

Circulation

JOURNAL OF THE AMERICAN HEART ASSOCIATION



Random Research

Lemuel A. Moyé

Circulation 2001;103;3150-3153

DOI: 10.1161/hc2501.090955

Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75214
Copyright © 2001 American Heart Association. All rights reserved. Print ISSN: 0009-7322. Online ISSN:
1524-4539

The online version of this article, along with updated information and services, is located on the
World Wide Web at:

<http://circ.ahajournals.org/cgi/content/full/103/25/3150>

Subscriptions: Information about subscribing to *Circulation* is online at
<http://circ.ahajournals.org/subscriptions/>

Permissions: Permissions & Rights Desk, Lippincott Williams & Wilkins, a division of Wolters Kluwer
Health, 351 West Camden Street, Baltimore, MD 21202-2436. Phone: 410-528-4050. Fax: 410-528-8550.

E-mail:
journalpermissions@lww.com

Reprints: Information about reprints can be found online at
<http://www.lww.com/reprints>

Random Research

Lemuel A. Moyé, MD, PhD

Abstract—Advances in computing have combined with the rapid dissemination of treatment discoveries for diseases of public health importance to create pressure for accelerated promulgation of promising research results to the medical community. The 2 recent examples of the US Carvedilol Heart Failure program and the Evaluation of Losartan In the Elderly (ELITE) study demonstrate the importance of the prospective nature of research design, as well as the consequences of its abandonment. This article explains in nonmathematical terms the rationale for the tenet “first say what you will do, then do what you said” in sample-based research. (*Circulation*. 2001;103:3150-3153.)

Key Words: trials ■ statistics ■ population ■ epidemiology

Clinical research programs that are designed to demonstrate a finding for one end point can provide surprising results for another. In 1996, the US Carvedilol Heart Failure program,¹ designed to examine the effect of carvedilol on morbidity in patients with congestive heart failure, demonstrated a 65% reduction in total mortality. Likewise, the Evaluation of Losartan in the Elderly (ELITE) study² was executed to compare the relative effects of losartan and captopril on the renal function of elderly patients with congestive heart failure and demonstrated a 46% reduction in total mortality associated with losartan. Each of these efforts was a double-blind, controlled clinical trial. In each study, the results represented findings of tremendous statistical significance. Nevertheless, pundits have suggested caution in the interpretation of these results,³⁻⁹ and the nonstatistically oriented reader is easily perplexed by the caution advocated by these experts. In each of these 2 circumstances, the findings are unmistakably plain, and the data clearly show an effect of therapy. Why should this data not be taken at face value? Why does the arbitrary insistence on the prospective end point statement weaken the findings of these trials? The purpose of this discussion is to describe in nonmathematical terms the rationale for the prospective nature of research.

The Sampling Nemesis

Ideally, the ambition of the researcher is to study the disease of interest in every patient in the total population. Faced with the impossibility of this task, the researcher compromises; she studies not the entire population, but rather a sample of the population. The researcher now has the resources to carry out this circumscribed research effort. In drawing only a sample, however, the researcher gives up a critical element in drawing research conclusions—certainty. She acknowledges that other equally diligent and motivated researchers can collect

their own independent samples. Because these samples contain different patients with different experiences and different data, how can we decide which one is most representative of the total population?

This difference in the data across samples is sampling variability, and its presence raises the very real possibility that the sample, merely through the play of chance, will mislead the researcher about the characteristics of the total population. This can happen even when the investigator uses modern sampling techniques to obtain a representative sample. In the end, the investigator has one and only one sample, and therefore, she can never be sure whether conclusions drawn from the sample data truly characterize the entire population.

