Published in final edited form as:

Clin Trials. 2010 October ; 7(5): 537–545. doi:10.1177/1740774510378695.

A threshold sample-enrichment approach in a clinical trial with heterogeneous subpopulations

Aiyi Liu, PhD,

Biostatistics and Bioinformatics Branch, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Rockville, MD 20852, U.S.A

Qizhai Li, PhD,

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Chunling Liu, PhD,

Biostatistics and Bioinformatics Branch, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Rockville, MD 20852, U.S.A

Kai F. Yu, PhD, and

Biostatistics and Bioinformatics Branch, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Rockville, MD 20852, U.S.A

Vivian W. Yuan, PhD

Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD 20993, U.S.A

Abstract

Background—Large comparative clinical trials usual target a wide-range of patients population in which subgroups exist according to certain patients' characteristics. Often, scientific knowledge or existing empirical data support the assumption that patients' improvement is larger among certain subgroups than the others. Such information can be used to design a more cost-effective clinical trial.

Purpose—The goal of the article is to use such information to design a more cost-effective clinical trial.

Method—A two-stage sample-enrichment design strategy is proposed that begins with enrollment from certain subgroup of patients and allows the trial to be terminated for futility in that subgroup.

Results—Simulation studies show that the two-stage sample-enrichment strategy is cost-effective if indeed the null hypothesis of no treatment improvement is true, as also so illustrated with data from a completed trial of calcium to prevent preeclampsia.

Limitations—Feasibility of the proposed enrichment design relies on the knowledge prior to the start of the trial that certain patients can benefit more than others from the treatment.

Conclusions—The two-stage sample-enrichment approach borrows strength from treatment heterogeneity among target patients in a large scale comparative clinical trial, and is more cost-effective if the treatment are of no difference.

Keywords

Sample size and power; stopping for futility; subgroup analysis; treatment heterogeneity

Introduction

In a large comparative (phase III) clinical trial, subgroups of patients, as categorized by certain characteristics, usually exist, and the improvement due to the treatment often varies among patients in different subgroups. As an example, Gordon et al. [1] reported a statistically significant overall hazard ratio estimate from a randomized clinical trial in which women with ovarian cancer were treated with either pegylated liposomal doxorubicin or topotecan. For patients with platinum-sensitive disease, an even more significant estimate of hazard ratio was found. However, among patients with platinum-refractory disease, the hazard ratio was not significant. Using data from this trial, Song and Chi[2] demonstrated a two-stage procedure to test hypothesis concerning both the overall hazard ratio and subgroup-specific hazard ratio in the first stage followed by testing each individual hypothesis in the second stage. An earlier work of Follmann[3]described the PATHS (Prevention and Treatment of Hypertension Study) study (Cushman et al.[4]) in which interim data yielded different treatment effect estimates among two strata (80–89 mm Hg and 90–99 mm Hg diastolic blood pressure) of patients, and presented methods to control type I error rate when screening criterion was modified for the two strata.

When patients' characteristics are well defined, even before the trial starts, investigators may have good reasons to believe that patients in one group, say group *X*, will reveal better treatment outcomes than another group, say group *Y*, either based on scientific knowledge or empirical data from previous trials with similarly functioned drugs. Therefore, if the treatment does not look promising for patients in group *X*, then very unlikely it will show promising results for patients in group *Y*. If this is the case, we argue that a two-stage sample-enrichment trial strategy described below tends to be more ethical and cost-effective than the conventional approach which simultaneously enrolls and randomizes patients from both groups into treatment arms. At the first stage, only patients from group *X* are enrolled and randomized. Treatment difference is then estimated based on data from the first stage, and only if the observed difference is promising, will patients from both groups be enrolled and randomized for the remaining part of the trial. This strategy can be viewed as a trial with early stopping on futility (e.g. Shih, Quan and Li[5], and Lachin[6]). However, here, futility is only determined by data from the more promising group of patients. When the treatment shows no improvement within any groups, such design avoids further recruiting of patients from either group and is thus cost-effective, and more ethically sound.

The CPEP (Calcium to Prevent Preeclampsia) trial serves well as a motivating example. The trial was a randomized, double-blind clinical trial conducted by the Division of Epidemiology, Statistics and Prevention Research of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development from 1992 to 1995. The principal objective of the trial was to determine if calcium supplementation in healthy pregnant nulliparae reduces the incidence of preeclampsia. A total of 4589 healthy nulliparous women who were 13 to 21 weeks pregnant were randomized to receive daily treatment with either 2g of elemental calcium or placebo for the remainder of their pregnancies. The rationale, design and methods of the trial were reported in Levine et al.[7] and the main finding that calcium supplementation did not prevent preeclampsia in healthy nulliparous women was reported in Levine et al.[8]. It is believed that, if calcium supplementation indeed can prevent preeclampsia, then the reduction in preeclampsia incidence is expected to be higher among the healthy nulliparae who normally have low calcium intake (Group *X*) than among

those who normally have higher calcium intake (Group Y). In Section 4, we shall use the data from the completed trial to demonstrate that, had this information been used to design the trial as a two-stage sample-enrichment trial, the trial could have been terminated with much fewer enrollments.

