ELSEVIER

Trials within Trials: Confirmatory Subgroup Analyses in Controlled Clinical Experiments

Lemuel A. Moyé, MD, PhD, and Anita Deswal, MD

University of Texas School of Public Health, Houston, Texas (L.A.M.), and Veterans Administration Medical Center and Baylor College of Medicine, Houston, Texas (A.D.)

ABSTRACT: Subgroup analyses remain a popular and necessary component of controlled clinical trials. However, lack of prospective specification, inadequate sample size, inability to maintain power, and the cumulative effect of sampling error can complicate their interpretation. This article demonstrates that clinical trial design tools that would allow the medical community to draw confirmatory and not just exploratory conclusions from specific subgroup evaluations are available to methodologists. Distinct from the use of a treatment by subgroup interaction term, this methodology provides an evaluation of the effect of an intervention within a particular subgroup stratum prospectively declared to be of interest to the investigators. The necessary prespecification of stratum-specific type I error rates, when combined with (1) a stratum-specific event rate in the subgroup, (2) a stratum-specific primary endpoint, (3) a stratum-specific endpoint precision, and/or (4) a stratum-specific efficacy, satisfies the requirements for a subgroup stratum's "stand-alone" interpretation at the trial's conclusion. *Control Clin Trials* 2001;22:605–619 © Elsevier Science Inc. 2001

KEY WORDS: *Subgroups, prospective design, alpha allocation*

INTRODUCTION

There may be no better maxim for guiding the interpretation of subgroup analyses as currently executed in clinical trials than "Look, but don't touch." Research investigators are trained to be thorough in their evaluation of the intervention's impact in a controlled clinical trial. This has naturally evolved into the interpretation of this impact not just in the entire cohort, but in subcohorts as well, searching for heterogeneity of effect. Commonly evaluated subgroup analyses are the evaluation of the controlled clinical trial's intervention in patients with certain demographic characteristics (e.g., women) or in patients with certain biochemical characteristics (e.g., patients with low-density lipoprotein [LDL] cholesterol levels < 125 mg/dL). Some clinical trials report these results both in the article announcing the trial's overall results [1–4] and separately [5–7]. Such subgroup analyses can provide leading information about an

Address reprint requests to: Lemuel A. Moyé, MD, PhD, University of Texas School of Public Health, RAS Building E815, 1200 Herman Pressler, Houston, TX 77030 (lmoye@sph.uth.tmc.edu).

Received March 5, 2001; accepted August 3, 2001.

Control Clin Trials 22:605–619 (2001)

655 Avenue of the Americas, New York, NY 10010

0197-2456/01/$–see front matter
PII S0197-2456(01)00180-5

unanticipated benefit or hazard of an intervention being evaluated in the clinical trial.

However useful and provocative these results can be, it is well established that subgroup analyses are often misleading [8–11]. Although the medical community often rejects such findings, attributing their effects to sampling error, it continues to be tantalized by spectacular subgroup findings from clinical trials, as seen most recently [12, 13]. Assmann et al. [14] have demonstrated how commonly subgroup analysis is misused. Articles in both the clinical [15] and methodologic literature point out that accepting subgroup analyses as confirmatory, independent of the findings of the trial for the entire cohort, is hazardous.

One of the complications of subgroup analyses is that these evaluations focus on heterogeneity of the intervention effect among the subgroup strata. Although this is often a correct approach [16, 17], there are nevertheless circumstances where the presence of this heterogeneity is not the question. In these cases, issue lies in demonstrating efficacy within a single subgroup stratum (e.g., patients randomized from a particular country in a multinational study, or the effect of the intervention in patients whose prognosis is particularly grim).

The purpose of this article is to identify those circumstances in which the estimation of therapy effect size in a subgroup stratum will be confirmatory rather than exploratory while adhering to the tenets of Yusuf et al. [18]. The rationale section will provide the justification for the use of prospective devices to improve the protection of the subgroup's results from sampling error. This will be followed by the delineation of five strategies to strengthen the identification of a stratum-specific efficacy.

RATIONALE

We will assume throughout this article that a randomized, controlled clinical trial to study an intervention's effectiveness is the vehicle in which the subgroup will be analyzed. We will also assume that each subgroup stratum in which efficacy is to be examined has been announced prospectively and that the subgroup is a proper subgroup [18]. We will in addition assume that the prospectively planned clinical trial is executed concordantly (i.e., the experiment is executed according to its protocol) [19]. Thus, in this environment, the estimates the trial provides of the effectiveness of the intervention are trustworthy and need only have appropriately low levels of type I and type II error to produce a confirmatory evaluation of the intervention's effect in the subgroup.

