STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:767–779

Alpha calculus in clinical trials: considerations and commentary for the new millennium

Lemuel A. Moyé*,†

University of Texas School of Public Health, Ruell A. Stallones Building, 1200 Herman Pressler, Houston, Texas 77030, U.S.A.

SUMMARY

Regardless of whether a statistician believes in letting a data set speak for itself through nominal *p*-values or believes in strict alpha conservation, the interpretation of experiments which are negative for the primary endpoint but positive for secondary endpoints is the source of some angst. The purpose of this paper is to apply the notion of prospective alpha allocation in clinical trials to this difficult circumstance. An argument is presented for differentiating between the alpha for the experiment ('experimental alpha' or α_E) and the alpha for the primary endpoint (primary alpha, or α_P) and notation is presented which succinctly describes the findings of a clinical trial in terms of its conclusions. Capping α_E at 0.10 and α_P at 0.05 conserves sample size and preserves consistency with the strength of evidence for the primary endpoint of clinical trials. In addition, a case is presented for the well defined circumstances in which a trial which did not reject the null hypothesis for the primary endpoint but does reject the null hypothesis for at least one of the secondary endpoints may be considered positive in a manner consistent with conservative alpha management. Copyright © 2000 John Wiley & Sons, Ltd.

INTRODUCTION

The *p*-value concept for the interpretation of clinical experiments is a 20th century innovation which is the subject of much discussion and confusion. Although useful in an experiment with two treatment arms and one clinical endpoint, the evolution of research programs has led to clinical experiments of ever increasing complexity (for example, trials with more than two therapy assignment arms and trials with multiple endpoints) with attendant difficulty in *p*-value interpretation. Multiple comparisons [1,2] and global hypothesis tests [3] have all been developed to aid in the interpretation of clinical experiments with multiple treatment arms and multiple endpoints. However, because of either resultant alpha thresholds which are too low for investigators to accept, or comprehension difficulties for the non-statisticians who must make policy decisions involving the experiment's results, these erudite solutions are often not invoked. Since easily understood evolutions in *p*-value interpretation have not kept apace with the increasing

* Correspondence to: Lemuel A. Moyé, University of Texas School of Public Health, Ruell A. Stallones Building, 1200 Herman Pressler, Houston, Texas 77030, U.S.A.
† E-mail: lmoye@utsph.sph.uth.tmc.edu

Received October 1998
Accepted February 1999

Copyright © 2000 John Wiley & Sons, Ltd.

complexity of clinical research programs, the best wisdom still seems to be as stated by Friedman *et al*. [4]. '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'.

Perhaps one of the more frustrating activities in the interpretation of a research program is the assessment of a clinical trial which is negative for the primary endpoint but has positive findings for either secondary endpoints or in *post hoc* analyses. This contentious setting has been the focus of sharp and spirited debate recently [5,6]. In some unfortunate cases, investigators allow the negative primary endpoint to languish in scientific backwaters as findings for positive endpoints (sometimes prospectively stated, sometimes not) are shoved to the forefront. Others would argue that a conservative alpha strategy throws out the 'baby' of scientific progress with the 'bath water' of alpha hypersensitivity [5,7]. The interpretation of these experiments has been, and perhaps should be, contentious. Part of the difficulty in interpreting trials which are negative for the primary endpoint but positive in secondary endpoint analyses or *post hoc* analyses is the terminology of clinical trial interpretation. Words such as 'positive' or 'negative' are useful but coarse. The purpose of this paper is to advance the notion of a prospective alpha allocation scheme (PAAS), introducing trial result descriptions in terms of trial endpoints. The combination of a prospective allocation argument and new terminology provides the setting in which an experiment which did not reject the null hypothesis for the primary endpoint but does reject the null hypothesis for secondary endpoints may be considered positive. In addition, an argument is made for differentiating between the type 1 error of the experiment (experimental alpha α_E) and the total type 1 error for the primary endpoint(s), α_P. Capping α_E at 0.10 and α_P at 0.05 serves the useful purposes of conserving sample size, and preserving consistency with past standards of strength of evidence for the primary endpoint of clinical trials.

THE PAAS SYSTEM

An adaptation of the Bonferroni procedure [1], the concept recently developed [8] is termed the prospective alpha allocation scheme (PAAS). The philosophy and mathematics of the PAAS system and nomenclature developed here are explored in the two-tailed testing environment. However, these procedures are directly applicable to situations in which-one tailed testing is deemed appropriate. PAAS takes advantage of the investigators" authority to prospectively set multiple alpha levels at distinctly different levels for each of the experiment"s prospectively stated endpoints. Assume that in a prospectively designed experiment, there is one primary endpoint and one secondary endpoint. Let the total alpha available for the experiment be prospectively specified by the investigators as α_E, (say $\alpha_E = 0.05$). Further assume that the endpoints are independent of each other (the complication of dependency is presented in the discussion). Begin by setting α_P as the alpha error prespecified for the primary endpoint, and compute α_S for the secondary endpoint as

$$\alpha_S = 1 - \frac{1 - \alpha_E}{1 - \alpha_P}$$

This computation is taken from a Bonferroni argument [1]. Using a recent formulation [8], we can generalize to n_p primary endpoints and n_s secondary endpoints

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

Copyright © 2000 John Wiley & Sons, Ltd.

