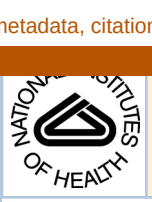

Published in final edited form as:

J Biopharm Stat. 2012 ; 22(4): 758–772. doi:10.1080/10543406.2010.528103.

Flexible Analytical methods for Adding a Treatment Arm Mid-study to an Ongoing Clinical Trial

Jordan J. Elm, PhD, **Yuko Y. Palesch, PhD**, **Gary G. Koch, PhD**, **Vanessa Hinson, MD**, **Bernard Ravina, MD**, and **Wenle Zhao, PhD**

Abstract

It is not uncommon to have experimental drugs under different stages of development for a given disease area. Methods are proposed for use when another treatment arm is to be added mid-study to an ongoing clinical trial. Monte Carlo simulation was used to compare potential analytical approaches for pairwise comparisons through a difference in means in independent normal populations including 1.) a linear model adjusting for the design change (stage effect), 2.) pooling data across the stages, or 3.) the use of an adaptive combination test. In the presence of intra-stage correlation (or a non-ignorable fixed stage effect), simply pooling the data will result in a loss of power and will inflate the type I error rate. The linear model approach is more powerful, but the adaptive methods allow for flexibility (re-estimating sample size). The flexibility to add a treatment arm to an ongoing trial may result in cost savings as treatments that become ready for testing can be added to ongoing studies.

Introduction

Analytical methods for adding a new treatment arm to an ongoing clinical trial have not been addressed in the literature. Consider the scenario of a randomized, double-blind parallel arm clinical trial of treatment A versus placebo. This study may be large and long-term. At some point after randomization has begun, but prior to the end of enrollment, a new treatment B showing promise is identified. Investigators and Sponsor desire to add treatment B to the ongoing study in order to reduce the number of placebo subjects that would be needed if two separate clinical trials were to be conducted. Thus, without re-randomizing previously enrolled subjects, the decision is made to randomize all new patients to one of three arms: A (with placebo for B), B (with placebo for A), or placebo for both A and B.

One example where such a scenario may be applied is the NINDS Exploratory Trials in Parkinson's Disease (NET-PD) initiative. This program funds a series of clinical trials of potentially disease modifying agents(NINDS NET-PD Investigators, 2006; NINDS NET-PD Investigators, 2007). These agents (believed to impact different biological mechanisms of action) are at different stages of development; some require Phase II testing, while others are ready for Phase III testing. In Parkinson's disease, a definitive Phase III trial is costly and requires five or more years of follow-up to evaluate improvement in clinical progression. This work was originally motivated by the possibility of adding active arms to an ongoing Phase III randomized trial of creatine versus placebo in Parkinson's disease (the LS-1 study).

When more than one promising treatment are available for a Phase III clinical trial, conducting a multi-arm study is more efficient than conducting separate studies of each

Correspondence to: Jordan J Elm, PhD, Medical University of South Carolina, Division of Biostatistics & Epidemiology, 135 Cannon St., Suite 303, PO Box 250835, Charleston, SC 29425, Phone (843) 876-1605, elmj@musc.edu.

intervention; less placebo patients are needed and savings on infrastructure costs (such as coordinating centers) can be expected. However, delays in the drug supply chain for one drug, or the need to obtain supporting clinical or pre-clinical data can make it more difficult to start a multi-armed clinical trial compared to separate studies of each intervention. For example, one drug may not have adequate preliminary data (although it is being used in practice) compared to the other(s). Once a safety profile has been established, the investigators wish to add this drug to an ongoing trial, lest the trial become obsolete, in that it does not reflect real-world practice, before it is completed.

There are other situations in which, for external reasons, it may be practical to begin a clinical trial with the possibility that another arm may be added mid-course. One such example is a multi-dose study where for safety reasons, investigators do not wish to include a higher dose arm at the outset before a lower dose has been administered(Peace & Koch, 1993). Another example is the multi-arm clinical trial of the effectiveness of several antipsychotic drugs in patients with schizophrenia (the CATIE trial). At the outset of the CATIE trial, ziprasidone was pending regulatory approval. A ziprasidone arm was added after approximately 40% of patients had been enrolled, once it had been approved by the FDA (Lieberman et al., 2005). One can envision other examples in comparative effectiveness trials, in which new agents (biologics) become available mid-course of an ongoing trial.

When an arm is added to an ongoing trial there are several statistical considerations. Here, we focus on the family-wise type I error rate, power, sample size, and the choice of analytical methods. It is assumed that it is possible to ensure adequate blinding, that re-randomization of existing subjects cannot and will not be done, and the optimal allocation ratio will be applied. (The allocation scheme would be unequal after the new treatment arm is added and would minimize the time to total enrollment.) In this paper, analytical methods for both single-stage and group sequential designs are addressed for this novel scenario. The power and type I error rate are compared for several analytical methods for a design with a fixed sample size. We will restrict our attention to the case in which the main interest in multi-arm studies is to identify any and all drugs that are better than placebo (not to identify the best drug) as is the case for the NET-PD project. Note the decision to add treatment B is driven by external considerations independent of any impression for performance of A in the trial.

Methodology

Two-stage design with fixed sample size

The choice of test statistic to be applied depends on the original design of the comparison of treatment A versus placebo, and the potential for a cohort or stage effect. Let $y_{1A}, y_{2A},\ldots.$ and $y_{1P}, y_{2P},\ldots.$ be sequences of independent observations receiving treatment A and placebo, respectively, in a two armed clinical trial. Restricting our attention to the test of normal means, assume their respective means are μ_A and μ_P with variance unknown. The one-sided null hypothesis of interest for the comparison of two normal means with variance unknown is $H_A : \theta_A \leq 0$, where $\mu_A - \mu_P = \theta_A$ and a positive difference represents a treatment benefit. Later, a new treatment B is added ($y_{1B}, y_{2B},\ldots.$), where the null hypothesis of interest is : $H_B : \theta_B \leq 0$. Now we have the new overall null hypothesis H_{AB}: $H_A \cap H_B$.

