Perspectives

The Fragility of Cardiovascular Clinical Trial Results

LEMUEL A. MOYÉ, MD, PhD,* ANITA DESWAL, MD†

Houston, Texas

ABSTRACT

Background: Clinical trials that have their prospective analysis plan altered are difficult to interpret.

Methods and Results: After providing 4 examples of problematic trial results that have had their findings reversed, the necessity of a fixed research protocol is developed. Investigators generally wish to extend the results from their research sample to the larger population; however, this delicate extension is complicated by the presence of sampling error. No computational or statistical tools can remove sampling error—the most that researchers can do is to provide to the medical and regulatory communities a measure of the distorting effect that sampling error can produce. Investigators accomplish this by providing an estimate of how likely it is that the population produced a misleading sample for them to study. However, studies in which the data determine the analysis plan damage these estimators. When they are damaged, these estimators produce untrustworthy assessments of the degree to which the study results reflect the population findings.

Conclusions: The way to avoid these complications is to design the experiment carefully, then carefully execute the experiment as it was designed.

Key Words: Statistics, estimators, epidemiology, prospective design, clinical trials.

The prospectively designed, randomized, double-blind clinical trial is held as the most advanced research tool to assess an intervention's effect in clinical medicine. Often meticulously planned, requiring hundreds of researchers, thousands of patients, and millions of dollars, these research enterprises can provide important new information about the safety and efficacy of medical and or surgical management of patients. Pharmaceutical companies, regulatory agencies, managed care organizations, and private physicians look to these research devices as the conduit through which this critical, new information for treating patients is transmitted. However, there has been a curious trend in recent clinical trial reports studying congestive heart failure (CHF). Although circumstances in which two clinical trials have been carried out to assess the same intervention are rare, the results from these trial pairs, when available, have not been consistent. This has been the case with the evaluations of vesnarinone, losartan, and amlodipine. In each of these three circumstances a pair of clinical trials was carried out sequentially. In each circumstance, the first clinical trial suggested a benefit, followed by a second clinical trial that reversed or nullified the result of the first experiment. In these circumstances, the nonstatistical

*From the *University of Texas School of Public Health, Houston, Texas, and †Winters Center for Heart Failure Research and Houston Center for Quality of Care and Utilization Studies, Veterans Administration Medical Center and Baylor College of Medicine, Houston, Texas.*

Manuscript received February 15, 2002; revised manuscript received May 15, 2002; revised manuscript accepted May 30, 2002.

Reprint requests: Lemuel A. Moyé, MD, PhD, University of Texas School of Public Health, RAS Building E815, 1200 Herman Pressler, Houston, TX 77030.

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

1071-9164/02/0804-0010$35.00/0
doi:10.1054/jcaf.2002.126917

reader of these studies is easily perplexed, and even the most ardent supporter of clinical trial methodology cannot avoid a moment of pause and disquiet. After a brief review of three examples, a likely reason for these paradoxes is provided. In the process, simple rules will be described that authors and readers should consider in judging the indelibility of a clinical trial result.

Vesnarinone

In an initial study of an oral positive inotrope, vesnarinone,[1] patients with symptomatic left ventricular systolic dysfunction were randomized to receive either placebo, 60 vesnarinone, or 120 mg vesnarinone per day, in addition to conventional heart failure therapy. The primary outcome measure was a combined endpoint of all cause mortality and major cardiovascular morbidity at 6 months; a secondary endpoint was all-cause mortality. Although the study was originally designed to randomize 150 patients to each of the three treatment arms, one arm was terminated early with a corresponding increase in the number of patients in the remaining two arms. At the trial's conclusion, the administration of 60 mg vesnarinone was associated with both a 50% risk reduction in the primary endpoint of all cause mortality and major cardiovascular morbidity (95% confidence interval [CI] 20–69; $P = .003$), as well as a 62% reduction in the secondary endpoint of all cause mortality (13 deaths in the vesnarinone group and 33 in the placebo group; 95% CI, 28–80; $P = .002$). Concerns about the small size of this first trial and adverse events led to the second study, the Vesnarinone Trial (VEST).[2] This study randomized 3,833 patients with CHF, New York Heart Association (NYHA) functional class III or IV, and left ventricular ejection fraction ≤30% to either placebo, 30 mg, or 60 mg vesnarinone per day and followed these patients for a maximum of 70 weeks. The primary endpoint of the study was all-cause mortality; however, the mortality rate was observed to be higher in each of the 30 mg (21.0%) or 60 mg (22.9%) groups than in the placebo group (18.9%), and the time to death was significantly shorter in the 60-mg vesnarinone group than in the placebo group ($P = .02$). The first trial, which demonstrated a mortality benefit for 60 mg vesnarinone, was reversed by the second trial, which demonstrated a mortality hazard of this same dose. In addition the beneficial findings for a reduction in the primary endpoint of the first study were also not replicated in VEST.

