STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:785–793

Commentary on 'Alpha calculus in clinical trials: considerations and commentary for the new millennium'[†‡]

Robert T. O'Neill

Food and Drug Administration, Center for Drug Evaluation and Review, Office of Biostatistics, 5600 Fishers Lane, Parklawn Building, Room 15B-45, Rockville, MD 20857, U.S.A.

Dr Moyé should be congratulated for tackling a difficult and long standing statistical issue in the design, analysis and interpretation of clinical trials and for proposing, as a solution, a strategy to control the statistical uncertainty of conclusions derived from clinical trials which utilize multiple endpoints to evaluate treatment effects. Moyé's proposal concentrates our attention on primary and secondary endpoints as two separate categories of endpoints, on a terminology for the interpretation of clinical trial outcomes as positive, negative or inconclusive, and on a prospective alpha (type I) allocation scheme (PAAS) which describes trial results in terms of each of the prospectively determined trial endpoints. Lastly, Moyé argues for distinguishing between the type I error for the overall experiment and the type I error allocated to primary endpoint evaluation. His proposal is placed in the context of interpreting the overall results of a clinical trial as 'positive', for a variety of scenarios, including the situation when trial results are judged 'not significant' for the primary endpoints, but 'significant' for the secondary endpoints, where 'significant or not' is determined according to whether the observed p-value is above or below a pre-set level.

Dr Moyé would also argue against the inappropriate use, reporting and interpretation of p-values for multiple endpoint interpretation including situations when investigators report only endpoints that are determined to be 'statistically significant' and are silent on the results for all other endpoints or on the trial planner's initial rank ordering of the importance of these comparisons. As a solution, Moyé recommends 'full disclosure' in advance of all endpoint comparisons, full disclosure in advance of the type I levels each endpoint will be held to, and full disclosure at the reporting stage of all endpoint comparisons planned, including categorizing endpoint results as positive, negative and inconclusive.

I applaud Dr Moyé's strong emphasis on the need for prospective planning of endpoints and for a prospective consideration of how to control type I error collectively for all endpoint comparisons of clinical interest. His position is in the spirit of others (Pocock *et al.* [1], Westfall *et al.* [2]) who have also argued for prospective weighting of the importance of each endpoint in a multiple clinical endpoint situation, as it is well known by clinical trial practitioners that *post hoc* interpretations can take on many inventive explanations which seriously distort any control of the type I error and invite false conclusions.

[†] This article is a US Government work and is in the Public domain in the United States

[‡] The opinions expressed in this paper are the professional views of the author and do not necessarily reflect the official position of the U.S. Food and Drug Administration.

Published in 2000 by John Wiley & Sons, Ltd.

Clinical trial planners have a choice between two basic approaches for dealing with multiple endpoints. The first approach is the class of global statistical tests that assume that each endpoint will contribute in some positive way to the overall effect and that do not attempt to identify which of the individual endpoints are impacted by treatment but rather draws strength from the collective contribution of each endpoint by utilizing a single composite of all of them. The second approach is the general class of tests designed to identify which individual endpoints out of a multiplicity of endpoints is impacted by treatment. These methods have received extensive discussion (Zhang *et al.* [3] and Sankoh *et al.* [4]) and may provide alternative solutions to the problem posed by Moyé, though he did not explore this issue. I will not discuss Moyé's proposals in this context. My commentary will focus on what I judge to be the well intentioned but deficient aspects of the PAAS scheme, of the notation and nomenclature proposed and the scheme for allocation of type I error. In particular, I will address problems his proposals raise rather than solve.

1. THE DISTINCTION BETWEEN PRIMARY AND SECONDARY ENDPOINTS: WHAT MEDICAL CONSIDERATIONS SHOULD BE FEATURED IN THEIR DEFINITION AND CHOICE

At the outset, it is important to recognize that the designation of multiple endpoints as primary or secondary is mainly a clinical consideration within the context of the objectives of a clinical trial and not a statistical issue. The statistical framework helps to plan the parsimony of endpoints that will characterize the disease, to quantify the clinically important outcomes and measures of treatment effect and to bring order to the quantification of statistical uncertainty/certainty of the multiplicity of possible results so that the trial results are interpretable and defensible. We need to recognize the limitations of the clinical and statistical perspectives in desiring a satisfactory solution under all possible circumstances to the problems posed by Moyé.