Linear Model Method

There are several methods that could be used to test $H_{AB} : H_A \cap H_B$. We will use the terminology stage 1 to refer to the time period prior to the design change and stage 2 to refer to the time period after the design change. A linear model adjusting for a stage/cohort effect

could be applied. The linear model of interest is $y_{ijk} = \mu + \tau_j + c_k + e_{ijk}$ for j=A,B,P for treatments; k=1,2 for cohorts; i = 1,2,...., = n_{jk} for patients in cohorts. Also, $n_{A1} = n_{P1} = n_1$ with $n_{B1} = 0$ and $n_{A2} = n_{P2} = (n - n_1)$, $n_{B2} = n$. The e_{ijk} are iid $N(0, \sigma_e^2)$ random errors. The c_k can be fixed or independent $N(0, \sigma_c^2)$ for k=1,2. If c_k are fixed then this model is an ANOVA, where by convention $\tau_P = 0$ and $c_1 = 0$. In the random effects case, $var(y_{ijk}) = \sigma_e^2 + \sigma_c^2$, $cov(y_{ijk}, y_{i'j'k}) = \sigma_c^2$, where $i'j' \neq ij$ for subjects in cohort k. Also, $cov(y_{ijk}, y_{i'j\,i'k'}) = 0$ where $k \neq k'$.

For the linear model with the fixed cohort effect (ANOVA) where

$$\underset{\sim}{\bar{y}} = (\bar{y}_{*A1,}\bar{y}_{*A2,}\bar{y}_{*B2,}\bar{y}_{*P1,}\bar{y}_{*P2})' \text{ where } \bar{y}_{*jk} = \frac{\sum_{i=1}^{n_{jk}} y_{ijk}}{n_{jk}}$$

$$\underset{\sim}{d} = \begin{pmatrix} \bar{y}_{*A1} - \bar{y}_{*P1} \\ \bar{y}_{*A2} - \bar{y}_{*P2} \\ \bar{y}_{*B2} - \bar{y}_{*P2} \end{pmatrix} = (d_{AP,1}\, d_{AP,2}\, d_{BP,2})', E(\underset{\sim}{d}) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \tau_A - \tau_P \\ \tau_B - \tau_P \end{pmatrix} = \underset{\sim}{X}_d \underset{\sim}{\tau}$$

$$\text{var}(\underset{\sim}{d}) = \sigma_e^2 \begin{pmatrix} 2/n_1 & 0 & 0 \\ 0 & 2/(n-n_1) & 1/(n-n_1) \\ 0 & 1/(n-n_1) & [\frac{1}{n} + \frac{1}{(n-n_1)}] \end{pmatrix} = \sigma_e^2 \underset{\sim}{V}_d$$

In a weighted regression, let $\widehat{\underset{\sim}{\tau}} = (\underset{\sim}{X}_d' V_d^{-1} \underset{\sim}{X}_d)^{-1} \underset{\sim}{X}_d' \underset{\sim}{d}$. Then it can be shown that

$var(\widehat{\underset{\sim}{\tau}}) = (\underset{\sim}{X}_d' V_d^{-1} \underset{\sim}{X}_d)^{-1} \sigma_e^2 = \sigma_e^2 \begin{pmatrix} \frac{2}{n} & \frac{1}{n} \\ \frac{1}{n} & \frac{4n-3n_1}{2n(n-n_1)} \end{pmatrix}$.

Of note, $var(\hat{\tau}_B - \hat{\tau}_P) < var(d_{BP,2})$. Thus, for the B versus P treatment comparison, the variance of the linear model (in the fixed effects case) is smaller than variance of the second cohort alone even though no patients receive treatment B in stage 1. This can be explained by the Rao-Blackwell theorem since, $\widehat{\tau}_B - \widehat{\tau}_P = d_{BP,2} + \frac{n_1}{2n}(d_{AP,1} - d_{AP,2})$.

For the linear model with the random cohort effect $c_k \sim N(0, \sigma_c^2)$

$$E(\underset{\sim}{\bar{y}}) = \begin{pmatrix} \mu_A \\ \mu_A \\ \mu_B \\ \mu_P \\ \mu_P \end{pmatrix} \text{var}(\underset{\sim}{\bar{y}}) = \begin{pmatrix} \sigma_c^2 + \frac{\sigma_e^2}{n_1} & 0 & 0 & \sigma_c^2 & 0 \\ 0 & \sigma_c^2 + \frac{\sigma_e^2}{(n-n_1)} & \sigma_c^2 & 0 & \sigma_c^2 \\ 0 & \sigma_c^2 & \sigma_c^2 + \frac{\sigma_e^2}{n} & 0 & \sigma_c^2 \\ \sigma_c^2 & 0 & 0 & \sigma_c^2 + \frac{\sigma_e^2}{n_1} & 0 \\ 0 & \sigma_c^2 & \sigma_c^2 & 0 & \sigma_c^2 + \frac{\sigma_e^2}{(n-n_1)} \end{pmatrix}.$$

$$var(\bar{y}_{*A*} - \bar{y}_{*P*}) = \frac{2\sigma_e^2}{n}, var(\bar{y}_{*B2} - \bar{y}_{*P*}) = \frac{2\sigma_e^2}{n} + 2\frac{n_1^2}{n^2}\sigma_c^2.$$

Pooled Data Method

Alternatively, one could ignore that the subjects randomized to placebo in stage 1 did not receive the placebo for drug B, and then naively pool the data across stages. In the pooled analysis, the usual two-sample t-statistic for independent samples with pooled estimate of variance is performed. The pooled estimate of variance (across treatment group and cohort) can be found using the estimates of variance s^2_{*A*}, s^2_{*P*}, and s^2_{*B2}.

$s^2_{*A*}=\sum_{k=1}^{2}\sum_{i=1}^{n_{Ak}}(y_{iAk}-\bar{y}_{*A*})^2/(n-1)$ has $E(s^2_{*A*})=\sigma_e^2+\frac{n_1(n-n_1)}{n(n-1)}(c_1-c_2)^2$.

Likewise, for s^2_{*P*}. $s^2_{*B2}=\sum_{i=1}^{n}(y_{iB2}-\bar{y}_{*B2})^2/(n-1)$ has $E(s^2_{*B2})=\sigma_e^2$.

