Journal of Biopharmaceutical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lbps20

Some Controversial Multiple Testing Problems in Regulatory Applications

H. M. James Hung [a] & Sue-Jane Wang [b]
[a] Division of Biometrics I, OB/OTS/CDER, FDA, Silver Spring , Maryland, USA
[b] Office of Biostatistics, OTS/CDER, FDA, Silver Spring , Maryland, USA
Published online: 07 Jan 2009.

To cite this article: H. M. James Hung & Sue-Jane Wang (2009) Some Controversial Multiple Testing Problems in Regulatory Applications, Journal of Biopharmaceutical Statistics, 19:1, 1-11, DOI: 10.1080/10543400802541693

To link to this article: http://dx.doi.org/10.1080/10543400802541693

Journal of Biopharmaceutical Statistics, 19: 1–11, 2009

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400802541693

Taylor & Francis
Taylor & Francis Group

SOME CONTROVERSIAL MULTIPLE TESTING PROBLEMS IN REGULATORY APPLICATIONS

H. M. James Hung[1] and Sue-Jane Wang[2]
[1]*Division of Biometrics I, OB/OTS/CDER, FDA, Silver Spring, Maryland, USA*
[2]*Office of Biostatistics, OTS/CDER, FDA, Silver Spring, Maryland, USA*

Multiple testing problems in regulatory applications are often more challenging than the problems of handling a set of mathematical symbols representing multiple null hypotheses under testing. In the union-intersection setting, it is important to define a family of null hypotheses relevant to the clinical questions at issue. The distinction between primary endpoint and secondary endpoint needs to be considered properly in different clinical applications. Without proper consideration, the widely used sequential gate keeping strategies often impose too many logical restrictions to make sense, particularly to deal with the problem of testing multiple doses and multiple endpoints, the problem of testing a composite endpoint and its component endpoints, and the problem of testing superiority and noninferiority in the presence of multiple endpoints. Partitioning the null hypotheses involved in closed testing into clinical relevant orderings or sets can be a viable alternative to resolving the illogical problems requiring more attention from clinical trialists in defining the clinical hypotheses or clinical question(s) at the design stage.

In the intersection-union setting there is little room for alleviating the stringency of the requirement that each endpoint must meet the same intended alpha level, unless the parameter space under the null hypothesis can be substantially restricted. Such restriction often requires insurmountable justification and usually cannot be supported by the internal data. Thus, a possible remedial approach to alleviate the possible conservatism as a result of this requirement is a group-sequential design strategy that starts with a conservative sample size planning and then utilizes an alpha spending function to possibly reach the conclusion early.

Key Words: Composite endpoint; Group-sequential; Intersection-union; Noninferiority; Sequential gate keeping; Studywise; Superiority; Union-intersection.

1. INTRODUCTION

In the late phase of drug development, a statistical hypothesis under testing in a clinical trial is driven by a clinical question or hypothesis, particularly for beneficial effects, which a test treatment is anticipated to yield. A frequently asked clinical question is: "will the test treatment reduce morbidity or mortality and/or relieve undesirable symptoms?" A clinical benefit is often characterized by a set of

Received June 11, 2008; Accepted October 10, 2008
Address correspondence to H. M. James Hung, Division of Biometrics I, OB/OTS/CDER, FDA, 10903 New Hampshire Ave, Bldg 21 Room 4616, HFD-710 Silver Spring, MD 20993-0002, USA; E-mail: hsienming.hung@fda.hhs.gov

correlated response (or efficacy) variables. The correlations among the variables are unknown, but may be estimable from external or internal data.

To prove a therapeutic benefit may require that at least one or h of the K efficacy variables ($h \leq K$) under consideration show statistically significant results in favor of the test treatment. The relevant type I error rates associated with such statistical significance tests need to be controlled adequately. In regulatory practices, however, it is often not quite clear exactly what type I error rate needs to be controlled, from a consideration of the entire clinical testing program. Should it be per hypothesis error, per study error, or even per therapeutic indication error? A widely used criterion is, at the very least, to control the studywise type I error rate in each trial that is taken for a formal statistical inference. In addition, it is usually required that the statistically significant findings be replicated in at least two trials. Identification of positive findings entails that each concerned study meet a proper control of the relevant "familywise" or "studywise" type I error rate if multiple statistical hypotheses are tested in that study. Another dimension is the proper description of the proven clinical benefits of the test treatment in labeling. In many disease areas, the labeling process calls for inclusion of endpoint(s) that shows clinical benefits with significant two-sided p-value(s) (e.g., <0.05 or $\ll 0.05$).