We will first give detailed description of the two-stage sample-enrichment trial strategy and presents methods to control the Type I error rate and preserve power for testing group-specific treatment effects. Details are then given on testing an overall treatment improvement indexed as a weighted average of the group-specific treatment effects. Simulation results are presented to investigate the characteristics of the tests. The methods are exemplified the strategy using the CPEP trial data.

Fixed sample size approach

Consider a clinical trial involving two subpopulations, $X \sim N(\mu_X, 1)$ and $Y \sim N(\mu_Y, 1)$, where μ_X and μ_Y measure the treatment difference between the experimental treatment arm and the standard treatment arm for the two subpopulations, respectively. The null hypothesis is $H_0: \mu_X = 0$ and $\mu_Y = 0$, that is, there is no difference between the two treatment arms for patients from either subpopulations. For simplicity we consider one-sided alternative hypothesis, $H_1: \mu_X > 0$ or $\mu_Y > 0$, that is, treatment difference exists in favor of the experimental arm for at least one subpopulation.

In a fixed sample size trial a sample of $X_1, \dots, X_n$ from the X -population $N(\mu_X, 1)$ and a sample of $Y_1, \dots, Y_m$ from the Y -population $N(\mu_Y, 1)$ are collected. The null hypothesis is

$$\begin{aligned} & \text{rejected if } \sqrt{n}\widehat{\mu}_X > c_X \text{ or } \sqrt{m}\widehat{\mu}_Y > c_Y, \text{ where } \widehat{\mu}_X = \sum_{i=1}^n X_i/n \text{ and } \widehat{\mu}_Y = \sum_{j=1}^m Y_j/m \text{ are the two} \\ & \text{sample means. The two critical values } c_X \text{ and } c_Y \text{ are chosen so that the overall type I error} \\ & \text{rate is controlled at a level of } \alpha. \text{ Because observations from the two populations are} \\ & \text{independent, we have} \end{aligned}$$

$$\begin{aligned} \alpha &= P_{H_0} \{ \sqrt{n}\widehat{\mu}_X > c_X, \text{ or } \sqrt{m}\widehat{\mu}_Y > c_Y \} \\ &= P_{H_0} \{ \sqrt{n}\widehat{\mu}_X > c_X \} + P_{H_0} \{ \sqrt{m}\widehat{\mu}_Y > c_Y \} - P_{H_0} \{ \sqrt{n}\widehat{\mu}_X > c_X \} P_{H_0} \{ \sqrt{m}\widehat{\mu}_Y > c_Y \} \\ &= \alpha_X + \alpha_Y - \alpha_X \alpha_Y, \end{aligned} \quad (1)$$

where $\alpha_X = P_{H_0} \{ \sqrt{n}\widehat{\mu}_X > c_X \} = \Phi(c_X)$ and $\alpha_Y = P_{H_0} \{ \sqrt{m}\widehat{\mu}_Y > c_Y \} = \Phi(c_Y)$, representing the type I error rates for testing the individual group-specific null hypothesis, $H_{0X}: \mu_X = 0$ and $H_{0Y}: \mu_Y = 0$, respectively, when the same critical values are used. Throughout, ϕ and Φ are the standard normal density and distribution functions, respectively, and $\Phi = 1 - \Phi$. Thus, for given α_X and α_Y satisfying Eq. (1), the critical values c_X and c_Y are given by $c_X = \Phi^{-1}(1 - \alpha_X)$ and $c_Y = \Phi^{-1}(1 - \alpha_Y)$.

If there is no preference given to one subpopulation over the other in controlling the type I error rates, we may set $\alpha_X = \alpha_Y$. For generality, let $\alpha_X = \omega \alpha_Y$ ($\omega > 0$), then solving Eq. (1) and noting that $0 < \alpha_X, \alpha_Y < 1$, we have

$$\alpha_X = \frac{1 + \omega - \sqrt{(1 + \omega)^2 - 4\omega\alpha}}{2}, \quad \alpha_Y = \frac{1 + \omega - \sqrt{(1 + \omega)^2 - 4\omega\alpha}}{2\omega}.$$

In particular, if $\omega = 1$ then $\alpha_X = \alpha_Y = 1 - \sqrt{1 - \alpha}$. Thus, $c_X = c_Y = \Phi^{-1}(\sqrt{1 - \alpha})$.