Losartan

The Evaluation of Losartan in the Elderly Study (ELITE) I[3] was a prospective double-blind study that randomized 722 elderly patients with symptomatic heart failure to either losartan (an angiotensin II type I receptor antagonist) or the angiotensin-converting enzyme (ACE) inhibitor captopril. The patients were followed for 48 weeks, with the primary objective of comparing persistent increases in serum creatinine in the two treatment groups. At the end of the trial the percentage of patients with an increase in serum creatinine was observed to be the same in these two treatment arms (10.5% in each group; risk reduction 2% [95% CI, −51 to 36]; $P = .63$). However, analysis for one of the prespecified secondary endpoints, all-cause mortality, demonstrated an unexpected risk reduction in mortality with losartan versus captopril (4.8% *versus* 8.7%; 17 *versus* 32 deaths; risk reduction 46% [95% CI, 5–69]; $P = .035$). A second larger study ELITE II,[4] was then carried out in a similar group of patients to confirm the superiority of losartan over captopril in improving survival in patients with heart failure. In this second study, 3,152 patients were randomized to losartan or captopril with a median follow-up of 1.5 years. However, the primary endpoint of cumulative all-cause mortality rate was not significantly different between the losartan and captopril groups (17.7% *versus* 15.9%; 280 *versus* 250 deaths; hazard ratio 1.13 [95.7% CI, 0.95–1.35]; $P = .16$). Thus losartan did not prove to be superior to captopril in improving survival in the elderly with heart failure, as had been suggested by ELITE I.

Amlodipine

The Prospective Randomized Amoldipine Survival Evaluation (PRAISE) trial[5] was designed with the primary objective of assessing the long-term effect of the calcium channel blocker amlodipine on morbidity and mortality in patients with advanced heart failure. This study randomized 1,153 patients with CHF (NYHA functional class IIIB or IV, and left ventricular ejection fraction <30%) to either amlodipine or placebo, and followed them for a maximum of 33 months. Because it was expected that amlodipine might have different effects by heart failure etiology, the randomization was stratified into ischemic (n = 732) or nonischemic (n = 421) causes of CHF. In the overall cohort there was no significant difference in the occurrence of the primary endpoint (combined risk of all-cause mortality and cardiovascular hospitalization) between the amlodipine and placebo groups (39% *versus* 42%, 9% reduction [95% CI, −24 to 10]; $P = .31$). The secondary endpoint of all-cause mortality was also not significantly different between the amlodipine and placebo groups for the overall cohort (33% *versus* 38%, 16% reduction [95% CI, −31 to 2]; $P = .07$). However, a subgroup analysis revealed that treatment with amlodipine reduced the frequency of the primary endpoint (58 fatal or nonfatal

events in the amlodipine group and 78 in the placebo group, 31% risk reduction [95% CI, 2–51% reduction], $P = .04$) and secondary endpoint (74 deaths in the placebo group and 45 in the amlodipine group, 46% reduction in risk in the amlodipine group [95% CI, 21–63% reduction]; $P < .001$) in patients with nonischemic dilated cardiomyopathy. 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. A second trial, PRAISE 2[6] was then conducted to verify the beneficial effect on mortality seen in the subgroup analysis in PRAISE 1. This trial, although focusing on only patients with heart failure of nonischemic origin, was similar in design to PRAISE 1. PRAISE 2 randomized 1,650 patients to either amlodipine or placebo, following them for up to 4 years. However, unlike in PRAISE 1, PRAISE 2 did not find a difference in mortality between the two groups (33.7% in the amlodipine arm and 31.7% in the placebo arm; risk ratio 1.09; $P = .28$). Thus the marked mortality benefit with amlodipine seen in the subgroup analysis in PRAISE 1 was not confirmed in PRAISE 2.

