Walter Offen (Chair)
Eli Lilly and Company

Christy Chuang-Stein
Pfizer

Alex Dmitrienko
Eli Lilly and Company

Gary Littman
Wyeth

Jeff Maca
Novartis

Laura Meyerson
Biogen-Idec

Robb Muirhead
Pfizer

Paul Stryszak
Schering Plough

Alex Boddy
Sanofi-Aventis

Kun Chen
Bayer

Kati Copley-Merriman
Pfizer

Willard Dere
Amgen

Sam Givens
Hoffman-LaRoche

David Hall
Boehringer-Ingelheim

David Henry
Bristol Myers Squibb

Joseph D. Jackson
Bristol Myers Squibb

Alok Krishen
GlaxoSmithKline

Thomas Liu
Amgen

Steve Ryder
Pfizer

A. J. Sankoh
Sanofi-Aventis

Julia Wang
J&J PRD

Chyon-Hwa Yeh
Procter & Gamble

Key Words
Bayesian approach; Mixed Bayesian/frequentist approach; Multiple endpoint expert team; Multiplicity adjustment; Restricted null space; Reverse multiplicity

Correspondence Address
Walter W. Offen, Global Statistical Sciences, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285-2233 (e-mail: offen_walter_w@lilly.com).

Multiple Co-primary Endpoints: Medical and Statistical Solutions
A Report From the Multiple Endpoints Expert Team of the Pharmaceutical Research and Manufacturers of America

There are quite a few disorders for which regulatory agencies have required a treatment to demonstrate a statistically significant effect on multiple endpoints, each at the one-sided 2.5% level, before accepting the treatment's efficacy for the disorders. Depending on the correlation among the endpoints, this requirement could lead to a substantial reduction in the study's power to conclude the efficacy of a treatment. To investigate the prevalence of this requirement and propose possible solutions, a multiple-disciplinary Multiple Endpoints Expert Team sponsored by Pharmaceutical Research and Manufacturers of America was formed in November 2003. The team recognized early that many researchers were not fully aware of the implications of requiring multiple co-primary endpoints. The team proposes possible solutions from both the medical and the statistical perspectives. The optimal solution is to reduce the number of multiple co-primary endpoints. If after careful considerations, multiple co-primary endpoints remain a scientific requirement, the team proposes statistical solutions and encourages that regulatory agencies be receptive to approaches that adopt modest upward adjustments of the nominal significance levels for testing individual endpoints. Finally, the team hopes that this report will draw more attention to the problem of multiple co-primary endpoints and stimulate further research.

INTRODUCTION

Most human diseases are characterized by multiple measures, including signs, symptoms, quantitative measurements, and patient-reported outcomes. Migraine, for example, is characterized by moderate-to-severe headache pain that is frequently accompanied by nausea, photophobia, and phonophobia. Arthritis patients experience not only pain, but also swelling and stiffness in their joints. Alzheimer's disease is characterized by poor cognition and disorderly behavior or deficits in activities of daily living. Other disorders manifesting through multiple measures include depression, multiple sclerosis, psoriasis, and lupus erythrematosus. Generally, a clinically meaningful improvement in these disorders is assessed by improvements in multiple measures. As a result, when an intervention is assessed for its effect on these disorders in a clinical trial setting, the effect is typically examined via multiple endpoints that describe the state of, or measure the change in, the multiple measures.

It is customary to classify efficacy endpoints as primary or secondary when evaluating an intervention's efficacy. *Primary endpoints* describe how the most important aspects of the disease are affected by the intervention. In addition to primary endpoints, an intervention's effect on the secondary endpoints can further help prescribing physicians in identifying suitable treatments for their patients. While it may be clinically desirable or necessary to consider multiple endpoints as primary, making statistical decisions based on multiple primary endpoints could have a substantial impact on the probabilities associated with erroneous decisions.

For clarity, we differentiate between two types of multiplicity of the primary endpoints. The

first case is when an intervention is deemed efficacious if it improves on at least one of the multiple primary endpoints. The second case is when an intervention is deemed efficacious *only if* it improves on all of the multiple primary endpoints. For simplicity, we call the multiple endpoints in the first case *alternative* primary endpoints. The word *alternative* is used to indicate that each primary endpoint is an alternative to other primary endpoints in determining the efficacy of the intervention. The multiple primary endpoints in the second case is called multiple *co-primary* endpoints to represent the simultaneous improvements required of the intervention. It is the latter on which we focus.

Much statistical literature has been devoted to the case of alternative primary endpoints (eg, Ref. 1). The central issue there is to control the false-positive rate at the study level since there are many chances to declare efficacy. A similar issue occurs when a sponsor is interested in claiming statistical significance of selected secondary endpoints in the Clinical Studies section of the label. A common approach to handle this traditional multiplicity problem is to adjust the significance level downward for individual testings so that the overall false-positive rate can be maintained at a desirable level (2).

By comparison, the implications of multiple co-primary endpoints are less recognized. The regulatory position in this case is to test each primary endpoint at the (two-sided) 5% level if 5% is the allowable studywise false-positive rate (3,4). Since this situation requires statistical significance on all primary endpoints, with superiority of study drug over placebo for all endpoints, we call this a *reverse* multiplicity problem. The word *reverse* is used to differentiate this situation from the one discussed in the preceding paragraph. Since a two-sided 5% level translates to a 2.5% level when testing in the desirable direction, without loss of generality, we focus on one-sided alternatives for the remainder of the article.

The 2.5% level discussed is required by regulatory agencies regardless of how many endpoints are on the co-primary list. The reason for using the 2.5% significance level for individual tests, rooted in the desire to control the chance of erroneously concluding efficacy, is given in the next section. This desire does come with a price. The price is in the form of study power and sample size, which are also described in the next section.

Reverse multiplicity exists with many common disorders. For example, diseases that have required two co-primary endpoints include Alzheimer's disease, asthma, chronic obstructive pulmonary disease, skin aging, fracture healing, male pattern baldness, and organ transplantation. There are disorders that have required three or more co-primary endpoints. These include migraine, sleep disorders, osteoarthritis, acne, and glaucoma. Vaccines present a special case, often requiring 10 or more co-primary immunogenicity endpoints. Table 1 gives a list of disorders known to us for which regulatory agencies have required multiple co-primary endpoints when assessing the effect of an intervention. As clinical measurements become more sophisticated with new technology and as more multifaceted diseases are studied, we are likely to encounter additional situations for which an intervention will be required to demonstrate a statistically significant effect on multiple co-primary endpoints to obtain marketing authorization.