The definition and choice of a primary endpoint depends upon the knowledge of the disease at the trial initiation and upon how well one can clinically characterize the anticipated benefit. Despite best efforts at defining a single primary endpoint to characterize benefit, there may be sufficient medical uncertainty in capturing the primary benefit of the treatment that a need arises to measure other endpoints. One should distinguish between the situation where several endpoints are measured and included in a trial because it is clinically expected that any one of them alone would characterize a primary benefit of the treatment and the situation where several endpoints are measured because some subset of them would lend confirmation to the clinical benefit measured by a primary endpoint, or lend additional biological/physiological or pharmacological plausibility to the expected effects on other endpoints measured. The former concept highlights each endpoint that would carry enough clinical importance and credibility by itself that its demonstration alone would be all that is needed to claim a treatment benefit, despite no other endpoints confirming their sensitivity to treatment. The latter concept relegates non-primary endpoints to a confirmatory role, where though of interest, it is not necessary that they be sensitive to treatment and therefore 'significant' for the clinical experiment to be judged a success. If they are sensitive to treatment, it is the additional confirmatory role they play when the primary endpoint is also sensitive to treatment. When the primary endpoint is not sensitive, and the secondary endpoint is,

it should be justification for a follow-up study to address that clinical question if deemed important.

My point for distinguishing between primary and secondary endpoints is to examine the rationale for how Moye's proposal relates to allocating type I error among primary and secondary endpoints, in contrast to all K endpoints, or to a prespecified subset of K endpoints. As O'Brien and Geller [5] state, misperceptions of the relative merits of various statistical tests for efficacy with multiple endpoints have developed perhaps because of a focus on the procedures themselves, without sufficient consideration to a careful specification of the medical questions that the analysis is intended to answer. Analytic procedures are thus sometimes inappropriately matched to medical questions. We should consider where Moyé's PAAS proposal fits into this concern.

In clinical trial planning, generally there are three relevant considerations for deciding on endpoints. The first is choice of the clinically relevant benefit to the patient of the test treatment including the characterization and quantification of that benefit. The second is the anticipated sensitivity of the endpoint(s) to characterize that benefit to the treatment. The third is the tailoring of the study design to demonstrate this anticipated endpoint sensitivity to treatment. As Moyé has suggested, one could employ a variety of strategies, as for example, rank ordering in advance, all candidate endpoints, or perhaps a collection of them, in order of their clinical meaningfulness. However, the most meaningful clinical event may be the one for which it is most difficult to empirically demonstrate a treatment effect as it may be least sensitive to the treatment given the sample size of the trial. Take mortality as an endpoint, for example. The difficulty arises because the weight one wishes to allocate to mortality can either be high because it is clinically the most important, or low because it has the least power of demonstrating a treatment benefit. The trade-off solution, *a priori*, is not obvious. When mortality is the candidate endpoint for becoming the primary or secondary endpoint, arbitrarily, there is no easy solution to the weighting nor is there a clear logic to calling the endpoint primary or secondary.

When mortality, the endpoint indicative of a major clinical benefit, is relegated to secondary endpoint status only because it is expected to be insensitive to the proposed treatment or that the comparison will be under-powered to detect the benefit because of trial sample size, then the trial objective turns to controlling 'the surprise factor' for this unique endpoint. A similar argument could be made for any low incidence endpoint for which there is strong clinical interest but recognized insensitivity to treatment benefit for the proposed study size.

I claim that Moyé's scheme allows for artificial categorization of major clinical endpoints into one of two categories and for an artificial attempt to control their separate type I errors by placing them into these two categories. Moreover, when endpoints are reasonably correlated and essentially describe the same aspects of treatment response, it seems counterproductive to place them in separate designations. Doing so can create some illogical scenarios, as I will demonstrate in Section 2.