When multiple hypotheses are tested in a single trial, the issues with multiple comparisons arise, particularly in the union-intersection setting; e.g., asserting that the trial yields a positive finding is based on at least one test. There are many viable methods for handling multiple comparisons in the literature. For some widely used methods, the readers are referred to Hochberg and Tamhane (1987), Hsu (1996), and the articles cited therein. In other types of applications, to prove a benefit of the test treatment requires that all K efficacy variables show statistically significantly favorable results. As a similar scenario, in combination drug development, a typical requirement is that the combination drug be demonstrated superior to each component drug. For this type of application (this is the so-called intersection-union setting), the usual concern of multiple comparisons in the union-intersection setting is irrelevant. However, many argue that the requirement that each test be statistically significant at the two-tailed 5% level is too stringent or too conservative, and that there is a way to alleviate the conservatism, e.g., by performing each test at an alpha level higher than 5%.

In this paper we shall visit a few commonly encountered controversial multiple testing problems under the union-intersection setting in Section 2 and under the intersection-union setting in Section 3 in regulatory applications. Concluding remarks follow.

2. SOME CONTROVERSIAL SEQUENTIAL GATE KEEPING STRATEGIES

As articulated above, for a pivotal clinical trial that has any implication on a regulatory decision regarding clinical benefits of a test treatment, statistical inference hinges on an adequate adjustment for multiple testing in the union-intersection setting. A widely used strategy is often referred to as a sequential gate keeping strategy. By this strategy, the primary study endpoint is always tested first and consumes the entire intended level of alpha. If the primary endpoint fails to achieve

statistical significance, the study is regarded as a failure and no secondary endpoint will be tested. Otherwise, the study will be declared "statistically positive" and the positive findings of the primary endpoint will then be subjected to clinical interpretation. In addition, the secondary endpoints will be tested subsequently according to a prespecified hierarchical order, applying the same alpha level as that for testing the primary endpoint to each step of the secondary endpoint testing, until the nominal p-value exceeds this alpha level and the sequential testing will stop. This sequential gate keeping strategy is attractive in practice because the primary endpoint often addresses the most important study question and thus deserves consumption of the entire alpha. In addition, it allows for testing the secondary endpoint at the same alpha level, provided that the primary endpoint reaches statistical significance. When the success or failure of the study is driven by the primary endpoint, this sequential strategy makes logical sense.

The sequential gate keeping strategy is a close testing procedure (Marcus et al., 1976), and hence it controls the studywise type I error rate strongly in the most common typical scenario, wherein the primary endpoint is tested once and each secondary endpoint is tested at most once. The studywise type I error rate is often defined as the maximum familywise type I error rate associated with testing the family of the relevant null hypotheses and all their intersections. The essential feature of this strategy is that all the possible tests must be lined up in a single chain. This feature imposes logical restrictions that can make the strategy senseless, at least, the following three scenarios.

Multiple Doses and Multiple Endpoints

Consider the simplest scenario that two doses (high and low) of a test drug are tested against a placebo arm on a single primary endpoint and a single secondary endpoint that are relevant to a labeling claim. For the single primary endpoint, applying one of the strong-control multiple comparison procedures, such as the Dunnett (1955) procedure, one can be assured that either or both doses can be asserted to beat the placebo arm statistically with confidence that the probability of making such a false conclusion is properly controlled at the desired alpha level. However, the sequential gate keeping strategy, starting with the application of this type of multiple comparison procedures (e.g., Dunnett procedure) to the primary endpoint for the two doses, would require that both doses beat the placebo arm in order to test the secondary endpoint for either dose or both doses. If the high dose beats the placebo arm on the primary endpoint, but the low dose does not, then this sequential strategy cannot allow for testing the high dose on the secondary endpoint at the same alpha used for testing the primary endpoint. This requirement is controversial in practice, because it does not make common sense for the following reason. If there is sufficient evidence to conclude that the study is positive and the high dose is effective with respect to the placebo, then why does the secondary endpoint testing for this dose have to be subjected to proving that the low dose is also effective with respect to the primary endpoint? The sequential gate keeping strategy as described breaks down for this problem.