In sample-based research, researchers cede the high ground of certainty. We do insist, however, that bounds be placed on the size of the sampling errors that can be produced from sample-based conclusions. These sampling errors are commonly known as type I and type II errors. If the researcher identifies the effect of therapy in his sample, the researcher must prepare an answer to the question, “How likely is it that the total population, in which there was no therapy effect, produced a sample that experienced a therapy effect?” This is the type I error, addressed by the *P* value. The reverse sampling error, in which the total population does experience a therapy effect but produces a sample with no therapy effect, is the type II error. These are solely measurements of sampling error and must be combined with the sample size, effect size (eg, relative risk), and the effect size variability in order to successfully convey the results of the research. However, in order for the medical community to successfully draw a conclusion about the pertinence of these findings to the larger population from which the sample was taken, these estimates

From the University of Texas School of Public Health, Houston, Tex.

Guest Editor for this article was Robert M. Califf, MD, Duke Clinical Research Institute, Durham, NC.

Correspondence to Lemuel A. Moyé, MD, PhD, University of Texas School of Public Health, RAS Building E-815, 1200 Herman Pressler, Houston, TX 77030. E-mail lmoye@utsph.sph.uth.tmc.edu
(*Circulation*. 2001;103:3150-3153.)

© 2001 American Heart Association, Inc.

Circulation is available at <http://www.circulationaha.org>

must be accurate. They are accurate only when the sole source of variability in the experiment is the data.

If executed correctly, a research program will obtain its sample of data randomly and will lead to accurate statistical computations. When more than the data is random, however, the underlying assumptions are no longer in play, and the statistical computations are corrupted—they are incorrect and cannot be corrected. Consider the following example.

Troubled Enthusiasm and End Points

An enthusiastic young researcher, Dr C, is interested in demonstrating the progressive deterioration in heart function observed in patients with congestive heart failure. She designs a simple research program to examine the change in left ventricular end-diastolic volume (EDV) over time. Dr C recruits a sample of patients and follows them for 2 years, measuring heart function at baseline and at the study's end using echocardiography. Although Dr C is focused on the change in EDV, she also collects other measurements of left ventricular dysfunction, such as end-systolic volume (ESV), stroke volume, and cardiac output. At the study's conclusion, Dr C discovers, to her surprise and horror, that there has not been an important increase in EDV. However, ESV has increased substantially. She therefore decides to change the end point for the program from EDV to ESV and reports the latter findings.

Many researchers would have no problem with Dr C's EDV-to-ESV end point switch, arguing that each end point is a measurement of the same underlying physiology and pathophysiology. Why should Dr C be criticized for making an initial wrong guess about the best end point to choose? Because she had the insight to measure several different indicators of left ventricular function, perhaps she should be commended for (1) her foresight in measuring ESV and (2) her courage in raising the significant ESV result to a prominent place in her report. Others among us would be uncomfortable with the end point change but may be uncertain as to exactly what the problem is. We might say that the decision to change the end point was "data driven." Well, what's so wrong with that? Aren't the results of any study data driven?

Random Research Versus Anchored Research

What is wrong with Dr C's analysis is not that the data are random—this we expect. What is wrong is that the research effort itself is random. The initial idea was to execute a fixed, anchored research plan that would accept random data. If this were the case, statisticians and epidemiologists would know how to compute relative risks with standard errors, confidence intervals, and *P* values accurately. In this familiar setting, the random data component would be appropriately channeled to type I and type II error for the medical community's interpretation.

The research as Dr C chose to execute it has a new, random component. The choice of the end point was a random selection—the data dictated which end point would be chosen. Of course, the data always contribute to the end point, but, in this unfortunate case, the data chose

the end point. Because the data are random, the end point selection process is now random. A different sample might have led to a different choice of end point. This new source of variation wrecks the standard computations of relative risks, standard deviations, confidence intervals, and type I/II errors because the underlying assumptions have altered. These quantities are still computed, but the computations can no longer be trusted.

There are additional problems with changing the end point from EDV to ESV. Magnitude of effect, variability of the estimate, and required sample size would be different, as well. Although logistically important, these concerns are not the focus of this discussion.

