STATISTICS IN MEDICINE
*Statist. Med.* 2000; **19**:781–784

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

Gary G. Koch

*Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, NC 27599-7400, U.S.A.*

The paper by Moyé [1] provides useful discussion for some important statistical issues concerning the possibly complicated ways in which multiple comparisons across primary and secondary endpoints can affect the results from clinical trials. How to balance tolerable inflation of type I error against a more extensive structure for evaluating success or failure of a clinical trial is the principal question. As suggested by Moyé, a reasonable way to address this question is to use a larger experimentwise significance level ($\alpha_E = 0.10$) for a prespecified set of primary and secondary endpoints with the traditional $\alpha_P = 0.05$ (two-sided) maintained for the primary endpoint. Through such a structure for confirmatory inference, the reporting of success or failure for the results of a clinical trial would include both primary and secondary endpoints, and Moyé illustrates a convenient notation for this purpose [1]. Such reporting would also agree with the suggestions in Davis [2].

A very important point in the paper by Moyé is that *a priori* specification of the secondary endpoints in a structure that controls the experimentwise significance level $\alpha_E$ is necessary for enabling confirmatory inferences concerning their results. Moreover, this point applies nearly as strongly to studies with positive results for primary endpoints as those with negative results. The relevant considerations for a situation without such structure for $\alpha_E$ are that favourable results for secondary endpoints are only interpretable as descriptively supportive when the primary endpoint is positive, whereas they have an exploratory hypothesis generating nature when the primary endpoint is negative. The only way for making confirmatory inference possible for a secondary endpoint is with its inclusion in a formal statistical testing procedure, and such a method for confirmatory inference must have a reasonable level of control for the experimentwise significance level (for example, $\alpha_E = 0.10$) in order to be convincing.

An important property for a planned testing procedure for secondary endpoints is usefully high power (for example, $> 0.70$) for potentially realistic alternatives [3–5]. This property is important because it restricts the scope of secondary endpoints for confirmatory inference to a possibly small set (for example, 1 to 5 members) for which high power might have some documentation through computations or simulations with respect to information from previous studies. In view of this underlying consideration for *a priori* specification for secondary endpoints, a natural question is why not have a secondary endpoint as primary if prior information suggested higher power for it. In cases where such a secondary endpoint has substantial clinical relevance, having it become primary and having a primary endpoint with weaker or more uncertain power become secondary can be a useful strategy [3, 4]. However, secondary endpoints typically have less clinical relevance than a primary endpoint, and whether they have higher power does not have a role

for modifying their priority. In these cases, the use of $\alpha_P = 0.05$ for the primary endpoint and $\alpha_E = 0.10$ for the experimentwise significance level makes confirmatory inferences possible for prespecified secondary endpoints with high power in clinical trials which did not demonstrate statistical significance at $\alpha_P = 0.05$ for the primary endpoint.

The suggestion of $\alpha_E = 0.10$ for the experimentwise significance level and $\alpha_P = 0.05$ for the primary endpoint is probably most useful for clinical trials for disorders which are difficult to evaluate through one or more endpoints, and/or do not have clearly effective treatments, and/or have an active control group. In these situations, inferential evaluation for both primary and secondary endpoints can enable more informative interpretation of the results of the clinical trial. The use of $\alpha_E = 0.10$ is probably not needed for placebo controlled clinical trials to show that a new medicine has beneficial effects on a well understood and dominant endpoint for a particular disorder which already has two or more effective medicines, particularly when all of these treatments belong to the same class. In this situation, significance at $\alpha_E = \alpha_P = 0.05$ for the primary endpoint is a prerequisite for the meaningful evaluation of secondary endpoints. Moreover, for any secondary endpoints which are objectives of confirmatory inference in these clinical trials, *a priori* specification remains necessary and so does a method for controlling the significance level $\alpha_S$ of their statistical tests.

