ELSEVIER

P-Value Interpretation and Alpha Allocation in Clinical Trials

LEMUEL A. MOYÉ, MD, PHD

PURPOSE: Although much value has been placed on type I error event probabilities in clinical trials, interpretive difficulties often arise that are directly related to clinical trial complexity. Deviations of the trial execution from its protocol, the presence of multiple treatment arms, and the inclusion of multiple end points complicate the interpretation of an experiment's reported alpha level. The purpose of this manuscript is to formulate the discussion of *P* values (and power for studies showing no significant differences) on the basis of the event whose relative frequency they represent.
METHODS: Experimental discordance (discrepancies between the protocol's directives and the experiment's execution) is linked to difficulty in alpha and beta interpretation. Mild experimental discordance leads to an acceptable adjustment for alpha or beta, while severe discordance results in their corruption.
RESULTS: Finally, guidelines are provided for allocating type I error among a collection of end points in a prospectively designed, randomized controlled clinical trial.
CONCLUSIONS: When considering secondary end point inclusion in clinical trials, investigators should increase the sample size to preserve the type I error rates at acceptable levels.
Ann Epidemiol 1998;8:351–357.

KEY WORDS: Type I Error, Experimental Discordance, Alpha Allocation, Clinical Trials.

INTRODUCTION

Investigators are obligated to design and execute their research programs to protect both the patients included in the study and the population to whom the experiments will be generalized. This later responsibility has become complicated in experiments that measure multiple outcomes (1–3) or can have multiple treatment arms (4). In addition, investigators perform post hoc analyses that were not prospectively planned (5–7). Even in simulation experiments, p-value interpretation can be challenging (8). However, investigators are nevertheless obligated to design and execute their research programs so that the type I and type II errors are easily interpreted.

The purpose of this manuscript is to provide guidelines for allocating type I error among a collection of end points in a prospectively designed experimental program. Experimental discordance is defined, and a distinction is made between an adjustment of alpha and its corruption.

FOUNDATION OF ALPHA LEVELS IN EXPERIMENTAL SETTINGS

Clinical investigators have two obligations in clinical trials: patient protection and population protection. Investigators share the responsibility to ensure that the results of their research do not mislead. This responsibility is administered through controlling sampling error (*P* values and type II error rates) and ensuring that the research program is executed as planned.

The alpha level is a probability and, in order to understand it, we must understand the event whose relative frequency it represents. Alpha levels are traditionally described as conditional probabilities, e.g., the probability that the test statistic falls in the critical region given the condition that the null hypothesis is true. However, the motivation of this mathematical definition lies in the use of population sampling. The strategy of drawing a random sample from the population at large makes the experiment executable, since logistics preclude including all patients in the population in the trial. However, the investigators must pay a price for this executability; that price is the recognition that another sample with a different set of subjects drawn from the same population might yield different results. The variability of samples generated by a population and the variability of results produced by these samples is sampling error. Although populations can generate many different samples, under correct sampling plans, most of the samples will be representative of the population. However, despite the investigators' best efforts, the population may have "dealt them a bad hand," i.e., provided a sample whose findings of efficacy will not reflect the findings in the population at large. This sampling error is dangerous to the integrity of the experiment and, if handled inappropriately, critically weakens the investigators' abilities to generalize their find-

From the Department of Biometry, University of Texas School of Public Health, Houston, Texas.
Address reprint requests to: Dr. Lemuel A. Moyé, E815, University of Texas School of Public Health, RAS Building, 1200 Herman Pressler, Houston, Texas 77030.
Received May 20, 1997; accepted January 5, 1998.

1047-2797/98/$–see front matter
PII S1047-2797(98)00003-9

AEP Vol. 8, No. 6
August 1998: 351–357

Selected Abbreviations and Acronyms

NSABP = national surgical adjuvant breast and bowel project
CHF = congestive heart failure
PTCA = percutaneous transluminal coronary angioplasty

ings to the larger population from which the sample was drawn. The *P* value handles this sampling error by linking the experiment's outcome to the types of samples the population may have produced. Assume an experiment is executed and results in a 15% efficacy for the intervention. The alpha level is the probability that the population from which the sample of the experiment was randomly drawn derives no therapeutic benefit but misleads the investigators by producing an unrepresentative sample with 15% efficacy. We term this event the type I error event. The likelihood of this event is measured by the *p*-value and is what we hope to minimize.