The reader should note that the problem introduced by random research is not one of sloppy calculations. On the contrary, great effort is expended on these computations, using state-of-the-art statistical packages. When a statistician is asked to consider the development of a test statistic for the change in EDV, he commonly begins by saying, "Let *x* represent the EDV change. Then *x* has the following probability distribution. . . ." From these statements, estimators of effect size, standard errors, confidence intervals, and probability value are available to effectively convey the strength of evidence the data contains. However, this paradigm falls apart when *x* is not prospectively and arbitrarily chosen but is instead selected by the data, which itself contains sampling error. In this second (random) paradigm, there is a new probability distribution that governs the selection of the EDV variable itself. This change in assumption results in a more complicated second paradigm in which our commonly used, familiar estimators are no longer the best. Therefore, estimators developed for the first (arbitrary, prospective) paradigm are useless when the paradigm has shifted to the second (random) one. The only protection against this dilemma is the prospective specification of the end points in complete detail, leaving nothing to chance. This is the central motivation for the tenet "first say what you will do, then do what you said" in research.

This dilemma is resolved at once if the researcher is able to study the entire population. If a researcher is interested in identifying the effect of therapy in a community hospital for a short time period with no desire to extend the findings to another hospital or to a future time, there is no concern for estimate trustworthiness. As long as the conclusions will be applied only to those patients in the hospital at that point in time (ie, the sample equals the total population), the notion of prospective determination, from a sampling point of view, is not necessary. However, the freedom gained in the end point selection process by studying the entire population is counterbalanced by the inability to generalize.

Clinical Trials With Random End Points

The application of this principle of prospective planning to clinical research is not complicated. If the decisions to select the end point and report the data are made on the basis of the data itself (as opposed to a clear, prospective statement in the research protocol), the research is random, the assumptions

underlying the computations are altered, and our commonly used statistical estimates are corrupted and uncorrectable.

The US Carvedilol Heart Failure program and the ELITE study have each been afflicted with the same source of debilitating variability. In each of these efforts, morbidity end points were chosen prospectively. In each case, these end points were upstaged by other findings. In ELITE, the authors stated that "the study demonstrated that losartan reduced mortality compared with captopril."² It is important in this discussion to focus on the purpose of the implication. There is no doubt that the use of carvedilol is associated with reduced mortality in the US Carvedilol Heart Failure program sample. Analogously, ELITE clearly demonstrated that in its sample of patients, losartan was associated with fewer deaths. However, because many findings in a research sample cannot be extended to the larger population, the scientific community must carefully consider which findings should be extended and which should be characterized as inconclusive. Statistical measurements are useful tools in this analysis. The random research component in both of these studies makes these tools unreliable.

Additional investigations involving carvedilol are underway. The Carvedilol Prospective Randomized Cumulative Survival (COPERNICUS) trial is a multicenter, randomized, double-blind, placebo-controlled study to determine the effect of carvedilol on mortality in patients with severe chronic heart failure. This study has the prospectively selected primary end point of total mortality and will permit a more definitive conclusion about the magnitude of the effects of carvedilol than did the US Carvedilol Heart Failure study.

In the case of ELITE, additional information is available. The Losartan Heart Failure Survival Study (ELITE II)¹⁰ examined the effect of losartan and captopril in 3152 patients aged ≥ 60 years with NYHA Class II-IV heart failure. This study, designed to formally test the hypothesis that losartan could reduce mortality, concluded that there was no mortality effect in the population. Possible reasons for the differences in the findings between ELITE II and the original ELITE could be difference in age (the mean age in ELITE II was lower) and differences in the use of concomitant medications. In addition, more patients had NYHA Class III-IV heart failure in ELITE II than in the earlier ELITE. However, the explanation for the differences in the studies' results should begin with consideration of the sampling error. Measurements of sampling error are useless in the original ELITE. In ELITE II, the end point selection was fixed and not subject to change after an examination of the data. Thus, the research is well anchored and the statistical error measurements are accurate in ELITE II.

With most clinical investigations, we do not have the luxury of an ELITE II to point out the erroneous conclusions of prior studies. It is more likely that we will have only one study on which to base our opinions. The best we can do in such cases is to ensure that the measurements of statistical error are reliable. In random research, no matter

how carefully calculated, these measurements cannot be trusted and must be discarded.