In each of these three examples, two prospectively designed, randomized, controlled clinical trials were executed. For each of these three pairs of studies, the first trial, carefully designed, executed at great expense, and thoroughly analyzed, identified a finding the investigators claimed was positive. In each of these studies, the investigator, recognizing the fragility of the findings from these initial studies, argued for an obtained funding for a second confirmatory study. Fortunately, they were persuasive in these arguments and a secondary study was executed. However the nullification (and, in the case of vesnarinone, the outright reversal) of each of these results by the corresponding follow-up experiments, also prospectively designed, randomized controlled clinical trials, produced confusion in the cardiovascular community. How can well-designed trials produce such disparate conclusions?

Fraught with Fragility

The major contribution of clinical trials is their ability to control the use of a therapy so that the effect of that intervention can be isolated and relatively easily measured. However, for this exercise to produce a product of lasting value, the findings in the research must be translated from the relatively small sample to a much larger population of thousands (and often millions) of patients. The heart failure population in the United States numbers more than 4 million patients and is capable of producing an almost innumerable number of samples. Because these samples contain different patients, with different life experiences, the lessons learned about CHF that apply to the population will vary from sample to sample. Unfortunately, by and large, researchers are restrained from studying many samples; they instead must choose one. And, although it is reasonable to conclude that any sample produced by the population holds part of the truth about that population, it is unreasonable to think that every fact embedded in that sample directly represents a true population finding. Unfortunately, despite the intense efforts of trialists, we can offer no guarantee to the medical community that the findings of a rigorously executed program on a single sample of patients will represent the truth about the population.

We view the population only through the spectacles of sampling error. No computational or statistical tools can remove these glasses—the best that researchers can provide to the viewing medical community is a measure of the distortion that sampling error produces. They do this by providing an estimate of how likely it is that the population produced a misleading sample for them to study. These statistical measures of error (known as type I and type II errors) provide estimates of the magnitude of sampling error. Despite the longstanding controversy that swirls around the use of these concepts,[7–19] the research, regulatory, and medical communities have come to rely on these error estimates. However, the utility of these tools is tightly circumscribed by the reliability of these error estimates. These estimates themselves can be easily damaged, and, when damaged, can provide misleading assessments. This was the case in each of the vesnarinone, ELITE, and PRAISE research programs.

Untrustworthy Estimators

When investigators competently report the effects produced by the studied intervention in the clinical research program, the event rates (eg, 4.8%), the risk reduction (eg, 46%), the CI (eg, 5–69), and the P value (eg, .035) are measured using state-of-the-art estimators that have been developed over many years by epidemiologists and biostatisticians. However, these estimators are reliable only under very clear assumptions. When those assumptions are broken, the estimators on which we rely can become volatile and misleading. Unfortunately, it is all too easy to convert trustworthy estimators into untrustworthy ones.

In traditional or fixed research, the protocol of the experiment (containing such details as the dose, duration, and analysis plan) is fixed—the only random component of the research effort is the data. Our commonly used estimators of effect size and sampling error serve us well in this circumstance for which they were designed. The alternative research environment is a random one.[20] In

this distorted environment, the data, which are random, are allowed to determine the analysis plan (e.g., when a secondary endpoint is raised to prominence by its unanticipated findings [Vesnarinone and ELITE 1] or a subgroup analysis bears particular fruit and receives a dominant place in a manuscript [PRAISE 1]). This it a hallmark of random research—the analysis plan changes as the experiment progresses. In this research environment, the common underlying assumptions are no longer valid, the statistical computations are corrupted[21] and the *P* value loses its meaning. The classic estimators, so reliable in fixed research protocols, were never designed to function in these random circumstances. Like blind guides, they become disoriented and mislead the medical community in the process.

