

CardiologyRounds™

AS PRESENTED IN THE ROUNDS OF THE CARDIOVASCULAR DIVISION
OF BRIGHAM AND WOMEN'S HOSPITAL, BOSTON, MASSACHUSETTS

Statistical Methods in Clinical Trials: Using Your Allocated Statistical Firepower

By LEMUEL A. MOYÉ, M.D., Ph.D.

A two-front war

When designing a clinical trial, investigators commonly feel that they are fighting (or are caught in the middle of) a two-front war. One front is driven by the requirement that the research effort should be productive, the other by statistical concerns.

It is the desire of all involved that the clinical trial bear a rich bounty of valuable, reproducible results. In order to comply with the need to be efficient, investigators genuinely want to evaluate all of the available information that can be collected on every patient in their research cohort. This natural tendency to use the dataset from the clinical trial to its fullest produces many statistical hypothesis tests. In addition, there are fundamentally sound motivations for the inclusion of dose-response analyses, the assessment of the effect of therapy on different but related endpoints, and the evaluation of possible mechanisms of therapeutic action.

However, the drive to conduct expansive, inquisitive analyses is inhibited by statistical concerns that make up the second front of the battle in the heart and mind of the investigator. Statistical concerns arise because investigators make a definite, though not formally recognized, compromise when performing research. The desire to study an entire population of patients in complete detail cannot be achieved. For example, a clinical trial studying a new heart failure therapy may wish to recruit all heart failure patients. However, logistical, financial, and ethical concerns preclude this effort. Therefore, we choose not to study the entire population of patients, but instead make a compromise and take a small sample from the large population. The process of drawing a sample helps on the one hand, but hurts on the other. The availability of the sample offers the ability to carry out an executable research program on a relatively small number of subjects; however, the same process hurts by removing the ability to identify population effects with certainty. The same population can produce different samples and, since these samples contain different patients with different experiences, the results from each sample are different. Which sample is right? Sometimes the population produces – just through the play of chance and the random aggregation of events – a sample that does not accurately reflect the therapy-outcome relationship of the true overall population, even though the sample was selected randomly.

We cannot directly observe or measure sampling error when examining a randomized participant, or entering that patient's lab results into a database, or computing a sample relative risk; we observe only the data. However, sampling error, like gravity, goes unseen, but has powerful effects. The central contribution of statistics to health research in general, and to controlled clinical trials in particular, is their guidance on research result interpretation in the presence of sampling error. The successful application of statistics within clinical trials is not to remove sampling error; instead, sampling error is appropriately channeled into our estimators of effect size, standard errors, confidence intervals, p values, and power. However, this sampling error segregation is only successful when the underlying assumptions that go into these estimators have been satisfied.

One critical assumption is that the experiment is executed concordantly (ie, in accordance with its prospectively written protocol). The estimators are reliable only if they are produced within a research environment in which the only source of variability is the sample-to-sample variability of the endpoint data. This assumption is violated by, for example, the process of data dredging. In this circumstance, the sample doesn't just provide data for the endpoint, but instead exceeds this contribution by actually selecting the endpoint (eg, choosing the only endpoint with a small p value). In this circumstance, our commonly used estimators are no longer valid; they have become newly distorted by this additional source of endpoint selection variability. If too much credence is given to what these disoriented estimators tell us,



BRIGHAM AND
WOMEN'S HOSPITAL

A Teaching Hospital of

HARVARD MEDICAL SCHOOL

Cardiovascular Division (Clinical)