Table I. PAAS table: alpha allocation and expenditure in a clinical experiment

Experimental alpha	Allocation	Expenditure
Alpha allocated for all primary endpoints	α_E	
P1		α_{P1}
P2		α_{P2}
	.	.
	.	.
	.	.
Pn_p		α_{Pnp}
Alpha allocated for all secondary endpoints		
S1	α_{S1}	
S2	α_{S2}	
	.	.
	.	.
	.	.
Sn_s		α_{Sns}

Thus levels of type I error are chosen and computed from (1), allowing completion of the allocation column of the PAAS alpha allocation table, Table I.

The investigators select α_E and the level of alpha errors for each of the primary and secondary endpoints prior to the experiment. Because PAAS is prospectively applied, it should only be used for prospectively determined endpoints (that is, formal hypothesis testing), and not the less formal exploratory analyses designed to identify relationships not anticipated at the experiment's inception. Since the alpha levels are chosen in accordance with equation (1), total alpha error is conserved in α_E. At the trial's inception, a PAAS table is constructed and becomes part of the experiment's protocol. Upon the experiment's conclusion, the expenditure column is completed and the finished table promulgated with the study results.

As an example, consider a clinical trial investigating the effect of a single intervention to reduce morbidity and mortality in a population of patients with congestive heart failure. There are two treatment arms: control and intervention. The primary endpoint of the experiment is total mortality, but interest remains in examining the effect of therapy on changes in left ventricular ejection fraction (secondary endpoint S1), worsening heart failure (S2), and total hospitalizations for heart failure (S3). Using the above scheme, the investigators choose to allocate 0.05 alpha (that is, set $\alpha_E = 0.05$), distributing it among the primary and secondary endpoints. Only endpoints which have adequate power can be interpreted as negative. Without adequate power, the hypothesis test for the endpoint can only be interpreted as uninformative if its test statistic does not fall in the critical region (see Table II).

The investigators allocate 0.035 of the total alpha to the primary endpoint (that is, $\alpha_P = 0.035$), distributing the remaining 0.015 alpha equally among the secondary endpoints. This distribution ensures that the total type I error for the trial is 0.05, meeting the investigators obligation for community protection from type I error commission [6,8].

Copyright © 2000 John Wiley & Sons, Ltd.

Table II.

	Alpha allocation		Scenario 1 Alpha expenditure	Scenario 2 Alpha expenditure	Scenario 3 Alpha expenditure
Total	0.050				
Alpha for primary endpoints	0.035		0.020	0.040	0.080
Alpha for secondary endpoints	0.015				
Secondary endpoints:					
S1		0.005	0.070	0.070	0.070
S2		0.005	0.080	0.010	0.010
S3		0.005	0.100	0.001	0.001

We may examine interesting implications of this alpha allocation decision through the consideration of three clinical scenarios, each scenario reflecting a hypothetical experimental result. Assume the hypothesis tests for the primary endpoint in each of these scenarios is adequately powered. In scenario 1, the experiment produced a p-value for total mortality of 0.020. In the customary manner of interpreting clinical trials, this finding would be considered a positive result. Also, under PAAS, the p-value of 0.020 is less than the alpha prospectively allocated for the primary endpoint, and the study would be considered positive for the primary endpoint. Consideration of the large p-values for each of the secondary endpoints S1, S2 and S3 supports the conclusion that the study is either negative (if the endpoint hypothesis tests are adequately powered), or uninformative (if the endpoints are underpowered), although the experiment as a whole would most assuredly be interpreted as positive based on the finding for the primary endpoint by either the customary procedure or PAAS.