When composite endpoints, which are composites of any primary and/or secondary endpoints, are included as one of the multiple endpoints, a special class of problems are created for which PAAS provides no guidance. In some examples discussed by Moyé, mortality is one of the individual endpoints, and it is included also as a composite endpoint along with non-fatal stroke or myocardial infarction. In this situation, not all endpoints are observable as individual endpoints as they are competing risks for others. For example, non-fatal stroke does not censor mortality, but mortality censors non-fatal stroke. Pepe [6] and colleagues have considered this problem.

Table I.

Endpoint number	Alpha allocation PPPP	PSSS	Observed *p*-value
1	0.0125	0.05	0.03
2	0.0125	0.017	0.015
3	0.0125	0.017	0.10
4	0.0125	0.017	0.20

2. THE PAAS SCHEME INVITES ILLOGICAL CONCLUSIONS AS A CONSEQUENCE OF RAISING THE EXPERIMENTWISE ERROR

According to the PAAS scheme, the type I error for the primary endpoint(s) is set separate from the type I error for the overall experiment, the latter necessarily needing to be larger than that for the primary endpoints to account for the secondary endpoint comparisons. The type I error for primary endpoints is capped at, say, 0.05, and for the overall experiment at, say, 0.10. Planning the appropriate sample size of a clinical trial is based on determining that sample size which provides a lower bound on the probability of detecting a treatment benefit on any one of the primary endpoints given an assumed effect size for each, so the chosen sample size is based upon the statistical power against specified alternatives for each primary endpoint.

Suppose there are four endpoints and all of them can be considered primary from the clinical perspective. To implement the PAAS scheme, assume that one has a choice between two endpoint strategies, calling the first PPPP and the second PSSS. Note that all four endpoints are clinically capable of being judged primary but three of them are placed into the secondary category in the PSSS scheme. Using the strategy PPPP one would power the trial in such a way as to maintain a minimum power for each primary endpoint and where the type I error would be equally allocated over four primary endpoints (Table I).

In the strategy PSSS, the study is sized to maintain power for the single primary endpoint P, and since type I error is allocated only to one primary endpoint, the sample size needed in the PSSS strategy would most likely be much lower than that in the PPPP strategy.

Assume that at completion of the trial, the observed *p*-value for endpoint 1 is 0.03, which would mean according to the PAAS decision rule that the trial fails using strategy PPPP but wins using strategy PSSS even though the trial was sized much larger for strategy PPPP. The cause of this dilemma is that PPPP can only use a type I error of 0.05 but PSSS can use an error of 0.10 for the same trial to be considered a win.

Thus a different conclusion from the trial is reached solely on the basis of moving primary endpoints from one PAAS scheme to another. As discussed in Section 1, how much flexibility, if any, should there be to move an endpoint which could classify as primary, to secondary, just on the basis of the chances of success according to a prespecified alpha allocation.

3. CONCLUDING FOR EACH ENDPOINT WHETHER THE RESULT IS POSITIVE, NEGATIVE OR INCONCLUSIVE

Another component of the PAAS scheme is the use of notation and a nomenclature that, at the completion of the trial, describes the statistical outcome for each endpoint, calling it a positive,

negative or inconclusive result. Moyé suggests defining a 'positive' experiment as one in which any of the prospectively defined endpoints has its respective *p*-value less than the prospectively allocated alpha level. He distinguishes between findings determined positive for primary and secondary endpoints. Two concepts are involved. The first is the *a priori* power of the study for each endpoint against the *a priori* alternative hypothesis for each endpoint and the second is the conditional power for each endpoint for its *a priori* alternative given the data in hand. The concept of conditional power, as it is used for stochastic curtailment of clinical trials, seems to have a place in the interpretation of results (Lan and Wittes [7]). For example, the concept of conditional power has been used to address whether a clinical trial should terminate early because the chance of success at completion is small. The conditional power is calculated in this situation under the null hypothesis using the proportion of the current sample size (information) relative to that needed for the planned power at trial completion, but the conditional power can be calculated under the assumed or observed treatment effect, or any assumed treatment effect (alternative hypotheses). Experience indicates that it is difficult enough to choose a realistic treatment effect size for a single primary endpoint and to power the trial accordingly, without the additional complexities introduced by planning for realistic treatment effect sizes for all primary endpoints and all secondary endpoints, especially when it is impossible for a fixed sample size to simultaneously control power for each endpoint at those chosen alternatives. Moyé has not addressed this issue in a way that illustrates the difficulty in using his PAAS notation and nomenclature in a practical way.