Michelle Albert, MD
Elliott Antman, MD
Donald S. Baim, MD
Joshua Beckman, MD
Gavin Blake, MD
Charles M. Blatt, MD
Eugene Braunwald, MD
Christopher Cannon, MD
Ming Hui Chen, MD
Michael Chin, MD, PhD
Mark Creager, MD
Victor Dzau, MD
Elazer Edelman, MD, PhD
Andrew Eisenhauer, MD
Laurence Epstein, MD
James Fang, MD
Mark Feinberg, MD
Jonas Galper, MD, PhD
Peter Ganz, MD
J. Michael Gaziano, MD
Marie Gerhard-Herman, MD
Robert Giugliano, MD
Michael Givertz, MD
Samuel Z. Goldhaber, MD
Thomas B. Graboys, MD
Howard Hartley, MD
Carolyn Ho, MD
Mukesh Jain, MD
John Jarcho, MD
Paula Johnson, MD
Ralph Kelly, MD
Scott Kinlay, MD
Jamil Kirdar, MD
James Kirshenbaum, MD
Gideon Koren, MD
Richard Kuntz, MD
Raymond Kwong, MD
Michael J. Landzberg, MD
Richard Lee, MD
Eldrin Lewis, MD
James Liao, MD
Peter Libby, MD (Division Chief)
Leonard Lilly, MD
Bernard Lown, MD
William Maisel, MD
Thomas Michel, MD, PhD
David Morrow, MD
Karen Moulton, MD
Gilbert Mudge, MD
Anju Nohria, MD
Patrick O'Gara, MD
Marc A. Pfeffer, MD, PhD (Editor)
Jorge Plutzky, MD
Jeffrey Popma, MD
Shmuel Ravid, MD
Frederic Resnic, MD
Paul Ridker, MD
Thomas Rocco, MD
Campbell Rogers, MD
Maria Rupnick, MD, PhD
Arthur Sasahara, MD
S. Dinakar Satti, MD
Jay Schneider, MD
Christine Seidman, MD
Andrew Selwyn, MD
Daniel Simon, MD
Laurence Sloss, MD
Kyoko Soejima, MD
Regina Sohn, MD
Scott Solomon, MD
Lynne Stevenson, MD
William Stevenson, MD
Peter Stone, MD
Michael Sweeney, MD
Frederick Welt, MD
Justina Wu, MD

Brigham and Women's Hospital

Fax: (617) 732-5291 Website: www.heartdoc.org

The editorial content of *Cardiology Rounds* is determined solely by the Cardiovascular Division of Brigham and Women's Hospital. This publication is made possible by an educational grant.

then we as investigators can lose our way,² as the following example demonstrates.

Case history 1: ELITE I / II

The use of angiotensin converting enzyme (ACE) inhibitors has increased dramatically since the 1980s. First approved as a treatment for hypertension, their use expanded into treatment of other cardiovascular disease, specifically the treatment of congestive heart failure (CHF). However, these effective ACE inhibitors were also associated with undesirable adverse events; the most common were cough, angioedema, and hypotension. As a response to this undesirable profile, angiotensin II type 1 receptor blockers were developed. It was hoped that this newer class of agents would be safer than the original ACE inhibitors, while continuing to confer a survival benefit for patients with CHF. In order to compare the relative safety of angiotensin II type 1 receptor blockers to that of ACE inhibitors, the Evaluation of Losartan in the Elderly Study (ELITE) I³ was designed. The goal of ELITE I was to compare the effectiveness of the angiotensin II type 1 receptor blocker, losartan, to the ACE inhibitor, captopril, in a randomized, double-blind, clinical trial. The primary endpoint of this study was a safety measure that was prospectively defined as an increase in serum creatinine by 0.3 mg/dL or more above baseline.

This study randomized 722 patients and followed them in a double-blind fashion for 48 weeks. Just prior to the end of the study, an additional endpoint was added. This measure was the composite endpoint of death and/or admission for heart failure. At the conclusion of ELITE I, the investigators determined that the increase in serum creatinine was the same in the 2 treatment arms (10.5 % in each group; risk reduction 2% [95% CI, -51 to 36], $p=0.63$). However, the findings for the new composite endpoint were tantalizing. Death and/or heart failure admission occurred in 33 of the 352 losartan-treated participants, while 49 events occurred in the 370 captopril-treated patient group. This translated into a risk reduction of 32% (95% CI, -0.04 to 0.55; $p=0.075$) and suggested that there was a benefit attributable to losartan. The investigators, emboldened by these findings, deconstructed the composite endpoint and discovered that 17 deaths occurred in the losartan group and 32 deaths in the captopril group, a result that produced a risk reduction of 46% (95% CI, 5 to 69; $p=0.035$). These deaths with an “apparent significant p value” received the principle emphasis in the discussion section of the manuscript, at the expense of minimizing the role of the primary safety endpoint. Although the need to repeat the trial was mentioned in the abstract, the balance of the discussion focused on the reduced mortality rate of losartan. According to the authors, “This study demonstrated that losartan reduced mortality compared with captopril; whether the apparent mortality advantage for losartan over captopril holds true for other ACE inhibitors requires further study.” While it was clear that this was the case within the ELITE I sample, the investigators implied that these findings generalized to the population of elderly patients with CHF. Others even went so far as to attempt to explain the mechanism for the reduction in sudden death observed in ELITE I.^{4,5}