For the pooled analysis of treatment A versus P, $E(\bar{y}_{*A*}-\bar{y}_{*P*})=\tau_A-\tau_P$ and $E(s^2_{*A*}+s^2_{*P*})/n=\frac{2\sigma_e^2}{n}+\frac{2n_1(n-n_1)}{n^2(n-1)}(c_1-c_2)^2$, whereas $var(\bar{y}_{*A*}-\bar{y}_{*P*})=\frac{2\sigma_e^2}{n}$. Hence, the pooled estimate of variance from the pooled cohort analysis overestimates the variance when fixed cohort effects are present. Thus the pooled analysis would have lower than nominal type I error and somewhat reduced power as cohort effects are larger (compared to an analysis that adjusts for the cohort effect).

For the pooled analysis of treatment B versus P, the following estimator is used $\tau^*_B=\bar{y}_{*B2}-(n_1\bar{y}_{*P1}+(n-n_1)\bar{y}_{*P2})/n$. $E(\tau^*_B)=(\tau_B-\tau_P)-\frac{n_1}{n}(c_1-c_2)$, Hence τ^*_B is a biased estimator. Also $E(s^2_{*B2}+s^2_{*P*})/n=\frac{2\sigma_e^2}{n}+\frac{n_1(n-n_1)}{n^2(n-1)}(c_1-c_2)^2$ whereas $var(\tau_B)=\frac{4n-3n_1}{2n(n-n_1)}$, so the pooled estimate of variance also somewhat overestimates the variance when fixed cohort effects are present.

If c_1, c_2 are independent $N(0,\sigma_c^2)$ then $E(\bar{y}_{*A*}-\bar{y}_{*P*})=\tau_A-\tau_P$ and $E(\tau^*_B)=\tau_B-\tau_P$ there is no longer bias. Nevertheless, for A versus P comparison $E(s^2_{*A*}+s^2_{*P*})/n=\frac{2\sigma_e^2}{n}+\frac{4n_1(n-n_1)}{n^2(n-1)}\sigma_c^2$, whereas $var(\bar{y}_{*A*}-\bar{y}_{*P*})=\frac{2\sigma_e^2}{n}$, and so the pooled analysis of A versus P would still lead to overestimation of variance and slightly lower than nominal type I error and slightly reduced power.

For the B versus P comparison, $E(s^2_{*B2}+s^2_{*P*})/n=\frac{2\sigma_e^2}{n}+\frac{2n_1(n-n_1)}{n^2(n-1)}\sigma_c^2$, whereas $var(\bar{y}_{*B2}-\bar{y}_{*P*})=\frac{2\sigma_e^2}{n}+\frac{2n_1^2}{n^2}\sigma_c^2$, and so variance is substantially underestimated, which can lead to inflated type I error.

Adaptive Procedure

A third approach is to apply an adaptive combination rule for the data from the two stages. By introducing an additional treatment arm, all methods have a penalty from the need to adjust for two comparisons. Moreover, as shown later, the penalty in the context of the fixed sample size is somewhat more for the adaptive methods, although in other paradigms they will offset this penalty with added flexibility.

Given another arm is added mid-study, multiple comparison procedures must be utilized in order to control the family-wise error rate (FWE) at the pre-specified rate. Although it has

been argued that those performing multi-armed studies are penalized by having to maintain the FWE rate at alpha while two separate studies could each be performed at level alpha(O'Brien, 1983), we will restrict our attention to methods for controlling the FWE rate strongly, as this is standard practice for confirmatory clinical trials. Although all pairwise comparisons or many-to-one comparisons could be of interest, we will restrict our attention to many-to-one comparisons. A simple approach valid for any test is the Bonferroni-adjustment, although other approaches such as Holm(Holm, 1976), Hochberg(Hochberg, 1988), or stepwise closed testing methods(Hochberg & Tamhane, 1987; Marcus, Peritz, & Gabriel, 1976) may be less conservative. For the linear model adjusting for stage, closed testing may be done by performing an overall F-test with 2 degrees of freedom and then stepping down to test each pairwise comparison, if only 1 arm is added.

When an arm is added to an ongoing trial, a change has been made to the study design. The primary hypothesis changes from a pairwise test of A versus placebo to a global (intersection) hypothesis of two pairwise tests (or to an F-test). An adaptive design may be well suited for this application, as a way of splitting the study into two stages (before (k=1) and after (k=2) introduction of the new arm).

Let $d_{AP,1} = (\bar{y}_{*A1} - \bar{y}_{*P1})$, $d_{AP,2} = (\bar{y}_{*A2} - \bar{y}_{*P2})$, $d_{BP,2} = (\bar{y}_{*B2} - \bar{y}_{*P2})$, where $\bar{y}_{*jk} = \frac{\sum_{i=1}^{n_{jk}} y_{ijk}}{n_{jk}}$. Then $s^2_{**1} = \sum_{j=A,P}\sum_{i=1}^{n_1}(y_{ij1} - \bar{y}_{*j1})^2/2(n_1 - 1)$ and $s^2_{**2} = \sum_{j=A,B,P}\sum_{i=1}^{n_{j2}}(y_{ij2} - \bar{y}_{*j2})^2/(3n - 2n_1 - 3)$. For the two adaptive methods that will be considered in this paper, use is made of $d_{AP,1}$ and s^2_{**1} when k=1, and $d_{AP,2}$, $d_{BP,2}$ and s^2_{**2} when k=2 to form two-sample t-statistics for each stage. Then the stagewise p-values of the t-statistics are used in an adaptive combination test.

Adaptive combination tests, such as the two-stage Inverse Chi-square (Fisher's) combination test (ICHI) and weighted inverse normal combination test(Mosteller & Bush, 1954) (INORM), can be used to combine data across the two stages(Bauer & Kohne, 1994). The combination rule must be pre-specified at or before the time of the design change (or any study unblinding).

If Inverse Chi-square (Fisher's) combination test (ICHI) is specified as the adaptive combination test, then H_{AB} is rejected if

$$C(p_1, p_2) = p_1 p_2 \leq u_\alpha = \exp[-\frac{1}{2}\chi^2_4(1-\alpha)]$$

where p_1 is the p-value from stage 1 and p_2 is the p-value from stage 2 and $\chi^2_4(1-\alpha)$ is the (1−α)-quantile of the central χ^2 distribution with 4 degrees of freedom(Bauer & Kohne, 1994).

