The Perils and Promises of Multiple Endpoints in Clinical Trials.

Lemuel A. Moyé, M.D., Ph.D.*

Short Title – Moyé Multiple Endpoints in Clinical Trials

Total word count 2634 sans abstracts, references

*University of Texas School of Public Health, Houston, Texas

Please address correspondence to
Lemuel A. Moyé, M.D., Ph.D.
University of Texas School of Public Health
RAS Building E-815
1200 Herman Pressler
Houston, Texas 77030
Email lmoye@utsph.sph.uth.tmc.edu
Voice 713-500-9518
Fax 713-500-9530

Abstract

This manuscript explains how to interpret multiple endpoints in clinical trials, and provides the thought process by which clinical investigators can embed multiple endpoints into their studies to protect those endpoints' clear, unambiguous interpretations. The justification for secondary endpoints in clinical trials has a solid basis in both efficiency and epidemiology. Unfortunately, interpretation of these additional endpoints is often controversial. This is especially true when the primary endpoint of a clinical trial returns a null finding but at least one secondary endpoint of that trial is positive. The prospective declaration of endpoints, coupled with the intelligent selection of *a priori* alpha levels for each endpoint, regardless of the endpoint's primary or secondary ranking provides a clear solution to the secondary endpoint interpretation dilemma. The implementation of this approach requires that 1)the investigators appreciate the role of type I error not solely as a obstacle to manuscript publication or regulatory approval, but as a community protection issue 2) the statistical significance of each endpoint be based on the level the investigators set for it before the trial and 3) the execution of the trial be per protocol (i.e. concordant trial execution). Although historically under-utilized, this procedure will generate appropriate discussion among investigators at the trial's inception, is easily implemented, and permits a clinical trial to be considered positive based on a secondary endpoint even though the primary endpoint may not reach statistical significance. Two suggested uses of this approach in clinical cardiology with discussion are provided.

Condensed Abstract

This manuscript explains how to interpret multiple endpoints in clinical trials, and provides the thought process by which clinical investigators can embed multiple endpoints into their studies to protect those endpoints' clear, unambiguous interpretations. This procedure will generate appropriate discussion among investigators at the trial's inception, is easily implemented, and permits a clinical trial to be considered positive based on a secondary endpoint even though the primary endpoint may not reach statistical significance. Two suggested uses of this approach in clinical cardiology with discussion are provided.

Introduction

A well considered collection of secondary endpoints can strengthen the persuasive power of a clinical trial by generating a cohesive set of results and by providing insight into the mode of action of the study medication. Furthermore secondary endpoints improve the logistical efficiency of the trial design since the cost of the trial is only marginally increased by their inclusion. Astute clinical investigator, having understood the reasons for the use of secondary endpoints in clinical trials, can therefore be forgiven for their startled reaction to the reluctance of biostatisticians, epidemiologists, and regulators to accept the positive findings of these secondary endpoints on their face value. When it comes to the interpretation of secondary endpoints, very different rationale are used in their interpretation than in their inception. This is especially painful in the circumstance where the primary endpoint is null, but a secondary finding is positive as in ELITE [1] or the US Carvedilol Program [2-7]. In these circumstances, the influence of the secondary endpoint is explicitly and purposefully negated precisely when its impact is needed most by the investigators.

The purpose of this manuscript is to outline a straightforward approach to the evaluation of secondary endpoints in clinical trials.

What Does it Mean for the Population

While designing their experiment, researchers face two choices in deciding which patients they should plan to admit. One choice is to accept every patient who is in the population into the trial – clearly impossible. The other choice is to select a tiny sample from the population, study that small sample, and then believe that what they observe in that sample represents the truth from the population (Figure 1).