It does not seem to be appreciated that when one plans a trial for a targeted statistical power, say of 90 per cent, against a specific posited treatment effect (for example, alternative hypothesis D), and for a pre-specified error, say 5 per cent, that if the planning assumptions are correct, not only should 90 per cent of the anticipated *p*-values at completion of the trial be less than or equal to 0.05, but that a very large percentage of these anticipated *p*-values should be very much smaller than 0.05. In fact, for test statistics that are normally distributed and for two-sided tests of hypothesis, 50 per cent of the anticipated *p*-values should be below 0.001 for a 90 per cent powered study (Hung *et al.* [8], Goodman [9]). Thus, interpreting *p*-values at the conclusion of a study, in the absence of knowing what the planned power of the trial against a planned alternative was for that endpoint, and in fact, what the power of the completed study actually was for an anticipated effect size in that endpoint is uninformative. It is impossible under hypothesis testing theory to use the observed *p*-value alone to distinguish between the traditional hypothesis testing conclusion of 'negative' or 'inconclusive'. It is impossible on the basis of only a *p*-value judged to be above or below 0.05, or any other prespecified error to decide if the clinical benefit observed in an endpoint is 'negative'. The concept is more complex than that, entailing estimation considerations, balancing power and other issues. This has important consequences for implementation of the PAAS strategy and for the use of a terminology for each endpoint that is 'positive', 'negative' or 'inconclusive'.

4. IS THE NEYMAN–PEARSON THEORY OF HYPOTHESIS TESTING OF PRESPECIFIED HYPOTHESES SUFFICIENT TO DEAL WITH THE PRIMARY–SECONDARY ENDPOINT ISSUE? BALANCING TYPE I AND TYPE II (POWER) FOR EACH ENDPOINT

A number of statisticians and clinical trialists have taken the position that a statistical significance test used as decision rule for the acceptance or rejection of a hypothesis does not do justice to the

complexity of clinical trial analysis and that the role of analysis can more broadly be viewed as summarizing data, estimating effects, and quantifying the weight of evidence (Cutler *et al.* [10], Simon [11] and Ingelfinger *et al.* [12]). Simon [11] argues that the hypothesis testing approach is not really an adequate framework for medical decision making because it forces decisions on the user where there is insufficient evidence, such as when the observed differences are not of practical importance, or for when the size of the study is too small to believe the result. Ingelfinger *et al.* [12] argue that we should not equate *p*-values or the results of hypothesis testing with decisions. The result of a hypothesis test might be best described as a conclusion, rather than a decision, to emphasize that the results of hypothesis tests are another way of reporting data.

Despite these criticisms, the Neyman–Pearson theory, which draws a dichotomy between hypothesis testing and hypothesis generation, does help protect clinical trials against the all too often approach of data dredging for data driven suggestive findings. The Neyman–Pearson theory appropriately focuses attention at the planning and design stage on avoiding a variety of unspecified hypotheses in advance of observing the data, but the concept of power has a role both in planning and in interpreting a clinical study (Inglefinger *et al.* [12]), especially in terms of the power at the conclusion of the trial against the planned effect.

Having said this, can the Neyman–Pearson theory be more accommodating to meet the spirit of Moyés proposals. We will advance some options.

5. CONTROLLING AND MINIMIZING TYPE II: THE THEORY OF HYPOTHESIS TESTS

Like the *p*-value, power is a probability, but power involves two hypotheses (a null and an alternative) and a significance test, while *p*-values require only one hypothesis (a null) which is usually set at zero effect but could be set at a minimal clinical effect of interest (Inglefinger *et al.* [12], Spiegelhalter *et al.* [13]). It is useful to ask the question, under traditional hypothesis testing assumptions used in clinical trials, whether one may change either type I or type II after the protocol is implemented? This idea has been proposed by Lehman [14]. Here is why I raise this issue.

The use of the Neyman–Pearson theory of hypothesis testing for both planning the size of a clinical trial and for analysis of endpoints and for interpretation of uncertainty faces difficulties which have yet to be resolved by any one methodological approach or solution. Controlling error in the face of multiplicity of endpoints solely through allocation of alpha is one of the areas of difficulty and there should be some flexibility for situations which demand it.