Because of the prevalence of the reverse multiplicity problem and the expected increase in the use of multiple co-primary endpoints by regulatory agencies, a multiple-disciplinary Multiple Endpoints Expert Team (MEET) was formed in November 2003. MEET, consisting of statisticians, clinicians, and outcome research scientists, was charged to address both the clinical and statistical challenges brought on by multiple co-primary endpoints. The team, chaired by Walter Offen, was sponsored by the Biostatistics and Data Management Technical Group of the Pharmaceutical Research and Manufacturers of America (PhRMA). MEET also had the endorsement of the Clinical Leadership Committee of PhRMA. The team discussed options from both the medical and the statistical perspectives. The number one recommendation from the team is to see whether a single primary endpoint could

List of Diseases (Along With Endpoints and Correlation Estimates) for Which Regulatory Agencies Have Required Two or More Co-primary Endpoints

TABLE 1

Disease	Endpoints	Correlation*
Migraine	A. Pain free at 2 hours B. Nausea at 2 hours C. Photosensitivity at 2 hours D. Phonosensitivity at 2 hours	B C D A M M M B L L C M
Acute pain (single-day multiple doses)	A. Patient global assessment B. Sum of pain intensity change over 24 hours C. Sum of pain relief over 24 hours	B C A M H B H
Alzheimer's disease	A. Alzheimer's Disease Assessment Scale–Cognitive (ADAS-Cog) B. Clinician Interview-Based Impression of Change (CIBIC)	B A L
Fibromyalgia	A. Pain reduction B. Patient Global Improvement C. Health Outcome Measure	B C A M M B M
Low back pain	A. Pain intensity (VAS) B. Functional status C. Patient global assessment	B C A M H B M
Osteoarthritis	A. Pain scale B. Patient global assessment C. Function (eg, HRQOL)	B C A H H B H
Asthma, chronic obstructive pulmonary disease	A. FEV1 (currently accepted as the single primary endpoint at FDA) B. Symptomatic benefit (from CHMP Points to Consider)	B A L
Erectile dysfunction	A. IIEF Erectile Function Domain (sum of 6) B. SEP question 2 (% attempts with successful insertion) C. SEP question 3 (% attempts with successful intercourse)	B C A H H B H
Symptom modifying for osteoarthritis	A. Pain B. Functional disability (from CHMP Points to Consider)	B A M
Menopausal symptoms	A. Hot flash count at week 4 B. Hot flash severity score at week 4 C. Hot flash count at week 12 D. Hot flash severity score at week 12	B C D A M H M B L M C M
Fracture healing	A. Radiographic healing B. Functional/clinical endpoint	NA
Acne†	A. Physician global assessment B. Inflammatory lesion count C. Noninflammatory lesion count D. Total	NA
Male pattern baldness	A. Hair count B. Patient assessment	NA
Glaucoma	Nine endpoints: Intraocular pressure at weeks 2, 6, and 12 for three timepoints: 8 AM, 10 AM, and 4 PM	H (among the nine measurements)
Vaginal atrophy (VA)	A. Subject self-assessed most bothersome VA symptom B. Vaginal pH C. Percentage of vaginal parabasal cells from the maturation index D. Percentage of vaginal superficial cells from the MI	B C D A L L L B L L C L

TABLE 1

Continued

Organ transplantation	A. Composite—biopsy-proven acute rejection, graft loss, death at 6 months B. Composite—graft loss or death at 12 months	B A L-M
Primary biliary cirrhosis	A. Cholate injury B. Portal inflammation C. Piecemeal necrosis D. Fibrosis	B C D A L M L B M M C M
Benign prostatic hyperplasia	A. Peak flow rate B. Symptom score	B A L
Multiple sclerosis‡	A. Relapse rate at 1 year B. Disability at 2 years	B A M
Vaccines	Depending on specific diseases of interest, efficacy endpoints may include A. Incidence rate of disease (eg, chickenpox) B. A composite endpoint of disease incidence and severity (eg, herpes zoster vaccine) Immunogenicity endpoints may include A. Percentage of subjects achieving a threshold level of immune response B. Geometric means of immune responses C. Responses to different serotypes in vaccines with multiple serotypes or responses to different components in a combination or concomitant use vaccine trial	Correlations are usually low to moderate; in many cases, success of study usually requires success in all endpoints

VAS, Visual Analog Scale; HRQOL, health-related quality of life; FEV1, forced expiratory volume in 1 second; CHMP, Committee for Medicinal Products for Human Use; IIEF, International Index of Erectile Dysfunction; SEP, Sexual Encounter Profile; NA, not applicable.
*L = 0 to 0.35; M = 0.35 to 0.65; H = 0.65 to 1.
†Not all four must achieve statistical significance; for example, it has been proposed that statistical significance on PG and NLC, but no worsening on ILC, is sufficient.
‡If endpoint (A) is highly compelling, then a single endpoint is acceptable, but in any case, a 2-year study is required to evaluate 2-year effect on disability.

be identified or a composite developed from the medical perspective. If this is not possible, the team recommends considering statistical solutions. Results from the team's deliberations are given in the third and fourth sections. The section on statistical options contains most of the technical detail supporting the statistical solutions. For readers who are less interested in the technical detail, we suggest that they browse through this section and the first portion of the section on statistical implications, and move on to the Discussion section, where we offer additional comments and suggestions.

IMPLICATIONS OF CO-PRIMARY ENDPOINTS

In this section, we discuss clinical and statistical implications of reverse multiplicity. For convenience, our discussion is limited to endpoints that may be treated as continuous. The points raised in the following discussion are also relevant to other types of endpoints. Without loss of generality, we assume that we are comparing a new investigational drug to a placebo, and that high values represent a more favorable outcome than low values for all endpoints.

CLINICAL IMPLICATIONS

Clinical considerations should always drive the requirements for an intervention to qualify as an efficacious treatment. Ideally, the effectiveness decision could be based on one endpoint. Unfortunately, there are clinical settings in which multiple co-primary endpoints are necessary. When this happens, one would hope that the identified co-primary endpoints are equally im-

portant to the assessment of the treatment efficacy. In other words, the endpoints should be interchangeable in the sense that the conclusion on an intervention's efficacy would not change if findings were switched between endpoints. For example, if two endpoints favored the test drug with respective (two-sided) *P* values of .01 and .06, the conclusion would be the same regardless of which endpoint was associated with the smaller *P* value. If the co-primary endpoints do not satisfy the above description, then clinical input should be sought to identify a subset of endpoints that are truly co-primary.

In addition to interchangeability, the clinical rationale for adopting multiple co-primary endpoints should be clear and should not be due to experts' inability to choose among several endpoints. Furthermore, we acknowledge that consideration of benefit-to-risk assessment of the study drug involves both primary and secondary endpoints, along with numerous key safety endpoints.

STATISTICAL IMPLICATIONS

Assume there are J co-primary endpoints in a randomized trial. Let $\{X_j, j = 1,\ldots, J\}$ represent the observations on the J endpoints of an individual in the placebo group and $\{Y_j, j = 1,\ldots, J\}$ the corresponding observations of an individual in the group receiving the new drug. To simplify the notations, we omit the subscript denoting subject and assume equal sample size n for the two groups. In addition, we assume that $(X_1,\ldots,X_J)$ has a multivariate distribution with mean μ_x and a known covariance matrix $\Sigma = (\sigma_{ij})$. Similarly, we assume that $(Y_1,\ldots,Y_J)$ has a multivariate distribution with mean μ_y and a covariance matrix Σ. Let $\Delta = \mu_y - \mu_x = (\Delta_1,\ldots, \Delta_J)$. If the new drug is better than the placebo on all J endpoints, then Δ_j will be greater than 0 for all j.