We propose and illustrate a detailed evaluation of the design for the subgroup analysis under consideration during the trial's design phase. This prospective specification can go further than (1) the announcement that a particular subgroup evaluation is to take place at the trial's end and (2) provide the inclusion/exclusion criteria for the subgroup. In addition to these statements, the investigators have the authority to specify (1) type I and type II error levels for the subgroup stratum of interest, (2) control group event rates in the subgroup stratum (or in the case of continuous outcome, the standard deviation of the endpoint measurement), and (3) the level of therapeutic effectiveness (efficacy of the intervention) in the stratum. The evaluation of each of

Table 1 Prospective Allocation of Alpha

Analysis	Prospective Alpha Allocation
Total cohort	α_1
Subgroup analysis	α_2

these quantities is a necessity when designing the trial to consider the effect of the intervention in the entire cohort. We advocate their explicit consideration when designing a subgroup stratum-specific analysis. Furthermore, the characteristics of a subgroup stratum can lead to conclusions about the values of these quantities that are different than those conclusions used to design the trial for the entire cohort. Thus at the initiation of the trial, two confirmatory plans are deployed. The first is the plan for the evaluation of the effect of therapy for the entire cohort of the controlled clinical trial. The second is the confirmatory plan for the evolution of the intervention within the subgroup stratum. Each plan is predicated on its own defensible assumptions for statistical errors, endpoint choice, endpoint event rate, and intervention efficacy.

Strategy 1: Apply Different Type I Error Levels Prospectively

Recent work [20, 21] has demonstrated the advantages of the allocation of type I error across different prospectively stated hypotheses, not in equal fractions but in differently sized, hypothesis-dependent portions. This same procedure can be used to carry out a prospectively defined subgroup analysis, prospectively choosing the type I error for the evaluation. For example, consider a controlled clinical trial in which an intervention is being evaluated for its effect on the prospectively specified endpoint of the study. The investigators have an a priori interest in evaluating not just the effect of the intervention in the entire cohort but, in addition, are interested in evaluating the effect of the intervention in a proper subgroup stratum. To make confirmatory statements for each of the two analyses while keeping the overall type I error level low, alpha levels are prospectively selected for each of the two analyses (Table 1).

The probability of at least one type I error is $1 - (1 - \alpha_1)(1 - \alpha_2)$ [17],[1] and the investigators plan to invest no more than

$$T_1(i) = 1 - \prod_{j=1}^{2} [1 - \alpha_j],$$

where $T_1(i)$ is the upper bound on the type I error acceptable to the investigators at the initiation of the study.[2] Assume the experiment is conducted concordantly (i.e., according to its protocol). At its conclusion, two p-values are produced: one for the total cohort analysis, p_1, and the second for the subgroup analysis, p_2 (Table 2).

[1]This assumes that the hypothesis tests are independent [22]. Dependence among the tests, when explicitly portrayed, can lead to additional alpha error savings.

[2]Other evaluations are possible of course, including nonprospectively identified data analyses, but these are only exploratory, and results produced from them cannot be extended to the population.

Table 2 Comparison of Prospective and Observed Type I Error

Analysis	Prospective Allocation	Observed p-Value
Total cohort	α_1	p_1
Subgroup analysis	α_2	p_2

Since the analyses were prospectively designed and the experiment was conducted according to its protocol, the formula by which type I error is accumulated over these prospectively determined hypothesis tests is the upper bound of the type I error at the termination of the study, $T_1(t)$, where

$$T_1(t) = 1 - \prod_{j=1}^{2} [1 - \min(\alpha_j, p_j)] \tag{1}$$

In Eq. (1), α_j is the prospectively allocated type I error for the jth prospectively determined hypothesis test for the trial and p_j is the p-value for that test that is computed at the end of the trial. It is important to note that α_j is the maximum type I error to be expended for the jth hypothesis test. The experiment is positive (Table 2) if $T_1(t) < T_1(i)$ or

$$1 - \prod_{j=1}^{2} [1 - \min(\alpha_j, p_j)] < 1 - \prod_{j=1}^{2} [1 - \alpha_j] \tag{2}$$

This can occur if $p_1 < \alpha_1$ and/or $p_2 < \alpha_2$.

One outcome of this experiment is especially provocative: the possibility that the trial is positive when the primary analysis for the trial is null[3] but the subgroup analysis is positive. The key to understanding this outcome's interpretation is to note that since a maximum of type I error was prospectively set for the two analyses prospectively, this ceiling limits the alpha accumulation for each analysis. In the circumstance where $p_1 > \alpha_1$ for the primary analysis, the finding for the primary endpoint is null. However, although all of α_1 was expended, no more than α_1 is spent, since only α_1 was prospectively set aside. The primary endpoint finding is null at the α_1 level. This leaves α_2 for the prospectively identified subgroup analysis, and because less than this is expended for the result of this analysis at the trial's conclusion, a significant finding was identified for the subgroup analysis. This is a legitimate conclusion, but only when each of the two analyses are prospectively identified with a priori alpha allocations and the experiment is executed per protocol [19]. T_1 may be approximated by T_1*

$$T_1 \approx T_1^* = \sum_{j=1}^{2} \min(\alpha_j, p_j) \tag{3}$$

Thus, when there are $j = 1, 2, 3, \ldots, k$ prespecified hypothesis tests in a clinical trial, each with alpha α_j assigned during the design phase of the trial, each pro-