Within the framework of a clinical trial, where one endpoint is chosen, where the size of the clinical trial is planned on the basis of controlling the error at a prespecified level (say 0.05), one only hopes to maximize the power or correspondingly minimize the type II error for a selected expected endpoint benefit. Once sample size is set in advance, the type II error cannot be controlled at the analysis stage in the same manner that the type I error can be, unless one allows for retrospective changing of the type I error. This is an important consideration, at the heart of current efforts for sample size re-estimation in clinical trials to maintain planned power (Proschan and Hunsberger [15]), when attempting to interpret the magnitude of the observed *p*-value and for contrasting these *p*-values across endpoints with different *a-priori* powers, and therefore type II errors. There is no way, as Moyé proposes, of classifying results of endpoints as negative or inconclusive because the theory does not help one do so.

Table II.

Endpoint number	Percentiles of the p-value (one- sided upper) for the specified power and allocated α					
	25th	50th	75th	90th	Power	Two-sided α allocated
1-primary	0.00005	0.00006	0.005	0.025	90%	0.05
2-secondary	0.0003	0.003	0.018	0.069	70%	0.025
3-secondary	0.0084	0.043	0.15	0.33	30%	0.025

6. A MODIFIED PROPOSAL: INCLUDE THE EXPECTED PERCENTILES OF THE *P*-VALUES FOR THE PROTOCOL PLANNED TREATMENT EFFECT FOR EACH ENDPOINT, USING THE ADJUSTED ERROR FOR EACH PAAS ENDPOINT

Since the PAAS strategy places such major emphasis on type I error and not on the power of the endpoint comparisons or of the overall procedure, it might be informative to know at the planning stage what the expected range of p-values are for a planned endpoint comparison, given the sample size chosen for the trial and the posited treatment effect on each endpoint. Consider an example of three endpoints, one primary and two secondary, where the correlation structure is ignored and where a two-sided test of hyothesis would be carried out for each endpoint. The overall trial is powered for 90 per cent against a posited effect size for the primary endpoint, and as a result of this chosen sample size, the power for each of the secondary endpoints at posited effect sizes is 70 per cent and 30 per cent respectively, for each of the secondary endpoints. Table II illustrates the range of expected p-values (Hung *et al.* [8]) for this PAAS for two situations. The first situation is when the overall experiment is controlled at $\alpha = 0.10$ and the second situation is when the overall experiment is controlled at $\alpha = 0.05$.

For the first situation, for the primary endpoint, assuming that all the planning assumptions were correct, one should expect that the median one-sided upper tailed p-value for the primary endpoint would be 0.0006, meaning that 50 per cent of the anticipated p-values should be equal to or less than 0.0006. On the other hand, one should expect that the median p-value for the secondary endpoint with 30 per cent power is 0.043, meaning that 50 per cent of the anticipated p-values should be greater than 0.043. If one did observe a two-sided p-value of 0.025 (equivalently, one-sided p-value of 0.0125) for the third underpowered endpoint, one might question its credibility since it should not likely occur, given that the planning assumptions are correct, or the observed treatment effect for this endpoint is much larger than expected and consequently, it needs to be justified in a follow-up study.

For the second situation, Table III indicates the median p-values one should expect if the planning assumptions are correct.

7. HOW DOES THE PAAS CONCEPT APPLY TO INTERPRETING RESULTS OF TWO OR MORE STUDIES WITH MULTIPLE ENDPOINTS

To illustrate concepts, Moyé refers to several examples of clinical trials which have employed multiple endpoints. One of his references [16] is to the spirited debate which dealt with a much more complex problem of multiple studies of multiple endpoints. It is not clear that the PAAS

Table III.

Endpoint number	Percentiles of the p-value (one-sided-upper) for the specified power and allocated α					
	25th	50th	75th	90th	Power	Two-sided α allocated
1-primary	0.00001	0.0002	0.002	0.0125	90%	0.025
2-secondary	0.0001	0.001	0.009	0.041	70%	0.0125
3-secondary	0.005	0.028	0.108	0.264	30%	0.0125

scheme would help at all in evaluating collective evidence from multiple independent studies. The issues of bias, of 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 each endpoint, of clinically unexpected findings in a direction not otherwise planned, all complicate the interpretation of a collection of studies and PAAS may not help in evaluating such a situation.