The circumstances are different in scenario 2. Under the customary 0.05 rule, scenario 2 reveals a positive finding for the primary endpoint. In addition, many workers would feel comfortable with the determination that S3 is positive ($p = 0.001$), and perhaps, would advocate that S2 is positive ($p = 0.010$) as well. A line of reasoning would be that, if the endpoints (and the alpha accumulation) were considered in the order P1, S3, S2 and S1, an alpha accumulation argument would support the positive interpretation of S3. However, this line of reasoning offers no protection from considering other endpoint orderings. For example, one could consider the endpoints (and accumulate alpha) in the order P1, S1, S2, S3. In this case the consideration of P1 and S1 has already consumed alpha well in excess of 0.05 ($1 - (1 - 0.04) \times (1 - 0.07) = 0.107$, computed from equation (1)). With no investigator sponsored prospective statement, we have no guide on how to accumulate alpha. If one is not concerned with type I error accumulation, then S3 is positive. On the other hand, some consideration for the magnitude of α_E (admittedly *post hoc*) leads to a negative interpretation (if adequately powered) of endpoint S3. Thus, determining endpoint significance requires focusing not just on the magnitude of the p-values but also on the order in which one considers the secondary endpoints! This is an unsatisfactory state of affairs. Although ranking endpoints in a clinically meaningful hierarchy of endpoints is often useful, this ranking must occur prospectively, before the experiment is executed. However, a prospective alpha allocation scheme would lead to order independent, unambiguous results. The p-value of the primary endpoint (0.040) exceeds the alpha allocated for it. The adequately powered significance test associated with it would be interpreted as negative under PAAS. However, the small

Copyright © 2000 John Wiley & Sons, Ltd.

p-value for endpoint S3 (0.001) would allow a positive conclusion for S3 since it is less than the 0.005 allocated to it. Thus, since the *p*-value for S3 is less than the alpha allocation, the study would be interpreted under PAAS as positive, even though the primary endpoint finding was negative. Note that the conclusion from both the PAAS and customary procedures for *p*-value interpretation is that the study is positive, albeit for different reasons. It would be usefully to distinguish between these modes of positively.

Scenario 3 causes much concern and its pattern represents a common experimental motif. Here, the *p*-value for the primary endpoint is greater than the allocated alpha and would be interpreted as negative (again, assuming adequate power). This is also the conclusion of the customary 0.05 rule for the primary endpoint. The experiment is negative and customarily, to the chagrin of the investigators, this is where the interpretation ends since many workers understandably reject the notion of a positive trial with a negative primary endpoint. In this view, the primary endpoint of the trial is considered supreme; it is the axis around which the trial revolves. Generally, all alpha is expended on the primary endpoint and the conclusion of this single hypothesis test determines whether the trial is a success or a failure. With no prospective alpha allocation scheme, the common conclusion would be that scenario 3 was negative. However, use of the PAAS admits an additional possibility. Scenario 3 would be considered a positive result in the absence of a primary endpoint *p*-value less than the allocated alpha, because the *p*-value for the secondary endpoint S3 was less than the *a priori* defined alpha allocation. The prospective alpha allocation scheme admits the possibility of a positive trial with a negative primary endpoint finding, at the same time conserving type I error.

This suggests that a more useful definition of a positive experiment is an experiment in which any prospectively defined endpoint whose *p*-value is less than the prospectively allocated alpha level. However, in describing results, it would be useful to unambiguously distinguish between a clinical trial which was positive for the primary endpoint and a trial which was positive based on secondary endpoint findings.

NOTATION

Both scenario 2 and scenario 3 present circumstances in which a trial with adequate power for the primary endpoint had a negative finding for the primary endpoint, but would nevertheless be positive under PAAS. If trial results as represented by these two scenarios are to be accepted as positive, it would be helpful to have notation and nomenclature to differentiate these types of results from other trial results more traditionally deemed positive. That nomenclature is developed here.

Consider a clinical trial with exactly one primary endpoint and exactly one secondary endpoint. For each of these two endpoints, a hypothesis test is executed and interpreted, and for each endpoint we will conclude that the test is either positive (the *p*-value from the hypothesis test is less than the allocated alpha), negative (the *p*-value from the hypothesis test is greater than the allocated alpha, and the test was adequately powered) or inconclusive (the *p*-value from the hypothesis test is greater than the allocated alpha, but the test had insufficient power). Then describe the findings of that clinical trial as P_aS_b where the subscript a denotes the conclusion from the primary hypothesis test, and the subscript b denotes the conclusion from the hypothesis test of the secondary endpoint. The values of each of a and b can be p(positive), n(negative) or i(inconclusive). With this notation, a clinical trial which is positive for the primary endpoint and

Copyright © 2000 John Wiley & Sons, Ltd.

positive for the secondary endpoint would be denoted as P_pS_p. Analogously, a P_nS_n trial is a trial in which each of the endpoints were found to be negative. A P_pS_I trial has a positive primary endpoint and an inconclusive finding for the secondary endpoint due to inadequate power.

Of course, clinical trials often have more than one secondary endpoint and sometimes more than one primary endpoint. We can embed this multiplicity of findings for these multiple endpoints into this notation stipulating by P_p if at least one of the primary endpoints is positive and P_n if all of the primary endpoints in the trial are negative. Denote the finding that some of the primary endpoints are negative and the remaining ones uninformative as $P_{ni.}$ For example, consider a trial with two primary endpoints and two secondary endpoints. If one of the primary endpoints was positive and the other negative, and, among the secondary endpoints, one was negative and one inconclusive, that trial's results would be summarized as P_pS_{ni}.