To alleviate this controversy with logical restrictions, a number of parallel gate keeping strategies, such as Dmitrienko et al. (2003) and Dmitrienko et al. (2007), are developed in the literature. The essence of these new approaches is that some extent

of alpha allocation would be in place to multiple chains when they are present. Nonetheless, the most fundamental question is whether the studywise type I error rate needs to be tied in with testing secondary endpoints, particularly when the success or failure of the study is determined purely by the primary endpoint. When the primary endpoint has demonstrated that the test drug at an identified dose(s) is efficacious, the study is already declared a win. From this standpoint, the studywise type I error rate seems to be associated only with any false assertions regarding the primary endpoint. The inference regarding the secondary endpoints is conditional only from the standpoint that the study is positive on the primary endpoint.

Composite Endpoint

Composite endpoints are widely used for studying treatment effects on mortal or morbid outcomes (e.g., death, myocardial infarction, and stroke) as a whole, in disease trials such as cardiovascular or renal trials. The composite endpoints are often tested using time to event methods. Such a global assessment appears to avoid the problem of multiple testing, so long as there is no formal testing for each component of the composite endpoint. However, if the test treatment demonstrates a statistically significantly favorable effect on a composite endpoint, it is almost always necessary to describe the treatment effect on each component. Consequently, some type of statistical analysis of the time to each component is often performed with the nominal p-value, the effect estimate, and the confidence interval being reported. These statistics will directly or indirectly have inferential implications.

The component endpoints in some sense form a cluster. Intuitively, after the composite of these component endpoints achieves statistical significance, the test treatment demonstrates a favorable effect on the composite endpoint; thus, a strong control of type I error rates associated with testing the components appears to be sufficient. Also quite appealing is the sequential test strategy that starts with testing the targeted composite endpoint as the primary endpoint and then tests the component endpoints using a strong control multiple testing method after the composite endpoint reaches statistical significance. However, for testing another secondary endpoint, it will be very difficult to expand this sequential test strategy in order to incorporate this additional secondary endpoint in the sequential testing chain. If this secondary endpoint is tested after the cluster of the component endpoints, the chain will require that all the components meet statistical significance. Thus, to avoid a confrontation with this restriction, the components must be placed at the end of the chain. This is awkward. Why does testing of the component endpoints have to be conditional on a positive finding on that secondary endpoint? This is another scenario where the sequential gate keeping strategy that is statistically valid breaks down in terms of common sense.

Testing Superiority and NonInferiority for Multiple Endpoints

Consider an active control clinical trial where a test drug is compared with an active control drug for both superiority and well-defined noninferiority (i.e., with a prespecified well-defined noninferiority margin). For the primary endpoint (labeled E1), it is asymptotically (in the sense of large sample) valid to test at the same nominal alpha level for both superiority and noninferiority; see Morikawa

and Yoshida (1995), Dunnett and Tamhane (1997), and the relevant articles cited therein. That is, if the same 95% asymptotic confidence interval for the test drug versus the active control drug is used to test for both superiority and noninferiority, the overall type I error rate will be controlled at a 5% level.

