STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:795–799

Response to commentaries on 'Alpha calculus in clinical trials: considerations for the new millennium'

Lemuel A. Moyé

Both Dr Gary G. Koch and Dr Robert T. O.'Neill have provided insightful commentary on my manuscript 'Alpha calculus in clinical trials: considerations for the new millennium'. Indeed, each of their responses is well considered, offering important contributions in their own right to the philosophy to result interpretations in research endeavours in general, and to clinical trials in particular. It is both an honour and a challenge to respond to them.

To summarize, the PAAS system makes the following modifications to research interpretation:

1. Each endpoint (primary and secondary) which is central to the research issue and worthy of a hypothesis test should be prospectively specified and have an *a priori* alpha allocation.
2. Total alpha expended in the trial should be increased from 0.05 to 0.10, with no greater than 0.05 alpha allocated for the primary endpoint.
3. If 1 and 2 are followed, then a trial should be considered positive if it has a statistically significant finding on any endpoint that meets the criteria in 1. Thus, trials would be positive if they do not reach statistical significance in a hypothesis test on the primary endpoint, but are statistically significant on other endpoints under the assumptions of 1.

I also proposed a notation system to describe the results of trials in terms of their primary and secondary endpoint findings.

Dr Koch has correctly grasped one of the implications of the manuscript, that is, the blurring of the 'primary' and 'secondary' endpoint designations currently used. Perhaps the evolution of clinical trial design has progressed so that the key differentiation between endpoints is not that they are 'primary' or 'secondary', but that they are either prospectively identified with alpha allocation or not. This is a natural extension of the PAAS arguments. Dr Koch also clearly elaborates the importance of choosing a concordant collection of endpoints in a research effort if a system such as the PAAS is to be used. The selection of the prospectively identified endpoint constellation should be guided by the epidemiological principles as provided by Bradford Hill [1], which include strength of association, specificity, coherence, biologic plausibility and consistency. Clinical trial findings are more persuasive if the results for the main endpoint are substantiated by findings for the additional endpoints, thereby allowing the group of study endpoints to 'speak with one voice' about the effect of the intervention. The coherence of these endpoints adds greatly to the cogency of the trial's contention for therapy benefit.

I would be quite remiss if my response did not reflect the strong, common theme that runs through the manuscript, and through each of the commentaries of Dr Koch and Dr O'Neill on

Table I. PAAS system – concern for a total mortality effect; $\alpha_E = 0.100$

Endpoint	Experimental alpha = 0.100	
Primary endpoint alpha	0.050	
Primary endpoint – worsening congestive heart failure		0.050
Secondary endpoints alpha	0.050	
Exercise tolerance		0.045
Total mortality		0.005

the importance of prospective thought in research design. It is difficult to overestimate the persuasive power of prospectively identified endpoint results in clinical trials, while the findings of non-prospectively stated analyses are quickly and easily vitiated.

Dr O'Neill and Dr Koch have each raised the issue of the importance of statistical power in research interpretation, which suggests to me that my comments in the manuscript involving this issue could and should have been stronger. Certainly the PAAS system will only adequately consider a decision in which the test statistic is not in the critical region as negative only when there is adequate power, as is stated in the discussion section of the manuscript. In fact, this is why the subscript designation 'i' is included in the nomenclature proposed, whose use would denote that the finding for the endpoint in question is uniformative. In addition, requiring the report of power for each endpoint on which results will be reported would be a helpful step in the interpretation of study results. The notation suggested in the manuscript could be extended to report power in the summary of results. Thus a $P^{0.03}{}_{P}S^{0.60}{}_{I}$ trial would be an experiment in which the primary endpoint was positive at the 0.03 level, but the secondary endpoint was uninformative since the power was only 60 per cent. Dr Koch has also pointed out that some secondary endpoints are adequately powered, but have less clinical relevance. Such endpoints which are less clinically meaningful perhaps would not need any prospective alpha allocation and their results should be considered only in an 'exploratory' light. The notion of prospective allocation is most useful if the endpoints are directly relevant to the clinical hypothesis, justifying their inclusion in prospectively planned statistical hypotheses.

Dr O'Neill has raised several additional concerns about the PAAS system I have proposed. His comments reflect a thorough understanding of the strengths and weaknesses of the PAAS system. However, he does raise issues which I believe represent paradoxes to him but are not seen as such by the author. These must be flushed out in detail.

Dr O'Neill points out that the designation of a primary endpoint is mainly a clinical consideration, and, quite eloquently, reminds us that the three rationale for the determination of which endpoint should be primary are (i) the choice of a clinically relevant benefit, (ii) the sensitivity of the endpoint to the therapy and (iii) designing the trial for the efficient, precise capture of the endpoint information. Dr O'Neill also points out that there can be some difficulty in applying these criteria to an endpoint which is relevant but underpowered. As an example, he offers the problem of where to place total mortality in the hierarchy of endpoints in the design of a clinical trial where total mortality is clinically relevant, but conclusions based on it would be hampered by low power if it is chosen as the primary endpoint. Dr O'Neill suggests that the trade-off solution is not obvious. My view is that an underpowered endpoint cannot be a serious contender for the primary endpoint, so that endpoint must be placed elsewhere in the hierarchy.