No investigator (the author included) can resist the opportunity to use a data set to address questions which were not prospectively stated. Often, a new advance may suggest a question that the investigators did not know to ask at the trial's inception (for example, drawing blood from patients at baseline and storing the blood; future analyses would allow the implementation of technology to relate DNA findings to clinical endpoint occurrence). In addition, journal editors and reviewers sometimes ask for additional non-prospectively identified analyses to support the reviewed manuscript's thesis. We may include the conclusions of such hypothesis generating endeavours by adding an H_c at the end of the trial designation. Thus, a trial which is negative for all of the primary endpoints, negative or uninformative on secondary endpoints and positive for some hypothesis generating effort would be designated $P_nS_{ni}H_p$.

An advantage of this classification is that it differentiates positive trials which are positive for the primary endpoint from those trials which are negative for the positive endpoint but positive for secondary endpoints. We can now consider the impact of prior alpha allocation. With no prior alpha allocation, a P_nS_p trial would not be considered positive. This is because, in the absence of a prospective statement by the investigators, a reasonable path of analysis and alpha accumulation begins with the primary endpoint. Assess the statistical significance of the test statistic, then proceed through the secondary and tertiary prospectively stated endpoints, accumulating alpha until the maximum tolerable limit is achieved. Thus, if the maximum alpha allocated is 0.05 for the study and this type I error is exceeded for the primary endpoint, the *p*-values for secondary endpoints cannot contribute to the argument of intervention benefit since the allocated alpha has been exceeded, requiring the P_nS_p study be considered negative. However, as we have seen from scenarios 2 and 3, the prospective alpha allocation can produce a P_nS_p result which should be considered a positive trial, since the positive findings for the secondary endpoint occur without exceeding the α_E cap. It must be clearly recognized that as long as alpha allocation is provided prospectively, the P_nS_p experiment deserves no pejorative appellation. It should not be considered a second class result. This represents an important interpretative change in the evaluation of clinical trial results.

Examples of the use of this notation appear in Table III. For example, in SAVE, the results were positive for the primary endpoint of total mortality. The secondary endpoints of the trial included hospitalization for heart failure, worsening heart failure, recurrent myocardial infarction, and deterioration in ejection fraction. Since several of these secondary endpoints were positive, the designation for the Survival and Ventricular Enlargement Trial would be SAVE-P_pS_p. An asterisk is used to identify a harmful effect (for example, CAST–$P^*_P\,S^*_P$).

Copyright © 2000 John Wiley & Sons, Ltd.

Table III. Classification of a selection of clinical trials by endpoint findings

Clinical trial	Findings	Classification
SAVE – Survival and Ventricular Enlargement [9]	PEP: Positive for total mortality SEP: Positive for hospitalization for CHF Positive for worsening CHF Negative for protocol defined myocardial infarction	P_pS_p
CARE – Cholesterol and Recurrent Events [10]	PEP: Positive for CHD death/MI SEP: Positive for revascularization Positive for stroke	P_pS_p
SHEP – Systolic Hypentension in the Elderly [11]	PEP: Positive for fatal and non-fatal stroke SEP: Positive for myocardial infarction Positive for revascularization Positive for congestive heart failure	P_pS_p
NitroDur – Post-infarction nitrate paste use [12]	PEP: Positive for change in end systolic volume ESVI SEP: Negative for post-trial change in ESVI	P_pS_n
CAST [13]		$P_P^*S_p^*$
LRC(Lipid Research Clinics) [14]	PEP: Negative for reduction in CHD death/myocardial infarction SEP: Positive	P_nS_p
Linet *et al.* – magnetic fields [15]	PEP: Negative for association between magnetic field proximity and acute lymphoblastic leukaemia matched analysis SEP Negative for association between magnetic field proximity and acute lymphoblastic leukaemia unmatched analysis	P_nS_n
Hayes *et al.* – cardiac function and pacemakers [16]	PEP: Negative for pacemaker disruption with normal cell phone use SEP: Positive for pacemaker disruption with unusal cell phone position	P_nS_p

PEP denotes primary endpoint. SEP denotes secondary endpoint.

THE SAMPLE SIZE STRAIT-JACKET

One difficulty in the implementation of PAAS is the implication of α_P as input to the sample size computation. If $\alpha_E = 0.05$ and $\alpha_P < \alpha_E$, the sample size based on this smaller α_P will be larger. Consider an experiment designed to detect the effect of an intervention on a primary endpoint of total mortality, and a secondary endpoint of total hospitalizations. The investigators intend to compute the sample size of the trial based on the primary endpoint, planning to achieve a 20 per cent reduction in total mortality from a control group total mortality rate of 15 per cent with 80 per cent power. The standard procedure would be to compute the sample size based on a formula such as

$$N = \frac{2\,[p_1(1-p_1) + p_2(1-p_2)]\,[Z_{1-\alpha/2} - Z_\beta]^2}{[p_1 - p_2]^2} \tag{2}$$

Copyright © 2000 John Wiley & Sons, Ltd.