Figure 1: The Difficulty with Inferring the Samples Results
To the Population

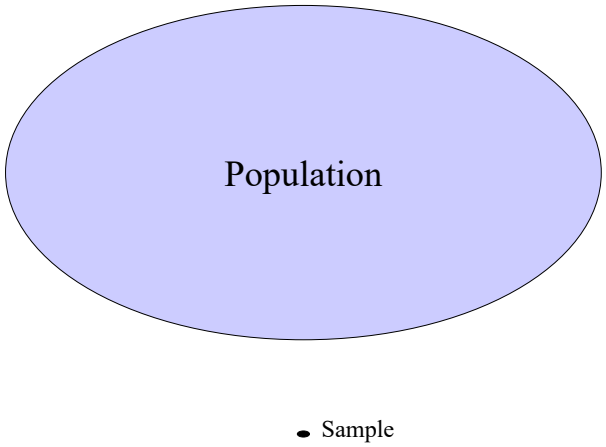

Figure one reveals that it is asking too much of random chance that every observed relationship in the sample reflect what the truth is about the population. Items in the sample aggregate randomly, and we cannot accept each of these aggregates as a reflection of "population-truth". In order for us to learn anything about the population from the sample, we must at least be able to measure the sample to sample variability (termed sampling error) associated with it. Statistics cannot remove sampling error, but it can provide a measure of whether a positive finding in a sample is due to noise or is a reflection of the truth in the population. This is the type I error.

Since so much of the information in the sample is likely to be background noise, trial designers focus on one sub-component of the information in the sample. That one component is the study question. They ask the question about the intervention-disease relationship first (prospective design), then execute the experiment according to that design, insuring that the measures of magnitude of the relationship, its standard error, and the type I and type II errors* are trustworthy. However, if too many questions are asked of the research sample, the likelihood that they contain too much background noise begins to grow. This is reflected in inflated type I

* Type I and type II errors are the estimates of sampling error that measure whether sampling error produced the findings the investigator observed in the population

and type II errors. Just like the probability of at least one tail in the successive flips of a coin increases as the coin is repeatedly tossed, the probability of at least one false answer increases if the sample is queried again and again for information which will be used to reflect population-truth. In a concordantly executed* experiment, the analysis for one endpoint may results in a *p* value of 0.05; when two questions are asked, the probability that at least one of the answers is not population-truth = 1 - (1 - 0.05)(1 - 0.05) = 0.0975. The likelihood that at least one of those conclusions is wrong becomes too large too quickly. We don't know which of the endpoint findings is wrong – we can only say that the likelihood that at least one of them is wrong has become too large.

Thus there are two problems with multiple endpoints as commonly used. The first is that they may not have been determined before the experiment began. This identification of endpoints after the fact makes the analysis uninterpretable because the endpoint was determined based on the data, which of course contains an important random component[8]. Secondly, type I error accumulates rapidly as the number of conclusions drawn from a sample increases. Multiple endpoint selection must be prospective, and there must be adequate type I error protection. However, if clinical investigators are to appreciate the significance of assigning type I error prospectively for endpoints in clinical trials, they must rise above the notion that type I error is simply a regulatory hurdle or obstacle to publication that they must overcome. The type I error level set by these scientists is a statement about the risk they are willing to accept in coming to the wrong conclusion that sample effectiveness translates directly to population effectiveness.

* Concordant execution simply means that the experiment was executed according to its protocol.

Satisfactory Secondary Endpoint Interpretation

Since the type I error is an important community protection device, it must be kept to an acceptably low level. This responsibility requires both a clear understanding of and a tight rein on type I error rates. Letting investigators choose their own *post hoc* analysis plan ("let the data speak for themselves"), when there is not a prospective statement for alpha allocation is unacceptable because the process is random[8]. The Bonferroni approach (i.e. dividing the total type I error by the number of hypothesis tests to be executed) [9] and its adaptations [10], commonly used for secondary endpoints does not work well as generally applied, since the type I error threshold for each test decreases to an un-usable low level quickly for each additional test. In addition, concerns about type I error conservation as currently practiced assure that research efforts with null findings for the primary endpoint, but positive findings on secondary endpoints will be considered null trials, a matter of great frustration to investigators who resist being compelled to place all of their "alpha eggs" in one primary endpoint "basket".