To the investigators’ credit, ELITE II⁶ was executed to confirm the superiority of losartan over captopril in improving survival in patients with heart failure. The primary endpoint in ELITE II was total mortality, a decision that required recruitment of 3152 patients, almost 5 times the number of patients recruited for ELITE I. These patients were followed for 18 months, almost twice as long as the duration of follow-up in ELITE I. At the

conclusion of ELITE II, the cumulative all-cause mortality rate was not significantly different between the losartan and captopril groups (280 deaths in the losartan group vs. 250 deaths in the captopril group, 17.7% vs. 15.9%; hazard ratio 1:13; 95% CI, 0.95 to 1.35, $p=0.16$). In fact, there was a trend to excess mortality in the losartan group. Thus, losartan did not confer a mortality benefit in the elderly with CHF when compared to captopril, as suggested by ELITE I. The investigators conceded, “More likely, the superiority of losartan to captopril in reducing mortality, mainly due to decreasing sudden cardiac death, seen in ELITE I should be taken as a chance finding.”⁶

Even though great effort went into the computation of estimators of effect size, standard error, confidence intervals, and p values in ELITE I, these estimators were not designed to work in an environment where an endpoint is altered based on the data of the research effort. Data are, of course, expected to contribute to the endpoint analysis but, in the case of ELITE I, the data chose the analysis (would the mortality effect seen in ELITE I have received such high prominence if losartan and captopril were associated with equivalent cumulative mortality rates?). When data choose the analysis plan, the findings in the sample about the data-driven endpoint are no longer reliable indicators of the effect that occurs in the population.

Type I error propagation

The issue raised by multiple testing is one of propagation of type I error. Since there is a chance that the sample will provide a false answer to the question asked by the investigator, the likelihood that at least one false answer will be provided increases with the number of questions asked (just as the probability of obtaining at least one head in a sequence of tosses of a coin grows as we continue to flip the coin). This error is critical in assessing the impact of the intervention, studied in a clinical trial, on the population from which the clinical trial patients were obtained. Since the intervention will produce adverse events and will likely have a financial cost associated with it, there must be some benefit that balances these disadvantages.

The familywise error rate (ξ) is the likelihood of making at least one type I error among all of the hypothesis tests that the investigators carry out. The symbol, ξ , conveys to the investigators the likelihood that the therapy will not be effective in the population, (ie, that treated patients in the population will experience adverse events of the medication and pay the financial cost of the therapy, but will not retain its benefits). The familywise error rate must be accurately measured and tightly controlled to insure that its level is kept to a minimum.

Thus, it may seem that investigators are in an intolerable position. As they design their clinical trial, they are quite naturally motivated to answer all of the relevant issues that their dataset can address, but they are simultaneously tightly bound by compelling statistical arguments to abstain from this desire and address only the primary pre-specified question. Thirsting for answers to their other scientific questions, they may feel like the parched man who, when at last he comes upon a fresh mountain river, responds with amazement when he is given only a tiny cup to collect the water that he requires. What water is caught in the small statistical thimble, while refreshing, does not satisfy, and he watches in amazement and disappointment as the rest of the (data) stream flows by unused.

Loosening the shackles

Investigators often feel as though the statistical concerns that are contained in the careful design of their clinical trial

have “put handcuffs” on them. The following discussion elaborates a wealth of design tools available to the trial investigator that can provide some of the keys to release (or at least loosen) their shackles. We begin with the assertion that no team of investigators should be denied or discouraged from analyzing any component of the dataset they desire. The unique combination of inquisitiveness, insight, and intuition that investigators possess should be encouraged, not repressed. However, it is best if their analysis is triaged so that interpretation of the results is clear. There are two levels to this triage process: planning, and error control.

- The first level of triage addresses the question of whether the analysis will be prospectively planned or data-driven. The major advantage of prospectively planned analyses is that the estimates of effect size, confidence intervals, standard errors, and p values are trustworthy. Non-prospectively planned, exploratory results should be carried out and reported, but they must be clearly labeled as hypothesis-generating, requiring confirmation before they can be integrated into the fund of knowledge of the medical and regulatory communities.

- The second level of triage analysis, during the design phase of the clinical trial, is carried out among the prospectively planned analyses, dividing them into primary analyses or secondary analyses. Primary analyses are those upon which the conclusions of the trial rest. Each of the primary analyses will have a prospectively set type I error level attached to it in such a way that the familywise error does not exceed the community accepted level (traditionally 0.05). The trial will be seen as positive, null (no finding of benefit or harm), or negative (harmful result) based on the results of the primary analyses.