Now suppose that a secondary endpoint (labeled E2) will be tested for noninferiority at the same alpha level as that for testing E1, after E1 demonstrates that the test drug is superior (abbreviated by *S*) or noninferior (abbreviated by NI) to the control. If the pre-condition for testing E2 is that E1 demonstrates NI and no superiority testing for E1 is performed, then this sequential gate keeping strategy performed at the same nominal alpha at each step will have a strong control of the 'studywise' type I error rate at alpha. However, there is a problem when the precondition for testing E2 is that E1 demonstrates NI and testing E1 for *S* is also performed. Although there is only one testing procedure (i.e., the same 95% confidence interval) for simultaneously or sequentially testing *S* and NI for E1, such a sequential gate keeping strategy applied to testing E2 after E1 that shows *S* or NI, but not both, may not have a proper control of the studywise type I error rate. For instance, the probability of falsely asserting superiority for E1 or falsely asserting noninferiority for E2 can double the intended alpha level. If we start with testing NI for E1 at alpha, we can only consider one of the following chains where the same alpha is applied to each conditional step:

Chain 1: Test NI for E1 $\rightarrow$ test *S* for E1 $\rightarrow$ test NI for E2
Chain 2: Test NI for E1 $\rightarrow$ test NI for E2 $\rightarrow$ test *S* for E1

in order to have a proper control of the studywise type I error rate associated with the three tests, NI for E1, *S* for E1, and NI for E2 at alpha. Either chain performed singly and exclusively is statistically valid. However, why does testing superiority for E1 have to be conditional on the finding from testing NI for E2 and vice versa? Thus, neither chain is logically sensible. No sequential gate keeping strategy that meets logical or common sense can be derived for this problem.

An important remark is in order. The mathematical constraints that are illogical in the union-intersection problems might be alleviated using the partitioning concept proposed by Hsu (1996). That is, all relevant null hypotheses involved in the closed testing hierarchy may be categorized into clinical relevant orderings or clinical relevant sets. For instance, in the multiple dose, multiple endpoint problem described above, if the main study goal is to identify an effective dose of the test drug, partitioning of the null hypotheses should be carefully considered to allow the sequential testing strategy to carry out within each dose path, without having to be conditional on the finding of the other dose. The partitioning approach requires more attention from clinical trialists in defining clinical hypotheses or clinical question(s) at the design stage and can avoid illogical close test constraints.

3. REQUIRING STATISTICAL SIGNIFICANCE OF ALL MULTIPLE ENDPOINTS

This is frequently referred to as a coprimary endpoint problem. The theory of statistical inference for joint statistical significance of multiple tests is well known from statistical literature, e.g., Lehmann (1952), Berger (1982), Laska and Meisner

(1989), and the relevant articles cited therein. For ease of presentation, let us consider a two endpoints situation. Let (δ_1, δ_2) denote the standardized effect sizes of the two response variables for the test drug versus the placebo arm. Suppose that a positive δ value represents a treatment benefit of interest. Under the intersection-union setting, the hypotheses for testing are $H_1 : \delta_1 > 0$ and $\delta_2 > 0$ versus $H_0 : \delta_1 \leq 0$ or $\delta_2 \leq 0$, which are tested by the test statistics denoted by (T_1, T_2). Assume that asymptotically, $(T_1, T_2) \sim N((n/2)^{1/2}(\delta_1, \delta_2), [1, 1, \rho])$, where ρ is the asymptotic correlation of the two tests and n is the per-group sample size.

On the boundary of H_0, the joint distribution of (T_1, T_2) has a mean parameter that can still range from 0 to ∞. The rejection region of H_0 is $[T_1 > C_1, T_2 > C_2]$, where C_1 and C_2 are the critical values to be determined. Obviously, it can be shown that the type I error probability associated with this rejection region is an increasing function of (δ_1, δ_2). Thus, without any restriction on the parameter space of (δ_1, δ_2) under H_0, it is necessary to evaluate the type I error probability at $(n/2)^{1/2}(\delta_1, \delta_2) = (0, \pm\infty)$ or $(\pm\infty, 0)$. Consequently, the maximum type I error probability is max $\{\Phi(-C_1), \Phi(-C_2)\}$, where Φ is the cumulative distribution function of the standard Gaussian distribution. In fact, this maximum is almost achievable at $(n/2)^{1/2}(\delta_1, \delta_2) = (0, \pm 3.5)$ or $(\pm 3.5, 0)$, as illustrated in Figure 1 of Hung et al. (1994). Hence, the α-level critical value is $C_1 = C_2 = z_\alpha$, where z_α is the upper αth percentile of the standard normal distribution. This means that the only way to control all possible type I error rates associated with such joint significance testing at 1-sided 2.5% level requires each endpoint to be significant at the 1-sided 2.5% level. This is also consistent with the convention that for drug labeling the statistical significance criterion for an individual endpoint is that its two-sided p-value is no greater than 5% (or one-sided p-value is no greater than 2.5%).