This difficulty has elicited discussion recently [11-16]. The major recommendations from this body of work are to 1)require that each primary and each of the secondary endpoints be prospectively chosen and 2)each of these endpoints have type I error attached in a prospective and reasoned fashion. This collection of procedures increases the rigor for the prospective statements concerning secondary endpoints, while permitting the straightforward interpretation of a research effort which is positive for secondary endpoints but in which the primary endpoint is not statistically significant. The following are possible uses of this strategy.

Example 1

Consider the following plans for a hypothetical, randomized, double blind clinical trial (MOROSE) designed to test the effect of a new treatment for heart failure. During the design phase of the trial, the investigators believe the most conclusive result from their study would be to demonstrate a decrease in total mortality. However, they are convinced, that, as desirable as such a result would be, it is unlikely that the trial would produce this finding. This is primarily because the anticipated cumulative mortality rate would be too low, requiring more patients than are available for the study. Thus, although the investigators' "clinical heart" wishes to choose total mortality as the primary endpoint, their "statistical conscience" demands the choice of a more frequently occurring morbidity endpoint in its place. The investigators choose exercise tolerance as the primary endpoint, and total mortality as the sole secondary endpoint. However, they are aware of prior experiments in which the primary endpoint was null, but another finding was positive, causing confusion in the trial's interpretation. Working to avoid this, the investigators choose to prospectively allocate alpha as follows.

Table 1: Prospective Alpha Allocation for a Trial with a Morbidity Primary Endpoint

	Allocated Alpha
Primary Endpoint – Exercise Tolerance	0.049
Secondary Endpoint 1 – Total Mortality	0.001

In this design, the trial will be positive if 1)the p value for the hypothesis test on exercise tolerance ≤ 0.049 or 2) there is a beneficial effect on total mortality with a p value < 0.001. Note, that, just because there is more than one endpoint does not require that alpha be allocated

equally to each of the endpoints. The investigators can allocate alpha in any proportion they choose – however, in order to have it interpretable, the allocation must be prospective.

Second, consider the following possible result of the trial. The *p* value for the primary endpoint at the trial's end is 0.10, and the *p* value for total mortality is 0< 0.001.[*] With no prior allocation for type I error, this result would be considered a null finding[†], and the trial would be considered "negative" in the face of the strong demonstration of a mortality benefit, because allocating alpha retroactively is inherently driven by the data results, and is uninterpretable. However, the simple tools of prospective alpha allocation leads to quite a different conclusion. Recall that only 0.049 is allocated for the primary endpoint; because the trial result provided a *p* value of 0.10, all of the 0.049 allocated is expended, but not more than that, since only 0.049 was prospectively set aside. This leaves the 0.001 for the secondary endpoint, and, because less than this is used by the trial for the secondary, total mortality endpoint, a significant finding was identified for total mortality. Because all alpha was apportioned prospectively, the interpretation of this trial is that it is positive, with a null finding for the primary endpoint and a positive finding for the secondary endpoint, which may be denoted as MOROSE— P_nS_p.[‡]

A final observation for this example focuses on the small magnitude of type I error set aside for a mortality benefit. It is extremely unlikely that the investigators may detect a mortality effect at that 0.001 level in a trial that enrolls only enough patients to have reasonable power to detect an effect on exercise tolerance. This tight criteria is by design. In this small trial, setup to detect an effect of the intervention on exercise tolerance, the finding of a mortality benefit will

---

[*] Clinical trial interpretation must include the joint consideration of sample size, effect size, its standard error, and the *p* value, not the *p* value alone. However, this manuscript focuses on p values because of its alpha allocation emphasis.

[†] Or negative finding if there is adequate power**.**

[‡] The findings of a clinical trial as P_aS_b where the subscript a denotes the conclusion from the primary hypothesis test, and the subscript c 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).

likely be based on a small number of deaths. A mild or moderate beneficial effect on mortality in a small number of patients would not make a persuasive argument that the study should be considered positive. Since the trial was not designed primarily with mortality in mind, the findings from the total mortality analysis should be overwhelming to persuade the medical community that a beneficial effect seen in this research sample truly reflects a finding that may be extended to the population. This is reflected in the low, 0.001 threshold.