A final illustration of the interpretative difficulties produced by trials that focus on endpoint analyses that were not declared as primary in the trial's prospective analysis plan is carvedilol. The US carvedilol program tested carvedilol against a placebo in a prospectively designed, double-blind, randomized, controlled clinical trial program. At the conclusion of approximately 1 year of follow-up, 31 deaths had occurred in 398 placebo group patients versus 22 deaths among 696 patients randomized to carvedilol (relative risk of 0.65; $P < .001$).[22] However, a US Food and Drug Administration Advisory Committee meeting focused on the fact that this program was not one clinical trial but a combined analysis of four different protocols, none of which had total mortality as a primary endpoint. This discovery produced a host of problems for the experiment's interpretation, and the advisory committee voted not to approve the drug for use in the CHF population. The interpretation of this discordant program was both complex and contentious.[23] In February 1997 these same investigators presented the results to the same committee, this time to apply for a claim that carvedilol reduced the incidence of the combined endpoint of morbidity and mortality in patients with CHF. There has been much discussion and debate about the manuscript published in the *New England Journal of Medicine* and the discussions that took place at these meetings.[24–28]

CAPRICORN (Carvedilol Post-Infarct Survival Control in LV Dysfunction)[29] was a subsequent study designed to clarify the relationship between carvedilol and total mortality. This study recruited 1,959 patients from 17 countries and 163 centers worldwide. Unfortunately, CAPRICORN had its own set of endpoint difficulties. The prespecified primary endpoint of CAPRICORN was all-cause mortality. However, the investigators changed this endpoint during the course of the trial to one that was a composite of all-cause mortality and cardiovascular hospitalizations. Also, the type I error level was reallocated so that the new composite endpoint had to have a *P* value < .045 and the all-cause mortality endpoint must have a *P* value < .005 to be considered statistically significant. However, despite this midtrial maneuvering, carvedilol failed to reach the threshold of significance for either of these endpoints. For the composite endpoint, carvedilol use was associated with a relative risk of 0.92 (95% CI, 0.80–1.07), and for total mortality the relative risk was 0.77 (95% CI, 0.60–0.98), $P = .03$. Thus the findings from the US carvedilol program were (at least from the sponsor's point of view) positive for a mortality benefit, but from CAPRICORN with its tortured analysis, the result was null, creating a substantial discrepancy.

COPERNICUS (Carvedilol Prospective Randomized Cumulative Survival Trial)[30] was an international study designed to look at the effect of carvedilol on total mortality in patients with advanced heart failure. This study was conducted in more than 300 medical centers in 21 countries and enrolled more than 2,200 patients with advanced heart failure. Patients were evaluated for up to 29 months. In COPERNICUS, patients treated with carvedilol showed a significantly lower mortality rate when compared to those treated with plabeco (11.4% *versus* 18.5%, respectively; 35% reduction in total mortality). It is difficult to compare the results of the COPERNICUS clinical trial with the US carvedilol program because the spectrum of heart failure seen in these two groups was quite different. This is an anticipated difficulty, and perhaps experienced cardiologists can deduce the appropriate metric for this comparison. However, any metric breaks down if the estimators on which this metric relies are untrustworthy.

What summary conclusions can we draw from these three clinical research programs? Analyses from the US Carvedilol program and CAPRICORN were carried out in a random analysis environment. Neither sheds useful light on the relationship between carvedilol use and total mortality. COPERNICUS was not plagued with these endpoint difficulties and should be interpreted as a positive study. We think the best conclusion to draw is that the only trial whose methodology allows a clear interpretation is COPERNICUS. Because the severity of heart failure was worse in the patients randomized in COPERNICUS than those in either the US Carvedilol program or CAPRICORN, the results of COPERNICUS cannot be extended to those patients with less severe heart failure who were recruited for the US Carvedilol program or for CAPRICORN.