The type I error probability associated with the joint testing each at one-sided 2.5% level has the lowest level of $(0.025)^2 = 0.000625$ at $(\delta_1, \delta_2) = (0, 0)$ and $\rho = 0$ in the domain of $\rho \geq 0$. The fallacy with controlling type I error only at $\delta_1 = \delta_2 = 0$ is

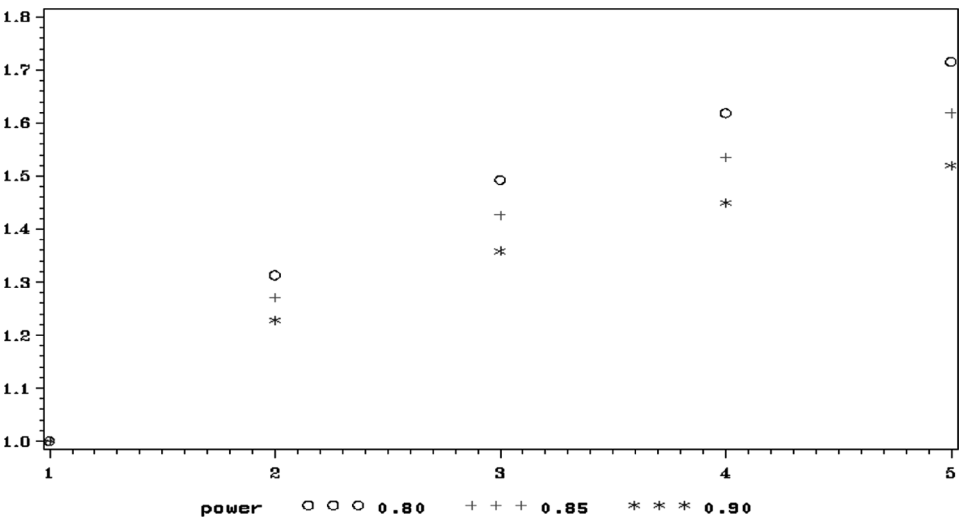

Figure 1 $n(m_1)^{-1}$ versus K.

that one can have sufficiently many endpoints with each having $p \approx 0.50$ (no signal at all), but that the type I error probability associated with joint significance testing ≤ 0.025 at all $\delta_k = 0$, i.e., $\Pr\{T_k \geq 0 \text{ for } k = 1, \ldots, K \mid \text{all } \delta_k = 0\} \leq 0.025$ for a sufficiently large K. Moreover, use of a critical value $< z_{0.025}$ (i.e., $p > 0.025$) for asserting treatment benefit requires justification for such a restriction on the null hypothesis, a fundamental change in the labeling rule, and a justification why a p-value substantially larger than 0.025 is significant that contradicts to the convention. Nevertheless, requiring each endpoint to be significant at 2.5% level can be conservative and demands a big price in sample size in most applications. Chuang-Stein et al. (2007) and Offen et al. (2007) recently proposed a few options for consideration to deal with this problem.

At each value of min (δ_1, δ_2), the statistical power for declaring joint statistical significance is lowest when $\delta_1 = \delta_2$. Thus, for the most conservative sample size planning, one could set

$$\Pr\{T_1 > z_\alpha, T_2 > z_\alpha \mid \delta_1 = \delta_2 = \delta, \rho = 0\} = 1 - \beta.$$

This will lead to a per-group sample size $n = 2(z_\alpha + z_\gamma)^2/\delta^2$, where $\gamma = 1 - (1 - \beta)^{1/2}$ and $1 - \beta$ is the power set for showing joint statistical significance. For each endpoint, the power level required for detecting the effect size δ is not much larger than $1 - \beta$ for $\beta \leq 0.20$. In fact, for $\beta \leq 0.20$, γ is approximately $\beta/2$.