Table IV.

	Alpha allocation
Total	0.050
Alpha for primary endpoints	0.030
Alpha for secondary endpoints	0.021
Secondary endpoints:	
S1	0.007
S2	0.007
S3	0.007

where N is the number of patients randomized to the placebo group plus the number of patients randomized to the active group, α = type I error, β = type II error, Z_c = the cth percentile from the standard normal probability distribution, p_1 = cumulative total mortality rate in the placebo group and p_2 = hypothesized total mortality rate in the active group. In this case, $p_1 = 0.15$, $p_2 = 0.12$ and N (the trial size) is 4060.

However, using PAAS, the investigators have Table IV.

In this setting, although $\alpha_E = 0.05$, $\alpha_P = 0.03$. The sample size from formula (2) based on α_P is 4699, an increase from the original sample size of 4060. Although this 16 per cent increase in sample size is the price the investigators must pay to provide the possibility of a positive finding on the secondary endpoint (earning the right to claim not just P_pS_p positivity but P_nS_p positivity), the sample size increase is substantial and the added financial and logistical burden is worrisome. Although there will be some increase in power for the secondary endpoints as well, this increase does not ensure adequate power for the secondary endpoints.

Although this difficulty is easily circumvented by increasing α_P to, say, 0.10, the scientific community and regulatory agencies would most likely raise concerns about this vitiation of the scientific evidence strength. However, the investigators could persuasively respond that the α_E they are accustom to spending on the primary endpoint must now be shared over primary and secondary endpoints. A point of compromise would be to allow an increase in α_E to 0.10, but to cap α_P at 0.05. Thus, the investigators would be free to construct an alpha allocation as in Table V.

In this circumstance, α_P is retained at 0.05, permitting adequate power for the primary endpoint with the original sample size of 4066. This recommendation permits the primary endpoint to be maintained at the 0.05 level of statistical significance, but the prospective alpha specification (0.053) is more lenient for the secondary endpoints, due to the increase in α_E from 0.05 to 0.10. The increase in α_E has permitted consistency in strength of evidence for the primary endpoint, expressed concern for the sample size, and still allowing ample possibility for a P_nS_p positive trial. This strategy may pose a problem for some, who may argue that the threshold for a positive trial has been reduced. However, in the PAAS framework, a positive trial is one in which the p-value for the prospectively delineated endpoint is less than the maximum alpha level permitted. In current practice, there is no consistent, satisfactory framework in which to consider P_nS_p positivity.

DISCUSSION

The development of p-values (the exact probability that a test statistic falls in the critical region under the null hypothesis) did not represent a great leap forward in statistical theory or in

Copyright © 2000 John Wiley & Sons, Ltd.

Table V.

	Alpha allocation	
Total	0.100	
Alpha for primary endpoints	0.050	
Alpha for secondary endpoints	0.053	
Secondary endpoints:		
S1		0.018
S2		0.018
S3		0.018

experimental practice. This movement instead resulted from the simple desire to sharpen the type I error bound of an experiment whose decision rule (that is, the *p*-value is less than a prespecified alpha level test statistic) was based on the Neyman-Pearson lemma [4]. While successful in simpler experiments with one endpoint, the interpretation of the *p*-value has become fraught with danger and controversy in an increasingly complex clinical trial environment, now replete with experiments containing multiple treatment arms and multiple endpoints.

Strategies for *p*-value interpretation have been offered in the past [4,17,19]. Here I advocate a change in alpha policy which encourages prospective alpha allocations for clinical experiments. This change in policy has two components. The first is to call for an improvement in the scientific community standard for prospective statements about alpha, requiring clear alpha allocations for primary and secondary endpoints. Secondly, there should be a differentiation between the experimental alpha, α_E, and the alpha allocated for the primary endpoint, α_P. I also suggest that α_E be set to a value greater than 0.05, say 0.10. This large value of α_E (fixing α_P at no greater than 0.05) permits investigators liberal alpha to distribute among secondary endpoints. Secondary endpoints add considerable strength to the findings of an experiment. If many endpoints are positive in the same direction, the trial is consistent, its several endpoints speaking with one voice. Investigators should not be penalized for including devices such as secondary endpoints which add to result coherency. Allowing a considerable difference between α_E and α_P permits ample opportunity for positive findings for secondary endpoints. The payoff for investigators is the new admissibility of P_nS_p positivity. Global omnibus tests of statistical significance can be useful in multiple testing situations. However, there are circumstances (for example, in the regulatory environment) when decisions must be made about individual endpoints. In this circumstance, an omnibus test is of less utility.