The Lure of Subsidiary Analyses

Subsidiary analyses (e.g., secondary endpoints and subgroup evaluations) in clinical trials are like fire—when used carefully and conservatively, they can add constructively to our fund of knowledge; however, their casual use (and interpretation) can do great damage. A

well-considered and carefully planned subsidiary analysis can strengthen the persuasive power of the experiment, provide useful data about the underlying biologic mechanism driving the interpretation's effect, and is cost-effective. However, these advantages collide with the observation that the cumulative type I error increases with the number of statistical evaluations. Because the type I error is an important assessment of the likelihood that the medical community will be misled by the results of the study simply because of sampling error, its level must be acceptably low. Letting investigators choose their own post hoc analysis plan ("let the data speak for themselves"), without a prospective analysis statement for alpha allocation, is particularly worrisome. Keeping in mind that it is not the sample results that are paramount, but what the sample results teach us about the population that is most important, uninterpretable findings from random research cannot be integrated into our scientific fund of knowledge and are therefore inconclusive. Recent work on secondary endpoints[31–35] and criteria for subgroup analyses[36] suggests methodologically sound procedures that address these dilemma. Multiple comparisons procedures—such as Bonferroni method[37–39] for hypothesis tests that are independent, or that of Westfall[30] in the more general circumstance—are available; however, one thing is clear. Type I errors are unacceptably high when subsidiary analysis results are proclaimed as positive when there was not prospective type I error statement.

Just as type I error levels are the focus of attention when a study is reported as positive, type II error levels must be reported when studies are reported as finding no significant effect (a null study). It is possible that a population in which the intervention has a clinically important effect may produce, just through the play of chance, a research sample in which the intervention is not seen to be effective. The medical community must be assured that this error has a low probability of occurrence. Therefore researchers should evaluate the probability of this error (type II error) when the prospectively declared endpoints are not statistically significant. Type II error level reports allow the medical community to separate null or true "no effect" findings of the intervention from merely uninformative findings, which take place when the type II error is too high (ie, low power).

Of course, uncontrolled sampling error is not the only cause for disparate results between clinical trials that are designed to examine the effect of the same intervention. Characteristics can be different between the patients randomized to two different heart failure clinical trials, if the effect of the intervention depends on the comorbidity of the patient population, the two trials can come to different conclusions about the effect of the intervention. Another factor to consider in attempting to explain differences between the findings of two clinical trials is the time-dependent nature of heart failure therapy. Treatment patterns for CHF are not static over time but dynamic. Because the therapy commonly used in patients changes over time, the effect of the intervention being testing can either be reduced or amplified by the background therapy with which the intervention is concomitantly used. Clinical trials in CHF instigated in the 21st century are carried out in patients who commonly are taking a combination of digitalis, diuretic, ACE inhibitor therapy, and beta blockers. This was not the background therapy of 10 years ago. Temporal changes in ongoing therapy for heart failure can make an important difference in the identification of a therapy effect. However, even to view these effects and to draw conclusions about the relationship between the intervention and clinical endpoints of heart failure, we are assuming that sampling error has been controlled enough for us to interpret results in different trials with different population bases. To "hear the music" the background noise must be reduced.

"Searchers" Versus "*Re*searchers"

Finally, the readership must develop a new skill of discrimination. Keeping in mind that the role of the investigator is not as a "searcher" who stumbles upon an unexpected finding but as a "*re*searcher" who confirms an a priori hypothesis with scientific rigor, readers of the peer-reviewed medical research literature must separate confirmatory from exploratory analyses. Confirmatory analyses are those for which there is a prospective specification of an analysis plan in complete detail, including type I error allocations, leaving nothing in the analysis plan to be determined later by the data. This is the best way to ensure that the estimators the investigators have provided are trustworthy. Data-driven protocol deviations are a klaxon for type I and type II error aberrations and can serve only to produce preliminary, exploratory evaluations. The need for the separation of confirmatory from exploratory work is well-known[42] and should not be used to discourage investigators from publishing the findings of their work. However, it would be useful if these investigators would segregate the confirmatory results (which are based on their a priori hypothesis testing strategies for which they have prospectively assigned type I error) from less rigorously defined analyses that are exploratory. These latter exploratory analyses raise important questions, but require follow-up studies to confirm or reverse them. As is the case with VEST, ELITE 2, and PRAISE 2, the exploratory findings that reached prominence in the first studies could not be confirmed.