Sampling Error and Multiple End Points

The role of sampling error certainly complicates the issue of multiple-end point interpretation. Clinical research programs, for carefully considered cost-effectiveness and epidemiological reasons, will often include multiple end points. The interpretation of these results requires great care. Of course, if the additional end points are chosen on the basis of the data observed by the trial, the results are corrupted and cannot be interpreted. However, even when all of the end points are prospectively chosen, there is still difficulty in their individual interpretation. This is especially vexing when a single primary end point for the study is not statistically significant, but at least one secondary end point is statistically significant.

This profound difficulty has elicited recent discussion.¹¹⁻¹⁵ The major recommendations from this body of work are (1) to require that each primary end point and each of the secondary end points have type I error attached in a prospective and reasoned fashion, (2) to separate the type I error allocated for the primary end point from that allocated for the experiment, and (3) to allow the total experimental type I error to be >0.05 while requiring that the type I error for the primary outcome be ≤ 0.05 . This collection of procedures increases the rigor for the prospective statements concerning secondary end points and allows for the straightforward interpretation of a research effort that is positive for secondary end points when the primary end point is not statistically significant. Although no consensus has yet been reached, important new dialogue is well underway.

Conclusions

Sampling error is a necessary evil in research programs. Left unchecked, unreduced, and unchanneled, it will wreck the research plan, making the study's conclusions uninterpretable. Vigilant investigators must remain on the lookout for sampling error's ability to corrupt the experiment. Nonprospective, random program execution will open the study to the corrosive effects of sampling error, turning the program into an uninterpretable husk. As we interpret a research program's *P* value, we should be on the alert for changes in trial execution. Such data-driven protocol deviations are klaxons for type I and II error contamination. Investigators should simply plan what they wish to do, then do precisely what they planned.

References

1. 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.
2. Pitt B, Segal R, Martinez FA, et al. Randomised trial of losartan versus captopril in patients over 65 with heart failure (Evaluation of Losartan in the Elderly Study, ELITE). *Lancet*. 1997;349:747-752.
3. Friedman L, Furberg C, DeMets D. *Fundamentals of Clinical Trials*. 3rd ed. St. Louis, MO: Mosby; 1996:307-308.

4. Meinert CL. *Clinical Trials: Design, Conduct, and Analysis*. New York, NY: Oxford University Press; 1986:214–215.
5. Moyé LA, Abernethy D. Carvedilol in patients with chronic heart failure. *N Engl J Med*. 1996;335:1318–1319. Letter.
6. Fisher LD. Carvedilol and the Food and Drug Administration (FDA) approval process: the FDA paradigm and reflections on hypothesis testing. *Control Clin Trials*. 1999;20:16–39.
7. Moyé LA. End-point interpretation in clinical trials: the case for discipline. *Control Clin Trials*. 1999;20:40–49.
8. Packer M, Cohn JN, Colucci WS. Response to Moyé and Abernethy. *N Engl J Med*. 1996;335:1318–1319.
9. Moyé LA. *Statistical Reasoning in Medicine: The Intuitive P value Primer*. New York, NY: Springer-Verlag; 2000. Chapters 7 and 8.
10. 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.
11. D’Agostino RB. Controlling alpha in a clinical trial: the case for secondary endpoints. *Stat Med*. 2000;19:763–766.
12. Moyé LA. Alpha calculus in clinical trials: considerations and commentary for the new millennium. *Stat Med*. 2000;19:767–779.
13. Koch GG. Discussion for ‘Alpha calculus in clinical trials: considerations and commentary for the new millennium’. *Stat Med*. 2000;19:781–784.
14. O’Neill RT. Commentary on ‘Alpha calculus in clinical trials: considerations and commentary for the new millennium’. *Stat Med*. 2000;19:785–793.
15. Moyé LA. Alpha calculus in clinical trials: considerations and commentary for the new millennium: rejoinder. *Stat Med*. 2000;19:767–779.