Capping α_P provides some protection for the scientific community and relief for the investigators. Although there is no theoretical justification for keeping α_P at the 0.05 level, the history of clinical trial significance thresholds nevertheless exerts considerable influence. Considering experiments as positive with $\alpha_P > 0.05$ would be inconsistent with previous work and weaken the strength of evidence standard. It may be difficult to integrate the findings from these experiments (considered by some to be vitiated) into the scientific fund of knowledge. Keeping α_P fixed at 0.05 maintains consistency with the past. In addition, since sample sizes are computed based on α_P, maintaining its level at 0.05 does not lead to sample size increases. Thus, investigators pay no penalty in sample size by using PAAS. By letting $0.05 = \alpha_P < \alpha_E$, the sample size strait-jacket which confined investigators even further in trial design has been relieved. This would not be the case if α_E were set at 0.05, and $\alpha_P < \alpha_E$.

Copyright © 2000 John Wiley & Sons, Ltd.

Moving from the definition of a positive trial as one which is positive for the primary endpoint to include trials which are negative for the primary endpoint but positive for secondary endpoints can only be seen as a relaxation of the criteria for positive trials. This, however, comes at the price of tightening secondary endpoint interpretation standards. In the author's experience, secondary endpoint decision rules are handled cursorily (if at all) in the design phase of a randomized controlled clinical trial. Only at the end of the experiment is there a scramble to 'put the right spin' on their interpretation. This manuscript rejects this approach to secondary endpoint management. Secondary endpoints would only be considered positive if an alpha allocation scheme (PAAS or another) were applied to them and rigorously followed for their interpretation. The customary standard for secondary endpoints is very weak – I suggest that the standard for secondary endpoint interpretation be increased, balancing this increase in standard with opening the door for a disciplined interpretation scheme admitting P_nS_p trials as positive.

A general rule for interpreting confirmatory trials (that is, those experiments which seek to confirm the findings of earlier studies) remains illusive. There is no standard context for interpreting these studies, although one would be hard pressed to interpret them in isolation, that is, as though the first study's result was not known. However, if there is were at least one study which demonstrated efficacy (either P_pS_n P_pS_p or P_nS_p) using a PAAS or other acceptable alpha tool, then a confirmatory trial would perhaps not be required to be P_p. The information from the first trial would make it easier to accept P_nS_p result as positive. This is consistent with the perspective that when multiple trials show the same pattern of exploratory findings, then their reproducibility through meta-analyses can become convincing enough to stimulate debate.

The propositions offered here have their basis in the pre-eminent need to protect the scientific and patient communities from dangerous type I errors [6,8]. An example of a harmful type I error would be the new availability to patients of ineffective compounds with non-trivial side-effects. However, the change in alpha policy advocated here is sensitive to the needs of statisticians, physician-scientist, regulatory agencies, industry and the patient community.

Unfortunately, students, applied statisticians and investigators often recoil from the notion of prospective alpha allocation. The concept of a *p*-value is so firmly entrenched that many of these practitioners have come to the conclusion that using *p*-values is letting the data speak for itself. They may believe that drawing conclusions based on highly partitioned prospective statements would blur the distinction between the experiment's objective scientific evidence and the investigators' intentions. After all, the experimental conclusions should reflect only the objective data and not the cleverness of the observer in 'guessing the right alpha allocation' before the experiment is conducted.

But how realistic is this concern? Consider the following experiment in which two investigators each reasonably allocate alpha for a clinical trial involving primary and secondary endpoints as in Table VI.

Each investigator has 0.10 alpha to allocate, and chooses to allocate it differently. The first investigator places 0.05 on the one primary endpoint, and distributes the remaining 0.053 alpha among the secondary endpoints. The second investigator places less alpha on the primary endpoint, demonstrating increased interest in the secondary endpoints. When the actual results are reported, at first glance, the investigators report the results differently. Investigator 1 would report the results as positive for the primary endpoint and negative for all secondary endpoints (P_pS_n). Investigator two reports a negative finding for the primary, but a positive finding for the secondary endpoint (P_nS_p). However, because of the prospective statement for alpha, each investigator is justified in calling the experiment positive. Reasonable alpha allocations produced

Copyright © 2000 John Wiley & Sons, Ltd.

Table VI.

	Investigator 1 Alpha allocation	Investigator 2 Alpha allocation	Actual results
Total	0.1	0.1	
Alpha for primary endpoints	0.05	0.03	0.04
Primary endpoints:			
P1	0.05	0.03	
P2			
P3			
Alpha for secondary endpoints	0.053	0.072	
Secondary endpoints:			
S1	0.018	0.04	0.03
S2	0.018	0.017	0.4
S3	0.018	0.017	0.4

from careful thought applied to the intervention–endpoint relationship will lead to coherent interpretations of the data set. Small differences in prior alpha allocation can lead to differences in the interpretation of a particular endpoint but are unlikely to lead to differences in the global interpretation of the experiment with a well chosen cadre of primary/secondary endpoints. However, each allocation should have a justification that is based on a consensus of other investigators and reviewers.