It is critical to note that a clinical trial can have more than one primary endpoint. If appropriately designed, the study can be judged as positive if any of those primary endpoints produce a p value less than the test specific alpha level for that hypothesis test. However, the medical community must examine all the confirmatory analyses in order to draw the correct conclusion from the experiment.

Secondary endpoints do not control the familywise error and each secondary analysis is typically interpreted at nominal 0.05 levels. Secondary analyses, when prospectively designed, produce trustworthy estimates of effect sizes and p values. However, because secondary analyses do not control the familywise error, the risk to the population is too great for confirmatory conclusions to be based upon them. Therefore, the role of secondary endpoints is to provide support for the primary endpoint findings and not to serve as independent, confirmatory analyses.

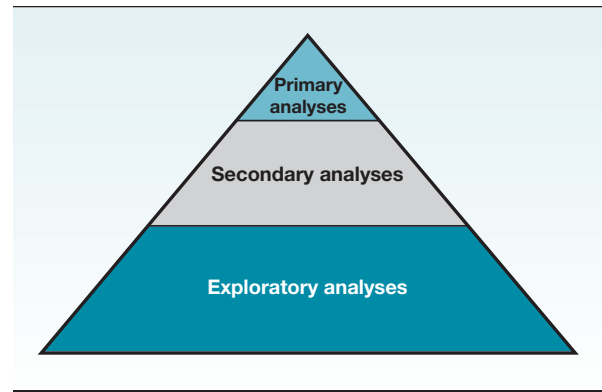
In the typical clinical trial, there are more exploratory analyses than there are prospectively declared analyses, and more secondary analyses than there are primary analyses (Figure 1). This is consistent with the statement that a small number of key questions should be addressed in a clinical trial, accompanied by careful deliberation on the necessity and extent of adjustment for multiple comparisons.⁷

Allocating type I error

Several tools are available to the investigator to allocate type I error among the primary analyses of their clinical trial.

The first is the unequal allocation of type I error. The Bonferroni procedure^{8,9} and its modern application^{10,11} provide equal allocation of the alpha error among the primary analyses. The application of this process typically produces type I error levels that are too small for some of the analyses. This, in turn,

Figure 1: The role and relative number of analyses in a clinical trial



generates sample sizes beyond the attainable. However, there is no theoretical requirement for the equal apportionment of type I error among different hypothesis tests, and investigators can allocate type I error selectively among the different primary analyses. The only rules they are obligated to follow are that the allocation be made prospectively and that the type I error levels be made to conserve the familywise error level, ξ .^{12,13}

Consider the work of the investigator during the design phase of his/her study. He/she has completed the endpoint triage system and has prospectively determined the K primary endpoints for which significance testing will be executed. The study is initially designed to have a familywise error level of ξ ; this rate is dispersed between the K primary endpoints with the j^{th} endpoint having test specific alpha α_j such that (using a Bonferroni-style approximation): $\xi \leq \sum_{j=1}^K \alpha_j$.

At the conclusion of the study, the investigator produces a p value for each of these hypothesis tests. Let p_j be the p value which is computed for the j^{th} primary endpoint, $j = 1, 2, 3, \dots, K$. Then, just as a familywise error level, ξ , is computed during the design phase of the trial, the observed familywise error level at the conclusion of the trial or posterior familywise error level, E , is computed based on the K different p values produced by the trial as:

$$E \leq \sum_{j=1}^K \min(p_j, \alpha_j)$$

where $\min(a, b)$ is the minimum of the numbers a and b . We need to be able to link comparisons of the *a priori* and posterior familywise error levels to the individual p values of the K primary endpoints. The study is positive* when the posterior familywise error level E is less than the *a priori* familywise error level ξ , or $E < \xi$. An alternative formulation is that the study is positive when:

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

This definition, in one sense, increases the complexity of the interpretation of the clinical trial. Although the study will be classified as either positive, negative, or null, in fact, with K primary endpoints, the investigators will report the findings for each of these prospective, confirmatory analyses. Thus, the medical community will learn not only that the study was positive, but also exactly what finding led to this positive description of the trial.

Since the familywise error level for the entire trial is typically fixed at 0.05, the test specific alpha levels for each of a large number of hypothesis tests become small very quickly.

*These comments assume that the clinical trial has been concordantly executed.