Comparing the new drug to the placebo based on the J co-primary endpoints is equivalent to testing the following hypothesis:

$$H_0: \Delta_j \le 0 \text{ for at least one } j$$
$$H_A: \Delta_j > 0 \text{ for all } j \qquad (1)$$

When Σ is known, the null space for testing H_0 versus H_A for the case of two co-primary endpoints is the shaded area in Figure 1. The null space includes all points on the x and y axes.

It can be easily seen that if one defines sub-hypotheses $H_{0,j}: \Delta_j \le 0$ and $H_{A,j}: \Delta_j > 0, j = 1,\ldots, J$, one can obtain the relationship in Eq. (2).

$$H_0 = \bigcup_{j=1}^{J} H_{0,j} \qquad H_A = \bigcap_{j=1}^{J} H_{A,j} \qquad (2)$$

The relationship in Eq. (2) means that one can test H_0 versus H_A using the intersection-union test (IUT) (5,6). In other words, testing H_0 versus H_A at the 2.5% level can be carried out by testing each $H_{0,j}$ versus $H_{A,j}$ at the 2.5% level. Under this approach, we reject H_0 at the 2.5% level only if all $H_{0,j}$'s are rejected at the same level.

The above is the basis for the current regulatory position; that is, when multiple co-primary endpoints are necessary to assess an intervention's efficacy, each endpoint should be evaluated at the significance level set for testing the hypothesis in Eq. (1). Eaton and Muirhead (7) showed that the likelihood ratio test for H_0 versus H_A is equivalent to the IUT under the multivariate normal distribution assumption.

Under the IUT, the false-positive error rate is preserved at the desirable level. In other words, the chance of declaring a new intervention to be efficacious when it is in fact ineffective on at least one of the co-primary endpoints is at most 2.5%. When there are two co-primary endpoints, it can be shown that the probability of rejecting H_0 over the complete null space as dis-

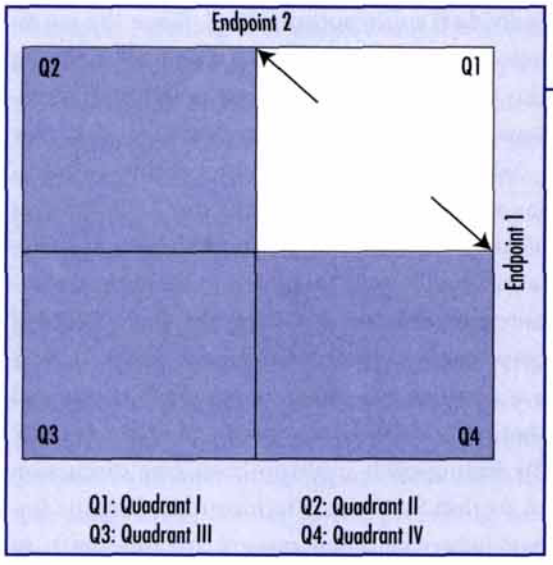

FIGURE 1

Complete null space when there are two co-primary endpoints. The complete null space is the shaded area in Q2, Q3, and Q4, including the x and y axes.

TABLE 2

Probability of Achieving Statistical Significance on All Primary Endpoints Under the Intersection-Union Test Approach, Assuming Equal Effect Size on All Endpoints and a Sample Size That Gives 80% Power to Detect the Effect Size for an Individual Endpoint

	No. of Co-primary Endpoints			
Correlation	2	3	4	9
0	64%	51%	41%	14%
0.2	66%	55%	47%	25%
0.5	69%	61%	56%	40%
0.8	73%	69%	66%	58%

played in Figure 1 is maximized at either $(\Delta_1, \Delta_2) = (0, \infty)$ or $(\Delta_1, \Delta_2) = (\infty, 0)$. The type I error in both cases is .025. In general, with J co-primary endpoints, the type I error takes its maximum on the boundaries of the J-dimensional null space, where $(J-1)$ of the coordinates takes the value of ∞, while the remaining one takes the value 0.

A natural follow-up question is what impact the IUT has on the overall power of the study. Intuitively, the overall power will be less than the smallest power for testing the individual endpoints. If the test statistics are completely independent, then the overall power is simply the product of the powers for testing each individual subhypothesis. If the test statistics are perfectly correlated, then the power for detecting the same (standardized) effect size on all endpoints is the same as the common power for detecting the same (standardized) effect size at the individual subhypothesis level. Since the correlation between test statistics is typically between the two extremes, the overall power will be between those at the two extremes.

Table 2 gives the overall probability of reaching statistical significance at the 2.5% level on all endpoints for four different choices of correlation coefficients (assumed to be the same between any two test statistics) and the number of co-primary endpoints between 2 and 9. Table 2 was constructed assuming equal effect size and therefore identical marginal power, set at 80%, for testing each subhypothesis (see discussion in section Statistical Options regarding the impact when marginal powers are different). In other words, results in Table 2 were obtained by evaluating the probability $Pr(Z_1 > 1.96,\ Z_2 > 1.96,\ \ldots,\ Z_J > 1.96 \mid \Delta_j / \sqrt{\sigma_{jj}} = \text{constant})$, where Z_j is the test statistic for testing $H_{0,j}$ versus $H_{A,j}$. These findings are consistent with those presented by Kong et al. (8).

As can be seen in Table 2, the power could be substantially less than 80% in some cases. For example, when there are three co-primary endpoints and the correlation among the test statistics is 0.2, the overall power for detecting the effect size corresponding to an 80% power at the individual subhypothesis level is only 55%. Note that the impact is reduced if one can power each individual endpoint at greater than 80%.

The correlation coefficients in Table 2 refer to those between test statistics. It can be shown that the degree of correlation between test statistics is generally consistent with that between endpoints. If the correlations between endpoints are the same among treatment groups, then the correlation between test statistics is equal to the correlation between endpoints. We assume this to be the case for the remainder of this article.

Results in Table 2 show how requiring statistical significance on multiple endpoints under the IUT could lead to a loss of power at the study level. One way to keep the overall power at a desirable level, say 80%, is to increase the sample size. Under similar assumptions as those for Table 2, Table 3 provides the extent the sample size needs to be increased to maintain the power at 80% for the trial. For example, when there is only one primary endpoint, 100 patients per

TABLE 3

Sample Size Multiplier to Maintain 80% Overall Power for the Clinical Trial with Multiple Co-primary Endpoints, Assuming Equal Effect Size for All Endpoints and a Base Sample Size That Has an 80% Power to Detect the Effect Size at the Individual Endpoint Level

	No. of Co-primary Endpoints			
Correlation	2	3	4	9
0	1.31	1.49	1.62	1.96
0.2	1.29	1.46	1.58	1.91
0.5	1.25	1.39	1.49	1.74
0.8	1.17	1.27	1.32	1.48

group are needed to have an 80% power to detect an effect size 0.4 ($=\Delta/\sigma$) using a one-sided test at the 2.5% level, whereas 146 patients per group (a 46% increase) are needed to have an 80% power to reach statistical significance on all three endpoints if the effect size is 0.4 for all three endpoints and the correlation coefficient between the endpoints is around 0.2. The sample size multiplier in Table 3 increases with the number of co-primary endpoints and the decrease in correlation. The increase could be substantial in some cases.