Example 2

Consider the hypothetical trial CLOCK, designed to examine the effect of a calcium channel blocking agent in a randomized double blind, placebo controlled clinical trial to reduce mortal and morbid events in patients with hypertensive heart disease. The investigators are interested in both all cause mortality and cardiovascular mortality. All cause mortality is somewhat less problematic than cause specific cardiovascular mortality, because cause specific findings are dependent on the coding scheme. Alternatively, cause specific mortality occurs at a lower incidence rate than total mortality, a fact that may be somewhat offset by the fact that there may be greater efficacy of the intervention when measured against cardiovascular mortality. The investigators also have an interest in the symptoms or signs of CHF based on the treating investigator's investigator opinion about CHF status. After consideration of these issues, the investigators for ClOCK allocate type I error prospectively as follows.

Table 2: Prospective Alpha Allocation for a Trial with a Mortal Primary Endpoint

Endpoint	Allocated Alpha
Primary Endpoint – Total Mortality	0.034
Secondary Endpoint 1-CV Mortality	0.015
Secondary Endpoint 2 –Symptoms/Signs CHF	0.001

Alpha is allocated unequally to the two mortal endpoints, with only a slight residual for the morbid endpoint. There is no theoretical difficulty with unequal apportionment of alpha for the two mortal endpoints. The necessary ingredient is not equity, but the prospective nature of the allocation. In this prospective allocation, the trial might be considered positive for findings on either of the primary endpoint or any of the secondary endpoints. There would be a sample size increase (from 4,066 to 4,545 for a trial designed to detect 20% reduction in the total mortality rate of 15% with 80% power). The additional 479 patients earns for the investigators the ability to have a positive trial based on findings for any of the three prospectively determined endpoints.

Conclusions

The principle of prospective research planning, well embedded in clinical trial methodology and necessary for clear trial interpretation, is the guide we need to steer us over the difficult and hazardous multiple endpoint terrain. The literature is replete with simple and complicated strategies for the interpretation of multiple endpoints [17-24].

As a reader, when evaluating the utility of multiple endpoints in clinical trials, one must directly ask if the secondary endpoints were designed prospectively into the study, and was the study executed according to the protocol. Ignore secondary endpoint findings where they were not prospectively identified – they are useful exploratory tools, but that is all. Discordant trial execution introduces a dangerous random component into an experiment, making its analysis difficult to interpret. While it may be clear what the analysis of this type of trial reveals about the sample of patients it examined, it is difficult to see what this experiment has revealed about the population from which the sample was derived. If the endpoints were added retrospectively,

they provide a useful exploratory analysis, but they cannot be seen as providing an answer that can be applied to the population.

If the endpoints were designed prospectively and the research effort was executed concordantly, then the estimates reflecting the impact of the intervention and sampling error are trustworthy. For each endpoint, jointly examine the number of patients who had the endpoint, the effect size and its standard error (and confidence interval if supplied), as these provide the magnitude of the effect of the intervention on that endpoint. If the effect seen in the sample is beneficial, then, in order to learn whether that effect can be reasonably thought to apply to the population at large, examine the *p* value[*], but beware of type I error accumulation. Accumulated type I error means that the population is not likely to see the benefit of the therapy demonstrated in the sample. One guarantee of this is whether the investigators set appropriate bounds on the type I error for each endpoint in the design phase of the trial, a determination which should be reported in the methodology section. If they did not, then marginal *p*-values at the end of the trial will accumulate type I error rapidly and, with the risk of making a type I error being too high. These findings should also be discounted.

The clearest evaluation of multiple endpoints is when 1) each of the multiple endpoints is declared prospectively 2) type I error is prospectively allocated to each endpoint and 3) the experiment is executed according to its protocol. In these circumstances, the estimates of effect size, confidence intervals, and *p* values are trustworthy, and an endpoint by endpoint examination comparing the *p* value for that endpoint to the alpha level that was prospectively allocated plainly demonstrates for which endpoints are the risks of mistakenly inferring a benefit to the population becomes too great. Under the circumstance of prospective endpoint declaration

[*] If the finding for the endpoint is negative, examine the power.

and alpha allocation, this study should be viewed as a positive study with the nomenclature as described [12].