Even if the researcher triages the analyses, focusing on a relatively small number of primary analyses in which alpha is prospectively allocated, and then allocates alpha differentially among these primary analyses, he/she commonly faces the difficulty that the prospectively-set alpha levels for some of the primary analyses will be very small. The recognition that some of these primary analyses will be dependent upon each other and that this dependence will lead to some type I error level conservation has been the focus of considerable attention; much of the statistical literature is devoted to the multiple analysis issue. Several approaches can be taken to incorporate dependence between endpoints.¹⁴⁻¹⁷ Clearly, the more complicated the dependency relationship between subsets of the primary analyses, the more complex the prospective alpha allocation procedure can be.

A simple development within reach of investigators would proceed as follows. Consider that there are two hypothesis tests. Let T_1 be the event that a type I error occurs on the first experiment with probability α_1 , and T_2 be the event of a type I error occurring on the second experiment, with probability α_2 . Then, a measure, termed D, is developed that will reflect the relationship between the occurrence of T_1 and T_2 . We would like D to have a minimum of 0 and a maximum of 1. It is useful to let $D = 0$ when the statistical hypothesis tests are independent, and the occurrence T_1 does not reflect one way or another the occurrence of T_2 , ie, $P[T_2 = 0|T_1 = 0] = 1 - \alpha_2$. Analogously, $D = 1$ will denote perfect dependence, ie, the case in which the conditional probability of interest $P[T_2 = 0|T_1 = 0] = 1$. Then we can write D in terms of the conditional probability:

$$D = \sqrt{1 - \frac{(1 - P[T_2 = 0|T_1 = 0])}{\alpha_2}}$$

More usefully, we can solve for α_2 , the test specific type I error level for the second hypothesis test as a function of the test specific type I error for the first hypothesis test α_1 ($\alpha_1 \geq \alpha_2$), the familywise error rate ξ , and the parameter D as:

$$\alpha_2 = \min \left[\alpha_1, \frac{\xi - \alpha_1}{(1 - \alpha_1)(1 - D^2)} \right]$$

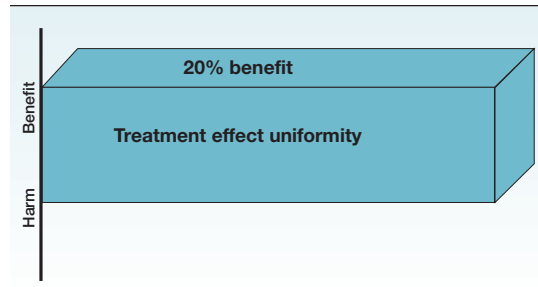
The notion of dependence between statistical hypothesis tests is especially useful when consideration is given to the role of combined endpoints in clinical trials. A formal examination of the use of combined endpoints reveals that these complicated implements are most effective when:

- the combined endpoint is clinically relevant
- it is cohesive (ie, the combined endpoint's components measure related, but distinct, aspects of the same disease process)
 - its component endpoints are each ascertained precisely and with superior quality
 - its component endpoints are each reported.

Subgroups

The purpose of studying a sample is to learn something about the population of patients from which the sample is derived. However, the fact that different samples can be drawn from the same population and yield different results (this is called sample-to-sample variability or sampling error) causes confusion in trying to extend sample

Figure 2: The common view of a therapy effect in a clinical trial — uniformity

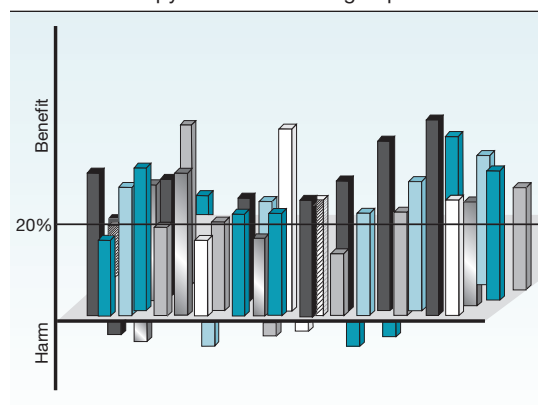


results to the population. As an illustration, consider the result of a hypothetical clinical trial in which the investigators report that the effect of therapy on the prospectively defined endpoint of total mortality produces a 20% benefit in the cumulative mortality rate. The first reaction is to believe that all collections of patients in the active group were beneficiaries of this 20% benefit (Figure 2). However, the actual results of the trial appear to be quite different (Figure 3).