With α_X and α_Y specified, the power of the test at μ_X and μ_Y is given by

$$\beta(\mu_X, \mu_Y; n, m) = \bar{\Phi}(c_X - \sqrt{n}\mu_X) + \bar{\Phi}(c_Y - \sqrt{m}\mu_Y) - \bar{\Phi}(c_X - \sqrt{n}\mu_X)\bar{\Phi}(c_Y - \sqrt{m}\mu_Y). \quad (2)$$

The sample sizes m and n are then determined by (2) so that the power requirement is met.

Two-stage sample-enrichment approach

The two-stage sample-enrichment approach applies the treatments to a sample of patients from one subpopulation first, and if data show promising treatment effect, i.e. the estimated treatment improvement exceeds certain prespecified threshold, then the treatments will be extended to all subpopulations, including the one already in the trial. Otherwise, the trial will stop for futility. To statistically formalize the concept, let $X_1, \dots, X_{n_1}$ be a sample from

$N(\mu_X, 1)$, where the sample size n_1 is a prespecified integer. If $\sqrt{n_1}\hat{\mu}_{1X} = \sum_{i=1}^{n_1} X_i / \sqrt{n_1} \leq c$, then the trial will be terminated and the experimental drug will be claimed to be of no treatment difference from the standard, under the assumption that $\mu_X \geq \mu_Y$. Otherwise, we will continue to observe $X_{n_1+1}, \dots, X_n$ from $N(\mu_X, 1)$ and $Y_1, \dots, Y_m$ from $N(\mu_Y, 1)$, and reject H_0 if $\sqrt{n}\hat{\mu}_X > c_X$ or $\sqrt{m}\hat{\mu}_Y > c_Y$, where the estimated means are derived using all available data.

Note that the null hypothesis can be rejected only if the test passes the threshold in the first stage. The type I error of such a two-stage test is thus given by

$$\begin{aligned} \alpha &= P_{H_0} \{ \sqrt{n_1}\hat{\mu}_{1X} > c, \text{ and } \{ \sqrt{n}\hat{\mu}_X > c_X, \text{ or } \sqrt{m}\hat{\mu}_Y > c_Y \} \} \\ &= P_{H_0} \{ \sqrt{n_1}\hat{\mu}_{1X} > c, \sqrt{n}\hat{\mu}_X > c_X \} + P_{H_0} \{ \sqrt{n_1}\hat{\mu}_{1X} > c, \sqrt{m}\hat{\mu}_Y > c_Y \} - P_{H_0} \{ \sqrt{n_1}\hat{\mu}_{1X} > c, \sqrt{n}\hat{\mu}_X > c_X, \sqrt{m}\hat{\mu}_Y > c_Y \} \\ &= P_{H_0} \{ \sqrt{n_1}\hat{\mu}_{1X} > c, \sqrt{n}\hat{\mu}_X > c_X \} + P_{H_0} \{ \sqrt{n_1}\hat{\mu}_{1X} > c \} P_{H_0} \{ \sqrt{m}\hat{\mu}_Y > c_Y \} - P_{H_0} \{ \sqrt{n_1}\hat{\mu}_{1X} > c, \sqrt{n}\hat{\mu}_X > c_X \} P_{H_0} \{ \sqrt{m}\hat{\mu}_Y > c_Y \} \\ &= \alpha_X + \alpha_Y - \alpha_X \alpha_Y / \bar{\Phi}(c), \end{aligned} \quad (3)$$

where $\alpha_X = P_{H_0} \{ \sqrt{n_1}\hat{\mu}_{1X} > c, \sqrt{n}\hat{\mu}_X > c_X \}$ and $\alpha_Y = P_{H_0} \{ \sqrt{n_1}\hat{\mu}_{1X} > c, \sqrt{m}\hat{\mu}_Y > c_Y \}$.

The two quantities, α_X and α_Y , can similarly be viewed as the corresponding type I error rate that under the sample-enrichment design the group-specific null hypothesis is rejected when the null hypothesis is true. If we still set $\alpha_X = \omega \alpha_Y$ ($\omega > 0$), then, noting that $\max\{\alpha, \alpha_X, \alpha_Y\} < \bar{\Phi}(c)$, we have

$$\alpha_X = \frac{(1+\omega)\bar{\Phi}(c) - \sqrt{(1+\omega)^2\bar{\Phi}^2(c) - 4\omega\alpha\bar{\Phi}(c)}}{2},$$

and

$$\alpha_Y = \frac{(1+\omega)\bar{\Phi}(c) - \sqrt{(1+\omega)^2\bar{\Phi}^2(c) - 4\omega\alpha\bar{\Phi}(c)}}{2\omega},$$

which, when $\omega = 1$, reduce to $\alpha_X = \alpha_Y = \bar{\Phi}(c) - \sqrt{\bar{\Phi}^2(c) - \alpha\bar{\Phi}(c)}$.