From our experience, the correlations between co-primary endpoints can be as low as 0.2 and as high as 0.8. Along with the list of diseases requiring co-primary endpoints, Table 1 also provides estimated correlations we have seen in actual trials. Without loss of generality, we assume correlations are nonnegative. Correlations in Table 1 are classified as low, medium, and high, with low correlations in the range of (0, 0.35), medium correlations in the range of (0.35, 0.65), and high correlations in the range of (0.65, 1). Correlations based on historical data are useful in helping us better plan the sample size needed for a study so that the overall power can achieve the desirable level. Even so, one might want to consider adjusting sample size based on an interim estimate of the correlations among co-primary endpoints. In this regard, sample size reestimation to account for correlations between co-primary endpoints is similar to that associated with resizing a study based on an interim estimate for the variability in the endpoints.

Even though we focus on reverse multiplicity as a result of multiple co-primary endpoints in this article, it is of interest to note that reverse multiplicity arises whenever multiple analyses must all achieve statistical significance for a trial to be considered a success. An example is when sensitivity analyses are required to "confirm" the primary results. Another example is to require significance with more than one approach to handling missing data (eg, last observation carried forward as well as multiple imputations). A recent example relates to the ICH E14 Draft Guidelines (June 2004) (9) on the clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs. The proposed requirement for a successful "thorough QT/QTc study" is that the time-matched mean difference between drug and placebo satisfies a prespecified noninferiority criterion for all time points when QT/QTc are measured. Needing to establish noninferiority at each time point leads to a reverse multiplicity situation. The traditional regulatory requirement for approval of a combination therapy, while justifiable, is also a reverse multiplicity situation (10,11). The latter requires that the combination is superior to each individual component.

MEDICAL OPTIONS

In this section, we consider situations when the clinical question of primary interest can probably be adequately addressed by a single primary endpoint and offer some options from the medical perspective. Similar to the preceding section, the focus is on situations when regulatory

agencies mandate statistical significance on multiple co-primary endpoints.

When there exists an accepted, clinically meaningful endpoint that, in Temple's (12) term and echoed by a National Institutes of Health working group (13), "measures directly how a patient feels, functions, or survives," there will be no problem of selecting the primary endpoint. In certain situations when multiple co-primary endpoints are currently required by regulatory agencies, one might be able to obtain input from the clinical and patient community on the relative importance of the required co-primary endpoints. If this were to lead to defining a single primary endpoint, with the remaining endpoints becoming key secondary endpoints, then the statistical problem resulting from reverse multiplicity disappears.

A compromise between maintaining the co-primary endpoints and demoting all but one to become secondary endpoints is to require that at least a "trend" be observed for each of the key secondary endpoints, while maintaining the one-sided .025 α level for the single primary endpoint. *Trend* may be defined in a number of different ways. One possibility is to test the key secondary endpoints at a higher significance level (eg, one-sided 5% level). Choice for the single primary endpoint in this setting might be to create a composite endpoint as defined by O'Brien (14).

Migraine offers a case in study. At present, in addition to pain relief, symptom relief from nausea, photophobia, and phonophobia are considered important measures of clinical effect. When migraine patients were queried concerning their preferences for therapeutic effects, they ranked aspects of pain management as more important (15). Based on patient feedback, it is reasonable to suggest selecting pain relief as the primary endpoint and treating relief of other symptoms as secondary. This is consistent with the guidelines issued by the International Headache Society (16), stating, "Percentage of patients pain-free at 2 h, before any rescue medication, should usually be the primary measure of efficacy" (section 1.3.2 of the guidelines document). Under the society's recommendations, the four co-primary endpoints are not interchangeable, failing the requirement in our section on Clinical Implications. To ensure that a new migraine therapy also demonstrates some evidence of efficacy on the relief of secondary symptoms, one could require a trend for these be established, as discussed briefly above.

An approach that has been successfully used by the American College of Rheumatology (ACR) is to combine several endpoints into a single composite endpoint. The composite endpoint ACR20, a binary endpoint often used in arthritis studies, is derived through simultaneous consideration of joint counts and several categorical assessments. The development and continuous refinement of ACR20 could serve as a working model to other disorders that currently require multiple co-primary endpoints.

Another useful application is to combine multiple event-based endpoints into a single composite endpoint and analyze the data as "time to the first event among events included in the composite," recognizing that some types of events are more common than others. Such composite endpoints have become a well-accepted approach in some therapeutic areas, such as cardiology (17,18). Composite endpoints that include relevant events can provide increased statistical power while retaining a meaningful clinical interpretation.

Sankoh, D'Agostino, and Huque (19) examined some practical clinical decision-making scenarios for the selection and analysis of efficacy outcome measures in clinical trials with inherent multiplicity components. They considered situations for which statistical significance needs to be demonstrated (at the prespecified significance level) for all primary endpoints, statistical significance needs to be demonstrated for the majority of the primary endpoints, and statistical significance needs to be demonstrated for one or more of the primary endpoints. Depending on the situations, Sankoh et al. (19) discussed appropriate statistical strategies with a focus on controlling the study-level type I error rate. Forming composite endpoints and requiring the results across pri-

mary endpoints to be consistent are among the approaches advocated.

There are cases for which the best way to measure an intervention's effect is simply unknown. If the candidate endpoints are equally meaningful from the clinical perspective, then it should be reasonable to choose one of the potential endpoints as primary. Alternatively, one should be able to proceed without fixing any particular endpoint in advance as long as one addresses how the false-positive error rate will be controlled. The latter transforms the multiple endpoint requirements into the alternative multiplicity problem that could be handled by approaches in Ref. 20.

The discussion above focuses on avoiding simultaneous testing of multiple hypotheses. There may remain cases for which it is not possible or appropriate to have a single primary endpoint. In addition to the disease itself, the lack of a single primary endpoint can arise in at least two ways: as a result of confounding or because it is necessary to use an endpoint that is not well accepted. The latter could be an unproven surrogate or a new and yet-to-be-validated scale.

Confounding is a concern when the drug may affect the endpoint through some (unintended) mechanism other than the expected one. Subjective measures of a patient's condition are particularly vulnerable in this regard. For example, in treatment of benign prostatic hyperplasia, the American Urological Association symptom scale is a well accepted and clinically meaningful measure of the disease. However, regulatory authorities currently require that studies using this endpoint also demonstrate an effect on urinary flow rate with the hope that the flow rate data could validate that symptomatic improvements are a result of effects on the prostate and not through some alternative mechanism. For example, a sleep aid would probably improve the score by reducing frequency of night urination but could not be considered to be treating the disease. Confounding remains a legitimate concern. Nevertheless, we argue that in some cases such mechanistic concerns could be addressed in specific studies without the need for repeated validation in every clinical trial, particularly for classes of drugs that are already well understood. When more knowledge of the mode of action is available, it should be possible to have the symptoms as the primary endpoint with the urinary flow rate as the secondary.

When using an unproven endpoint as a major endpoint, requiring significance on multiple endpoints may improve our confidence in the overall treatment effect. Again, with the passage of time, we hope that the accumulated knowledge of the unknown endpoints will alleviate the use of multiple endpoints for the same disorder in the future. Meanwhile, one could consider applying the approaches discussed in the next section to handle the challenge brought on by reverse multiplicity.