EXPERIMENTAL CONCORDANCE VS. EXPERIMENTAL DISCORDANCE

The *P* value of an experiment is relevant to the scientific community if the *P* value is the final repository of the experiment's sampling error, i.e., it is the result of a well-defined, prospectively fixed experiment whose only random component is the data itself. This is not the case when the investigators allow sampling error to affect the conduct of an experiment. If the data generated by the experiment are allowed to influence the experiment (i.e, leading the investigators to change the experiment's end point), then the experiment has an unanticipated random component. Since the random data have been allowed to transmit randomness to the analysis plan and the research program's results, the *P* value is meaningless since it is the result not only of random data, but also of a random analysis plan. The *P* value is germane only if it represents a population-based sampling event of interest to the scientific community. This occurs when the experiment is executed according to protocol, defined here as experimental concordance. With experimental concordance, the type I error event of the experiment answers the question raised by the research community (as stated in the protocol). When the data alter the experiment's execution, and discrepancies are created between the protocol's plan of operation and the experiment's actual execution (experimental discordance), the type I error event for the experiment may not be the type I error event directed by the protocol. If the discordance is mild, the experiment's type I error event remains pertinent to the medical community and the experiment's *P* value may be adjusted. However, if there is severe or profound experimental discordance, the experimental *P* value may be of little value. Severe experimental discordance can unfortunately be lethal to the interpretation of the results of a research program because it produces a random experiment with an uninterpretable corrupt *P* value. In a negative study, in which attention often focuses on power, the beta error is corrupted as a result of experimental discordance.

MILD EXPERIMENTAL DISCORDANCE: PROGRAM *P* VALUE ADJUSTMENT

Most research programs contain some experimental discordance. I define mild experimental discordance as the presence of discrepancies between the protocol and the actual program execution that still allow the research program's type I error rate to remain relevant to the scientific community. In this circumstance, the experimental *P* value can be adjusted easily.

Example 1: Mild Discordance

Consider the protocol for a prospectively designed, double-blind, randomized controlled clinical trial to assess an intervention for improving survival in patients with myocardial infarction. This experiment requires 2182 patients to demonstrate with 80% power and a two-sided alpha level of 0.05 a 20% relative reduction in total mortality from a cumulative placebo event rate of 25%. The per protocol type I error event is that the population with a cumulative event rate of 25% derives no efficacy from the therapy of interest, but misleads the investigators by producing an unrepresentative sample of 2182 patients in which 20% efficacy is demonstrated. However, suppose only 2000 patients are randomized, representing a discrepancy with the experimental plan. Our intuition tells us that this experiment (if positive) it still interpretable, since the type I error event of the experiment (the population derives no efficacy, but misleads us by producing a sample of 2000 patients in which a 20% efficacy is seen) is so close to the type I error event of the protocol as to render the trial's type I error event still meaningful and relevant. The protocol *P* value is replaced by the trial *P* value through use of the trial sample size in the computation of the test statistic and resultant type I error event rate. This example is typically considered as raising the issue of power. Of course, the power of the experiment is reduced, but if the trial is positive, the similarity of the experimental and protocol type I error event keeps the interpretation of the *P* value clear.

Example 2: Mild Discordance

Consider a research program designed to assess the impact of various nonpharmacologic and pharmacologic measures to control blood pressure. In order to have the widest possible generalizability of these findings, the investigators state

in the protocol their prospective plan to select a simple random sample 60% of which will consist of patients of African-American descent. The investigators realize that they will to have adequate power to carry out formal hypothesis tests in this subset, but they desire to have a credible argument for extending the findings of the trial to the African-American population. However, during the execution of the trial, despite the investigators' best efforts, their sample is only 40% African-American. Although many readers would decry the shortfall, in general there would be no objections to extending the findings of the experiment to African-American patients. The difference between the protocol type I error event (which concerned a sample in which 60% of its participants were African-American) and the executed type I error event (a sample for which 40% of its members were of African-American descent) is small.

SEVERE DISCORDANCE

Severe experimental discordance describes the situation where the experimental type I or type II error event is so different from the protocol type I error event rate that the experiment is rendered meaningless and irrelevant to the medical community. This unfortunate state of affairs can be produced by the flawed execution of a well-written protocol. The situation can be so complicated that, in the case of a positive study, the value of the experimental type I error cannot be computed without controversy. I use the term alpha corruption to signify that the protocol type I error cannot be approximated by the experimental type I error. In a negative study, where the issue of power is the critical consideration, beta corruption is the relevant term. Experiments that lead to the corruption of alpha and beta are essentially useless to the research community.

