482		0.0504, 0.36		0.0873			0.0366
400, 1/3	0		0		0		0.0533		0.0520		0.0542, 0.36		0.0909			0.0427
800, 1/3	0		0		0		0.0477		0.0489		0.0475, 0.36		0.0844			0.0359
200, 1/2	0		0		0		0.0505		0.0516		0.0521, 0.36		0.0901			0.0394
400, 1/2	0		0		0		0.0494		0.0510		0.0496, 0.35		0.0876			0.0402
800, 1/2	0		0		0		0.0524		0.0528		0.0513, 0.37		0.0876			0.0384
200, 2/3	0		0		0		0.0450		0.0441		0.0469, 0.36		0.0863			0.0351
400, 2/3	0		0		0		0.0535		0.0524		0.0527, 0.36		0.0898			0.0403
800, 2/3	0		0		0		0.0499		0.0489		0.0500, 0.36		0.0841			0.0373
400, 1/2	0		0		0.40		0.8371		0.8242		0.8408 , 0.95		0.8202			0.7907
400, 1/2	0.35		0.20		0		0.8305		0.8369		0.8258, 0.69		0.8105			0.8044
400, 1/2	0.30		0.30		0.30		0.7910		0.8202		0.8140, 0.36		0.7696			0.7874
400, 2/3	0		0		0.40		0.8306		0.8206		0.8315 , 0.98		0.8138			0.7865
400, 2/3	0.35		0.20		0		0.8246		0.8247		0.8215, 0.74		0.7989			0.7879
400, 2/3	0.30		0.30		0.30		0.8036		0.8213		0.8183, 0.36		0.7826			0.7834

* Note: $\hat{\pi}$ denotes the proportion of times $Z'(N_1 + 1, N)$ is chosen to be $Z_t^*(N_1 + 1, N)$ instead of $\max Z(N_1 + 1, N)$ under the CRP approach.

Table 2

Summary statistics for dyspnea frequency in SOLVD. Higher scores indicate more frequent dyspnea. Only participants who completed each survey are counted here.

	Placebo, Mean	$N_{\text{Plac}} = 517$ Std Dev.	Enalapril Mean	$N_{\text{Enal}} = 538$ Std Dev.
Baseline	2.56	1.88	2.60	1.89
6 Weeks	2.21	1.89	2.07	1.85
1 Year	2.18	1.87	2.00	1.86
2 Year	2.10	1.89	1.99	1.85

Table 3

Test statistics for dyspnea frequency in SOLVD computed for follow-up periods

	Follow-up Period		
	6 Week	1 Year	2 Year
Only First Stage Data: $(1, N_1)$	$Z_{6w}(1, 422) = 1.139$	$Z_{1y}(1, 422) = 0.518$	$Z_{2y}(1, 422) = 0.255$
Only Second Stage Data: $(N_1 + 1, N)$	$Z_{6w}(423, 1055) = 1.426$	$Z_{1y}(423, 1055) = 2.088$	$Z_{2y}(423, 1055) = 1.475$
All Data: $(1, N)$	$Z_{6w}(1, 1055) = 1.826$	$Z_{1y}(1, 1055) = 1.957$	$Z_{2y}(1, 1055) = 1.307$

Table 4

Observed statistics and significance thresholds for dyspnea frequency in SOLVD

	A1 Joint $Z_t^*(\mathbf{1}, N_t)$ $Z_t^*(\mathbf{1}, N)$	A2 Joint $Z_t^*(\mathbf{1}, N_t)$ $\max Z_{t_k}(\mathbf{1}, N)$	A3 CRP $Z_t^*(\mathbf{1}, N_t)$ $Z'(N_1 + \mathbf{1}, N) = \max Z_{t_k}(N_1 + \mathbf{1}, N)$	A4 Simple 2 Study $Z_t^*(\mathbf{1}, N_t)$ $Z_t^*(N_1 + \mathbf{1}, N)$	A5 Bonferroni $Z_t^*(\mathbf{1}, N_t)$ $\max Z_{t_k}(\mathbf{1}, N)$
Stage 1 Observed Stat.	1.139	1.139	1.139	1.139	1.139
Stage 1 Threshold	3.046	3.046	3.046	2.128	3.071
Stage 2 Observed Stat.	1.826	1.957	2.088	1.426	1.957
Stage 2 Threshold	1.958	2.067	1.998	1.645	2.141