STATISTICAL OPTIONS

The best approach to address the problem of reverse multiplicity is to identify a single primary endpoint. If this is not possible, then one must consider statistical solutions. As stated in the second section, the standard statistical approach, based on both the intersection-union and the likelihood ratio principles, leads to testing each endpoint at the level allowed for testing the hypotheses in Eq. (1). The impact of the IUT on power and therefore on sample size is discussed in detail in the second section.

It should be pointed out that reverse multiplicity might not pose a significant hardship when (*a*) the co-primary endpoints are highly correlated (eg, correlation coefficients above 0.9) or (*b*) the effect size (either clinically meaningful or anticipated) is much smaller for one endpoint (eg, 50% smaller) compared to the rest and the sample size is based on detecting the smallest effect size. In the latter case, powering for the smallest effect size leads to overpower for the other endpoints. This is a natural consequence if the endpoint with the smallest effect size is essential to determining the efficacy of a new treatment. Substantially different treatment effect on endpoints may occur when there are two co-primary endpoints and perhaps occasionally when there are three co-primary endpoints. For more than three primary endpoints,

it is hard to expect an intervention to have a consistently high effect on all co-primary endpoints except for one. As for the first instance of high correlation among the endpoints, when endpoints are highly correlated, one should be allowed to pick one endpoint as primary or create a composite endpoint as defined by O'Brien (14) since the endpoints are likely to be measuring the same things, therefore completely eliminating the problem of reverse multiplicity.

Because of the above consideration, we focus in this section on situations for which the correlations among endpoints are at most medium and the effect size is expected to be comparable across endpoints. These are the situations for which the IUT results in the greatest reduction in the overall study power when compared to the powers for testing individual endpoints. The effort in this section is to introduce statistical approaches that might increase the overall power of the trials while keeping in mind the need to control the type I error rate in some fashion. Some of the proposed methods require thinking that is different from the traditional frequentist considerations. It is our hope that the discussion in this section can invigorate more research on statistical solutions to the problem of reverse multiplicity.

STATISTICAL INFERENCES BASED ON A RESTRICTED NULL SPACE

The conservatism of the IUT partially comes from the need to control the maximum false-positive rate over the entire null space. This maximum often corresponds to an unrealistic situation in which the new treatment has no effect on one endpoint and unusually large effect on the other endpoints. One way to reduce the conservatism of the IUT is to restrict the null space to a more realistic space. When doing this, the null hypothesis given in Eq. (1) is changed to

$$H_0:(\Delta_1, \Delta_2, \ldots, \Delta_J) \in H_0^R,$$
$$\text{where } H_0^R \text{ is the restricted null space}$$
$$H_A:\Delta_j > 0 \text{ for all } j \qquad (3)$$

Offen and Helterbrand (20) proposed to restrict the null space to the no-effect null space that consists of all points in the third quadrant in Figure 1 and the bordering x and y axes. Even though this approach may allow one to increase greatly the significance level for testing individual hypotheses, restricting the null space this way ignores all situations for which the new treatment might have some effect on some endpoints but not on all endpoints.

A more reasonable approach is to include in the null space only those cases that are realistic or unacceptable from a clinical perspective. Consider a clinical trial with two normally distributed co-primary endpoints. Under the normality assumption, we compare treatments on each endpoint using a Z-test and a one-sided significance level of .025. Now, consider points in the complete null space in Figure 1 with coordinates that are no more than M standard deviations from the origin. The region consisting of such points is displayed in Figure 2. Treating this region as the new (restricted) null space, the false-positive rate of the IUT is the maximum probability of rejecting the null hypothesis over this region. The choice of M will be case specific. However, since we are focusing on situations for which the new treatment is not expected to have a dramatically different effect on the multiple co-primary endpoints, values between 0.5 and 1.0 will be reasonable choices for M in general.

Figure 3 describes the relationship among the maximum type I error rate, correlation between the two endpoints, and the maximum coordinate (M) from the origin. Sample size per group for Figure 3 is fixed at 64, which is the sample size needed to have 80% power to detect an effect of 0.5 for one endpoint at the 2.5% level. We choose two maximum coordinates of 0.5 and 0.8 to define the restricted null space. As expected, the maximum type I error rate increases with the maximum distance from the origin.

The power under this restricted null space will increase since the critical value is smaller than the traditional 1.96 (for $\alpha = .05$). However, the increase is inconsequential for sample sizes greater than 50/group, which is how most confirmatory phase III trials are sized, because the critical value is very close to 1.96.

The choice of the maximum coordinates to define the restricted null space should be based on the disease and what can be realistically expected of treatments for the disease, keeping in mind that a treatment with an effect size of 0.5 is generally considered "moderately" effective, while a treatment with an effect size of 0.8 is generally considered "highly" effective.

The idea of a restricted null space is not new. Patel (21) considered a restricted null space when comparing the efficacy of a combination therapy to that of its components. Concerned that the type I error rate could be inflated if the assumption about the restricted null space is wrong, Snapinn and Sarkar (22) proposed an alternative that would result in a penalty to the sponsor if their assumption about the null space is wrong. The penalty is in the form of a nonmonotone rejection region, as we discuss in the Bayesian Approach section. Since we focus on cases for which the available data do not suggest great disparities in the treatment effect on multiple primary endpoints, restricting the null space could be a reasonable approach for such cases.

The restricted null space described in Figure 2 can be easily generalized to the case in which the effect sizes for individual endpoints are thought to be unequal. For example, in the case of two primary endpoints, one could consider rectangular (ie, not square) areas in quadrants II, III, and IV (Figure 2) and consider different critical values for decisions regarding different endpoints. The different critical values are determined so that the maximum type I error over the new restricted null space is controlled at the 2.5% level. One can also consider nonrectangular areas in quadrants II, III, and IV and a decision rule that combines results on the endpoints in a nonlinear fashion to signal a trade-

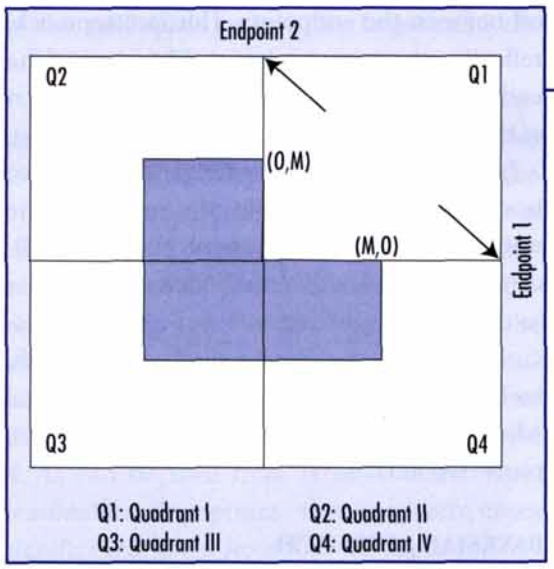

FIGURE 2

The restricted null space defined by points in the complete null space with coordinates that are no more than M units from both the x and y axes.