An example of severe experimental discordance would be the following clinical trial designed to assess the effect of an intervention on patients who have established heart failure at the time of randomization. Patients are randomized to one of placebo or active intervention. The endpoint of the trial is progression of heart failure as measured by change in background medication status (e.g., increase in diuretic use or the addition of angiotensin-converting enzyme inhibitor therapy) at the trial's end. According to the protocol, 482 patients are required to detect a 30% relative reduction with 80% power and two-sided alpha error rate of 0.05 from a 40% cumulative incidence rate of progression of heart failure. Type I error event would occur if the population of heart failure patients with a cumulative heart failure progression rate of 40% derived no benefit from therapy but misleads the investigators by producing an unrepresentative sample of 482 patients who demonstrate 30% efficacy. However, during the trial's execution, 60% of the patients have missing medical records, allowing end point computation on only the remainder. In this case, there would be disagreement on the computation of the trial's type I error since different assumptions would lead to different p-values. The experimental type I error event would occur if the population of heart failure patients with a cumulative heart failure progression rate of 40% derived no benefit from therapy but produced an unrepresentative sample of 482 patients in which 1) only 40% have available CHF progression information and 2) a 30% efficacy on the remaining 60%. This event is of little interest to the scientific community. The alpha corruption has profoundly blurred any clear interpretation of the study.

A more problematic case would be a two-armed clinical trial that has a statistically significant P value for efficacy against a total mortality end point but that violates the protocol by losing 15% of its randomized patients to follow-up. The implications of this discordance are substantial because the follow-up losses blur our view of the therapy efficacy within the sample. The degree of discordance here depends on the size of the P value. If the P value remained below the threshold of significance when we assume all lost patients assigned to the placebo group were alive and all lost patients in the active group died, we conclude that the discordance is mild because the worst implications of the losses to follow-up do not vitiate the results, and the type I error event is still relevant to the scientific community. However, if the P value changes in significance, we may say that the discordance is severe and the alpha error is corrupted.

As a final example of discordance, consider the findings of the NSABP protocol B-06 (9, 10), in which 99 ineligible patients were deliberately randomized with falsified data. To assess the experimental discordance produced by this event, we first examine of the trial results by excluding these 99 patients. If their exclusion moved the P value across the significance threshold, the inclusion of these patients produces unacceptable discordance. However, a second relevant question on the interpretation of the trial must be addressed since the presence of fraudulent data admits the possibility of dishonest behavior elsewhere in the trial apparatus. However, the audit of Christian and colleagues (11) may be considered an investigation of the degree of discordance. Since the protocol discrepancies identified were small in number, the degree of discordance was mild.

IMPORTANCE OF PROSPECTIVE EXPERIMENTAL DESIGN

Successful experiments protect experimental concordance. This concordance is easily lost if data collected during the experiment are allowed to affect decisions in an unplanned manner concerning the experiment's outcome, thus allowing sampling error to besmirch protocol-mandated procedures.

AEP Vol. 8, No. 6
August 1998: 351–357

Since sampling error is a necessary evil in clinical trials, it must be handled with care, ensuring that it does not contaminate the experiment. The one acceptable repository for sampling error is the type I and type II event probabilities.

The correct containment of sampling error is assured by preventing trial execution decisions from being affected by the data of the experiment, thereby insulating the experiment from the sampling error. A superior procedure that helps to assure ease of alpha and beta interpretation is the investment of investigator time into the development of the research protocol (e.g., knowing the population from which the sample is to be randomized and having the patience to carry out a pilot study to ascertain what is unknown but required for the successful execution of the trial). Once the protocol is written and accepted, investigators must insist on nothing less than its rigorous execution. The protocol is the rule book of the trial, and it must be strictly adhered to by all.

Protocol development requires the clear statement of end points. Since this has implications for the type I error event rate, the following is presented as a guideline to investigators in prospectively setting alpha levels.