The power of the test for rejecting either null hypothesis is given by

$$\beta(\mu_X, \mu_Y; n, m) = \Phi(c_Y - \sqrt{m}\mu_Y) \int_{c - \sqrt{n_1}\mu_X}^{\infty} \bar{\Phi}\left(\frac{c_X - \sqrt{n}\mu_X - \sqrt{\kappa}w}{\sqrt{1 - \kappa}}\right) \varphi(w) dw + \bar{\Phi}(c - \sqrt{\kappa n}\mu_X) \bar{\Phi}(c_Y - \sqrt{m}\mu_Y), \quad (4)$$

where $\kappa = n_1/n$ is the ratio of the first-stage sample size to the total sample size for the X -population.

Note that (4) reduces to (3) when $\mu_X = \mu_Y = 0$. The critical values, c_X and c_Y , and the sample sizes, n and m , need to be chosen to satisfy error requirements. Once the threshold value c for stopping after the first stage is determined, then critical values c_X and c_Y can be obtained from the following equations

$$\alpha_Y = \bar{\Phi}(c) \bar{\Phi}(c_Y), \alpha_X = \int_c^{\infty} \bar{\Phi}\left(\frac{c_X - \sqrt{\kappa}w}{\sqrt{1 - \kappa}}\right) \varphi(w) dw. \quad (5)$$

Thus $c_Y = \Phi^{-1}(1 - \alpha_Y / \bar{\Phi}(c))$. The value of c_X , however, can only be computed numerically from the second equation in (5).

(A brief derivation of (4) and (5) is provided in the Appendix.)

The size n_1 should be reasonably chosen so that a convincing decision can be made at the end of the first stage. Note that stopping at the first stage requires that $\widehat{\mu}_{1X} \leq c / \sqrt{n_1}$. Thus for a given n_1 , the threshold value c can be chosen so that $c / \sqrt{n_1}$ is a proportion of the smallest meaningful treatment improvement the investigators expect. For example, $c / \sqrt{n_1}$ can be set to be the smallest meaningful treatment difference, or half of the difference, etc.

Alternatively, c can be determined by controlling at certain level $\gamma (> \alpha)$ the error probability,

$P_{H_0}\{\sqrt{n_1}\widehat{\mu}_{1X} > c\} = \bar{\Phi}(c)$, of expanding the trial to include the Y -population when it should not be so. This gives $c = \Phi^{-1}(1 - \gamma)$. If n_1 is not pre-specified, then n_1 and c can be jointly determined by controlling additionally the error probability,

$P_{\mu_X}\{\sqrt{n_1}\widehat{\mu}_{1X} \leq c\} = \Phi(c - \sqrt{n_1}\mu_X)$, of stopping the trial at the first stage when the true treatment improvement is μ_X for the X -population.

Testing the overall treatment effects

A clinical trial is more often based on testing hypotheses concerning the overall treatment effect, usually in the form of a weighted average of the group-specific treatment effects, i.e. $\mu = \pi\mu_X + (1 - \pi)\mu_Y$, where $0 < \pi < 1$. In general the weight π can be chosen as the sample proportion or the prevalence of the subpopulation X . The hypotheses being tested are then

$$H_0: \mu = 0 \quad \text{versus} \quad H_1: \mu > 0.$$

With fixed sample sizes m and n that satisfy the error requirements, we reject the null hypothesis with significance level α if

$$\frac{\hat{\mu}}{\sqrt{\frac{\pi^2}{n} + \frac{(1-\pi)^2}{m}}} > \Phi^{-1}(1 - \alpha),$$

where $\hat{\mu} = \pi\hat{\mu}_X + (1 - \pi)\hat{\mu}_Y$. The power of the test as a function of μ_X and μ_Y is given by

$$\beta(\mu_X, \mu_Y; n, m) = \bar{\Phi} \left(Z_{1-\alpha} - \frac{\mu}{\sqrt{\frac{\pi^2}{n} + \frac{(1-\pi)^2}{m}}} \right). \quad (6)$$

It is noticed that the power of the fixed-size test for a weighted mean depends on μ_X and μ_Y through their weighted average μ . Thus for a fixed value of μ the test has the same power on the line $\{(\mu_X, \mu_Y): \pi\mu_X + (1 - \pi)\mu_Y = \mu\}$ in the parameter space.

Now consider testing the hypotheses with the two-stage sample-enrichment strategy. Again the trial will be terminated if $\sqrt{n_1}\hat{\mu}_{1X} \leq c$, and the experimental drug will be claimed of no treatment improvement. If $\sqrt{n_1}\hat{\mu}_{1X} > c$, we then expand the trial to include both populations and observe $X_{n_1+1}, \dots, X_n$ from $N(\mu_X, 1)$ and $Y_1, \dots, Y_m$ from $N(\mu_Y, 1)$, and reject H_0 if $\hat{\mu} / \sqrt{\frac{\pi^2}{n} + \frac{(1-\pi)^2}{m}} > C$.