Another consideration for $\alpha_E = 0.10$ and $\alpha_P = 0.05$ is how to evaluate the pattern of observed nominal $p$-values for the primary and secondary endpoints. For this purpose, re-sampling principles like those in Westfall and Young [6] can be useful. As an example, suppose nominal $p_P = 0.08$ for the primary endpoint and nominal $p_S = 0.03$ for the secondary endpoint. Relative to the global null hypothesis of no differences between treatments for both the primary and secondary endpoint, one can generate 10000 re-randomizations and determine the prevalences for such events as (i) $p_P \leqslant 0.05$ or $p_S \leqslant 0.03$, (ii) $p_P \leqslant 0.05$ or $p_S \leqslant 0.05$, (iii) $p_P \leqslant 0.08$ and $p_S \leqslant 0.03$ and (iv) $p_P \leqslant 0.08$ and $p_S \leqslant 0.08$. The prevalence for (i) would be the global $p$-value relative to $\alpha_E = 0.10$ through a prespecified decision rule which focused on significance for the primary endpoint at $\alpha_P = 0.05$ or significance for the secondary endpoint through $p_S \leqslant \alpha_S$ for $\alpha_S$ such that at most 10 per cent of the re-randomizations would have $p_P \leqslant 0.05$ or $p_S \leqslant \alpha_S$ (under the global null hypothesis). The prevalence for (ii) would be the global $p$-value for a prespecified decision rule which focused on significance for the primary endpoint at nominal $\alpha_P = 0.05$ or significance for the secondary endpoint at nominal $\alpha_S = 0.05$, and for positively correlated endpoints, its value would be less than $\alpha_E = 0.10$. The prevalences in (iii) and (iv) pertain to prespecified decision rules which require favourable results for both the primary and secondary endpoints to contradict the global null hypothesis [3], and one can note that they are usually smaller than the nominal $p$-values for either of their components. One can further note that when a prespecified decision rule such as (i)–(iv) contradicts the global null hypothesis, any interpretation for the specific endpoints requires corresponding subsequent evaluation through closed testing procedures [3–5] in order to be convincing. For the case of two endpoints, this assessment could be at $\alpha_P = 0.05$ for the primary endpoint and at $\alpha_S$ from (i) for the secondary endpoint, although their statistical significance would be at $\alpha_E = 0.10$ from an experimentwise perspective. Thus, if $\alpha_P = 0.05$ and $\alpha_S = 0.08$ corresponded to $\alpha_E = 0.10$ for (i), then each of the patterns of nominal $p$-values in (i)–(iv) would contradict the global null hypothesis and the null hypothesis for the secondary endpoint; (i) or (ii) would additionally contradict the null hypothesis for the primary endpoint, and all contradictions are at the $\alpha_E \leqslant 0.10$ experimentwise significance level.

How to apply $\alpha_E = 0.10$ with $\alpha_P = 0.05$ in an effective way to situations with two or more endpoints needs careful planning. If all secondary endpoints are at least moderately positively

 

correlated and at least moderately able to detect treatment differences, one can usefully apply the methods of O'Brien [7] and Lehmacher *et al.* [8] to composite endpoints for which the components are the respective secondary endpoints (for example, rankings of the secondary endpoints are averaged within patients to form the composite endpoint). Such methods have the advantage that all assessments can be at the 0.05 level, but for significance to apply to a particular endpoint, it must apply (via closed testing principles) to all composites (for example, one-way, two-way, three-way, etc.) which contain that endpoint [8]. However, one should recognize that this method may not be useful in situations where the composite endpoints have unsatisfactory power because there are no treatment differences for one or more of the secondary endpoints [3–5].

Another potential method of interest for the previously stated situation could be based on the procedures of Hochberg [9, 10], given that the endpoints were non-negatively correlated. With three endpoints, such a method would significantly contradict the global null hypothesis of equal treatment effects for all three endpoints if they all had nominal $p \leqslant \alpha_{S3} = 0.05$ or if two of the three endpoints had nominal $p \leqslant \alpha_{S2} = 0.025$ (with the third having nominal $p > \alpha_{S3} = 0.05$), or if one of the three secondary endpoints had nominal $p < \alpha_{S1} = 0.0167$ (with the other two having nominal $p > \alpha_{S2} = 0.025$). For this specification, statistical significance at the $\alpha_S = 0.05$ level applies as well to the separate endpoints which provide the basis for the contradiction of the global null hypothesis.