In the following, we compare the most conservative sample size plan with two sample size plans.

(S1) Plan sample size for detecting the endpoint with min $(\delta_1, \delta_2) = \delta$ at power $1 - \beta$. This will lead to a per-group sample size $m_1 = 2(z_\alpha + z_\beta)^2/\delta^2$, assuming one endpoint has a larger effect size than the other (so the power for showing joint significance is not much smaller than $1 - \beta$).

(S2) Plan sample size for detecting the endpoint with max $(\delta_1, \delta_2) = \delta$ at power $1 - \beta$ after Bonferroni adjustment. This would lead to a per-group sample size $m_2 = 2(z_{\alpha/2} + z_\beta)^2/\delta^2$. This is a frequently used sample size plan for showing that at least one endpoint is statistically significant, using the most conservative adjustment for multiple comparisons.

Note that the *power* referenced in S1 and S2 pertains only to each single endpoint.

The comparison can be generalized to the situation of $K > 2$ endpoints. Let n be the per-group sample size required for showing that all K endpoints are significant at α level and power $1 - \beta$ (assuming $\delta_1 = \cdots = \delta_K = \delta$), and they are statistically independent; let m_1 be the per-group sample size required for detecting the single endpoint with $\min(\delta_1, \ldots, \delta_K) = \delta$ at α level and power $1 - \beta$; let m_2 be the per-group sample size required for detecting the single endpoint with max $(\delta_1, \ldots, \delta_K) = \delta$ at α/K level (Bonferroni adjustment) and power $1 - \beta$; then we have

$$\frac{n}{m_1} = \left[\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}\left(\sqrt[K]{1-\beta}\right)}{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}\right]^2$$

$$\frac{n}{m_2} = \left[\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}\left(\sqrt[K]{1-\beta}\right)}{\Phi^{-1}(1-\alpha/K) + \Phi^{-1}(1-\beta)}\right]^2$$

Figures 1 and 2 illustrate the relationship between $n(m_1)^{-1}$ and K and the relationship between $n(m_2)^{-1}$ and K at $\alpha = 2.5\%$. It can be seen that when K is not large, the degree of conservatism is not large.

Even with five endpoints, the most conservative sample size for showing each significant at the 2.5% level is approximately 50% more than the sample size for showing that the single endpoint with the smallest effect size is significant and approximately 7% larger than the sample size for showing that at least one endpoint is statistically significant under the Bonferroni's adjustment, when the power is 90%.

Group Sequential Design Strategy

The most conservative sample size planning arguably may substantially overpower the study because of planning under the most pessimistic assumption about ρ and the relationship of the δ's. One remedial approach is the use of the group sequential design that allows interim termination for futility or for sufficient evidence of joint statistical significance of the endpoints. Asymptotically, the maximum type I error probability can be expressed as

$$\max \Pr\{T_1 > C \text{ and } T_2 > C \mid H_0\} = \Pr\{Z > C\} = \Phi(-C),$$

where

$$\begin{aligned} Z &= \pi Z_1 + (1-\pi)Z_2 \\ \pi &= 1 \quad \text{if } n^{1/2}(\delta_1 - \delta_2) \to \infty, \\ &\quad\ 0 \quad \text{if } n^{1/2}(\delta_1 - \delta_2) \to -\infty \\ (Z_1, Z_2) &\sim N((0,0), [1, 1, \rho]). \end{aligned}$$