8. CONCLUSIONS

Professor Moyé has proposed an interesting approach to addressing the problem of interpreting primary and secondary endpoint results. His insistence on prospective planning of the clinical endpoints, on a prospective consideration of which endpoints and comparisons deserve careful statistical interpretation, and especially on separate conclusions for primary and secondary endpoints, is novel and may be useful in some situations. The prospective differential allocation of type I error to both primary and secondary endpoints is appealing, especially if one were to mainly consider secondary endpoints as addressing a secondary and separate question for which a lesser threshold of statistical certainty was acceptable. In many clinical trial settings, this approach may be confining and in some situations illogical as pointed out in Section 2. I agree with Pocock *et al.* [1] that there is no unique, optimal strategy for the use of significance testing when analysing multiple endpoints in clinical trials. This is borne out by long standing discussions of knowledgeable clinical trialists (Cutler *et al.* [10]) by proposed solutions from frequentists and Bayesians (Spiegelhalter *et al.* [13], Westfall *et al.* [2]), and by many of the closed testing procedures now in the literature (Zhang *et al.* [3]). There is also a need to distinguish between decision making and drawing a conclusion from the reported p-value of a clinical trial (Ingelfinger *et al.* [12]). The strict hypothesis testing framework does have limitations.

This is not to say that we should de-emphasize the importance of a prospective commitment to what was planned and expected and for preservation of the principle of controlling erroneous conclusions in clinical trials, especially the type I error concept which has served us well.

ACKNOWLEDGEMENT

I would like to thank Dr James Hung for useful discussions on this manuscript.

REFERENCES

1. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; **43**: 487–498.
2. Westfall PH, Krishen A, Young SS. Using prior information to allocate significance levels for multiple endpoints. *Statistics in Medicine* 1998; **17**: 2107–2119.

3. Zhang J, Quan H, Ng J, Stepanavage ME. Some statistical methods for multiple endpoints in clinical trials. *Controlled Clinical Trials* 1997; **18**: 204–221.
4. Sankoh, AJ, Huque MF, Russell HK, D'Agostino RB. Global two-group multiple endpoint adjustment methods applied to clinical trials. *Drug Information Journal* 1999; **33**: 119–139.
5. O'Brien PC, Geller NL. Interpreting tests for efficacy in clinical trials with multiple endpoints. *Controlled Clinical Trials* 1997; **18**: 222–227.
6. Pepe MS. Inference for events with dependent risks in multiple endpoint studies. *Journal of American Statistical Association* 1991; **86**: 770–778.
7. Lan G, Wittes J. Data monitoring in complex clinical trials: which treatment is 'better"? *Journal of Statistical Planning and Inference* 1994; **42**: 241–255.
8. Hung HMJ, O'Neill RT, Bauer P, Köhne K. The behavior of the *P*-value when the alternative hypothesis is true. *Biometrics* 1997; **53**: 11–22.
9. Goodman SN. A comment on replication, *P*-values and evidence. *Statistics in Medicine* 1992; **11**: 875–879.
10. Cutler SJ, Greenhouse SW, Cornfield J, Schneiderman MA. The role of hypothesis testing in clinical trials. *Journal of Chronic Diseases* 1966; **19**: 857–882.
11. Simon R. Randomized clinical trials and research strategy. *Cancer Treatment Reports* 1988; **66**(5): 1083–1087.
12. Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH. *Biostatistics in clinical medicine*. MacMillan Publishing Co. New York, 1987.
13. Spiegelhalter DJ, Freedman LS. 'A predictive approach to selecting the size of a slinical trial, based on subjective clinical opinion. *Statistics in Medicine* 1986; **5**: 1–13.
14. Lehmann EL. Significance level and power. *Annals of Mathematical Statistics* 1958; **29**: 1167–1176.
15. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**: 315–1324.
16. Fisher L. Carvedilol and the FDA approval process: the FDA paradigm and reflections upon hypotheses testing. *Controlled Clinical Trials* 1999; **20**: 16–39.