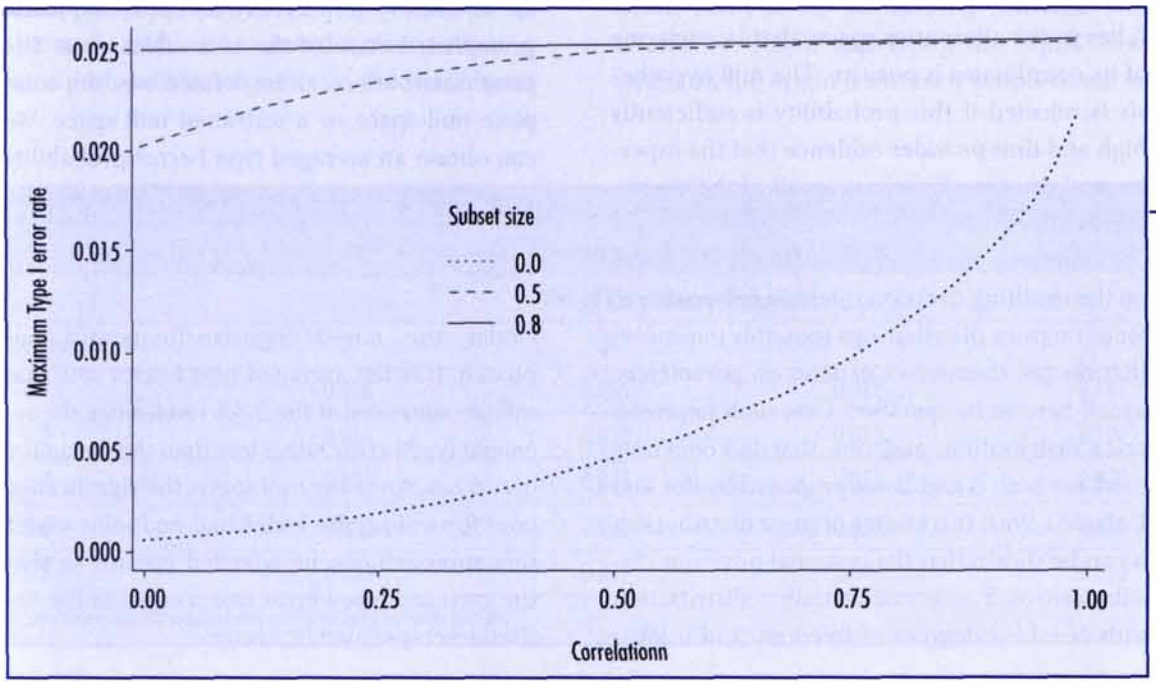

FIGURE 3

Relationship between the maximum type I error rate, correlation coefficient, and the subset size for two co-primary endpoints. The subset size corresponds to the maximum coordinate M used to define the restricted null space depicted in Figure 2.

off between the endpoints. This last approach, reflecting the noninterchangeable nature of the endpoints, will require much more clinical input at the design stage.

Type I error rate over the restricted null space is a function of the sample size, and hence the amount of upward adjustment on the significance level under this methodology decreases as the sample size increases. For a detailed discussion of the restricted null space approach, including the reasons why the significance level adjustment is a function of the sample size, please see Ref. 23.

BAYESIAN APPROACH

The statistical approaches to the reverse multiplicity problem introduced in the previous sections rely on frequentist arguments. In this section, we discuss a Bayesian approach. Some general discussion on the use of Bayesian approaches to draw statistical inference can be found in Casella and Berger (24).

Assume that the J co-primary endpoints are jointly normally distributed with a mean vector Δ and a covariance matrix Σ. Here, Δ is the vector of mean differences between the new treatment and the placebo. A Bayesian approach involves selecting a prior distribution for Δ and Σ. The Bayesian decision rule is based on the marginal posterior probability that the mean vector Δ lies in the alternative space; that is, each one of its coordinates is positive. The null hypothesis is rejected if this probability is sufficiently high and thus provides evidence that the experimental drug is efficacious on all of the co-primary endpoints.

To minimize the effect of subjective decisions on the resulting decision rule, it is reasonable to focus on prior distributions (possibly improper) that do not themselves depend on parameters, which have to be specified. One such improper prior distribution, and one that is commonly used, for both Δ and Σ was proposed by Box and Tiao (25). With this choice of prior distribution, it can be shown that the marginal posterior distribution of Σ is a multivariate t distribution with $N + 1 - J$ degrees of freedom, and a location vector that is proportional to the vector of sample mean differences and a parameter matrix that is proportional to the pooled sample covariance matrix. Here, $N = n_1 + n_2 - 2$, and n_1 and n_2 are the sample sizes in the new treatment and the placebo groups, respectively. Since the marginal posterior distribution of Σ is known, one can now construct the Bayesian decision rule.

It is important to point out that, unlike the frequentist IUT, the Bayesian solution depends on the sample covariance matrix and thus accounts for the correlation among the co-primary endpoints. For more on this approach in the reverse multiplicity problem, see the work of Grieve and Muirhead (26).

MIXED BAYESIAN/FREQUENTIST APPROACHES

Mixed Bayesian/frequentist tests are based on averaging the probability of finding significant treatment differences (frequentist concept) by a prior distribution assumed for the treatment effect (Bayesian concept) in the null space. Again, the case of two co-primary endpoints is used to illustrate this concept.

Let $\alpha(\Delta_1, \Delta_2)$ denote the probability of rejecting the null hypothesis (no treatment effect for at least one endpoint) when the true effect sizes are Δ_1 and Δ_2, respectively, and $p(\Delta_1, \Delta_2)$ is the prior distribution of the true effect sizes. The prior distribution can be defined over the complete null space or a restricted null space. We can obtain an averaged type I error probability by averaging $\alpha(\Delta_1, \Delta_2)$ using $p(\Delta_1, \Delta_2)$ as weight:

$$\alpha_{BAYES} = \int_{\text{Null Space}} \alpha(\Delta_1, \Delta_2) p(\Delta_1, \Delta_2) d\Delta_1 d\Delta_2 \qquad (4)$$

Under the mixed Bayesian/frequentist approach, it is the averaged type I error rate that will be controlled at the 2.5% level. Since the averaged type I error rate is less than the maximum of $\alpha(\Delta_1, \Delta_2)$ over the null space, the significance level for testing the individual endpoint under this approach can be adjusted upward so that the averaged type I error rate is equal to the desired level specified in advance.

When the described mixed Bayesian/frequentist approach is applied to the complete null space, it typically results in overly liberal decision rules. Consider, for example, a uniform prior over the complete null space. Since the function $\alpha(\Delta_1, \Delta_2)$ quickly converges to zero as both Δ_1 and Δ_2 in the null space move away from the x and y axes, the averaged type I error rate will be close to 0. For the averaged type I error rate to be near 2.5%, one needs to choose much higher significance levels for the individual test statistics. The associated decision rule is quite liberal and frequently yields significant results even when the P values from the IUT analyses are highly nonsignificant.

By comparison, mixed Bayesian/frequentist tests applied to a reasonably restricted null space do not suffer from this problem. The next two sections describe such applications. An example is to average over only the portion of the null space (or a reasonably restricted one) that borders the alternative space.