Similar to the derivation of (4) and (5) the power function of the test is found to be

$$\beta(\mu_X, \mu_Y; n, m) = \int_{c - \sqrt{n_1}\mu_X}^{\infty} \varphi(w) \bar{\Phi} \left(\frac{C \sqrt{\frac{\pi^2}{n} + \frac{(1-\pi)^2}{m}} - \mu - \pi w \sqrt{\frac{\pi^2}{n}}}{\sqrt{\frac{(1-\pi)^2}{n} + \frac{(1-\pi)^2}{m}}} \right) dw, \quad (7)$$

and thus the type I error of the test is given by

$$\alpha(\mu_X; n, m) = \int_{c - \sqrt{n_1}\mu_X}^{\infty} \varphi(w) \bar{\Phi} \left(\frac{C \sqrt{\frac{\pi^2}{n} + \frac{(1-\pi)^2}{m}} - \pi w \sqrt{\frac{\pi^2}{n}}}{\sqrt{\frac{(1-\pi)^2}{n} + \frac{(1-\pi)^2}{m}}} \right) dw. \quad (8)$$

Unlike the fixed-size test whose power only depends on μ , and whose type I error is parameter-free, we notice that the power of the two-stage sample-enrichment test depends on both μ and μ_X , and the type I error involves the one of the two mean parameters. This is not surprising since rejection of the null hypothesis relies on the estimate of μ_X from the first stage data. In general the sample sizes m and n are taken to satisfy $\sup_{\mu_X} \alpha(\mu_X; n, m) \leq \alpha$, the nominal significance level.

One complication with a weighted mean as the (overall) treatment effect is its interpretation if μ_X and μ_Y are in opposite directions. For instance equal weights ($\pi = 1/2$) and $\mu_X = -\mu_Y > 0$ results in $\mu = 0$, and it is then claimed of no treatment effects. However, when restricted to group X, a significant drug benefit can be claimed. If we assume that the treatment difference in the two populations falls in the same direction, i.e. $\mu_X \mu_Y \geq 0$, then $\mu = 0$ if and

only if $\mu_X = \mu_Y = 0$. In this case, the type I error of the two-stage enrichment design becomes parameter-free.

Numerical comparison of study designs

In the following we compare the fixed-size test and the sample-enrichment test, assuming both tests are at the same level of significance and have the same power under some alternative. We demonstrate that under the null hypothesis of no treatment improvement, the sample-enrichment test requires on average smaller sample sizes than the fixed-size test and thus is more cost-effective. To this end, we use the subscripts f and e to represent the fixed-size and sample-enrichment test, respectively. Thus the sample size for the fixed-size test is $N_f = n_f + m_f = n_f/\lambda$, and the (maximum) sample size for the sample-enrichment test is $N_{max} = n_e + m_e = n_e/\lambda$, where $\lambda = n_f/(n_f + m_f) = n_e/(n_e + m_e)$. The average sample size for the enrichment test at μ_X and μ_Y is

$$N_e = n_1 P\{\sqrt{n_1}\widehat{\mu}_{1X} \leq c\} + (n_e + m_e - n_1) P\{\sqrt{n_1}\widehat{\mu}_{1X} > c\} \\ = n_1 \Phi(c - \sqrt{n_1}\mu_X) + (n_e/\lambda - n_1) \Phi(c - \sqrt{n_1}\mu_X). \quad (9)$$

We consider testing H_0 with prespecified power at the alternative $\mu_X = 0.3$ and $\mu_Y = 0.2$. The sample sizes are obtained for each set of values with $\alpha = 0.025, 0.05$, power $1 - \beta = 0.80, 0.90$, error allocation ratio $\omega = \alpha_X/\alpha_Y = 0.5, 1.0, 2.0$, sample size ratio $\lambda = n_f/(n_f + m_f) = n_e/(n_e + m_e) = 0.4, 0.6, 0.8$, first-stage size ratio $\kappa = n_1/n_e = 0.5, 0.7$, and threshold $\gamma = 0.2, 0.3$. For the fixed-size test, the type I errors α_X and α_Y are obtained from Eq. (1) with specified values of ω and α . Subsequently the critical values c_X and c_Y are given by $c_X = \Phi^{-1}(1 - \alpha_X)$ and $c_Y = \Phi^{-1}(1 - \alpha_Y)$. The sample sizes n_f and m_f are then computed from Eq. (2) with specified values of $1 - \beta$ and λ . For the sample-enrichment test, the threshold c is given by $\Phi^{-1}(1 - \gamma)$ for a specified value of γ . The type I errors α_X and α_Y are then obtained from Eq. (3) with specified value of ω and α . The critical values c_X and c_Y are subsequently computed from Eq. (5). Note that the sample size n_1 for the first stage is given by κn_e . The sample sizes n_e and m_e are then computed from Eq. (4) with specified values of β , κ and λ . The average sample size is the computed from Eq. (9). For various settings, Table 1 presents the sample sizes and the average sample sizes of the enrichment tests for $\alpha = 0.05$ and $\beta = 0.9$. Clearly the average sample sizes are much smaller than the fixed sample sizes. Simulation results (data not shown) from others settings, and from testing the overall treatment improvement reveal the same conclusion.