[3]If the primary analysis was appropriately powered, the finding of $p > \alpha$, in the presence of adequate power, is described as a null finding. Without adequate power, this finding can only be characterized as uninformative.

ducing p_j as the p-value of the test statistic computed at the conclusion of the concordantly executed trial, then

$$T_1(t) = 1 - \prod_{j=1}^{k} [1 - \min(\alpha_j, p_j)] \approx \sum_{j=1}^{k} \min(\alpha_j, p_j) \tag{4}$$

and, as a simplifying approximation, a trial is positive when

$$\sum_{j=1}^{k} \min(\alpha_j, p_j) < \sum_{j=1}^{k} \alpha_j \tag{5}$$

While the calculations for the remainder of this article will be based on Eq. (1) and Eq. (2), the approximations contained in Eq. (3), Eq. (4), and Eq. (5) are also available.

As an illustration of this concept, consider the following plans for a hypothetical, randomized clinical trial that is being designed to determine the effect of therapy on patients with ischemic heart disease in reducing the risk of total mortality. During the design phase of the trial, the investigator believes the most conclusive result from their study would be to demonstrate a decrease in total mortality for the total cohort. However, she expresses a specific, prospectively stated interest in the identification of the effect of the clinical trial's intervention in the subgroup of patients who have experienced more than one myocardial infarction before they were randomized into the study. Thus, although the investigator's "clinical heart" wishes to choose the multiple infarct subgroup as the primary analysis, her "statistical conscience" demands that a larger cohort (that is also at risk of the endpoint and that stands to benefit from the therapy) be chosen. Thus, while prospectively identifying the effect of the intervention on total mortality in the full cohort as the primary analysis for the trial, she nevertheless wishes to retain the possibility of a positive finding for the multiple myocardial infarct subgroup of interest. She therefore makes the following prospective specifications for type I error (Table 3).

This produces a cumulative type I error of $T_1 = 1 - (1 - 0.045)(1 - 0.005) = 0.0498$. The investigator, in considering the consequences of this design, determines that the trial will be positive if (1) the p-value for the hypothesis test for the total cohort < 0.045 or (2) there is a beneficial effect in the multiple infarction subcohort with a p-value < 0.005.

Strategy 2: Choose a Separate, Subgroup-Specific Endpoint Prospectively

A second strategy in prospectively planned subgroup analysis is to prospectively choose a different endpoint for the subgroup than for the full cohort. With careful consideration, managing two separate endpoints (one for the overall cohort and a second for the subgroup) poses no great conceptual, logis-

Table 3 Prospective Alpha Allocation for a Trial with One Subgroup Analysis

Analysis	Allocated Alpha
Total cohort	0.045
Multiple infarctions subgroup	0.005

Table 4 Sample Size Computations for a Subgroup Specific Endpoint

Cohort	Alpha Level	Event Rate	Efficacy	Power	Sample Size
Total cohort	0.03	0.15	0.20	0.80	4706
Subgroup	0.02	0.30	0.20	0.80	2196

tical, or interpretative difficulty in a clinical trial. While it is certainly true that because of the lack of adequate prespecification and planning, the interpretation of multiple endpoints in controlled clinical trials has been immersed in confusion [12, 13], surprise [23, 24], and contention [25–30], the presence of lucid, detailed prospective statements can lead to clear interpretations. For example, consider a controlled clinical trial whose goal is to assess the effect of a randomly allocated intervention on the occurrence of the primary endpoint, fatal coronary heart disease (CHD) death. The investigators have an interest in the effect of the intervention that is to be the focus of the controlled clinical trial in a particular subgroup. They understand that, as traditionally planned and executed, the subgroup analysis carried out at the trial's conclusion would be exploratory. However, they have an important prospective interest in the findings in this subgroup stratum and wish to elevate the subgroup findings from exploratory to confirmatory.

However, upon first examination of this complex issue during the design phase of the trial, this goal appears unreachable. Since the overall trial scarcely has enough power (80%) for the primary endpoint, there will be inadequate power available in the smaller subgroups. The investigators, retaining their prospective wish to carry out a confirmatory subgroup analysis, therefore take the following prospective actions. First, they clearly state in the protocol during the design phase of the study their interest in examining the effect of the trial's intervention in this subgroup stratum. They then prospectively declare that the endpoint for this subgroup stratum's analysis will be the combined endpoint of fatal and nonfatal CHD death. The trial is designed as follows with the choice of the fatal/nonfatal endpoint component reflected in the event rate of the subgroup (Table 4).