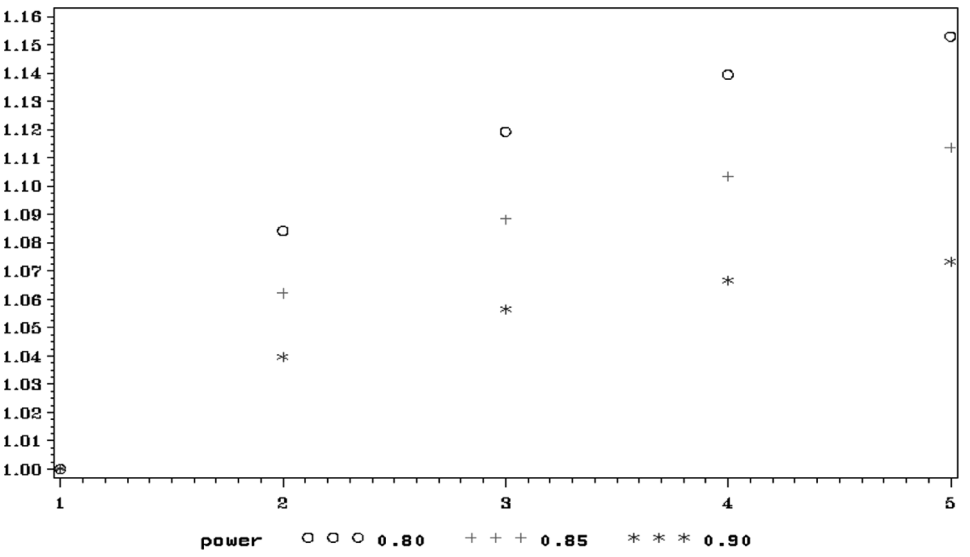

Figure 2 $n(m_2)^{-1}$ versus K.

Table 1 *N*-ratio for two endpoints with correlation of 0.5

	O'Brien-Fleming type		Pocock type	
(δ_1, δ_2)	P(rej H_0)	N ratio	P(rej H_0)	N ratio
(0, 0)	0.005	1.000	0.004	0.999
(.3, 0)	0.025	0.999	0.025	0.993
(.3, .3)	0.819	0.942	0.762	0.812
(.5, .3)	0.891	0.886	0.858	0.733
(.5, .5)	0.999	0.662	0.999	0.547

This expression is very useful for developing the strategy of repeated significance testing for H_0.

With a group sequential design we perform repeated significance testing at information times, $t_1, \ldots, t_m$ $(= 1)$, say, during the trial. Let $A_i = [\min(T_{1i}, T_{2i}) > C_i]$. max type I error probability

$$= \max \Pr\Big\{ \bigcup A_i \,|\, H_0 \Big\}$$
$$= \Pr\Big\{ \bigcup [Z_i \equiv \pi Z_{1i} + (1 - \pi) Z_{2i} > C_i] \Big\},$$

where $\cup$ is the union operator over $\{1, \ldots, m\}$. The process $\{Z_i : i = 1, \ldots, m\}$ is a standard Brownian motion process; thus, the rejection boundaries C_i can be generated as usual using Lan-DeMets alpha-spending method (1983).

Suppose that we have two endpoints with a correlation of 0.5 and that the most conservative sample size plan is made to detect $(\delta_1, \delta_2) = (0.3, 0.3)$ at $\alpha = 0.025$ and $\beta = 0.20$. The group sequential design allows an interim analysis at 50% information time. Let N-ratio be the ratio of the average sample size with the group sequential design to the most conservatively planned sample size. Table 1 presents the N-ratio under a variety of (δ_1, δ_2) under H_0 and H_1, based on the simulation results with 100,000 replications per run. The group sequential design can yield substantial savings on average sample size. For example, when $(\delta_1, \delta_2) =$ (.3, .3), if the O'Brien-Fleming type alpha spending function is used, the average sample size is 94% of the most conservative sample size. As the effect size is larger, the power is larger and so is sample size saving.

4. CONCLUDING REMARKS

In the late phase of drug development for registration, there generally is no maximum number of clinical trials that are considered for regulatory decision making. Indeed, we have seen a few applications in which more than ten clinical trials were conducted to study the possible clinical benefits of the test drugs. As such, finding a treatment effect in a sea of many negative studies clearly raises some type of post hoc multiple testing problem that is usually very difficult to address. The requirement of imposing a strong control on some type of *familywise* type I error rate in each trial seems to be a good minimum regulatory standard. The question, though, is how to define a relevant family of null hypotheses under testing in a given clinical context. For example, in one application area, identification of