If the weighted inverse normal (INORM) rule is specified as the adaptive combination test, then this can be written as

$$C(p_1, p_2) = w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2) \text{ where } \sum_{k=1}^{2} w_k^2 = 1 \text{ and } w_k = \left(\sum_{k=1}^{2} n_k\right)^{-1/2} \cdot \sqrt{n_k}$$

and n_k is the total number of observations at the k^{th} stage. Then the null intersection hypothesis H_{AB} is rejected at level alpha if $C(p_1,p_2) > \Phi^{-1}(1-\alpha)$ (Mosteller & Bush, 1954). It is advisable to perform one-sided rather than two-sided tests to avoid conflicting decisions when the intermediate test-statistics (for each stage) go in different directions.

The test of H_{AB} is as follows. We denote the intermediate p-values as $p_{k,m}$ for k=1, 2 stages and m=A, B, AB hypotheses. For stage 1, the H_A null hypothesis would be tested via the usual one-sided, two-sample t-test, using all observations from stage 1 to obtain the intermediate p-value $p_{1,A}$. Using the observations from stage 2 only, the intermediate p-value for H_A as the test of treatment A versus placebo is $p_{2,A}$ obtained via a one-sided, two-sample t-test. Likewise, the intermediate p-value $p_{2,B}$ for H_B (the test of treatment B versus placebo) is obtained. The stage 2 p-value for H_{AB} using Simes' method(Hochberg, 1988; Simes, 1986) is defined as $p_{2,AB} = min[2min(p_{2,A}, p_{2,B}), max(p_{2,A}, p_{2,B})]$. Finally, the intermediate p-value from stage 1 and the (multiplicity-adjusted) intermediate p-value from stage 2 form a combination test of $H_{AB}: H_A \cap H_B$.

Several authors have shown the usefulness of closed testing (closure) methods for controlling multiplicity in complex adaptive designs(Hommel, 2001; Kieser, Bauer, & Lehmacher, 1999). These methods are reviewed and examples are worked by Jennison and Turnbull(Jennison & Turnbull, 2007). Closed testing methods are applied to test H_A and H_B across stages. By closure methods, any individual hypothesis can be rejected at global level alpha if the set of all possible intersections is rejected at an appropriate alpha-level test. Thus, H_A can be rejected given H_{AB} is also rejected, where H_{AB} and H_A are both rejected via combination tests over stages 1 and 2. Similarly, H_B is rejected at global level alpha if H_{AB} and H_B are both rejected(Marcus et al., 1976), where H_B is tested using just the stage 2 data via $p_{2,B}$ (since there is no data for treatment B in stage 1).

Adaptive Procedure within Group Sequential setting

Assume the same design as previous, except now the null hypothesis H_A was originally planned to be tested at k interim looks using a group sequential approach to allow for early stopping in favor of the alternative hypothesis. Then, to add treatment arm B and test H_{AB}: $H_A \cap H_B$ via an adaptive combination test, one must select an adaptive test that incorporates the existing group sequential framework (Cui, Hung, & Wang, 1999; Kieser et al., 1999; Lehmacher & Wassmer, 1999). Optionally, the interim conditional power can be computed (either for the next interim look or for the whole study), and the sample size can be increased accordingly based on internal information about the effect size observed at an interim analysis.

Hypothetical Example of Adaptive Procedure within Group Sequential setting

Assume a phase III clinical trial of treatment A versus placebo is ongoing with an interim analysis planned after 50% of subjects have completed follow-up. Assuming Haybittle-Peto stopping boundaries (alpha=0.001 at interim, alpha=0.049 at final) for a one-sided test of alpha=0.05, the planned maximum sample size is 200 per group. After 40 subjects per group have been enrolled (20% of the total sample size), then the protocol is amended to include treatment B, and the planned sample size is increased to 600 total (200 per group) assuming the original design parameters for A versus placebo. At the first interim analysis (with 300 patients), the data are subset into stages 1 (with 80 patients, 40 receiving A, 40 receiving placebo) and 2 (with 220 patients, 60 receiving A, 60 receiving placebo, and 100 receiving B). Using the stage 1 data, the p-value for the test of H_A is $p_{1,A}$ =0.20. Given H_B is not available, the p-value for H_{AB} is $p_{1,AB} = p_{1,A}$ =0.20.

Using the stage 2 data at the first interim analysis, the p-value for H_A is $p_{2,A}$ =0.15 and for H_B is $p_{2,B}$ = 0.06. Using Simes' method the p-value for H_{AB} at stage 2 is $p_{2,AB}$ =0.12. Then a weighted inverse normal adaptive combination test for H_{AB} at the first interim analysis, using weights proportional to the original sample size, is

$C(p_{1,AB}, p_{2,AB}) = \sqrt{80/200} * \Phi^{-1}(1 - p_{1,AB}) + \sqrt{120/200} * \Phi^{-1}(1 - p_{2,AB}) = 1.442$ where 1.442 < 3 the Haybittle-Peto stopping boundary when the information time is 0.50 based on the original sample size.

Continuing the trial, using the data collected after the first interim analysis, the p-value for the test of H_A is $p_{3,A}$ =0.2 and for H_B is $p_{3,B}$ = 0.03. So the p-value for H_{AB} at look 2 is $p_{3,AB}$ =0.06. Then the combination test at the final analysis for H_{AB} is:

$C(p_{1,AB}, p_{2,AB}, p_{3,AB}) = \sqrt{200/400} * C(p_{1,AB}, p_{2,AB}) + \sqrt{200/400} * \Phi^{-1}(1 - p_{3,AB}) = 2.119$ where 2.119>1.65 where 1.65 is the Haybittle-Peto stopping boundary when 100% of the subjects have completed follow-up. Since the overall null hypothesis H_{AB} is rejected, we go on to test H_A and H_B, via combination tests. For : $H_A : \theta_A \leq 0$, we fail to reject the null hypothesis since the Z-statistic for the combination test, 1.539 is less than 1.65. For $H_B : \theta_B \leq 0$ we reject the null hypothesis, since 2.429 is greater than 1.65. In this example, treatment A fails, but treatment B is superior to placebo. For one arm to stop early, one would have to reject H_{AB} and then H_A or H_B.