A third strategy would be to use the methodology of Westfall and Young [6] to control the experimentwise significance level for the collection of secondary variables at $\alpha_S = 0.05$, thus ensuring $\alpha_E \leqslant 0.10$ for the primary endpoint and the secondary set taken as a whole. The strategy would incorporate correlations among the secondary variables to reduce the level of conservatism, and is comparable to the previously mentioned method of Lehmacher [8], but is based on the minP statistic [6, 11] instead of an O'Brien statistic [7]. The point here is that the specification of $\alpha_E = 0.10$ does not imply that the evaluation of secondary endpoints should use overly conservative methods. On the contrary, such evaluation is more effective when used with more powerful methods for managing multiplicity among endpoints. Also, recognition of $\alpha_E = 0.10$ clarifies the need for careful prospective planning concerning methods of analysis for secondary endpoints.

An extension of the scope of $\alpha_E = 0.10$ is confirmatory inference for treatment comparisons within one or more prespecified subgroups [3, 12]. The considerations for these assessments are similar to the ones previously stated for secondary endpoints. Most importantly, a subgroup should have some documentation of high power relative to information from previous studies in order to qualify for a secondary inferential role. Since comparisons within subgroups are non-negatively correlated with those for all patients and with one another, their assessment (for the global hypothesis of equal treatment effects for both all patients and for the prespecified subgroups) can have better power through resampling methods for decision rules which are analogous to the procedures of Hochberg [9]. These methods can also address null hypotheses for each specific subgroup through the application of closed testing principles.

In an overall sense, one should recognize that confirmatory clinical trials provide extensive information for multiple endpoints and multiple subgroups. Through randomization and procedures for objective data collection in their designs, they provide a scientifically sound basis for planned treatment comparisons. The specification of only one primary endpoint for a clinical trial is clearly an insufficient statistical use of its valuable information. Moyé's suggestion of $\alpha_E = 0.10$ and the corresponding requirement of better planning for the inferential analysis of secondary endpoints should substantially strengthen the statistical evaluation of confirmatory clinical trials.

 

ACKNOWLEDGEMENTS

The author would like to thank C.E. Davis and P.H. Westfall for helpful comments with respect to the revision of a previous version of this discussion.

REFERENCES

1. Moyé LA. Alpha calculus in clinical trials: considerations and commentary for the new millennium. *Statistics in Medicine* 2000; **19**: 767–779.
2. Davis CE. Secondary endpoints can be validly analyzed, even if the primary endpoint does not provide clear statistical significance. *Controlled Clinical Trials* 1997; **18**: 557–560.
3. Koch GG, Gansky SA. Statistical considerations for multiplicity in confirmatory protocols. *Drug Information Journal* 1996; **30**: 523–534.
4. Koch GG, Davis SM, Anderson RL. Methodological advances and plans for improving regulatory success for confirmatory studies. *Statistics in Medicine* 1998; **17**: 1675–1690.
5. Troendle JF, Legler JM. A comparison of one-sided methods to identify significant individual outcomes in a multiple outcome setting: stepwise tests or global tests with closed testing. *Statistics in Medicine* 1998; **17**: 1245–1260.
6. Westfall PH, Young SS. *Resampling-based Multiple Testing*: *Examples and Methods for p-Value Adjustment*. Wiley: New York, 1993.
7. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984; **40**: 1079–1087.
8. Lehmacher W, Wassmer G, Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* 1991; **47**: 511–532.
9. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**: 800–802.
10. Sarkar S, Chang CK. Simes' method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 1997; **92**: 1601–1608.
11. Westfall PH, Tobias RD, Rom D, Wolfinger RD, Hochberg Y. *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Institute Inc.: Cary, NC, 1999.
12. Koch GG. Discussion of '*p*-Value adjustments for subgroup analyses'. *Journal of Biopharmaceutical Statistics* 1997; **7**: 323–331.

 