With a cumulative CHD death rate of 0.15, 4706 are required to demonstrate a 20% reduction in the primary endpoint with 80% power. For the prospectively delineated subgroup evaluation, the greater event rate for the combined endpoint reveals that 2196 patients would be required to demonstrate a 20% reduction in the incidence of the combined endpoint with 80% power. The result of this prospective planning is a confirmatory subgroup analysis. Consider the interpretation of this concordantly executed trial if the following results are obtained (Table 5).

Table 5 Prospective Alpha Allocation for a Trial for Subgroup-Specific Endpoint

Prospective Analyses	Allocated Alpha	p-Value at Trial's End
Primary endpoint—CHD death	0.030	0.070
Secondary endpoint 1, subgroup 1—combined endpoint	0.020	0.005

Table 6 Sample Size Computations for Subgroup-Specific Event Rate

Cohort	Alpha Level	Event Rate	Efficacy	Power	Sample Size
Total	0.05	0.20	0.20	0.80	2894
Higher-risk subgroup	0.05	0.30	0.20	0.80	1717

Using Eq. (4) for $k = 2$, the total type I error $T_1(t)$ expended for the trial is $T_1(t) = 1 - (1 - 0.030)(1 - 0.005) = 0.035$, which is less than the total alpha allocated for the study. This trial's null findings for the primary endpoint but positive finding for the subgroup analysis requires that the trial should be considered positive. Note, however, that this concentrated statistical argument must be circumscribed by epidemiologic and clinical reasoning. The findings would only be generalized to the subgroup of the study, not to all members of the entire study cohort. In addition, the findings apply to the prospectively defined secondary endpoint of the subgroup stratum, not to the primary endpoint of the study. Also, the endpoint for the subgroup stratum must be carefully chosen so as to not change the scientific question. Here, the proposed stratum-specific endpoint in a particular subgroup strata is not a change in the scientific question, since the pathophysiology underlying the two endpoints are tightly intertwined. Certainly, however, the use of a combined endpoint will require increased work in collecting accurate information on the occurrence of the nonfatal myocardial infarctions in the subgroup stratum of interest.

Strategy 3: Recognize and Incorporate Greater Morbidity/Mortality in the Subgroup

Another promising strategy for evaluating the effect of an intervention on the experience of a prospectively defined subgroup in a controlled clinical trial begins with the recognition that some subgroup strata are unfortunately known to be at greater risk for the primary endpoint of the study. This greater risk often translates directly to the more common occurrence of that endpoint in the subgroup. This observation can be used to embed a confirmatory analysis into the controlled clinical trial for the prospectively delineated subgroup. As an example, consider the decisions facing a collection of investigators who are interested in determining the effect of antihypertensive therapy on reducing the incidence of total mortality in a cohort of patients. Planning this evaluation prospectively, the investigators compute that a sample size of 2894 patients is required to determine a 20% reduction in the 20% incidence of total mortality with 80% power and a two-sided alpha level of 0.05. However, they wish to be able to make a confirmatory statement about the effect of this therapy in higher-risk patients. Considering the experience of these higher-risk patients, the investigators realize that the cumulative primary endpoint rate in

Table 7 Sample Size Computations for Total Cohort and Higher-Risk Subgroup

Cohort	Alpha Level	Event Rate	Efficacy	Power	Sample Size
Total cohort	0.01	0.20	0.20	0.80	4307
Higher-risk subgroup	0.04	0.30	0.20	0.80	1834

Table 8 Sample Size Computations for End Systolic Volume Study—Study Cohort

Cohort	Alpha Level	Delta	Standard Deviation	Power	Sample Size
Total	0.05	8	25	0.80	307

this subgroup is not 20% but 30%. This suggests that 1717 patients would be required to draw confirmatory conclusions about the effect of therapy in the subgroup (Table 6).

However, there is a difficulty with this plan. With type I error allocated at the 0.05 level for each of the two tests, multiple testing concerns reveal that the probability of at least one type I error commission is $1 - (1 - 0.05)(1 - 0.05) = 0.098$, assuming independence. An alternative computation reveals that if the two-sided type I error expended for the overall cohort is 0.01 and the two-sided type I error of 0.04 allocated for the higher-risk subgroup, then 4307 patients are required for the total cohort evaluation (assuming 80% power to produce a 20% reduction in the total mortality rate) and 1834 patients required for the higher-risk component (Table 7).

This design required two unusual features. The first was the formal incorporation of the greater event rate of the higher-risk patients as a prospectively designed feature in this study. The second was the prospective adjustment in the type I alpha allocated for the two confirmatory hypothesis tests from the customary 0.05 level.