In addition, statisticians also remember that hypothesis testing did not begin with *p*-values, but with the Neyman–Pearson lemma. The notion of prospective alpha allocation gave statistical practitioners a way to make a decision about the tenability of a null hypothesis without simultaneously considering type I and type II error [3]. Movement to the *p*-value was not designed as a movement away from a prospective statement about hypothesis testing, but as movement toward improved type I precision. The use of a *p*-value could refine the statement of '$p < 0.05$' to the stronger '$p = 0.025$'. The purpose of the *p*-value was to sharpen the type I error estimate bound, not to reduce emphasis on prospective alpha control of decisions.

At best, reporting *p*-values presents an open display of the data, enabling readers to draw their own conclusions. However, we must recall that 'letting the data speak for themselves' is often not what occurs. Often those who argue that the data should speak for themselves are the first to tell us what decisions we should draw from the data. Unfortunately authors may emphasize significant findings, perhaps not even reporting some non-significant endpoints. Even if all endpoints are reported openly, authors and readers may still not appreciate the increased risk of an overall type I error rate [3]. It is unfortunately common to interpret a trial as positive if any endpoint has a treatment difference significant at the 5 per cent level, and there is a need to deter authors from such indiscriminate use of *p*-values [3]. These good investigators unfortunately sometimes pay the price when their labour produces a clinical trial with a P_nS_p result. With no prospective statement about alpha beyond the primary endpoint, this result must unfortunately be interpreted in the negative. The price we seem to have paid for the undisciplined interpretation of *p*-values is a wholesale movement away from prospective declarations about endpoints. Alpha allocation in complicated trials with multiple endpoints and multiple treatment arms is more necessary than in the simple experiments with one endpoint because of the alpha complexity. Wasteful use of *p*-values discards this discipline when it is most needed.

Copyright © 2000 John Wiley & Sons, Ltd.

The successful utilization of the P_aS_b system must incorporate the issue of power. When rejection of the null is not possible, the conclusion is based on the power. If the power is high, the result of the hypothesis test is negative. If the power is low, the 'conclusion' is only that the hypothesis test was uninformative or inconclusive. This is true for any prospectively stated endpoint. Of course the difficulty arises for secondary endpoints. Primary endpoints should always have adequate power. It is difficult to ensure adequate power for secondary endpoints. For example, the primary endpoint of a trial to test a randomly assigned intervention to reduce ischaemic heart disease may be a combination of fatal myocardial infarction or survival but non-fatal myocardial infarction. A reasonable secondary endpoint could be fatal myocardial infarction. However, the necessity of trial cost effectiveness often provides only the bare minimum of power (for example, 80 per cent) for the combined primary endpoint, ensuring the power for the fatal myocardial infarction endpoint which its lower cumulative incidence rate will be lower (for the same type I error rate and the same efficacy). Thus deciding whether the value of b in P_aS_b is n or i depends on the power of the secondary endpoint hypothesis test.

The development here subsumes the possibility of correlation between endpoints. Dependency of endpoints leads to a different computation than in (1) of alpha allocation, and, when there is dependency in the endpoints (1) can be much too conservative [20]. The presence of dependent hypothesis tests induced by endpoint set correlation can result in a generous alpha allocation. In these circumstances, the adjustment presented in the manuscript is an over-adjustment, leading to alpha levels lower than required. This consideration is dependency is admissible if (i) there is biologic plausibility for the nature of the dependency and (ii) the investigators make a reasonable prospective statement on the magnitude of the dependency. In such circumstances, the adjustment presented in the manuscript is too severe, leading to alpha levels lower than required, and the incorporation of a dependency argument can lead to an important saving in alpha allocation. However, the inclusion of such a dependency term (this is actually just the correlation coefficient) can lead to a substantial reduction in alpha expense. This presumes the nature of the dependency is clear, quantifiable and defensible.

In addition, the work of Westfall and Young [21] has demonstrated that corrections for multiple testing can accurately be made through either permutation type or boot strapping type resampling. These approaches offer substantial improvements over the usual Bonferroni type of adjustments because the dependence structures and other useful distributional characteristics are automatically incorporated into the analysis. However, the thrust of this manuscript's argument remains unchanged. There are several admissible ways of allocating alpha. One of them (based on an assessment of endpoint dependency) should be used. Also, this manuscript does not explicitly consider alpha spending function during interim monitoring. This important concept can be incorporated by using the determinations from the allocation column of the PAAS table as input to the alpha spending function approaches, using either sequential boundary procedures [22,23] or those of conditional power [24,25].