Simulation Study

Monte Carlo simulation was used to compare the power and type I error rate for testing H_A and H_B in a two-stage, fixed sample clinical trial. The following statistical analysis approaches were compared: 1. two-sample *t*-test with the data pooled across stages (POOL); 2. Linear model adjusting for a fixed stage/cohort effect (LIN); 3. Inverse Chi-Square (Fisher's) adaptive combination test (ICHI); 4. Weighted-Inverse Normal adaptive combination test (INORM).

Two samples were drawn from independent multivariate normal distributions corresponding to the two cohorts such that each y_{ijk} is normal with $E(y_{ijk}) = \mu_j$ for j=A, B, P for treatments, $var(y_{ijk}) = \sigma_e^2 + \sigma_c^2$, and $cov(y_{ijk}, y_{i'j'k}) = \sigma_c^2$, where $i'j' \neq ij$ for subjects in cohort k. The sample size per group was set to n=120 for a two sample *t*-test of H_A or H_B for detecting an effect size θ = 0.38 given σ^2 =1, α = 0.05(one-sided), and power=0.90. The data were simulated assuming different sizes of $\sigma_c^2 \left\{0, 14\theta, \frac{1}{2}\theta, \theta\right\}$ and t_P {0.1, 0.3, 0.5}, where $0 < t_P < 1$ is the time that the design change is made in terms of the fraction of the sample size.

In order to generate data from the multivariate normal distribution given above, the *drawnorm* function in STATA was used specifying the following structure for each cohort: $\underset{\sim}{Y} \sim N(\underset{\sim}{\mu}, \Sigma), \underset{\sim}{\mu}_{360\times 1} = (\mu_A \mathbf{1}'_{120} \mu_B \mathbf{1}'_{120} \mu_P \mathbf{1}'_{120})' \Sigma_{360\times 360} = \boldsymbol{I}_{360} + \sigma_c^2 \mathbf{1}_{360} \mathbf{1}'_{360}$ where $\mathbf{1}_r$ is a (r × 1) vector of ones and $\boldsymbol{I}_r$ is a (r × r) identity matrix. Using this function, 2 independent observations of y_1-y_{360} were generated defining a complete set of data from each cohort; y_1-y_{120} with mean μ_A, y_{121}-y_{240} with mean of μ_B, and y_{241}-y_{360} with mean of μ_P, and all with covariance of σ_c^2. The values of μ_A and μ_B were either 1 (under the null) or 1.38 (under the alternative) and the value of μ_P was 1. From these 2 independent cohorts of observations of y_1-y_{360}, observations were deleted as appropriate for the value of t_p under consideration, and the observations for y_{121}-y_{240} corresponding to observations under treatment B were deleted for stage 1.

This is the same as generating $t_P \times 2n = 2n_1$ random variables in cohort 1 (for treatment A and placebo) and $3n - 2n_1$ random variables in cohort 2 (for treatments A, B, and placebo).

For example if t_P =0.3, then for cohort 1, one could generate 72 random variables from a multivariate normal distribution where $\underset{\sim}{\mu}_{72\times1}=(\mu_A \mathbf{1}'_{36}\mu_P \mathbf{1}'_{36})'$ and $\Sigma_{72\times72}=\boldsymbol{I}_{72}+\sigma_c^2 \mathbf{1}_{72}\mathbf{1}'_{72}$. Likewise, for cohort 2, one could generate 288 random variables from a multivariate normal distribution where $\underset{\sim}{\mu}_{288\times1}=(\mu_A \mathbf{1}'_{84}\mu_B \mathbf{1}'_{120}\mu_P \mathbf{1}'_{84})'$ and $\Sigma_{288\times288}=\boldsymbol{I}_{288}+\sigma_c^2 \mathbf{1}_{288}\mathbf{1}'_{288}$.

One analysis was performed at the end of the study (with no interim analyses). For the adaptive combination tests, at the end of the study the data were subset into stages 1 and 2 prior to forming the combination test.

For all approaches closed testing multiple comparisons procedures were applied, as this approach is more powerful than single-step approaches, but do not provide confidence intervals(Westfall, Tobias, Rom, Wolfinger, & Hochberg, 1999). For consistency across methods, Simes' Modified Bonferroni procedure(Simes, 1986) was used to obtain an adjusted p-value for the global null hypothesis H_{AB}. All simulations were done in STATA. The number of replications was set at 6000. With 6000 replications, a two-sided 99% CI around the expected type I error rate of 0.05 will extend ±0.007, and a two-sided 99% CI around for the expected power of 0.90 will extend ±0.01.

Simulation Results

Table 1 gives the results for the empirical power to reject H_A and H_B when both treatments are superior to placebo ($\mu_A = \mu_B > \mu_P$), the power to reject H_A when only treatment A is superior to placebo ($\mu_A > \mu_B = \mu_P$), and the power to reject H_B when only treatment B is superior to placebo ($\mu_B > \mu_A = \mu_P$). The empirical power is defined as the proportion of times a hypothesis is rejected given the alternative hypothesis is true. Under the closed testing procedure, a particular hypothesis is rejected if all intersection hypotheses containing it are rejected (e.g. H_A is rejected if H_{AB} and H_A are rejected).

In general, when both treatments are superior ($\mu_A = \mu_B > \mu_P$) and when the time at which Arm B is added in terms of the fraction of the sample size is early (t_P=0.1), then all methods obtained nearly 90% power (the nominal level). As expected, when there is no stage effect (σ_c^2=0), then the pooled method has the highest power for testing both H_A and H_B. The linear model with a fixed cohort effect loses power as t_P increases for the test of H_B, but is still superior to the two combination tests. Fisher's combination test (ICHI) performs worst for both H_A and H_B. When the within stage (intracluster) covariance σ_c^2>0, then the pooled method performs worst for both H_A and H_B across all time points (t_P), while the linear model approach is superior. Averaging across all time points (t_P), the amount of covariance does not affect the power for the linear model or the combination tests, but power decreases as covariance increases for the pooled method.