The research effort reflects what appears to be important differences in the magnitude of treatment effects; differences that are mediated by subgroups. However, this non-uniformity is actually caused by the random aggregations of relatively small numbers of patients within a single sample. At first glance, it appears that the uniform mortality benefit has been replaced by a much more heterogeneous response. However, in reality, the 20% benefit has been well-disguised by the presence of background clutter that is produced by sampling error. The uniform 20% reduction in the total mortality effect is still there; the population from which the research sample was derived still experiences a 20% reduction in mortality. However, when that uniform effect is viewed from the prism of a small sample, the appearance of the effect is distorted. The subgroups appear to define different levels of responses. However, all that is happening is that the random selection mechanism causes individuals whose responses are similar to cluster together by chance alone. This random variability produces the differences seen in subgroups that are commonly described as a subgroup effect.

Consider the following example as an illustration of the heterogeneity of effect that is due solely to the random

Figure 3: Heterogeneity of treatment effects among subgroups. Each bar is the effect of the therapy in a different subgroup



aggregation of patients into small samples. A conference room chosen at random has a capacity of seating 60 observers. These 60 seats are divided by a central aisle, with 30 seats on both the left- and right-hand sides of the room. Sixty people seat themselves as they choose, distributing themselves in an unrestricted manner among the seats on each side of the room. When all are seated, the height of each person is measured and it is found that the average height is exactly 71 inches. Does that mean that the average height of those seated on the left side of the conference room will be 71 inches? No. The average height of those seated on the left-hand side of the conference room will be either <71 inches or >71 inches, but it will not be exactly 71 inches (since this average is based on only 30 of the 60 people). If the average height on the left side of the conference room is <71 inches, then those seated on the right-hand side will have an average height >71 inches. Is it fair to conclude that those who sit on the right-hand side of the conference room are in general taller than those who sit on the left? The simple, random aggregation and subaggregation of the observers has induced a subgroup effect that is based only on the play of chance.

A fine example of the misdirection that subgroup analysis can provide is the PRAISE/PRAISE-2 clinical trial pair. In the 1980s, the use of calcium channel blocking agents in patients with CHF was problematic. Initial studies suggested that patients with CHF experienced increased morbidity and mortality associated with these agents.¹⁸ However, additional developmental work on this class of medications proceeded. In the early 1990s, new calcium channel blocking agents appeared. The early data for these compounds suggested that their use may produce improvements in patients with CHF.

To formally evaluate this possibility, the Prospective Randomized Amlodipine Survival Evaluation (PRAISE)¹⁹ trial was designed. Its long-term objective was the assessment of the effect of the channel blocker, amlodipine, on morbidity and mortality in patients with advanced heart failure. The primary measurement in PRAISE was the composite endpoint of all-cause mortality and/or hospitalization.* The protocol also stipulated that there would be analyses carried out in the following subgroups of patients based on sex, baseline ejection fraction, NYHA class, serum sodium concentration, angina pectoris, and hypertension.

PRAISE began recruiting patients in March of 1992. Patients with CHF (NYHA functional class IIIb/IV and left ventricular ejection fraction <30%) were randomized to receive either amlodipine or placebo therapy. The investigators suspected that the effect of amlodipine might depend on the cause of the patient's CHF, so they stratified randomization† into two groups, patients with ischemic cardiomyopathy, and patients with non-ischemic cardiomyopathy.³ By the end of the recruiting period, the PRAISE investigators randomized 1153 patients and at the end of the study, they had been followed for a maximum of 33 months.

At the conclusion of PRAISE, the investigators determined that in the overall cohort, there was no significant difference in the occurrence of the primary endpoint

between the amlodipine and placebo groups (39% vs. 42%, 9% reduction [95% CI, -24 to 10], $p = 0.31$). The secondary endpoint of all-cause mortality was also not significantly different between the amlodipine and placebo groups for the overall cohort (33% vs. 38%, 16% reduction [95% CI, -31 to 2], $p = 0.07$).