There are other acceptable approaches as well. The use of a rank ordering strategy for endpoints is an important consideration, and can be useful in a hierarchy of events. However their utility is vastly improved with a clear prospective statement from the investigators including the precise decision path they intend to follow in the post trial analysis. This strategy also allows for an unambiguous interpretation of *p*-values and can fully incorporate the concept that $\alpha_p < \alpha_{E.}$ In addition, a useful Bayesian approach to this issue has been addressed by Westfall *et al.* [26] and the notion of posterior probabilities as a replacement for *p*-values remains attractive.

Copyright © 2000 John Wiley & Sons, Ltd.

Is the conservative approach to type I error always necessary? Certainly not. In pilot studies, and in endeavours which are exploratory, stringent alpha management is not required, and ultraconservatism may be counterproductive. However, in circumstances where a new standard of care is developed for a patient population, exposing these patients to compounds and interventions with significant side-effects, alpha management is essential.

REFERENCES

1. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**: 751–754.
2. Worsley KL. An improved Bonferroni inequality and applications. *Biometrika* 1982; **69**: 297–302.
3. Pocock SJ, Geller NL, Tsiasis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; **423**: 487–498.
4. Friedman L, Furberg C, DeMets D. *Fundamentals of Clinical Trials*, 3rd edn. Mosby, 1996; 308.
5. Fisher L. Carvedilol and the FDA approval process: the FDA paradigm and reflections upon hypothesis testing. *Controlled Clinical Trials*. 1999; **20**: 16–39.
6. Moyé LA. P-value interpretation in clinical trials. The case for discipline. *Controlled Clinical Trials*. 1999; **20**: 40–49.
7. Rothman RJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990; **1**: 43–46.
8. Moyé LA. P-value interpretation and alpha allocation in clinical trials. *Annals of Epidemiology* 1998; **8**: 351–357.
9. Pfeffer MA, Braunwald, E, Moyé LA *et al*. Effect of Captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction - results of the Survival and Ventricular Enlargement Trial. *New England Journal of Medicine* 1992; **327**(10): 669–677.
10. Sacks FM, Pfeffer MA, Moyé LA. The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *New England Journal of Medicine* 1996; **335**: 1001–9.
11. 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). *Journal of the American Medical Association* 1991; **265**(24).
12. Mahmarian JJ, Moyé LA, Chinoy DA, Sequeira RF, Habib GB, Henry WJ, Jain A, Chaitman BR, Weng CSW, Morales-Ballejo H, Pratt CM. Transdermal nitroglycerin patch therapy improves left ventricular function and prevents remodeling after acute myocardial infarction: results of a multicenter prospective randomized double-blind placebo controlled trial. *Circulation* 1998; **97**: 2017–2024.
13. The Lipid Research Clinics Coronary Primary Prevention trial results. *Journal of the Americal Meedical Association* 1984; **251**: 351–374.
14. Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecanide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New England Journal of Medicine* **321**: 227–233.
15. Linet MS. Hatch EE, Lkeinerman RA, Robison LL, Kaune WT, Freidman DR, Severson RK, Haines CM, Hartsock CT, Niwa S, Wacholder S, Tarone RE. Residental Exposure to Magnetic Fields and acute lymphoblastic leukemia in children. *New England Journal of Medicine* 1997; **2337**: 1–7.
16. Hayes DL, Wanbg PJ, Reynolds DW, Estes M, Griffith JL, Steffens RA, Carlo GL, Findlay GK, Johnson CM. Interference with cardiac pacemakers by cellular telephones. *New England Journal of Medicine* 1997; **36**: 1473–1479.
17. Miller, RG. *Simultaneous Statistical Inference*. : Springer-Verlag: New York, 1981.
18. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984; **40**: 1079–1087.
19. Gibbons JD, Pratt JW. P-values:interpretation and methodology. *American Statistician* 1975; **29**(1): 20–25.
20. Dubey SD. Adjustment of p-values for multiplicities of interconnecting symptoms. In *Statistics in the Pharmaceutical Industry*, 2nd edn, Buncher RC, and Tsay JY (eds). Marcel Dekkert Inc.: New York, 1994; 513–527.
21. Westfall PH, Young S. *p*-value adjustments for mulitple tests in multivariate binomial models. *Journal of the Americal Statistical Association*. 1989; **84**: 780–786.
22. Lan KKG, Demets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**: 659–663.
23. Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics* 1982;
24. Davis BR, Hardy RJ. Upper bounds for type I and type II error rates in conditional power calculations. *Communications in Statistics* 1990; **19**(10): 3571–3584.
25. Lan KKG, Wittes J. The b-value: a tool for monitoring data. *Biometrics* 1988; **44**: 579–585.
26. Westfall PH, Krishnen A, Young SS. Using prior information to allocate significance levels for multiple endpoints. *Statistics in Medicine* 1998; **17**: 2107–2119.

Copyright © 2000 John Wiley & Sons, Ltd.