If treatment A is superior and treatment B is null ($\mu_A > \mu_B = \mu_P$), the power to reject treatment A is somewhat reduced for all methods considered. This is because of the use of the closed testing procedure, such that in order to reject H_A, H_{AB} must also be rejected. If only treatment B is superior to placebo ($\mu_B > \mu_A = \mu_P$), then the loss of power to reject H_B is more pronounced for the adaptive combination tests considered here. The loss of power is demonstrated in Table 1 for these scenarios.

Table 2 shows the results of the familywise type I error rate for the four methods. For all methods the familywise type I error rate is controlled strongly when covariance σ_c^2=0. However, the *t*-test (with data pooled across cohorts) produces larger than nominal type I error when σ_c^2>0, as high as 0.22 when the design change is made after 50% of the sample

size for arms A and placebo have already been enrolled. The inflation of the familywise error rate is due to the treatment B versus placebo comparison, rather than the treatment A comparison.

Discussion

Adding an arm to an ongoing clinical trial provides a savings in sample size because fewer patients are randomized to placebo compared to two trials with separate placebo groups (A vs. placebo and B vs. placebo). When a new treatment arm (B) is added to an ongoing clinical trial (of A versus placebo), the placebo group will be different before and after the design change. The whole placebo group received placebo for drug A. However, this is not the case for the B versus placebo comparison. The placebo group in the first stage receives only the placebo for treatment A, while the placebo group in the second stage receives placebo for treatment A and treatment B. Thus, it is questionable whether it would be appropriate to pool the placebos from before and after the design change, since not all placebo subjects received the placebo for B. This may not be a concern when treatment B is an increased dose of treatment A (given the placebo looks the same and is taken in the same daily frequency). However, given the situation in which treatment B looks different than A, there may be ambiguity about the conclusions if the data are simply pooled. There is the potential that the original treatment versus placebo comparison could be perturbed due to enthusiasm over the new drug or a cohort effect(Feng, Shao, & Chow, 2007).

From a statistical standpoint, these simulations show that pooling the data without any adjustment for stage is not advisable. When the observations within stage are correlated, then pooling the data across stages will result in a loss of power for both treatment group comparisons, but particularly for the treatment B comparison. Under the random effects set up, for the A versus placebo comparison the pooled estimate of variance leads to an overestimation of variance, resulting in slightly lower than nominal type I error and slightly reduced power. The results of the simulation for the pooled t-test method for the B vs Placebo comparison seem paradoxical in that type I error is inflated (under the null) and power is reduced (under the alternative) as t_p and σ_c^2 increase. This is not a consequence of the closed testing procedure, because power is still reduced when there is no adjustment for multiple comparisons. Derivations suggest that power would be increased for the test of H_B because the pooled estimate of variance is an underestimate, but instead, the influence of the unaccounted for true variation is responsible for the marked decrease in power for H_B. If the data had been derived as a result of a fixed cohort effect (rather than random), then inflation of the type I error rate for the B versus placebo comparison would occur as a result of bias.

When both treatments are superior to placebo, the power for the linear model approach and the weighted-inverse normal adaptive test were only slightly less than the nominal level for the test of treatment A. However, both of these methods lose power for testing H_B as t_p increases, especially the adaptive test (INORM). Since the adaptive tests are not based on sufficient statistics, the observed loss of power was expected. In all cases Fisher's combination test is outperformed by the weighted-inverse normal combination test. Both the fixed-effect linear model method and the adaptive method control the type I error rate in the presence of a random stage effect.

At what point do the cost savings of introducing a new treatment arm into an ongoing trial rather than starting a new trial become negligible? These simulations did not consider the case when more than 50% of patients have already been enrolled. As we have seen, there is a loss of power by as high as 15 percentage points (75% rather than 90% power) for the test of H_B as the proportion of patients already enrolled increases to 50%, when an adaptive combination test is applied (regardless of the covariance).

These results show that the sample size for the treatment B versus placebo comparison would need to be increased in order to achieve the desired power. For the test of H_B, the linear model is more powerful, but adaptive allows for more flexibility to re-estimate the sample size mid-study. The reduction in power we observed in these simulations is partially due to the use of non-standard test statistics (adaptive methods), but primarily due to the use of unequal sample sizes since only the stage 2 placebo is used to test H_B. One solution to attain the desired power for the linear model or adaptive approaches is to enroll n new patients into the stage 2 placebo group. Further simulations (not shown) indicate that this will mitigate the loss of power for the INORM and linear model approaches as t_P increases. We can see that enrolling $t_P \times n + n$ subjects into the placebo group is still less than the $2 \times n$ subjects enrolled into placebo across two separate trials. However, if two separate trials are conducted, each will be designed with a higher type I error rate, and therefore the sample size for each is smaller than n. While it is undesirable to have a total number of subjects enrolled into placebo greater than either treatment arm, the probability of being allocated to the placebo group need not be greater than the chance of allocation to a treatment arm at any point in time. Further research is needed to adequately address approaches to increase the sample size, including the potential benefit of re-estimating the sample size. If an adaptive test is to be used, the sample size could be re-estimated using the observed effect size from stage 1 for the treatment A versus placebo comparison, without any risk of inflating the type I error rate.

There may be other considerations for selecting between methods. Adaptive methods began to be developed in the 1990s, but have only recently come into practical use. In general, mid-course design changes driven by internal information are more controversial than those driven by external information. Internal information refers to information within the study (such as the interim effect size or a sub-group analysis) that prompts a change in the design (change in sample size, change in target population, etc.). However, when the design change is made based on observed data, then the type I error for the final (pre-specified) analysis will be inflated. By applying adaptive methods (e.g. combination rule) then the pre-specified alpha can be constrained.

When external information (independent from the study) prompts a design change for a study, then the integrity of the study may be questioned because it is difficult to prove that internal information did not play a role in the design change. Thus adaptive methods are also suitable for this situation. In this case, adding a new treatment arm to an ongoing study is an example of a design change based on external information. In this context, the type I error probability will not be inflated due to an internal look at the data since the design change is externally-driven (not involving an unplanned look). However, if a new dose, or dose regimen, is added to an ongoing trial, it may be advisable to adjust for the change in design by a combination test, because it is difficult to show that inside knowledge of the treatment effect for the current dose, did not impact the decision to increase the dose, unless perhaps the treatment arm that is added is a lower dose. Trial integrity can be safeguarded with standard operating procedures for blinding and firewalls, such that it is documented when interim looks occur and by whom. This may allay concerns about the role of internal information in design changes.