Bayesian/Frequentist Test Based on Discrete Prior Probabilities. This section describes a simple method of discretizing the prior distribution and assigning probabilities to only selected points in the null space. For the case of two co-primary endpoints, the approach assigns mass probabilities only to (0,∞), (∞,0), and (0,0). The approach can be thought of as assigning prior probabilities to points on the boundaries of the null space to give the largest and the smallest type I error rates.

Chuang-Stein et al. (23) discussed the relationship between assigning equal probabilities to (0,∞), (∞,0), (0,0) and assigning a uniform prior to the restricted null space (Figure 2) in the case of two co-primary endpoints. This discussion can be extended to more than two co-primary endpoints. Under this approach, the adjusted significance levels for testing individual hypotheses for different correlations and different number of endpoints are given in Table 4. As can be seen from Table 4, under fairly reasonable assumptions, this approach raises significance levels from .025 to approximately between .030 and .10 for testing individual endpoints.

Mixed Bayesian/Frequentist Tests Based on Empirical Bayesian Arguments. Snapinn and Sarkar (22) considered an approach that augments the traditional rejection region when evaluating a combination therapy. In essence, the approach assigns the following weights to (0,∞), (∞,0), and (0,0):

$$c / 3,\ c / 3,\ \text{and}\ c / \left(\sqrt{\bar{d}_1^2 + \bar{d}_2^2}\right)^4 \qquad (5)$$

In Eq. (5), $(\bar{d}_1, \bar{d}_2)$ represents the standardized bivariate sample mean for comparing two treatments, and c is the normalizing constant. The weights on (0,∞), and (∞,0), are equal. The weight on the origin is inversely proportional to

TABLE 4

Adjusted Significance Levels Under the Bayesian/Frequentist Test Based on Discrete Prior Probabilities as a Function of the Number of Co-primary Endpoints and Correlations

	No. of Co-primary Endpoints			
Correlation	2	3	4	9
0	0.036	0.055	0.082	0.121
0.2	0.036	0.052	0.075	0.106
0.4	0.035	0.048	0.066	0.089
0.6	0.032	0.043	0.055	0.070
0.8	0.030	0.037	0.044	0.052

Simulations (300,000 samples) were used to evaluate the power function

distance of (0,0) from $(\bar{d}_1, \bar{d}_2)$ raised to the fourth power. Since the weights depend on the data, the rejection region will also depend on the data.

This approach expands the rejection region of the standard IUT and therefore leads to a slight increase in power. However, this augmentation is nonmonotone as mentioned in the section on statistical inferences. For example, for the case of two co-primary endpoints, individual *P* values of (.06, .06) might lead to the rejection of the null hypothesis in Eq. (1), while (.00006, .06) will not. Because of this property, we do not recommend this approach to address the problem of reverse multiplicity arising from multiple co-primary endpoints.

DISCUSSION

In this article, we point out that reverse multiplicity, when approached in the traditional way and tested with the IUT, could lead to a decrease in the study power or to an increased sample size to maintain the same power. A consequence of the reduced power is a higher regulatory hurdle to declare a new treatment to be efficacious. The optimal solution is to reduce multiple co-primary endpoints to a single primary endpoint. This can be achieved by selecting one endpoint to be the primary one or by creating a composite measure as discussed here. Forming a single composite endpoint would likely require some degree of consistency across the individual components. Alternatively, one could declare one to be the single primary endpoint and require a positive trend in the remaining endpoints. For diseases for which the list of co-primary endpoints cannot be reduced, we proposed statistical approaches that are positioned as alternatives to the IUT. The statistical options were designed primarily to increase the study power.

Insomnia is an example for which regulatory agencies acknowledge that the disorder has multiple components, and not all patients necessarily have all of them. There are drugs that reduce the time to the onset of persistent sleep (Sonata®) and drugs that treat both onset of sleep and the duration of sleep (Ambien®). Recently, we have seen the FDA moving in this direction with fibromyalgia. With fibromyalgia, FDA has expressed willingness to grant a claim of management of pain associated with fibromyalgia if a new treatment is successful in treating fibromyalgia-associated pain. On the other hand, if a new treatment can improve other dimensions of fibromyalgia in addition to pain, a claim of fibromyalgia treatment could be granted. We welcome this trend and hope such considerations could also be extended to other disorders mentioned in Table 1.

One might argue that if a study has two co-primary endpoints and the study is designed with at least 90% power for each endpoint, then the study should have at least an 81% power for testing the hypotheses in Eq. (1) under the IUT. In other words, the study still has a reasonable power despite the requirement of two co-primary endpoints. We would like to point out that a sponsor chooses the power of a study with great thought and strategy. In a disease for which a single large clinical study would be the only confirmatory phase III study, the sponsor would likely insist that the power of the study be at least 90%. So, the impact on sample size presented in the second section and Table 3 when using the IUT is real.

Currently, most sponsors choose to increase study sample size when facing the requirement of multiple co-primary endpoints to maintain the power at a desirable level. As demonstrated in Table 3, the increase can be substantial in some situations. In this article, we recommend considering statistical solutions when a single primary endpoint is not feasible. Even though sample sizes under the proposed approach will also be larger than those needed for a single primary endpoint, the increases are expected to be in general less than those required under the IUT approach. In this regard, we consider the frequentist/Bayesian approach a response to FDA's Critical Path Initiative (27) that espouses greater efficiency in clinical drug development through innovations.

During the PhRMA/FDA Multiple Endpoints Workshop held in Bethesda, Maryland, in October 2004, some important questions or sugges-

tions were raised that we would like to address here. A point was made that in some cases when we cannot move to a single primary endpoint, one could demonstrate a statistically significant effect on one or some of the endpoints in phase II, reducing the number of co-primary endpoints for phase III. This approach is certainly worthy of consideration if feasible but would require that the phase II trial protocol identify a single primary endpoint or multiple primary endpoints with appropriate multiplicity adjustments.

It was also expressed at this workshop that regulatory agencies would be concerned if sponsors could add their own co-primary endpoints to obtain an upward adjustment in the levels used to test individual endpoints. This concern can be addressed by adopting the policy that if sponsors add their own co-primary endpoints beyond those required by regulatory agencies, no adjustment will be allowed. This is reasonable because a motivation of the sponsor to add such co-primary endpoints might be that they have great faith that the study drug will achieve a high level of statistical significance on the endpoints they propose to add. In addition, sponsors could always adopt the gatekeeping strategy (28) to help get additional endpoints in the label.

Another concern expressed at the workshop was that if the FDA were to adopt an increase in the significance levels, one could have a study in which the null hypothesis in Eq. (1) (second section) is rejected at the increased significance level, but none of the individual endpoints is significant at the traditional one-sided 2.5% level. We do not feel this would lead to any scientific controversy. Decisions concerning treatment effect are made based on the strength of evidence. The conventional 2.5% level was chosen at a time when decisions were made based on a single endpoint. On the other hand, we are dealing with multiple co-primary endpoints in this article. When consistently strong evidence is available from multiple co-primary endpoints, the collective evidence becomes the basis to decide a treatment's efficacy for the disorder. The question is at what level should the "consistently strong" evidence be judged. The statistical approaches in the sections on inferences in a restricted null space and mixed Bayesian/frequentist approaches propose to use a slightly higher significance level when this requirement is applied to individual co-primary endpoints. In our opinion, the reasoning behind such approaches is quite logical.