The evaluation then turned to the etiology-specific CHF strata. PRAISE recruited 732 patients with an ischemic cause for their CHF and 421 patients with a non-ischemic cause. The analysis of the effect of therapy in these strata revealed that treatment with amlodipine reduced the frequency of primary and secondary endpoints in patients with non-ischemic dilated cardiomyopathy (58 fatal or nonfatal events in the amlodipine group and 78 in the placebo group, 31% risk reduction; 95% CI, 2 to 51; $p = 0.04$). Further evaluation of these events revealed that there were only 45 deaths in the amlodipine group and 74 deaths in the placebo group, representing a 46% reduction in risk in the amlodipine group (95% CI, 21 to 63 reduction; $p < 0.001$). Among the patients with ischemic heart disease, treatment with amlodipine did not affect the combined risk of morbidity and mortality or the risk of mortality from any cause. Again, the findings for a secondary endpoint overshadowed the primary endpoint's results.

A second trial, PRAISE-2²⁰ was then conducted to verify the beneficial effect on mortality seen in the subgroup analysis of patients with heart failure of non-ischemic etiology in PRAISE-1. This trial, while focusing only on patients with heart failure of non-ischemic origin, was otherwise similar in design to PRAISE-1. The PRAISE-2 investigators randomized 1650 patients to either amlodipine or placebo with up to 4 years follow-up. However, the results of PRAISE-2 were quite different from PRAISE-1. Unlike the first study, in PRAISE-2, there was no difference in mortality between the 2 groups (33.7% in the amlodipine arm and 31.7% in the placebo arm; odds ratio 1.09, $p = 0.28$). Thus, the marked mortality benefit seen in the subgroup analysis in PRAISE-1 for amlodipine was not confirmed in PRAISE-2. The PRAISE-1 investigators inappropriately raised a subgroup analysis to prominence outside of the statistical boundaries of the original design.

Evolution

Finally, we must remember that clinical trial standards are not static. Instead, they reflect the fluid sequence of inevitable and unstoppable progress. An examination of the clinical literature from 200 years ago reveals that healthcare research was primarily, if not exclusively, the evaluation of case reports. Research progressed from these accounts of individual circumstances to the appearance of the results of case series. The idea of a clinical trial that simultaneously combined both a contemporaneous control group and the use of randomization is relatively new, appearing only 60 years ago. During this evolutionary process, mistakes have been made. A fair criticism of clinical trial methodology would be that there has been an over reliance on p values and, unfortunately, confusion between confirmatory and exploratory analyses still reigns. Pocock²¹ has correctly pointed out that concerns for the multiplicity of type I error should be balanced. While we have not reached our destination, the development and maintenance of investigator discipline in research endeavors helps to insure that we stay on the right path.

*Hospitalization was defined as in hospital care for at least 24 hours for either acute pulmonary edema, severe hypoperfusion, acute myocardial infarction, or sustained hemodynamically destabilizing ventricular tachycardia or fibrillation.

†Stratified randomization is an adaptation of the random allocation to the therapy process.

References


1. Moyé LA. *Statistical Reasoning in Medicine – The Intuitive P value Primer*. NY: Springer-Verlag, 2000.
2. Moyé LA. Random research. *Circulation* 2001;103:3150-3.
3. Pitt B, Segal R, Martinez FA, et al. on behalf of the ELITE Study Investigators. Randomized trial of losartan vs. captopril in patients over 65 with heart failure. *Lancet* 1997;349:747-52.
4. Jensen BV, Nielsen SL. Correspondence: Losartan versus captopril in elderly patients with heart failure *Lancet* 1997;349:1473.
5. Fournier A, Achard JM, Fernandez LA. Correspondence: Losartan versus captopril in elderly patients with heart failure. *Lancet* 1997;349:1473-4.
6. 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;355:1582-87.
7. Proschan MA, Waclawiw MA. Practical guidelines for multiplicity adjustment in clinical trials. *Control Clin Trials* 2000;21:527-539.
8. Bonferroni CE. Il calcolo delle assicurazioni su gruppi di teste. In: Studi in Onore del Professore Salvatore Ortu Carboni, Rome, Italy. 1935:13-60.
9. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936;8:3-62.
10. Dunn OJ. Confidence intervals for the means of dependent, normally distributed variables. *J Am Stat Assoc* 1959;54:613-621.
11. Dunn OJ. Multiple comparisons among means. *J Am Stat Assoc* 1961;56:52-54.
12. Cook RJ, Farewell VT. Multiplicity consideration in the design and analysis of clinical trials. *JR Stat Soc A* 1996;159:93-110.
13. Moyé LA. P-Value interpretation and alpha allocation in clinical trials. *Ann Epidemiol* 1998;8:351-357.
14. Westfall PH, Young SS. *Resampling-based multiple testing: examples and methods for P-value adjustment*. New York: John Wiley and Sons; 1993.
15. Hochberg Y, Westfall PH. On some multiplicity problems and multiple comparison procedures in biostatistics. In: Sen PK, Rao CR. Eds. *Handbook of Statistics. Bioenvironmental and public health statistics*. Vol 18; New York: Elsevier; 2000:75-113.
16. James S. Approximate multinomial probabilities applied to correlated multiple endpoints in clinical trials. *Stat Med* 1991;11:23-1135.
17. Neuhauser M, Steinijans VW, Bretz F. The evaluation of multiple clinical endpoints with application to asthma. *Drug Inf J* 1999;33:471-477.
18. Multicenter Diltiazem Postinfarction Trial Research Group. The effect of diltiazem on mortality and reinfarction after myocardial infarction. *N Engl J Med* 1988;319:385-392.
19. 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;335:1107-14.
20. Packer M. Presentation of the results of the Prospective Randomized Amlodipine Survival Evaluation-2 Trial (PRAISE-2) at the American College of Cardiology Scientific Sessions. Anaheim, CA. March 15, 2000.
21. Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Control Clin Trials* 1997;18:530-545.