Any review of this issue must acknowledge the effort of the investigators in each of these studies. Had they not

appreciated the weak methodologic support for the findings from the first of each of these three pairs of studies, no additional confirmatory data would have been collected, and the heart failure community would have been left with the findings of the first vesnarinone studies, ELITE 1 and PRAISE 1, as the final research effort for these interventions. We should be encouraged that the heart failure community was willing to invest more precious resources into the execution of a more definitive study to get the correct results.

The readership would have been helped by the statement in each of these articles that the findings were exploratory in nature. The inclusion of a subsection of the results section, suitably titled, that deals exclusively with exploratory or hypothesis generating results would be very useful. Gladly, the investigators of all three research programs sought confirmation of the findings from the early trials, exerting the research discipline to carry out their additional studies. The rest of us in the medical community should now serve notice that this additional second effort must be the rule and not the exception when exploratory "discoveries" are announced. Although applauding these workers for their disciplined approach to these problems, we must exhort ourselves to seek confirmation of early findings in clinical trials whose job it is to raise and not answer questions.

Clinical trial interpretation requires judgment, and the balancing effort in which we all engage as we weigh a study's strengths against its weaknesses remains a central one. However, just as justice cannot prevail in the absence of the rule of law, causality determinations in clinical trials require the rules of methodology. The minimal rule that validates the estimators of the effect of the clinical trial's intervention is the presence of a prospective, fixed analysis plan. Allowing data-driven analyses to determine a clinical trial's results, however well-intentioned, strikes at the heart of the medical community's ability to generalize results from the research sample to the population at large.

References

1. Feldman AM, Bristow MR, Parmley WW, Carson PE, Pepine CJ, Gilbert EM, Srobeck JE, Hendriz GH, Powers ER, Bain RP, White BG, for the Vesnarinone Study Group: Effects of vesnarinone on morbidity and mortality in patients with heart failure. N Engl J Med 1993;329: 149–155
2. Cohn J, Goldstein SC, Greenberg BH, Lorell BH, Bourge RC, Jaski Be, Gottlieb SO, McGrew F. 3rd, DeMets DL, White BG, for the Vesnarinone Trial Investigators: A dose dependent increase in mortality seen with vesnarinone among patients with severe heart failure. N Engl J Med 1998;339:1810–1816
3. Pitt B, Segal R, Martinez FA, Meurers G, Cowley AJ, Thomas I, Deedwania PC, Ney DE, Snavely DB, Chang PI, 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
4. Pitt B, Poole-Wilson PA, Segal R, Martinez FA, Dickstein K, Camm AJ, Konstam MA, Riegger G, Klinger GH, Neaton J, Sharma D, Thiyagaraja B, on behalf of the ELITE II Investigators: Effect of losartan compared with captopril on mortality in patients with symptomatic heart failure: randomized trial—the losartan heart failure survivial study ELITE II. Lancet 2000;355:1582–1587
5. Packer M, O'Connor CM, Ghali JK, Pressler ML, Carson PE, Belkin RN, Miller AB, Neuberg GW, Frid D, Wertheimer JH, Cropp AB, DeMets DL, 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
6. 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.
7. Birnbaum A: On the foundations of statistical inference. J Amer Statist Assoc 1962;57:269–306
8. Berkson J: Experiences with tests of significance. A reply to R.A. Fisher. J Amer Statist Assoc 1942;37:242–246
9. Berkson J: Tests of significance considered as evidence. J Amer Statist Assoc 1942;37:335–345
10. Fisher RA: Response to Berkson. J Amer Statis Assoc 1942;37:103–104
11. Poole C: Beyond the confidence interval. Amer J Public Health 1987;77:195–199
12. Poole C: Feelings and frequencies: two kinds of probability in public health research. Amer J Public Health 1988;78:1531–1533
13. Gardner MJ, Altman DG: Confidence intervals rather than p values. Estimation rather than hypothesis testing. BMJ 1986;292:746–750
14. Fleiss JL: Significance tests have a role in epidemiologic research; reactions to A.M. Walker (Different Views). Amer J Public Health 1986;76:559–560
15. Fleiss JL: Confidence intervals vs. significance tests: quantitative interpretation (letter). Amer J Public Health 1986;76:587.
16. Fleiss JL: Dr. Fleiss response (letter). Amer J Public Health 1986;76:1033–1034
17. Walker AM: Reporting the results of epidemiologic studies. Amer J Public Health 1986;76:556–558
18. Walker AM: Significance tests [sic] represent consensus a and standard practice (letter) Amer J Public Health 1986;76:1033 (erratum 1986;76:1087)
19. Lang JM, Rothman KL, Cann CL: That confounded p value. Epidemiology 1998;9:7–8
20. Moyé L: Random research. Circulation 2001;103:3150–3153
21. Moyé LA: Statistical reasoning in medicine—The intuitive P value primer. Spingler-Verlag, New York, 2000