It is our experience that many researchers are not familiar with the impact of reverse multiplicity on clinical trials and drug development. We hope that this position paper will help promote the understanding and raise the awareness. In addition, we hope that our effort will help reduce the number of cases for which regulatory agencies require a new treatment to demonstrate a statistically significant effect on multiple co-primary endpoints, all at the 2.5% level, before accepting the new treatment as efficacious. If after careful considerations co-primary endpoints remain a requirement, then we hope that regulatory agencies are receptive to the idea of modest adjustments to the significance levels as have been introduced here.

Acknowledgments—We gratefully acknowledge Steve Snapinn (Amgen), Ivan Chan (Merck), Harry Cui (Pfizer), William Chang (AstraZeneca), and Marci Clark (Pfizer).

REFERENCES

1. Hsu JC. *Multiple Comparisons—Theory and Methods*. Chapman and Hall: London, UK; 1996.
2. Sankoh AJ, Huque MF, Dubey SD. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Stat Med*. 1997;16:2529–2542.
3. ICH E9. Statistical Principles for Clinical Trials. Step 5 guidance document. 2000. Available at: http://www.ich.org. Accessed November 11, 2006.
4. CPMP/EWP/908/99. Points to consider on multiplicity issues in clinical trials (adopted September 2002). Available at: http://www.emea.eu.int/pdfs/human/ewp/090899en.pdf. Accessed November 11, 2006.
5. Berger RL. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*. 1982; 24:295–300.
6. Laska EM, Meisner MJ. Testing whether the iden-

tified treatment is best. *Biometrics*. 1989;45: 1139–1151.
7. Eaton ML, Muirhead RJ. On a multiple endpoints problem. *J Stat Plann Infer*. In press.
8. Kong L, Kohberger RC, Koch GG. Type I error and power in noninferiority/equivalence trials with correlated multiple endpoints: an example from vaccine development trials. *J Biopharm Stat*. 2004;14:893–907.
9. *ICH E14. The Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-antiarrhythmic Drugs. Step 3 guidance document*. 2004. Available at: http://www.ich.org. Accessed November 11, 2006.
10. Leung HM, O'Neill RT. Statistical assessment of combination drugs—a regulatory view. In: Proceedings of the Annual Meeting of the American Statistical Association; August 7, 1986; Chicago, IL; Biopharmaceutical Subsection: 33–36.
11. Snapinn SM. Evaluating the efficacy of a combination therapy. *Stat Med*. 1987;6:657–665.
12. Temple RJ. A regulatory authorities opinion about surrogate endpoints. In: Nimmo WS, Tucker GT, eds. *Clinical Measurement in Drug Evaluation*. New York, NY: Wiley; 1995:322.
13. De Gruttola VG, Clax P, DeMets DL, et al. Considerations in the evaluation of surrogate endpoints in clinical trials: summary of a National Institutes of Health Workshop. *Control Clin Trials*. 2001;22: 485–502.
14. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics*. 1984;40: 1079–1087.
15. Caro G, Caro J, O'Brien JA, Anton S, Jackson J. Migraine therapy: development and testing of a patient preference questionnaire. *Headache*. 1998; 38:602–607.
16. International Headache Society Clinical Trials Subcommittee. Guidelines for controlled trials of drugs in migraine: second edition. *Cephalalgia*. 2000;20:765–786.
17. DeMets DL, Califf RM. Lessons learned from recent cardiovascular clinical trials: part I. *Circulations*. 2002;106:746–751.
18. Dahlof B, Devereux RB, Kjeldsen SE, et al. Cardiovascular morbidity and mortality in the Losartan Intervention for Endpoint reduction in hypertension study (LIFE): a randomised trial against atenolol. *Lancet*. 2002;359:995–1003.
19. Sankoh AJ, D'Agostino RB, Huque MF. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Stat Med*. 2003;22:3133–3150.
20. Offen WW, Helterbrand JD. Multiple comparison adjustments when two or more co-primary endpoints must all be statistically significant. In: Proceedings of the Annual Meeting of the American Statistical Association; August 7, 1986; Chicago, IL; Biopharmaceutical Subsection.
21. Patel HI. Comparison of treatments in a combination therapy trial. *J Biopharm Stat*. 1991;1:171–183.
22. Snapinn SM, Sarkar SK. Assessing the superiority of a combination drug with a specific alternative. *J Biopharm Stat*. 1996;6:241–251.
23. Chuang-Stein C, Stryszak P, Dmitrienko A, Offen W. Challenge of multiple co-primary endpoints: new approache. *Stat Med*. In press.
24. Casella G, Berger RL. *Statistical Inference*. 2nd ed. St. Paul, MN: Brooks/Cole; 2001.
25. Box GEP, Tiao GC. *Bayesian Inference in Statistical Analysis*. New York: Wiley Classics Library; 1992. [Original work published 1973.]
26. Grieve A, Muirhead RJ. A Bayesian approach to the multiple endpoints problem. Unpublished manuscript.
27. Food and Drug Administration. The Critical Path to New Medical Products. March 2004. Available at: http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html. Accessed November 11, 2006.
28. Dmitrienko A, Offen W, Westfall P. Gatekeeping testing strategies in clinical trials. *Stat Med*. 2003;22:2387–2400.

Alex Boddy has disclosed that he is an employee of Sanofi-Aventis. Kati Copley-Merriman has disclosed that she is a shareholder of Pfizer Inc and Eli Lilly and an employee of Pfizer Inc. Willard Dere has disclosed that he is a shareholder of Amgen, Eli Lilly and Co., Merck, and Pfizer. Alex Dmitrienko has disclosed that he has received grants/research support from Eli Lilly and Co. Samuel E. Givens has disclosed that he has received financial or material support from Hoffmann-La Roche, Inc. David Hall has disclosed that he is an employee of Boehringer Ingelheim Pharmaceuticals. David H. Henry has disclosed that he is a shareholder of GlaxoSmithKline, Schering Plough, and Pfizer, and an employee of Bristol-Myers Squibb. Joseph D. Jackson has disclosed that he is a shareholder and an employee of Bristol-Myers Squibb. Alok Krishen has disclosed that he is an employee of GlaxoSmithKline. Gary Littman has disclosed that he is a consultant to Wyeth Research. Thomas Liu has disclosed that he is a shareholder of Amgen, Inc. Jeff Maca has disclosed that he is an employee of Novartis Pharmaceuticals. Robb Muirhead has disclosed that he is an employee and stockholder of Pfizer Inc. Laura Meyerson has disclosed that she is an employee of Biogen Idec. Walter Offen has disclosed that he is an employee and stockholder of Eli Lilly and Co. Steven Ryder has disclosed that he is a shareholder of Pfizer Inc and an employee of Pfizer Global R&D, Pfizer Inc. Paul Stryszak has disclosed that he is an employee of Schering-Plough. Julia Wang, Kun Chen, Chyon-Hwa Yeh, Abdul J. Sankoh, and Christy Chuang-Stein report no relationships to disclose.