PROSPECTIVE ALLOCATION OF ALPHA

The two conflicting forces of end point abundance (the desire to measure many different assessments at the end of the experiment) vs. interpretive parsimony (the alpha level and therefore the success of the trial rest on the interpretation of one and only one end point) bedevil investigators as they plan their experiments. Guidance for the selection of end points in clinical trials is available (12). Motivations for secondary and tertiary end points involve both epidemiological considerations (coherency within the trial and consistency across trials) and cost-efficiency. How can all of this information on nonprimary end points be interpreted when the medical and scientific community focus on the experimental P value for the primary endpoint of a trial? What is the correct interpretation of a trial which produces negative results for its primary end point, but has nominally statistically significant secondary end points? If a trial has two active arms and one placebo arm, must a single comparison take precedence? Various strategies for interpretation of the family of P values from an experiment and various multiple comparison procedures are available (12–18). The concept presented here is an adaptation of multiple testing, taking advantage of the ability to set the multiple comparison values at different levels. It should be used only for prospectively determined end points (i.e., formal hypothesis testing), as opposed to hypothesis generation, a more inquisitive investigation, designed to identify relationships not anticipated before the beginning of the experiment.

ALPHA ALLOCATION

Type I error accumulates with each executed hypothesis test and must be controlled by the investigators. By the prospective selection of the alpha levels, the investigators set the standard by which the trial will be judged. P values for post hoc analyses are uninterpretable within a clinical trial setting since (*i*) being data driven (as opposed to protocol driven), they are intertwined with sampling error, and (*ii*) many post hoc analyses can be performed with only the most favorable ones being promulgated, leading to hidden alpha accumulation. Post hoc testing produces severe experimental discordance with corrupt alpha levels. Research investigators avoid this serious limitation of analysis interpretation by choosing end points prospectively. However, they strengthen their experimental design further by also choosing the allocation of the type I error. Alpha should be allocated to protocol-specified hypotheses and protocol-specified subsgroup analyses. If the subgroup analysis is based on findings of other trials, alpha can still be allocated if (*i*) no data from the current experiment are used in the examination of alpha; and (*ii*) the additional subgroup analysis is added with appropriate changes in the decision path before the end of the study.

The following is a guide to investigators on the apportionment of alpha during the design phase of a research program. Commonly, the scientific community considers an alpha level for each hypothesis test to be examined in a research program, leaving the interpretation of the overall alpha to the reader who is attempting to interpret the significance of the study. Consider instead an experimental (or trial) alpha, α_E, representing the type I error in the experiment. The primary end point will have alpha associated with it (α_P), and secondary end points will have alpha associated with them (α_S). Since the goal is to set an upper bound for the experimental type I error, there are limitations placed on α_P and α_S. I begin by writing the probability of no type I error in the experiment as the product of the probability of no type I error on the primary end point and the probability of no type I error on the secondary endpoint.

$$1 - \alpha_E = (1 - \alpha_P)(1 - \alpha_S)$$

$$\alpha_E = 1 - (1 - \alpha_P)(1 - \alpha_S) \qquad (1)$$

The probability invoked here is that of the event "at least one success," described in Snedecor and Cochran (14). (This assumes independence between the type I error event for the primary end point and the type I error event for the secondary end point. Relaxing this assumption requires specific information about the nature of the dependence between primary and secondary end points, which will be trial specific. For the purpose of this discussion, I will assume independence.) Thus, α_E is the probability of making at least one type I error—an error on either the primary or

TABLE 1. Alpha allocation: $\alpha_E = 0.05$ (two-sided)

End point	Allocated alpha
Primary end point	
Total mortality	0.02000
Secondary end points	
Hospitalization for CHF[a]	0.01031
Progression of CHF	0.01031
Maximal O_2 consumption	0.01031

[a] CHF, congestive heart failure.

the secondary end point. This formula generalizes to n_P primary end points and n_S secondary end points.

$$\alpha_E = 1 - \left[\prod_{i=1}^{n_p} (1 - \alpha_{p,i})\right]\left[\prod_{j=1}^{n_s} (1 - \alpha_{s,j})\right]$$

This probability has its upper bound approximated by Bonferoni's inequality, but an exact treatment will be developed here. Several examples are provided for the use of this function. We will assume that all hypothesis testing is two-tailed.

Scenario 1

An experiment randomized patients to one of two treatment arms for the assessment of an intervention reducing total mortality. There are three secondary end points of equal weight (i.e., to be assessed at the same alpha levels).