There are situations where a regression approach (adjusting for stage/cohort) is perfectly acceptable. One such scenario is when the arm that is added is an active comparator. For instance a clinical trial may be initiated in the US, and shortly thereafter investigators may realize that in order to meet European Regulatory authority (EMEA) approval, the current study treatment would need to be shown non-inferior to the standard of care in Europe. Then, it may make sense to modify the ongoing trial to add this standard as another arm. In this case the investigators wish to show study treatment superiority versus placebo (for the

FDA) and non-inferiority of the study treatment versus the standard (for the EMEA). In this case, since the arm that is added is a standard and the rationale is clearly externally driven, there seems to be no need to adopt an adaptive analysis approach. An approach such as that given by Denne and Koch would be well suited for this scenario since the hypotheses (superiority, noninferiority) are nested(Denne & Koch, 2001). In the context of non-adaptive designs, Denne and Koch have shown that it is possible to test both non-inferiority and superiority sequentially without adjusting for multiplicity because they are nested hypotheses; closed testing methods are applied(Marcus et al., 1976).

When the decision is made to add an arm to an ongoing clinical trial, the original design considerations must be taken into account. According to Follman *et al.* there should be equal criteria used to evaluate each treatment arm(Follmann, Proschan, & Geller, 1994). In order to add a treatment arm, the protocol must be amended. The timing of the protocol re-design may affect the degree to which design changes can be made. It may be more difficult to re-design a trial after an interim analysis has been performed.

When adding another treatment arm, it is important to adjust the randomization allocation ratio to ensure that all three treatment arms complete enrollment at roughly the same time. Firstly and above all, this is necessary to ensure blinding, such that the last patients enrolled are not all receiving treatment B. Secondly, if the randomization allocation ratio is 1:1:1, then, once all patients are enrolled into treatment A, there is a possibility of a observing a third cohort set, who are all enrolled in treatment arm B, having different characteristics than those patients in stage 1 or stage 2.

These results suggest that the linear model method will always outperform the adaptive methods as t_P increases, even in group sequential setting, unless the effect size is smaller than expected for one or both treatment arms or the variance is larger than expected. In this case an adaptive method allows the sample size to be increased due to internal information observed at the first interim analysis, and thus the desired power can be attained.

In Parkinson's disease clinical trials, there is a chance that there will be increased enthusiasm about a new drug. Publicity concerning phase II studies for certain drugs can impact the rate of enrollment into a Phase III study of the same drug, and this is likely to have an impact on subjective self-reported outcome measures. For Parkinson's disease clinical trials the cohort effect may be a legitimate concern. Prior NET-PD studies have shown that changes in clinical practice over time have a major impact on outcome measures(NINDS NET-PD Investigators, 2006; NINDS NET-PD Investigators, 2007). Another important point is that t_P and the covariance within stage are likely to increase together. That is, the longer you wait to add the new treatment arm, the higher the t_P will be and the more likely things are to have changed (new standards of care, etc.) introducing cohort effects.

The NET-PD investigators continue to pursue Phase II trials of additional agents for the treatment of Parkinson's disease. If new agents become ready for Phase III testing by the NET-PD group, design modification to the current ongoing Phase III trial is a possibility. These results suggest that it would be inadvisable to simply pool the treatment groups across the stages, since the type I error will be inflated and power will decrease if the intra-stage correlation is greater than zero. In the presence of possible intra-stage correlation, the linear model approach is more powerful, but the adaptive method allows for more flexibility to re-estimate the sample size. Both analysis approaches (regression and adaptive) control the type I error rate when no internal study information is used in the decision to add a new treatment arm mid-study.

Acknowledgments

This work was supported by the NIH (National Institute of Neurological Disorders and Stroke) U01NS043127 and U01NS059041.

Reference List

Bauer P, Kohne K. Evaluation of Experiments with Adaptive Interim Analyses. Biometrics. 1994; 50:1029–1041. [PubMed: 7786985]

Cui L, Hung HMJ, Wang SJ. Modification of Sample Size in Group Sequential Clinical Trials. Biometrics. 1999; 55:853–857. [PubMed: 11315017]

Denne JS, Koch GG. Monitoring a clinical trial with multiple hypotheses concerning the treatment effect on a single primary endpoint. Stat Med. 2001; 20:2801–2812. [PubMed: 11568939]

Feng H, Shao J, Chow S. Group sequential test for clinical trials with moving patient population. J Biopharmaceutical Statistics. 2007; 17:1227–1238.

Fisher, R. Statistical Methods for Research Workers. London: Oliver and Boyd; 1932.

Follmann DA, Proschan MA, Geller NL. Monitoring Pairwise Comparisons in Multi-Armed Clinical Trials. Biometrics. 1994; 50:325–336. [PubMed: 8068834]

Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988; 75:800–802.

Hochberg, Y.; Tamhane, A. Multiple Comparison Procedures. New York: John Wiley & Sons; 1987. p. 55-56.

Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1976; 6:65–70.

Hommel G. Adaptive Modifications of Hypotheses After an Interim Analysis. Biometrical Journal. 2001; 43:581–589.

Jennison C, Turnbull B. Adaptive seamless designs: Selection and prospective testing of hypotheses. J Biopharmaceutical Statistics. 2007; 17:1135–1161.

Kieser M, Bauer P, Lehmacher W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. Biom.J. 1999; 41:261–277.

Lehmacher W, Wassmer G. Adaptive Sample Size Calculations in Group Sequential Trials. Biometrics. 1999; 55:1286–1290. [PubMed: 11315085]

Lieberman J, Stroup T, McEvoy J, Swartz M, Rosenheck R, Perkins D, et al. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. N Engl J Med. 2005; 353:1209–1223. [PubMed: 16172203]

Marcus R, Peritz E, Gabriel KR. On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. Biometrika. 1976; 63:655–660.

Mosteller, F.; Bush, R. Selected quantitative techniques. In: Lindzey, G., editor. Handbook of Social Psychology. Cambridge, Massachusetts: Addison-Wesley: 1954. p. 289-334.