22. 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
23. Transcript for the May 2, 1996 Cardiovascular and Renal Drugs Advisory Committee
24. Moyé LA, Abernathy D: Carvedilol in patients with chronic heart failure (letter). N Engl J Med 1996;335: 1318–1319
25. Packer M, Cohn JN, Colucci WS: Response to Moyé and Abernathy. N Engl J Med 1996;335:1318–1319
26. Fisher LD, Moyé LA: Carvedilol and the Food and Drug Administration approval process: an introduction. Contr Clin Trials 1999;20:1–15
27. Fisher L: Carvedilol and the FDA approval process: the FDA paradigm and reflections upon hypothses testing. Contr Clin Trials 1999;20:16–39
28. Moyé LA; *P* value interpretation in clinical trials: the case for discipline. *Controlled Clin Trials* 1999;20:40–49
29. The CAPRICORN Investigators: Effect of carvedilol on outcome after myocardial infarction in patients with left-ventricular dysfunction: the CAPRICORN randomised trial. *Lancet* 2001;357:1385–1390
30. Packer M, Coats AJS, Fowler MB, Katus HA, Krum H, Mohacsi P, Rouleau JL, Tendera M, Castaigne Al, Roecker EB, Schultz MK, DeMets DL, for the Carvedolol Prospective Randomized Cumulative Survival Study Group: Effect of carvedilol on survival in severe chronic heart failure. N Eng J Med 2001;344:1651–8.
31 D'Agostino RB: Controlling alpha in clinical trials: the case for secondary endpoints. Statist Med 2000;19:763–766
32 Moyé LA: Alpha calculus in clinical trials: considerations and commentary for the new millenium. Statist Med 2000;19:767–779
33 Koch GG: Discussion for 'Alpha calculus in clinical trials: considerations and commentary for the new millennium.' Statis Med 2000;19:781–784
34. O'Neill RT. Commentary on 'Alpha calculus in clinical trials: considerations and commentary for the new millennium' Statis Med 2000;19:785–793.
35. Moyé LA: Alpha calculus in clinical trials: considerations and commentary for the new millenium (Rejoinder). Statis Med 2000;19:767–779
36. 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
37. OJ Dunn: Confidence intervals for the means of dependent, normally distributed variables. Journal of the American Statistical Assocation 1959;54:613–621
38. Dunn OJ: Multiple comparisons among means. Journal of the American Statistical Assocation. 1961;56:52–54
39. Westfall PH, Young SS, Wright SP: An adjusting p values for multiplicity. Biometrics 1993;49:941–945
40. Friedman L, Furberg C, DeMets D: Fundamentals of clinical trials, 3rd ed. Springer-Verlag, New York, 1998.