The unusual approach to the design of the study, essentially requiring, *ceteris paribus*, greater strength of evidence for the total cohort than the subgroup, requires some comment. The sample size for the entire cohort has increased from 2897 to 4307, a 49% increase. However, the increase in the number of higher-risk patients required for a confirmatory analysis is much more modest, from 1717 to 1834 or 8%. Since the inclusion of subjects with a worse prognosis in clinical trial may be very difficult, a reasonable alternative to meet the confirmatory analysis requirements may very well be to increase the size of the cohort that is easiest to recruit.

An additional argument supporting the design featured in Table 7 is the relationship between the allocated type I error and the clinical trial event rate. For the entire cohort, the smaller alpha level is associated with the lower event rate, and the higher event rate in the higher-risk patients is associated in the design with the higher type I alpha level. In this scenario, since the higher event rate among the higher-risk patients might very well demand more aggressive treatment action, the investigator is therefore willing to accept weaker evidence of efficacy[4] to decide that a benefit accrues to the higher-risk population.

Strategy 4: Improve the Precision of the Endpoint Measurement in the Subgroup

In this fourth circumstance, the prospectively defined endpoint of the trial is a continuous outcome, such as a change in left ventricular end systolic volume (ESV). In this circumstance, as in the previous examples, the investigators have

[4]This assumes that there is no greater incidence of adverse events in higher risks associated with the clinical trial's intervention. Also relative risk must be considered when evaluating strength of evidence for efficacy.

a special prospectively stated interest in a proper subgroup stratum. If it is possible to derive greater precision in the evaluation of the continuous outcome measure in this subgroup of interest, the investigators will be able to produce acceptable type I and type II level control with a smaller achievable sample size. Consider the scenario in which the investigators are interested in determining the effect of therapy in reducing the rate of increase in ESV in patients who are suffering from congestive heart failure. The plan is to measure each patient's ESV at baseline, randomize the patient to receive either active therapy or placebo therapy, and then follow that patient until the end of the trial. At the trial's conclusion, ESV will be measured again, and for each patient, the difference Δ_{ESV} = ESV(final) − ESV(baseline) will be taken. It is expected that the Δ_{ESV} will be smaller in the patients treated with the intervention than in the control group patients. The investigator has a particular prospectively declared interest in demonstrating this effect, not just in the total cohort of patients who have congestive heart failure, but also in the particular proper subgroup of patients who have had at least two myocardial infarctions before they entered the study. The initial plans for the experiment require a sample size of 307 patients (Table 8).

Although the investigator would like to keep the sample size small, she would also like to provide confirmatory evidence that the therapy produces benefit in patients who have had multiple heart attacks. Since she anticipates that this subgroup would compose approximately 50% of the trial, she would not have adequate power for the design parameters as in Table 8. In addition, type I error concerns make the design issue more difficult, since alpha conservation requires a type I error level of less than 0.05 for the subgroup, decreasing the power even further for the same efficacy, standard deviation, and effect size. However, the investigator recognizes that there is a more precise instrument available that can be used to measure ESV. Although her budget will not permit this procedure to be used on all patients, she can carry out this more expensive determination in the smaller multiple-infarction subgroup. This newer, more expensive procedure will reduce the standard deviation of the difference in ESV from 25 to 20. She now computes the following sample sizes (Table 9).

The total cohort has increased from 307 to 392. In addition, the cost of the trial has increased since the more expensive instrument must be used to evaluate the baseline and follow-up studies on the 227 patients in the multiple myocardial infarction cohort. Also, the type I error for the total cohort evaluation has been reduced from 0.05 to 0.02. However, a confirmatory analysis is now available for the total cohort and the subgroup stratum of patients with multiple myocardial infarctions.

Table 9 Sample Size Computations for End Systolic Volume Study; Increased Endpoint Precision in the Subgroup

Cohort	Alpha Level	Delta	Standard Deviation	Power	Sample Size
Total	0.02	8	25	0.80	392
Multiple myocardial infarctions	0.03	8	20	0.80	227

Table 10 Sample Size Computations for the Total Cohort

Cohort	Alpha Level	Event Rate	Efficacy	Power	Sample Size
Total	0.05	0.15	0.20	0.80	4072

Strategy 5: Choose and Justify a Different Minimum Efficacy for the Subgroup

This final scenario evaluates the consequences of the prospective clinical trial design that assumes that the intervention's effectiveness in the subgroup is different from that in the total cohort. The detectable efficacy of a compound in a controlled clinical trial should be the minimal effectiveness of the intervention that is believed by the medical community to be clinically important, making the determination of intervention effectiveness a clinical as well as a statistical issue. It is logical to integrate the level of adverse events associated with this intervention into the efficacy determination. Essentially, the clinical question confronting the investigators as they consider efficacy levels is: What minimum benefit must the intervention produce that, when balanced against the risk of therapy, demonstrates the positive worth of the intervention? If a subgroup can be unfortunately anticipated to have a greater frequency of adverse events, the efficacy should be greater to offset this increased risk and thereby produce a favorable risk-benefit assessment for the intervention.