This hypothetical scenario is reflected in Table I of this response, in which investigators are interested in measuring the effect of a new therapy for congestive heart failure. They believe, through an examination of the criteria for endpoint determination that Dr O'Neill elaborated (clinical relevance, sensitivity of the endpoint to the therapy, and efficient use of the design to capture the endpoint), that worsening congestive heart failure should be the primary endpoint. There is also interest in carrying out a hypothesis test on the effect of the intervention on exercise tolerance. However, although the investigators are interested in an assessment of the effect of therapy on total mortality, they know that the trial will be underpowered to detect a statistically significant difference, given their understanding of the cumulative total mortality incidence rate and the expected ability of the medication to reduce this rate. However, if the investigators either underestimate the sensitivity of total mortality to the endpoint and/or underestimate the total mortality event rate, thereby driving the total mortality test statistic into the critical region (the 'surprise factor' of Dr O'Neill), they wish to be able to claim that the trial is positive. A PAAS allocation as in Table I is one way the investigators could effectively allocate alpha, keeping their primary endpoint as hospitalization for congestive heart failure, but being vigilant for the possibility of a beneficial effect on total mortality. If they were 'surprised' with a *p*-value of 0.005 or less for the prospectively defined endpoint of total mortality, they could claim that the trial was positive.

Dr O'Neill moves on to another apparent paradox, this time through a quite informative scenario depicted in Table I of his commentary. The reaction he has to different decisions being made which are based on the threshold of an experiment is that the PAAS system has allowed an illogical conclusion, since the experiment with the larger sample size is the one with the negative (assuming adequate power) result. Fortunately, this concern can be addressed at once. Placing aside the PAAS system for the moment, consider two decision rules, each based on the measurement of one and only one endpoint (that is, no secondary endpoints are considered for hypothesis testing). The first rule sets alpha prospectively at 0.0125 (similar to column 1, Table I of Dr O'Neill's response). The second sets alpha prospectively at 0.05 (similar to column 2 of the same table). If the *p*-value for the experiment in this hypothetical example is 0.03, then Dr O'Neill's 'paradox' remains, even in the absence of a PAAS strategy. The trial with the larger sample size is negative while the trial with the smaller sample size is positive.

It is therefore not PAAS which provides the 'paradox'. In fact, even the non-PAAS finding is not paradoxical. In scenario 1, the investigators have prospectively determined that the type I error level needs to be very low (0.0125) before they can proclaim the trial as positive; the investigators required greater strength of evidence. This greater standard required a larger sample size, but, of course, the larger sample size does not guarantee the *p*-value will fall in the critical region. When investigators require stronger evidence of benefit, they are less likely to have positive results than investigators that require weaker evidence, *ceteris parabus.*

In Dr O'Neill's Table I, PPPP appears to be the classic application of the Bonferroni rule, that is, the indiscriminant division of alpha equally among all endpoints. Thus, the type I error threshold is very low for each endpoint. The PSSS scenario is more in keeping with PAAS system. The PAAS system at its core is the prospective, differential dispersal of alpha, at different levels, across different endpoints. Dr O'Neill is quite correct in his observation that the relaxation of the total alpha error to 0.100 is the reason for the positive findings of the trial under PSSS. However, the strength of evidence for the primary endpoint has not been diluted – in order for the primary endpoint to be positive, the primary endpoint has to have a significance level less than the 0.05 level. This is the traditional threshold for statistical significance and remains unchanged here for the primary statistical hypothesis test of the study. Thus, although greater alpha ia available in the PAAS system, the traditional strength of evidence used to assess the primary endpoint remains at 0.05. What the scientific community gains by relaxing the overall alpha to 0.10 under PAAS is increased structure among the secondary endpoints which are worthy and relevant for hypothesis testing. Essentially, if the investigators are willing to discipline themselves by allocating alpha for the clinically relevant secondary endpoints, they will have additional type I error to allocate.

The contribution of PAAS is that is represents one step forward from the current, amorphous 'don't ask – don't tell' policy for multiple endpoint, *a priori* alpha allocations. Moreover, this step forward that does not require clinical trialists to accept ultra low alpha levels that often result from the application of Bonferroni's approximation. However, attempting to manipulate the prospective choice of an alpha level solely to 'win', that is, choosing alpha requirements with the sole motivation to get the test statistic for the primary endpoint into the critical region, denigrates the entire process, regardless of which system for alpha allocation is used. Any prospectively chosen alpha levels should be set based on the notion of community protection (2). Alpha designations based on preventing the exposure of the community of noxious placebo would add a much needed clinical dimension to what is commonly perceived as a sterile, artificial, mathematical metric. I fear that in this commentary, Dr O'Neill did not go far enough. Not only is the choice of an endpoint primarily a clinical decision, the choice of prospective alpha levels should be based on clinical reasoning and community protection concerns as well.