The CPEP trial revisited

Preeclampsia, a hypertensive disorder that occurs only in women during pregnancy, may affect almost any organ system in the body, causing eclampsia, strokes, pulmonary edema, renal failure, liver dysfunction, liver rupture, coagulopathy, hemolysis, placental abruption, and other complications. Due to its importance in public health, effectively preventing preeclampsia has become a major focus in obstetrical research.

Conducted at five university medical centers by the Division of Epidemiology, Statistics and Prevention Research of the Eunice Kennedy Shriver National Institute of Child Health and Human Development from 1992 to 1995, the CPEP (Calcium to Prevent Preeclampsia) trial aimed at providing a thorough evaluation of the effects of calcium supplementation for the prevention of preeclampsia in the United States.

The rationale, design and methods of the trial were reported in Levine et al.[7]. With equal allocation, the minimum total sample size was estimated to be 4500 women, taking into

consideration lost to follow-up or noncompliance. This sample size was sufficient to obtain 85% power to detect a 50% reduction in the true risk of preeclampsia among women perfectly compliant with calcium supplementation from incidence levels of at least 4% in the placebo group.

The main finding that calcium supplementation did not prevent preeclampsia in healthy nulliparous women was reported in Levine et al.[8]. At the end of the trial, a total of 4589 healthy women who were 13 to 21 weeks pregnant were equally randomized to receive daily treatment with either 2g of elemental calcium ($n = 2295$) or placebo ($n = 2294$) for the remainder of their pregnancies. Excluding 296 women from the analysis due to missing preeclampsia status, the study yielded 2143 women in the calcium group with 158 cases (7.37%) of preeclampsia, as compared to 168 cases (7.81%) out of 2150 women in the placebo group, a merely 5% reduction in preeclampsia incidence rate. It thus concluded that calcium supplementation did not significantly reduce the incidence of preeclampsia (P-value=0.6, two-sided χ^2 test).

Before entering the study, women may have different profile in accessing calcium supplementation, and it is believed that, if calcium supplementation indeed can help prevent preeclampsia, then the reduction in preeclampsia incidence is expected to be higher among the healthy nulliparae who normally have low calcium intake than among those who normally have high calcium intake. Indeed, this is well supported by data from the CPEP trial. Table 2 summarizes the distribution of enrolled women by treatment, calcium intake, and preeclampsia status, where the calcium intake status is based on a median cutpoint of 975mg in the past 24 hours prior to enrollment. (Women with missing status were excluded from the table.)

Among the low calcium intake group, calcium supplementation reduced preeclampsia incidence by nearly 10%, a much larger reduction as compared to only 1% among women who reported high calcium intake 24 hours prior to enrollment.

This information, however, was not considered at the time the CPEP trial was designed. We demonstrate below that, had this information been used to design the trial as a two-stage sample-enrichment trial, the trial could have been terminated with much fewer enrollments, and thus could have saved substantial resources. Suppose at the first stage, n_1 women who had reported low calcium intake in the past 24 hours were enrolled into the trial and then subsequently equally randomized to receive either calcium supplementation or placebo. These women were then followed up and their preeclampsia status were determined. If the observed percent reduction in preeclampsia incidence rate among women receiving calcium supplementation is larger than 25% (half of the expected 50% reduction for the study) from the incidence rate among women receiving placebo, then the trial would be expanded to women who reported high calcium intake 24 hours prior to enrollment. Otherwise the study will be terminated without further enrollment and calcium supplementation will be claimed to be of no help in reducing preeclampsia incidence. For various choices of n_1 , the percent reduction in preeclampsia incidence rate of calcium supplementation from placebo is presented in Table 3. Inclusion of women into the analysis is based on their enrollment date. For example, if $n_1 = 1000$, then we select the first 1000 women enrolled into the study who reported low calcium intake 24 hours prior to the enrollment.

It is noticed that for every choice of n_1 , the observed percent reduction in preeclampsia incidence rate by calcium supplementation is substantially less than the 25% threshold, and much smaller than the expected 50% reduction. The percent reduction is expected to be even smaller among women who reported high calcium intake in the past 24 hours. Therefore, if

the two-stage sample-enrichment design were used, the trial would be terminated after the first stage without further enrollment.