Consider a clinical trial in which the investigators are interested in demonstrating a 20% reduction in a clinical endpoint whose cumulative incidence over the course of the trial is estimated to be 15% in the control group. If the trial is to be powered at 80%, then 4072 patients are required for the study for a two-sided type I error level of 0.05 (Table 10).

If this cohort has a subgroup (e.g., the elderly) that has a greater frequency of adverse events reasonably believed to be associated with the intervention, a justification for the use of this drug would be the demonstration of greater efficacy in this elderly subgroup. If the investigators required 30% efficacy from this subgroup that is more likely to experience adverse events, the calculated, minimum sample size for the subgroup would be 2203 (Table 11).

In this prospective plan, the required sample size of the total cohort has increased from 4072 to 4706, an increase driven exclusively by the reduction in type I error for the total cohort evaluation from 0.05 to 0.03. This decrease in type I error was required to ensure that there is adequate alpha conservation when the confirmatory hypotheses are executed at the trial's conclusion. Using a type I error level of 0.02 for the prospectively defined subgroup analysis reveals that requiring an efficacy of 30% requires 2203 patients for this subgroup. With a subgroup of this size, the two clinical hypotheses to be carried out at the conclusion of the study would be confirmatory.

Table 11 Sample Size Computations for Clinical Trial with Different Efficacy Levels within the Subgroup of Elderly Patients and Alpha Level of 0.02

Cohort	Alpha Level	Event Rate	Efficacy	Power	Sample Size
Total	0.03	0.15	0.20	0.80	4706
Elderly	0.02	0.15	0.30	0.80	2203

DISCUSSION

Subgroup analyses is hazardous ground in clinical trial interpretations for well-understood reasons. Retrospectively considered, sometimes only casually planned, their conclusions, while descriptive of the findings in the sample, often times do not reveal the truth about the relationship in the larger population. The recent discussions in the literature concerning the wide variation in results by clinical center in the BHAT trial [31–33] is an illustration of the difficulty in interpreting these examinations. Currently, the findings of the PRAISE 1 study [12] that suggested that the subgroup of patients with congestive heart failure of a nonischemic etiology would benefit from amlodipine were not confirmed by a second study specifically designed to illustrate this same beneficial effect [13]. These examples demonstrate the propriety of sharply circumscribing subgroup interpretation. Indeed, current literature [11, 16–18, 20, 34, 35] recommends that, as currently incorporated in clinical trials, subgroup analyses interpretations are exploratory; they can suggest, but not confirm, a relationship in the population at large.

However, there are circumstances in which subgroup evaluations can produce confirmatory results that will stand on their own, separate from those of the overall cohort. These criteria, characterized by Yusuf et al. [18], are that the subgroups be prospectively specified, proper, and important consideration be given to type I and type II error. Unfortunately the appearance of these criteria have not led to a plethora of well-designed, prospective subgroup analyses with confirmatory evaluations at the study's end. This is in all likelihood due to the fact that, as currently designed, subgroup evaluations cannot meet these criteria for confirmatory evaluations, primarily, the requirement for type I/II error control. Thus, the growth of the use of subgroups as confirmatory tools has to some extent been stunted by the recognition that, as currently executed, one cannot reasonably construct a prospective clinical trial with an embedded, prospectively defined proper subgroup for which tight statistical control is provided for type I and type II statistical errors.

Oftentimes, subgroup analyses mean an assessment of an explicit term in a statistical model that directly measures the effect of treatment by subgroup interaction. However, not every useful subgroup evaluation need be based on an interaction effect that evaluates the heterogeneity of the effect of the interaction across subgroup strata. The methodology suggested herein is not a strategy that would supplant the interaction examination; we suggest a methodologic answer to a different question: Is there an explicit effect of the intervention in the prospectively defined subgroup stratum of interest?

An important component of the strategies recommended in this article is the selective levels of type I error. There is nothing but tradition that binds clinical trial methodologists to the 0.05 level of statistical significance. Recent work [21] has demonstrated the advantages of the allocation of type I error across hypothesis testing in clinical trials, in which the type I error is allocated not in equal components but in different sizes depending on the risks the investigators are willing to run to mistakenly conclude that there is an effect in the population based on the sample findings when there is no such effect in the population. However, this work has been predicated on the notion of independence of the executed hypothesis tests. The notion of dependent type I errors is a critical one. Gray [36] has explored this issue involving right-censored end-

point data. Our preliminary examination of this problem when there are two hypothesis tests has revealed that considerable savings in type I error can accrue when dependence between the hypothesis tests is taken into account. Further research is needed in this area. In addition, the weighted average of Senn [37], which focuses not on hypothesis testing as we do here but on estimation of effect size and several Bayesian approaches [38–40], has also been considered.