In this case α_E is the probability of making at least one type I error for either of the one primary or three secondary end points and is set as a maximum of 0.05. If we choose $\alpha_P = 0.02$, then we can find the available alpha for the secondary end points from equation 1 as

$$\alpha_S = 1 - \frac{1 - \alpha_E}{1 - \alpha_p} = 1 - \frac{(0.95)}{(0.98)} = 0.03061$$

So $\alpha_S = 0.03061$ is the available type I error for the family of secondary end points. Apportioning this equally, we find

$$1 - 0.03061 = (1 - \alpha_S^*)^3$$

and $\alpha_S^* = 0.01031$. An alpha allocation table assembled by the investigators and supplied prospectively in the experiment's protocol (Table 1) is an unambiguous statement of the investigators' plans for assessing the impact of the experimental intervention.

Scenario 2

An investigator is designing a clinical trial, with a placebo and two treatment arms A_1 and A_2. There is equal interest is testing A_1 against placebo and A_2 against placebo. For each of these tests, there is one primary end point, total mortality, and two secondary end points (intermittent claudication and unstable angina). Set $\alpha_E = 0.05$, to be divided equally between the two tests (A_1 versus placebo and A_2 versus placebo). We find

TABLE 2. Alpha allocation: $\alpha_E = 0.05$ (two-sided)

End point	Allocated alpha
A_1 vs. placebo comparison	0.02532
Primary end point	
Total mortality	0.02000
Secondary end points	
Hospitalization for CHF[a]	0.00272
Progression of CHF	0.00272
A_2 vs. placebo comparison	0.02532
Primary end point	
Total mortality	0.02000
Secondary end points	
Hospitalization for CHF	0.00272
Progression of CHF (medication status)	0.00272

[a] CHF, congestive heart failure.

$$\alpha_{A_1} = \alpha_{A_2} = 1 - (1 - \alpha_E)^{0.5} = 0.02532$$

If 0.02 of this is allowed for the primary end point, and the remainder is distributed equally across the secondary end points, we conclude

$$\alpha_S^* = 1 - \left(\frac{1 - 0.02532}{1 - 0.02}\right)^{0.5} = 0.00272$$

Since A_2 would be handled analogously, the allocation of alpha for the end points can be easily completed (Table 2). Note that this prospective alpha allocation would allow a study, negative for the primary endpoint to be positive if the findings were strong enough ($p < 0.00272$) for the secondary endpoints.

Scenario 3

Consider a two-armed trial testing the impact of an intervention on the primary end point of mortality and each of two secondary end points. The investigators give each of these secondary end points equal weight. Applying equation 1, the investigators compute an alpha allocation table (Table 3). Assume that the experiment is executed according to the protocol and that the significance of the end points

TABLE 3. Alpha allocation: $\alpha_E = 0.05$ (two-sided)

End point	Alpha allocated (design)	Alpha expended (execution)
Primary end point	0.02500	
Total mortality	0.02500	0.00100
Secondary end points	0.02564	
Hospitalization for CHF[a]	0.01290	0.02000
Progression of CHF (medication status)	0.01290	0.00400

[a] CHF, congestive heart failure.

AEP Vol. 8, No. 6
August 1998: 351–357

TABLE 4. Alpha allocation and dependency

Secondary end point	Primary end point: Type I error	Primary end point: No type I error	Total
Type I error	0.02	0.005	0.025
No type I error	0.005	0.97	0.975
Total	0.025	0.975	1

is assessed. The overall alpha expended in the experiment = $1 - (1 - 0.00100)(1 - 0.02000)(1 - 0.00400) = 0.02490$. However, although the trial is positive with less than 0.05 in alpha allocation spent, the finding for hospitalization for CHF must still be considered negative, a conclusion reaffirming the importance of the investigators' prospective statement on level of statistical significance for the trial end points.

The presence of dependent hypothesis tests induced by end point set correlation can result in a generous alpha allocation (15). In such circumstances, the adjustment presented here is an overadjustment, leading to alpha levels lower than required. The incorporation of a dependency argument can lead to an important savings in alpha allocation. This presumes that the nature of the dependency is clear, quantifable, and defensible. A clinical trial (examining the impact of integrilin in the immediate post-PTCA setting) that made prospective arguments for dependency among the collection of primary and secondary events was successfully presented and defended at the Federal Food and Drug Administration public hearing in February, 1997. This presentation led to the computation of an alpha based on both primary and secondary end points that kept the overall alpha spent below an acceptable upper bound.