an effective dose may be clinically more important than evaluation of secondary endpoints, whereas in another area the priority is reversed. The distinction between these two application scenarios is usually ignored in discussing control of the *studywise* type I error rate, and thus the methods applied work on the same set of mathematical symbols, namely, H_0's as the null hypotheses. Moreover, the distinction between the primary endpoint and the secondary endpoint may be important in many clinical contexts. If the study win criterion is dictated merely by the primary endpoint, then the studywise type I error rate seems to be more appropriately associated with the primary endpoint alone than all endpoints, particularly in these clinical contexts. If, however, the designation of primary versus secondary is considered mostly based on the feasibility to conduct the trial, then these endpoints might be of equal importance. Thus, in this setting, the studywise type I error rate arguably is associated with all these endpoints that should be considered together as a family. For these reasons, the pitfalls of the so-called sequential gate keeping strategy may need to be reconsidered, seeing that it imposes too many logical restrictions to be sensible in dealing with many application problems. A sensible alternative might be to define clinical relevant null hypotheses and the relevant orderings for regulatory decision by partitioning all null hypotheses into relevant versus irrelevant ordering sets accordingly.

For the intersection-union setting, joint statistical significance of endpoints at level α requires that each endpoint be statistically significant at level α, unless the parameter space under the null hypothesis can be substantially restricted. Such restriction frequently requires insurmountable justification and usually cannot be supported by the internal data. This requirement that each endpoint be statistically significant is also consistent with the usual practice that inclusion of a significant endpoint in drug label entails a two-sided p-value smaller than 0.05. A solution is to minimize the number of endpoints to be considered as the coprimary endpoints from the design. In addition, the statistical design may consider planning sample size under the most pessimistic scenario and then utilizing, if feasible, a group sequential method to alleviate the possible conservatism with such sample size planning.

Disclaimer

The views expressed in this paper are not necessarily those of the U.S. Food and Drug Administration.

REFERENCES

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics* 24:295–300.

Chuang-Stein, C., Stryszak, P., Dmitrienko, A., Offen, W. (2007). Challenge of multiple co-primary endpoints: a new approach. *Statistics in Medicine* 26:1181–1192.

Dmitrienko, A., Offen, W. W., Westfall, P. H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine* 22:2387–2400.

Dmitrienko, A., Wiens, B. L., Tamhane, A. C., Wang, X. (2007). Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in Medicine* 26:2465–2478.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 50:1096–1121.

Dunnett, C. W., Tamhane, A. C. (1997). Multiple testing to establish superiority-equivalence of a new treatment compared with kappa standard treatments. *Statistics in Medicine* 16:2489–2506.

Hochberg, Y., Tamhane, A. J. (1987). *Multiple Comparison Procedures*. New York: John Wiley & Sons.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. London: Chapman & Hall.

Hung, H. M. J., Chi, G. Y. H., Lipicky, R. L. (1994). On some statistical methods for analysis of combination drug studies. *Communications in Statistics—Theory and Methods* 23(2):361–376.

Lan, K. K. G., DeMets, D. L. (1983). Discrete sequential boundaries in clinical trials. *Biometrika* 70:659–663.

Laska, E. M., Meisner, M. (1989). Testing whether an identified treatment is best. *Biometrics* 45:1139–1151.

Lehmann, E. L. (1952). Testing multiparameter hypotheses. *Annals of Mathematical Statistics* 23:541–542.

Marcus, R., Peritz, E., Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 63:655–660.

Morikawa, T., Yoshida, M. (1995). A useful testing strategy in phase III trials: combined test of superiority and test of equivalence. *Journal of Biopharmaceutical Statistics* 5(3): 297–306.

Offen, W., Chuang-Stein, C., Dmitrienko, A., Littman, G., Maca, J., Meyerson, L., Muirhead, R., Stryszak, P., Boddy, A., Chen, K., Copley-Merriman, K., Dere, W., Givens, S., Hall, D., Henry, D., Jackson, J. D., Krishen, A., Liu, T., Ryder, S., Sankoh, A. J., Wang, J., Yeh, C. H. (2007). Multiple co-primary endpoints: medical and statistical solutions. *Drug Information Journal* 41:31–46.