Discussion

When heterogeneity exists among target patients in a large scale comparative clinical trial, certain group of patients often show more improvements from the treatment than other groups of patients. Taking this information into consideration we propose in the present paper a sample-enrichment design strategy to conduct the trial. The proposed design is more cost-effective if the treatment are of no difference.

Feasibility of the proposed enrichment design relies on the knowledge prior to the start of the trial that certain patients can benefit more than others from the treatment. While for many trials such knowledge may not be available, for many others it could be well validated and supported, especially with the rapid growing research in disease-associated genes and biomarkers.

The proposed sample-enrichment strategy can be viewed as an adaptive design in the sense that the inclusion criteria of patients are modified after the first stage to expand targeted patients population. In addition to cost-effectiveness, the proposed design allows investigators to use the data collected at the first stage to investigate other assumptions made at the beginning of the trial.

Acknowledgments

Research of AL, CL and KFY is supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health. Research of QL is partially supported by the Knowledge Innovation Program of the Chinese Academy of Sciences, No. 30465W0 and 30475V0. The opinions expressed in the article are not necessarily of the National Institutes of Health, nor the Food and Drug Administration. The authors thank Richard Levine and Cong Qian for their help with the CPEP trial data.

References

1. Gordon AN, Tonda M, Sun S, Rackoff W. Doxil study 3049 investigators. Long-term survival advantage for women treated with pegylated liposomal doxorubicin compared with topotecan in a phase 3 randomized study of recurrent and refractory epithelial ovarian cancer Gynecologic. *Oncology* 2004;95:1–8.
2. Song Y, Chi GYH. A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine* 2007;26:3535–3549. [PubMed: 17266164]
3. Follmann D. Adaptively changing subgroup proportions in clinical trials. *Statistica Sinica* 1997;7:1085–1102.
4. Cushman WC, Cutler JA, Bingham SF, Harford T, Hanna E, Dubbert P, Collins JF, Dufour M, Follman D, Allender PS. Prevention and treatment of hypertension study (PATHS): Rationale and design. *American Journal of Hypertension* 1994;7:814–823. [PubMed: 7811440]
5. Shih WJ, Quan H, Li G. Two-stage adaptive strategy for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine* 2004;23:2781–2798. [PubMed: 15344186]
6. Lachin JM. A review of methods for futility stopping based on conditional power. *Statistics in Medicine* 2005;24:2747–2764. [PubMed: 16134130]
7. Levine RJ, Esterlitz JR, Raymond EG, DerSimonian R, Hauth JC, Ben Curet L, Sibai BM, Catalano PM, Morris CD, Clemens JD, Ewell MG, Friedman SA, Goldenberg RL, Jacobson SL, Joffe GM. Trial of calcium for preeclampsia prevention (CPEP): Rationale, design, and methods. *Controlled Clinical Trials* 1996;17:442–469. [PubMed: 8932976]
8. Levine RJ, Hauth JC, Curet LB, Sibai BM, Catalano PM, Morris CD, DerSimonian R, Esterlitz JR, Raymond EG, Bild DE, Clemens JD, Cutler JA. Trial of calcium to prevent preeclampsia. *New England Journal of Medicine* 1997;337:69–76. [PubMed: 9211675]

Appendix: Proof of (4) and (5)

Similar to (3), the power function is

$$\begin{aligned}\beta(\mu_x, \mu_y; n, m) &= P\{\sqrt{n_1}\widehat{\mu}_{1X} > c, \text{ and } \{\sqrt{n}\widehat{\mu}_X > c_X, \text{ or } \sqrt{m}\widehat{\mu}_Y > c_Y\}\} \\ &= P\{\sqrt{n_1}\widehat{\mu}_{1X} > c, \sqrt{n}\widehat{\mu}_X > c_X\} \\ &\quad + P\{\sqrt{n_1}\widehat{\mu}_{1X} > c\}P\{\sqrt{m}\widehat{\mu}_Y > c_Y\} \\ &\quad - P\{\sqrt{n_1}\widehat{\mu}_{1X} > c, \sqrt{n}\widehat{\mu}_X > c_X\}P_{H_0}\{\sqrt{m}\widehat{\mu}_Y > c_Y\}.\end{aligned}$$

Let W_1 and W_2 be jointly normal with $W_i \sim N(\mu_i, 1)$ and correlation coefficient ρ . Then $W_2 - \rho W_1$ and W_1 are mutually independent, and thus $W_2|W_1 = w_1 \sim N(\mu_2 - \rho(\mu_1 - w_1), 1 - \rho^2)$.