As a further elaboration on the impact of a dependency argument in alpha allocation, consider the implications of correlation between the likelihood of a type I error for two end points (Table 4). In this case, the total alpha expended is $1 - 0.97 = 0.03$. By assuming that the test on the primary end point is independent of the test on the secondary end point, the type I error expended is $1 - (0.975)(0.975) = 0.0494$. When compared to the expenditure of 0.03 using the dependency argument, the incorporation of dependency led to a $100\ (0.0494 - 0.03)/0.0494 = 39.3\%$ savings.

DISCUSSION

Opinions on both the necessity and strategy of alpha allocation are diverse (12–18). The use of multiple comparison procedures based on the Bonferoni method (14) has invoked strong criticism from Rothman (18), who states that such tools trivialize the possible hypothesis-testing outcomes by reducing the maximal P value acceptable for a positive finding to a level that seems to preclude proclaiming any effect as positive. The tack taken in this manuscript allows experimenters the freedom to choose a priori the alpha level of each hypothesis but allows the levels of alpha to reflect the importance of the end point. Thus the investigators tailor their alpha selection to their experience with the intervention and target those hypothesis tests of the greatest clinical relevance, akin to the construction of a minimax rule. However, the type I error of the experiment should be conserved, since this strategy protects the population to whom the results will be extended from an excessive number of false-positive findings. This procedure is in contradistinction to work on partial null hypotheses, (i.e., minimizing alpha within each of a collection of subsets of hypothesis tests) which may not keep the experiments overall alpha below a prespecified level.

By enforcing experimental concordance the investigators ease the task of interpreting their research. However, rigor and discipline in experimental execution should not exclude the prospective determination of acceptable alpha error levels. This is a serious obligation of the investigator, since protection of both the population and the individual patient are the responsibility of clinical scientists. The result of the strategies suggested herein is that, in an organized fashion, investigators prospectively trace the path of alpha accumulation through the end points, aligning their end point choices with the restrictions of interpretive parsimony.

A consequence of the proposed approach is, that, since alpha is to be expended on secondary end points, less alpha can be expended on the primary end point in order to constrain the alpha of the experiment at an acceptable upper bound. Thus, experiments with secondary end points must pay a price for these end points' interpretation (an increased sample size for the α_P of < 0.05). Many workers will understandably react negatively to this consequence of the procedures proposed in this work. However, too often, no type I error is allocated for the secondary end points, but much interpretive weight is placed there when the experiment has concluded. If the secondary end point is to have an objective interpretation, this interpretation must occur in the context of the alpha expended.

Alpha-spending functions for interim analyses (19–22) are very useful, and any alpha allocation for the end-of-trial analysis must be reduced if alpha was allocated during the interim analyses.

There is wisdom in the comments from Friedman and colleagues (23) who state "it is more reasonable to calculate sample size based on one primary response variable comparison and be cautious in claiming significant results for other comparisons." However, the degree to which investigators violate this principle in interpreting trial results suggests that this wisdom is not well appreciated. Other strategies in multiple testing are also admissible and those clinical investigators who disagree with the findings of Friedman and

colleagues (23) would benefit from a structured approach to the allocation of alpha. However, this approach should be consistent with the investigators' responsibility to protect the population to whom their research will be generalized from excessive false-positive errors.

One might successfully argue that the admittedly conservative alpha allocation strategy is an inappropriate standard in the academic setting, where findings from secondary end points in pilot studies might be used to generate hypotheses for subsequent experiments, and there are important criticisms of the analysis path approach that I advocate. However, conservative allocation of alpha has the advantage of being disciplined, prospectively identified, and unambiguous in its interpretation. We must also keep in mind that regulatory agencies place weight on *P* value interpretation, and that, in rendering final judgement, the population at large often bears the brunt of type I errors. The alpha allocation approach advocated here has the virtue of being prospectively defined, quantitative, and easy to reproduce. However, it must also be noted that this same procedure also has the weakness of its virtues. By being quantitative, it is also formal and restrictive. In being prospectively defined, it is nonreactive to new information made available while the trial is in progress. An alternative approach would be to incorporate prior information about the relative likelihood of the research effort's ability to produce a positive finding and also to develop a loss function for drawing the incorrect conclusion. Bayes procedures, which incorporate the parameterization of prior information concerning an action and information from the experiment itself into a posterior decision rule using a credible region, have become of greater interest (24). Since much of clinical decision making is based on more than a single research program, and interpretation and subsequent actions are generally in the context of other information and other research, not solely based on a solitary *P* value, the Bayesian approach would offer epidemiologist a way to describe more accurately their own methods of implicitly integrating prior information into the consideration of the data at hand.