NINDS NET-PD Investigators. A randomized, double-blind, futility clinical trial of creatine and minocycline in early Parkinson disease. Neurology. 2006; 66:664–671. [PubMed: 16481597]

NINDS NET-PD Investigators. A Randomized Clinical Trial of Coenzyme Q10 and GPI-1485 in Early Parkinson's Disease. Neurology. 2007; 68:20–28. [PubMed: 17200487]

O'Brien P. The appropriateness of analysis of variance and multiple comparison procedures. Biometrics. 1983; 39:787–794. [PubMed: 6652209]

Peace K, Koch GG. Statistical methods for a three-period crossover design in which high dose cannot be used first. J Biopharm Stat. 1993; 3:103–16. [PubMed: 8485531]

Simes R. An improved Bonferroni procedure for multiple tests of significance. Biometrika. 1986; 73:751–754.

Westfall, P.; Tobias, R.; Rom, D.; Wolfinger, R.; Hochberg, Y. Multiple Comparisons and Multiple Tests Using the SAS system. Cary, NC: SAS Institute Inc; 1999. p. 149

Table 1

Simulated Power for tests of H_A and H_B (proportion of rejections after 6000 replications)

		$\mu_A = \mu_B > \mu_P$								$\mu_A > \mu_B = \mu_P$				$\mu_B > \mu_A = \mu_P$			
		POOL		**LIN**		**ICHI**		**INORM**		**POOL**	**LIN**	**ICHI**	**INORM**	**POOL**	**LIN**	**ICHI**	**INORM**
t_P	σ_c^2	H_A	H_B	H_A	H_B	H_A	H_B	H_A	H_B	H_A	H_A	H_A	H_A	H_B	H_B	H_B	H_B
0.1	0	0.893	0.891	0.893	0.886	0.857	0.853	0.894	0.877	0.827	0.828	0.793	0.827	0.831	0.819	0.701	0.775
	¼ θ	0.881	0.878	0.882	0.876	0.843	0.847	0.885	0.869	0.824	0.829	0.792	0.824	0.819	0.823	0.707	0.778
	½ θ	0.878	0.859	0.888	0.881	0.845	0.844	0.888	0.868	0.826	0.836	0.795	0.834	0.804	0.820	0.703	0.775
	θ	0.875	0.838	0.891	0.884	0.85	0.851	0.889	0.878	0.814	0.831	0.795	0.829	0.784	0.827	0.706	0.784
0.3	0	0.891	0.885	0.888	0.857	0.871	0.815	0.894	0.824	0.830	0.831	0.815	0.829	0.833	0.796	0.635	0.618
	¼ θ	0.872	0.8	0.882	0.86	0.863	0.818	0.886	0.829	0.821	0.827	0.815	0.824	0.753	0.792	0.638	0.628
	½ θ	0.866	0.758	0.889	0.861	0.87	0.819	0.892	0.83	0.829	0.837	0.825	0.836	0.704	0.793	0.635	0.622
	θ	0.85	0.718	0.888	0.857	0.873	0.819	0.892	0.829	0.800	0.830	0.813	0.830	0.661	0.806	0.645	0.635
0.5	0	0.882	0.889	0.877	0.823	0.871	0.763	0.888	0.764	0.839	0.840	0.829	0.837	0.831	0.750	0.541	0.419
	¼ θ	0.868	0.734	0.885	0.826	0.878	0.747	0.892	0.747	0.829	0.833	0.820	0.830	0.687	0.744	0.540	0.420
	½ θ	0.854	0.684	0.88	0.824	0.875	0.754	0.891	0.753	0.833	0.839	0.832	0.836	0.643	0.747	0.538	0.423
	θ	0.834	0.63	0.885	0.824	0.876	0.761	0.894	0.761	0.809	0.831	0.826	0.832	0.605	0.753	0.544	0.432

POOL= *t*-test (Pool Data across cohort)
LIN=Linear Model (adjusting for cohort as fixed effect)
ICHI= Inverse Chi-Square Combination test
INORM= Weighted Inverse Norm Comb Test
t_P is the time at which design change is made (as a proportion of total sample size enrolled). $H_A: \theta_A \leq 0$, $H_B: \theta_B \leq 0$ where $\theta_A = \mu_A - \mu_P$ and $\theta_B = \mu_B - \mu_P$. The results are given for different sizes of intra-stage covariance $\sigma_c^2 \left\{0, 14\theta, \frac{1}{2}\theta, \theta\right\}$ where θ=0.38 is the true effect size. Simulations for fixed sample size of 120/group.

Table 2

Family-Wise Type I error rates (results of 6000 simulations)

t_P	σ_c^2	POOL	LIN	ICHI	INORM
0.1	0	0.048	0.048	0.048	0.047
	¼ θ	0.054	0.050	0.047	0.049
	½ θ	0.052	0.045	0.044	0.044
	θ	0.065	0.049	0.048	0.049
0.3	0	0.048	0.048	0.048	0.047
	¼ θ	0.089	0.049	0.047	0.048
	½ θ	0.116	0.047	0.047	0.046
	θ	0.154	0.049	0.050	0.048
0.5	0	0.048	0.048	0.047	0.046
	¼ θ	0.134	0.046	0.045	0.045
	½ θ	0.174	0.046	0.048	0.045
	θ	0.215	0.047	0.046	0.046

POOL= *t*-test (Pool Data across cohort)
LIN=Linear Model (adjusting for cohort as fixed effect)
ICHI= Inverse Chi-Square Combination test
INORM= Weighted Inverse Norm Comb Test
NOTE: Family-Wise error rate is the probability of rejecting any true hypothesis from a family of hypotheses given all possible configurations of the null hypotheses ($\mu_A = \mu_B = \mu_P$, $\mu_A > \mu_B = \mu_P$ or $\mu_B > \mu_A = \mu_P$). Under closed testing procedure, a particular hypothesis is rejected if all intersection hypotheses containing it are rejected (e.g. H_A is rejected if H_{AB} and H_A are rejected). tp is the time at which design change is made (as a proportion of total sample size enrolled). The results are given for different sizes of intra-stage covariance $\sigma_c^2 \left\{0, 14\theta, \frac{1}{2}\theta, \theta\right\}$ where θ=0.38 is the true effect size.