$$\begin{aligned}P(W_1 > w_1, \\ W_2 > w_2) &= E\{P(W_1 > w_1, W_2 > w_2 | W_1)\} \\ &= E\left\{I(W_1 > w_1)\overline{\Phi}\left(\frac{w_2 - \mu_2 - \rho(W_1 - \mu_1)}{\sqrt{1 - \rho^2}}\right)\right\} \\ &= \int_{w_1 - \mu_1}^{\infty} \varphi(w)\overline{\Phi}\left(\frac{w_2 - \mu_2 - \rho w}{\sqrt{1 - \rho^2}}\right)dw.\end{aligned}$$

Now note that $\sqrt{n_1}\widehat{\mu}_{1X}$ and $\sqrt{n}\widehat{\mu}_X$ are jointly normal with correlation coefficient $\sqrt{\kappa}$, and means $\sqrt{n_1}\mu_X$ and $\sqrt{n}\mu_X$ respectively, and unit variances. Thus

$$P(\sqrt{n_1}\widehat{\mu}_{1X} > c, \sqrt{n}\widehat{\mu}_X > c_X) = \int_{c - \sqrt{n_1}\mu_X}^{\infty} \varphi(w)\overline{\Phi}\left(\frac{c_X - \sqrt{n}\mu_X - \sqrt{\kappa}w}{\sqrt{1 - \kappa}}\right)dw.$$

Equation (4) and (5) thus follow.

Table 1

Average sample size of sample-enrichment designs

(Type I error $\alpha = 0.05$, power $\beta = 0.9$)												
ω	λ	γ	κ	n_f	m_f	N_f	n_e	m_e	$N_e(H_0)$			
0.5	0.4	0.2	0.5	81	121	202	103	155	82			
0.5	0.4	0.2	0.7	81	121	202	80	121	74			
0.5	0.4	0.3	0.5	81	121	202	86	129	82			
0.5	0.4	0.3	0.7	81	121	202	74	111	76			
0.5	0.6	0.2	0.5	104	69	173	111	74	70			
0.5	0.6	0.2	0.7	104	69	173	92	62	69			
0.5	0.6	0.3	0.5	104	69	173	101	67	71			
0.5	0.6	0.3	0.7	104	69	173	91	61	71			
0.5	0.8	0.2	0.5	118	29	147	120	30	66			
0.5	0.8	0.2	0.7	118	29	147	105	26	70			
0.5	0.8	0.3	0.5	118	29	147	113	28	65			
0.5	0.8	0.3	0.7	118	29	147	107	27	70			
1.0	0.4	0.2	0.5	78	118	196	103	155	82			
1.0	0.4	0.2	0.7	78	118	196	80	121	74			
1.0	0.4	0.3	0.5	78	118	196	86	129	82			
1.0	0.4	0.3	0.7	78	118	196	74	111	76			
1.0	0.6	0.2	0.5	97	65	162	109	73	69			
1.0	0.6	0.2	0.7	97	65	162	91	60	68			
1.0	0.6	0.3	0.5	97	65	162	98	65	68			
1.0	0.6	0.3	0.7	97	65	162	89	59	69			
1.0	0.8	0.2	0.5	108	27	135	116	29	64			
1.0	0.8	0.2	0.7	108	27	135	100	25	67			
1.0	0.8	0.3	0.5	108	27	135	108	27	62			
1.0	0.8	0.3	0.7	108	27	135	101	25	66			
2.0	0.4	0.2	0.5	78	117	195	103	155	82			
2.0	0.4	0.2	0.7	78	117	195	81	122	75			
2.0	0.4	0.3	0.5	78	117	195	86	129	82			

(Type I error $\alpha = 0.05$, power $\beta = 0.9$)												
ω	λ	γ	κ	n_f	m_f	N_f	n_e	m_e	$N_e(H_0)$			
2.0	0.4	0.3	0.7	78	117	195	75	112	77			
2.0	0.6	0.2	0.5	94	62	156	109	73	69			
2.0	0.6	0.2	0.7	94	62	156	90	60	68			
2.0	0.6	0.3	0.5	94	62	156	97	64	68			
2.0	0.6	0.3	0.7	94	62	156	88	58	68			
2.0	0.8	0.2	0.5	102	25	127	114	28	63			
2.0	0.8	0.2	0.7	102	25	127	98	24	66			
2.0	0.8	0.3	0.5	102	25	127	104	26	60			
2.0	0.8	0.3	0.7	102	25	127	97	24	63			

Preeclampsia incidence by treatment and prior calcium intake

	Low intake			High intake		
	PE-free	PE	% incidence	PE-free	PE	% incidence
Calcium	1052	86	7.56	933	72	7.16
Placebo	1009	92	8.36	973	76	7.24

Table 3
% reduction in preeclampsia incidence among women who reported higher prior calcium intake in the CPEP trial

n_1						
	1000	1100	1200	1300	1400	1500
% reduction observed	12.5	9.42	5.82	11	13.69	6.17