This article provides no casual solution to the subgroup analysis issue in clinical trials. The illustrations provided here do not vitiate the need for a disciplined approach to confirmatory subgroup analysis; they amplify it. The planned subgroup evaluations must be considered very carefully. There must be a biologically plausible rationale that leads the investigators to focus on the response of the subgroup to the clinical trial's intervention. The investigators must give careful consideration of the initial type I error allocations, and the investigators must think through the possible implications of the trial's possible findings. As demonstrated in the illustrations of this article, the size of the subgroup is commonly on the order of 40–60% of the total cohort sample size for the confirmatory analyses to be executed successfully. In some cases, the size of the overall trial must be adjusted. These procedures certainly cannot be carried out for every subgroup of interest in the study. After careful study, one or perhaps two subgroups can have confirmatory analyses prospectively embedded in the trial. The remaining subgroup analyses can be traditionally executed and interpreted in an exploratory light. Also, interpretation of trial results must jointly consider the quality of prospective planning, the manner of trial execution (concordant or discordant), effect size with its standard error, confidence intervals, and p-values. The focus of this article is on the p-value component, but this focus does not detract from the primacy of the joint interpretation.

In addition, many subgroups may be misinterpreted because subgroup membership may merely be a surrogate for the true risk-determining or efficacy-determining characteristic. The investigator must consider this possible explanation for her subgroup-specific effect in her interpretation of the analysis.

In the planning stages of a clinical trial, the strategies outlined in this article can be combined. The investigators have the freedom, indeed they have the mandate, to choose the appropriate combinations of these strategies prospectively. For example, it is possible to simultaneously take advantage of an acknowledged greater event rate in the subgroup of interest and implement the use of a second combined endpoint, while simultaneously giving careful consideration to alternative alpha allocations, when designing the study. When each of these strategies is acknowledged and built into the study prospectively, followed by the experiment's concordant execution, a confirmatory subgroup evaluation is produced at the trial's conclusion. In fact, this effort will produce a prospectively designed "trial within a trial," with the subtrial having its own inclusion and exclusion criteria (subgroup definitions), prospective endpoint, and type I/type II error specification. Only the stratification of the therapy allocation within the subgroup is missing to complete the "trial within a trial" construction and is easily supplied.

A positive subgroup evaluation when the overall trial result is null and the treatment by subgroup interaction is negative is an unusual argument in the standard clinical trial paradigm. However, that paradigm has been altered in this article, changing the interpretation of this result. The prospective identifi-

cation of the subgroup of interest in concert with the apportionment of type I error between the overall cohort and this subgroup renders the positive subgroup evaluation interpretation appropriate because it is a prospective design that preserves type I error. In this prospective subgroup-designed clinical trial, the interaction analysis (notoriously underpowered in many major clinical trials) would not be executed because there is no interest in subgroup heterogeneity. One can focus on the subgroup finding with the significant result when the significance level was prospectively determined, the trial was concordantly executed, and type I error was conserved.

There are several questions that can be asked of subgroups. One is whether the response to the intervention differs by subgroup, a question best addressed using a treatment by subgroup interaction analysis. The methodology we propose addresses a different question: Is there an explicit effect of the intervention in the prospectively defined subgroup stratum of interest? Both are relevant questions, and investigators should choose carefully which of these questions is the most important to address in their scientific inquiry.

Dr. Deswal's effort was supported by V.A. Cooperative Studies Program Clinical Research Career Development Award (CRCD #712B).

REFERENCES

1. Pfeffer MA, Braunwald E, Moyé LA, et al. Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction—Results of the Survival and Ventricular Enlargement Trial. *N Eng J Med* 1992;327:669–677.
2. Sacks FM, Pfeffer MA, Moyé LA, et al. The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *N Engl J Med* 1996;335:1001–1009.
3. The SHEP Cooperative Research Group. Prevention of stroke by antihypertensive drug therapy in older persons with isolated systolic hypertension: Final results of the Systolic Hypertension in the Elderly Program (SHEP). *JAMA* 1999; 265:3255–3264.
4. The Long-Term Intervention with Pravastatin in Ischaemic Disease (LIPID) Study Group. Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. *N Engl J Med* 1998;339:1349–1357.
5. Moyé LA, Pfeffer MA, Wun CC, et al. Uniformity of captopril benefit in the post infarction population: Subgroup analysis in SAVE. *Eur Heart J* 1994;15(Suppl B):2–8.
6. Lewis SJ, Moyé LA, Sacks FM, et al. Effect of pravastatin on cardiovascular events in older patients with myocardial infarction and cholesterol levels in the average range. Results of the Cholesterol and Recurrent Events (CARE) trial. *Ann Intern Med* 1998;129:681–689.
7. Lewis SJ, Sacks FM, Mitchell JS, et al. Effect of pravastatin on cardiovascular events in women after myocardial infarction: The cholesterol and recurrent events (CARE) trial. *J Am Coll Cardiol* 1998;32:140–146.
8. Peto R, Collins R, Gray R. Large-scale randomized evidence: Large, simple trials and overviews of trials. *J Clin Epidemiol* 1995;48:23–40.
9. MRFIT Research Group. Multiple Risk Factor Intervention Trial: Risk factor changes and mortality. *JAMA* 1982;248:1465–1477.
10. ISIS-I Collaborative Group. Randomized trial of intravenous atenolol among 16027 cases of suspected acute myocardial infarction-ISIS-1. *Lancet* 1986;2:57–66.