This is the clearest and safest environment to interpret the findings from multiple endpoints in clinical trials, and it is an environment which the clinical investigator has complete control in creating. Clinical trial designers have great freedom in choosing the alpha allocation levels for endpoints, and need not be shackled by the Bonferroni approach. In this manuscript, several examples of the prospective allocation of alpha have been provided.

References

1. Pitt B, Segal R, Martinez FA, et. al. on behalf of the ELITE Study Investigators. Randomized trial of losartan versus captopril in patients over 65 with heart failure. Lancet 1997; 349:747-52.
2. Packer M., Bristow M.R. Cohn J.N. et al. (1996) The effect of carvedilol on morbidity and mortality in patients with chronic heart failure. *New England Journal of Medicine* 334:1349-55.
3. Moyé LA, Abernethy D. (1996) Carvedilol in Patients with Chronic Heart Failure (Letter). *New England Journal of Medicine* 335: 1318-1319.
4. Packer M., Cohn J.N., Colucci W.S. Response to Moyé and Abernethy (1996). *New England Journal of Medicine* 335:1318-1319.
5. Fisher LD, Moyé LA**.** Carvedilol and the Food and Drug Administration Approval Process: An Introduction. *Control Clin Trials* 1999,20:1-15.
6. Fisher L. (1999) Carvedilol and the FDA approval process: the FDA paradigm and reflections upon hypotheses testing. *Controlled Clinical Trials* 20:16-39.
7. Moyé LA. (1999) *P* Value Interpretation in Clinical Trials. The Case for Discipline. *Controlled Clinical Trials* 20:40-49.
8. Moyé LA. Random Research. *Circulation*.2001;103:3150-3.
9. Snedecor , G.W, and Cochran WG (1980) Statistical Methods, 7th Edition. Iowa State University Press.
10. Simes RJ (1986). An improved Bonferonni procedure for multiple tests of significance. *Biometrika* 1986;73,751-754.

11. Moyé, LA. P-Value Interpretation and Alpha Allocation in Clinical Trials. *Ann Epidemiol* 1998;8:351-357.
12. Moyé LA. Alpha Calculus in Clinical Trials: Considerations and Commentary for the New Millenium. *Statist. Med.*2000;19:767-779.
13. Moyé LA. Alpha Calculus in Clinical Trials: Considerations and Commentary for the New Millennium. Rejoinder *Statist. Med.*2000;19:767-779
14. D'Agostino RB. Controlling alpha in clinical trials; the case for secondary endpoints. *Statist Med* 2000;19:763-766.
15. Koch GG. Discussion for 'Alpha calculus in clinical trials: considerations and commentary for the new millennium' *Statis Med* 2000;19:781-784.
16. O'Neill RT. Commentary on 'Alpha calculus in clinical trials: considerations and commentary for the new millennium' *Statis Med* 2000;19:785-793.
17. Worsley, K.L.1982. An improved Bonferroni inequality and applications. *Biometrika* 69, 297-302.
18. Friedman L, Furberg C, and DeMets D. 1996.*Fundamentals of Clinical Trials 3rd edition*; Mosby.
19. Meinert CL. 1986. Clinical Trials Design, Conduct, and Analysis. New York. Oxford University Press.
20. Dowdy S. Wearden S. 1991.Statistics for Research. Second Edition. New York. John Wiley and Sons.
21. Dubey S.D. "Adjustment of p values for multiplicities of interconnecting symptoms" in *Statistics in the pharmaceutical industry 2nd Edition*. Editors Buncher R.C. and Tsay J.Y. New York, Marcel Dekker Inc.
22. Gnosh B.K. Sen P.K. Handbook of Sequential Analysis. New York. Marcel Dekker Inc.

23. Miller R.G. (1981).Simultaneous Statistical Inference 2nd Edition. New York. Springer-Verlag.

24. Rothman R.J. 1990 “No adjustments are needed for multiple comparisons” Epidemiology 1:43-46.