Both Dr Koch and Dr O'Neill raise the issue of correlation between endpoints and the impact of this correlation on the PAAS system, and are each quite correct in asserting the role of inter-endpoint correlation in reduced type I error. This manuscript and a previous one [2] discuss the notion of correlation in computing the type I error to be allocated, and I agree that the computation of alpha assuming correlated endpoints can lead to substantial type I error reductions. However, Dr O'Neill raises a somewhat different concern that requires attention. Choosing among correlated endpoints and placing them in a primary or secondary position, as Dr O'Neill has characterized the PAAS system, seems counterproductive to him. Since his concern about this is well stated, one can only conclude that he must have a very difficult time indeed with the current state of affairs, because artificial categorization of correlated endpoints is the norm. Commonly, one endpoint is chosen from among a collection of correlated endpoints as the primary endpoints of the study. All alpha is expended on it and the remaining correlated endpoints are essentially removed from formal statistical considerations. This traditional and prevalent approach reflects the ultimate in artificial categorization, and as an extreme example of a counter productive decision process must discomfit Dr O'Neill, yet it is PAAS which allows us to move away from this unfortunate extreme. Under PAAS, alpha can be applied prospectively and differentially to each of the correlated endpoints, allowing each endpoint to make a contribution to the decision process, avoiding the requirement of placing all alpha 'eggs' in one primary endpoint 'basket'. I would think that this would go a long way to palliate Dr O'Neill's concern, not aggravate it.

PAAS was designed to promote better prospective alpha structure among endpoints clarifying their interpretation in the end of a single trial; there are many important issue for which it was not designed. Dr O'Neill, in addressing the issues of censoring and inter-trial endpoint comparisons, has identified two of the latter. Dr O'Neill correctly points out that the PAAS system does not help with a conclusion about the effect of a secondary endpoint, for example, non-fatal stroke. Using non-fatal stroke as an endpoint would be suspect under almost any circumstance because of the censoring that occurs with mortal events, not just PAAS. Also, the issues of bias, informative and non-informative censoring of patients and endpoints, of missing data both of an informative and non-informative nature, of multiple competing risks represented by the endpoint, of clinically unexpected findings in a direction not otherwise planned – each of these complicates the interpretation of a collection of studies. This is not PAAS's doing – it is the nature of the experimental environment. Of course, the PAAS system, designed for intra-trial examinations of significance testing, will not remove all of the complexities of drawing conclusions in this turbulent experimental environment. However, in his fairly complete description of the complications in inter-trial interpretations, Dr O'Neill did not mention the additional complexity induced by *post hoc* analyses and decisions. This is course also complicates the interpretation of multiple clinical trials, and would be addressed by PAAS

Finally, each of my learned colleagues has referenced and suggested scholarly alternatives to the PAAS system, and the manuscript refers to several of these are well. Dr Koch has appropriately mentioned the well established approaches in the literature, including re-sampling methods of Westfall and Young. I agree with Dr Koch that these avenues should be explored, and perhaps we should work harder to incorporate them in clinical designs which are now in their design phase. Each suggested alternative is erudite – but each has the weakness of its strength. In being complicated and somewhat abstract, they are fairly incomprehensible to non-statisticians. We must keep sight of the fact that, in contemporary society, weighing the evidence from research efforts does not reside within the isolated purview of technical specialists. Today, data are reviewed by regulators, by legislators, by practising physicians, and, in these litigious times, by judges and juries as well. Many of these decision makers can understand the underlying principle of sample based research and the role of *p*-values in assessing the impact of sampling error on a research conclusion. Perhaps the experience of the commentators is different, but my sense is that non-quantitative experts do not understand complex algorithms based on either the 're-randomization computations' or '*p*-value percentiles' that have been counterproposed here.

This begs the question, how are these non-quantitative experts to draw conclusions in this complicated research environment? It is unfortunate that the legislators, judges, regulators and physicians who must draw conclusions from sample based research cannot patiently wait for the future blossoms of statistical research efforts while they tightly hold the sharp thorns of today's research controversies. We as statisticians must provide tools which are useful, informative, intuitive and consistent with the fundamental tenet of sample based research (first say what you plan to do, then do what you said). I think the PAAS system passes these tests.

REFERENCES

1. Hill BA. The environment and disease; association or causation? *Proceedings of the Royal Society of Medicine* 1965; **58**: 295–300.
2. Moyé, LA. P-value interpretation and alpha allocation in clinical trials. *Annals of Epidemiology* 1998; **8**: 351–357.