The author wishes to acknowledge the contributions of the reviewers in the preparation of this manuscript.

REFERENCES

1. Pfeffer MA, Braunwald E, Moyé LA, Basta L, Brown EJ, Cuddy TE, et al. Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction-results of the Survival and Ventricular Enlargement Trial. N Engl J Med. 1992; 327:669–677.
2. Sacks FM, Pfeffer MA, Moyé LA, Rouleau JL, Rutherford JD, Cole TG, et al. The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. N Engl J Med. 1996;335:1001–1009.
3. The SHEP Cooperative Research Group. Prevention of Stroke by Antihypertensive Drug Therapy in Older Persons with Isolated Systolic Hypertension: Final Results of the Systolic Hypertension in the Elderly Program (SHEP). JAMA. 1991;265.
4. Cohn JN, Archiband DG, Ziesche S, Franciosa JA, Harston WE, Tristani FE, et al. Effect of vasodilator therapy on mortality in chronic congestive heart failure. N Engl J Med. 1986;314:1547–1552.
5. Packer M, Bristow MR, Cohn JN, Colucci WS, Fowler MB, Gilbert EM, et al. The effect of carvedilol of morbidity and mortality in patients with chronic heart failure. N Eng J Med. 1996;334:1349–1355.
6. Pfeffer MA, Sevenson LW. Beta-adrenergic blockers and survival in heart failure (editorial). N Engl J Med. 1996;334:1396–1397.
7. Moyé LA, Abernethy D. Carvedilol in patients with chronic heart failure. N Engl J Med. 1997;335:1318–1319.
8. Hilsenbeck SG, Clar GM. Practical *P* values: Adjustment for optimally selected cutpoints. Stat Med. 1996;15:103–112.
9. Fisher B, Bauer M, Margolese R, et al. Five year results of a randomized clinical trial comparing total mastectomy and segmental mastectomy with or without radiation in the treatment of breast cancer. N Engl J Med. 1985;312:665–673.
10. Fisher B, Bauer M, Margolese R, et al. Eight year results of a randomized clinical trial comparing total mastectomy and segmental mastectomy with or without radiation in the treatment of breast cancer. N Engl J Med. 1989;320:822–828.
11. Christian MC, McCabe MS, Korn EL, Abrams JS, Kaplan RS, Friedman MA. The National Cancer Institute audit of the national surgical adjuvant breast and bowel project protocol B-06. N Engl J Med. 1995;333:1469–1474.
12. Meinert CL. Clinical trials design, conduct, and analysis. New York: Oxford University Press; 1986.
13. Dowdy S, Wearden S. Statistics for Research. 2nd ed. New York: John Wiley and Sons; 1991.
14. Snedecor GW, Cochran WG. Statistical methods. 7th ed. Ames, Iowa: Iowa State University Press; 1980:116.
15. Dubey SD. Adjustment of *P*-values for multiplicities of interconnecting symptoms. In: Buncher RC, Tsay JY, eds. Statistics in the Pharmaceutical Industry. 2nd ed. New York: Marcel Dekker, Inc.; 1994:513–527.
16. Gnosh BK, Sen PK. Handbook of Sequential Analysis. New York: Marcel Dekker, Inc.
17. Miller RG. Simultaneous Statistical Inference. 2nd ed. New York: Springer-Verlag; 1981.
18. Rothman RJ. No adjustments are needed for multiple comparisons. Epidemiology. 1990;1:43–46.
19. Davis, BR. Hardy RJ. Upper bounds for type I and type II error rates in conditional power calculations. Comm Statis. 1990;19:3571–3584.
20. Lan KKG, Demets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983;70:659–663.
21. Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. Comm Statis. 1982.
22. Lan KKG, Wittes J. The b-value: A tool for monitoring data. Biometrics. 1988;44:579–585.
23. Friedman L, Furberg C, DeMets D. Fundamentals of Clinical Trials. 3rd ed. St. Louis: Mosby; 1966:308.
24. Press, JS. Bayesian Statistics Principles, Models and Applications. New York: John Wiley and Sons; 1989.