11. Lee KL, McNeer F, Starmer CF, Harris PJ, Rosari RA. Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. *Circulation* 1980:61;508–515.
12. Packer M, O'Connor CM, Ghali JK, et al. for the Prospective Randomized Amlodipine Survival Evaluation Study Group. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. *N Engl J Med* 1996;335:1107–1114.
13. Packer M. Results of the Prospective Randomized Amlodipine Survival Evaluation-2 Trial (PRAISE-2), presentation at the American College of Cardiology Scientific Sessions, Anaheim, California, March 15, 2000.
14. Assmann S, Pocock S, Enos L, Kasten L. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355:1064–1069.
15. Bulpitt C. Subgroup analysis. *Lancet* 1988;2:31–34.
16. Simon R. Patient subsets and variation in therapeutic efficacy. *Br J Clin Pharmacol* 1982;14:473–482.
17. Pocock SJ. *Clinical Trials: A Practical Approach*. Chichester: John Wiley & Sons; 1983:213–215.
18. Yusuf S., Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;266:93–98.
19. Moyé LA. *P*-value interpretation and alpha allocation in clinical trials. *Ann Epidemiol* 1998;8:351–357.
20. Moyé LA. *Statistical Reasoning in Medicine. The Intuitive P Value Primer*. New York: Springer-Verlag; 2000.
21. Moyé LA. Alpha calculus in clinical trials: Considerations and commentary for the new millennium. *Stat Med* 2000;19:767–779.
22. Snedecor GW, Cochran WG. *Statistical Methods*. 7th ed. Ames, Iowa: Iowa State University Press; 1980.
23. Pitt B, Segal R, Martinez FA, et al. on behalf of the ELITE Study Investigators. Randomized trial of losartan versus captopril in patients over 65 with heart failure. *Lancet* 1997;349:747–752.
24. Pitt B, Poole-Wilson PA, Segal R, et al. Effect of losartan compared with captopril on mortality in patients with symptomatic heart failure: Randomized trial—The losartan heart failure survival study ELITE II. *Lancet* 2000;355:1582–1587.
25. Packer M, Bristow MR, Cohn JN, et al. The effect of carvedilol on morbidity and mortality in patients with chronic heart failure. *N Engl J Med* 1996;334:1349–1355.
26. Moyé LA, Abernethy D. Carvedilol in patients with chronic heart failure (Letter). *N Engl J Med* 1996;335:1318–1319.
27. Packer M, Cohn JN, Colucci WS. Response to Moyé and Abernethy. *N Engl J Med* 1996;335:1318–1319.
28. Fisher LD, Moyé LA. Carvedilol and the Food and Drug Administration approval process: An introduction. *Control Clin Trials* 1999;20:1–15.
29. Fisher L. Carvedilol and the FDA approval process: The FDA paradigm and reflections upon hypotheses testing. *Control Clin Trials* 1996;20:16–39.
30. Moyé LA. *P*-value interpretation in clinical trials. The case for discipline. *Control Clin Trials* 1999;20:40–49.
31. Horwitz RI, Singer B, Makuch RW, Viscoli CM. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. *J Clin Epidemiol* 1996;49:395–400.
32. Altman DG. Within trial variation—A false trial? *J Clin Epidemiol* 1998;51:301–303.
33. Feinstein AR. The problem of cogent subgroups: A clinicostatistical tragedy. *J Clin Epidemiol* 1998;51:297–299.
34. Friedman L, Furberg C, DeMets D. *Fundamentals of Clinical Trials*. 3rd ed. New York: Springer; 1996.

35. Meinert CL. *Clinical Trials Design, Conduct, and Analysis.* New York: Oxford University Press; 1986.
36. Gray R. A simultaneous inference procedure for clinical trials. *Communications in Statistics* 1987;16:499–510.
37. Senn S. Statistical issues in drug development. Chichester: John Wiley & Sons; 1997:132–136.
38. Davis CE, Leffingwell D. Empirical Bayes estimates of subgroup effects in clinical trials. *Control Clin Trials* 1990;11:37–42.
39. Donner A. A Bayesian approach for the interpretation of subgroup results in clinical trials. *J Chronic Dis* 1982;34:429–435.
40. Louis TA. Estimating a population of parameter values using Bayes and empirical Bayes methods. *J Am Stat Assoc* 1984;79:393–398.