Dr. Lemuel A. Moyé, M.D., Ph.D. is a physician and biostatistician at the University of Texas School of Public Health. He earned his medical degree at Indiana University Medical School in 1978 and completed a Ph.D. in Community Health Sciences – Biometry in 1987. He has worked as a clinician in Houston for 8 years and is Professor of Biometry at the University of Texas School of Public Health. He has been Co-Principal Investigator of two multinational clinical trials examining the relationship between innovative therapy and post-myocardial infarction survival (SAVE) and in the use of cholesterol-reducing agents prolonging post-myocardial infarction survival in patients with normal cholesterol levels (CARE). Dr. Moyé has had an active service as a consulting biostatistician to the Federal Food and Drug Administration, serving as biostatistician on the

CardioRenal Advisory Panel to the Federal Food and Drug Administration. He is currently a member of the Pharmacy Sciences Advisory Committee. He has authored over 100 manuscripts, with papers appearing in the *Journal of the American Medical Association*, the *New England Journal of Medicine*, *Statistics in Medicine*, and *Controlled Clinical Trials*. He is the sole author of a book entitled *Statistical Reasoning in Medicine: The Intuitive P value Primer* (Springer) and the co-author of the text *Difference Equations with Applications in Public Health* (Marcel Dekker). A third book, entitled *Multiple Analyses in Clinical Trials: Fundamentals for Investigators* (Springer), will be available in 2003.

Dr. Moyé has no disclosures to announce related to the enclosed CME program.



Harvard Medical School
Department of Continuing Education
and
Brigham and Women's Hospital
Cardiovascular Division



Present

**Cardiovascular Medicine:
Review and Update for the Practitioner**

May 4 - 8, 2003

in Boston at the
Fairmont Copley Plaza Hotel

Course Directors:
**Samuel Z. Goldhaber, MD Peter Libby, MD
Patrick O'Gara, MD**

To register or view course details online visit us at:
www.cme.hms.harvard.edu/cardiomedreview

For more information regarding registration:
Call: 617-384-8600, Monday-Friday, 10am to 4pm EST
Mail: Harvard MED-CME, P.O. Box 825, Boston, MA 02117-0825
Email: hms-cme@hms.harvard.edu

Tuition: \$795 (U.S.)
Early Registration (on or before December 31, 2002): \$595 (U.S.)
Trainees, Nurses, PAs, CPTs: \$395 (U.S.)

For information regarding course content, call: 617-732-7566

Brigham and Women's Hospital,
Cardiovascular Division website:
www.heartdoc.org

This publication is made possible by an educational grant from

Novartis Pharmaceuticals Corporation

© 2002 Brigham and Women's Hospital, Boston, Massachusetts, which is solely responsible for the contents. The opinions expressed in this publication do not necessarily reflect those of the publisher or sponsor, but rather are those of the author based on the available scientific literature. Publisher: **SNELL Medical Communication Inc.** in cooperation with Brigham and Women's Hospital, Boston, Massachusetts. TMCardiology Rounds is a Trade Mark of SNELL Medical Communication Inc. All rights reserved. The administration of any therapies discussed or referred to in *Cardiology Rounds* should always be consistent with the recognized prescribing information as required by the FDA. **SNELL Medical Communication Inc.** is committed to the development of superior Continuing Medical Education.