Anesthesia & Analgesia

Design and Analysis of Studies with Binary-Event Composite Endpoints: Guidelines for Anesthesia Research

Edward J. Mascha, PhD,*† and Daniel I. Sessler, MD†

Composite endpoints consisting of several binary events, such as distinct perioperative complications, are frequently chosen as the primary outcome in anesthesia studies (and in many other clinical specialties) because (1) no single outcome fully characterizes the disease or outcome of interest, and/or (2) individual outcomes are rare and statistical power would be inadequate for any single one. Interpreting a composite endpoint is challenging because components rarely meet the ideal criteria of having comparable clinical importance, frequency, and treatment effects. We suggest guidelines for forming composite endpoints and show advantages of newer versus conventional statistical methods for analyzing them. Components should be a parsimonious set of outcomes, which when taken together, well represent the disease of interest and are very plausibly related to the intervention. Adding components that are too narrow, redundant, or minimally influenced by the study intervention compromises interpretation of results and reduces power. We show that multivariate (i.e., multiple outcomes per patient) methods of analyzing a binary-event composite provide distinct advantages over standard methods such as any-versus-none, count of events, or evaluation of individual events. Multivariate methods can incorporate clinical importance weights, compensate for events occurring at varying frequencies, assess treatment effect heterogeneity, and are often more powerful than alternative statistical approaches. Methods are illustrated with an American College of Surgeons National Surgical Quality Improvement Program registry study that evaluated the effects of smoking on major perioperative outcomes, and with a clinical trial comparing the effects of crystalloids and colloids on major complications. Sample data files and SAS code are included for convenience. (Anesth Analg 2011;112:1461–71)

In comparative research of many specialties, including anesthesiology and perioperative medicine, the primary outcome is often a composite endpoint consisting of various binary events such as distinct major postoperative complications. Composites are typically chosen for 2 reasons: (1) they capture the disease of interest better than any single outcome, and (2) they are expected to increase power. Composite outcomes are particularly useful for diseases that are manifested in complex ways, and also those with unknown etiologies and thus no consensus on the most important efficacy endpoint.[1] Individual events such as major complications are often (happily) quite rare. Combining 2 or more usually increases the overall outcome incidence, thus improving power compared with a single outcome. Our goal in this article is to assist researchers in design and analysis of studies using composite endpoints.

Choice of the specific components for a composite is of paramount importance to a study's design and the interpretation of results, yet a consistent practice based on specific recommendations for how to formulate a composite is lacking.[2] For example, components should ideally have the same severity, frequency, and treatment effect in order for the composite endpoint and treatment effects on it to be easily interpreted. However, these criteria are rarely met.[3–6] In the "Forming a Composite Endpoint" section below, we therefore review these and other specific guidelines for choosing the components of a composite endpoint.

Additionally, the choice of statistical methods has important implications for both power and interpretation of results. Standard approaches include comparing groups on the collapsed composite of any-versus-none, the count of events, or individual component analyses. Newer multivariate (i.e., multiple outcomes per patient) methods[7–10] allow more flexibility by incorporating clinical severity weights, preventing the composite from being driven by the highest frequency component, and assessing the consistency of treatment effects across components. We show advantages of the newer statistical methods over standard choices in the "Methods for

From the Departments of *Quantitative Health Sciences, and †Outcomes Research, Cleveland Clinic, Cleveland, Ohio.

Accepted for publication February 10, 2011.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.anesthesia-analgesia.org).

Funding: Internal.

The authors declare no conflicts of interest.

Reprints will not be available from the authors.

Address correspondence to Edward J. Mascha, PhD, Departments of Quantitative Health Sciences, Cleveland Clinic, 9500 Euclid Ave., JJN-3, Cleveland, OH 44195. Address e-mail to maschae@ccf.org.

DOI: 10.1213/ANE.0b013e31821796d3

Analyzing a Binary-Event Composite Endpoint" section, and discuss "Factors Affecting Powers of Tests for a Binary-Event Composite Endpoint." Sample data are taken from a study of the effects of smoking on perioperative outcomes and a randomized trial of crystalloids versus colloids on postoperative complications.

FORMING A COMPOSITE ENDPOINT

A good composite endpoint should well capture the disease or health state of interest. For example, serious cardiac morbidity in a β-blocker trial is often assessed using a composite endpoint of all-cause mortality, myocardial infarction, and stroke. Investigators in the National Institute of Neurological Disorders and Stroke rt-PA stroke trial chose 4 neurological scales that were deemed by a panel of experts to best capture a patient's neurological status when taken together.[11] Perioperative medicine interventions are often expected to affect several organ systems or disease areas; a clear specification of the target of the intervention is thus a good first step in choosing the most appropriate outcome components. For example, for an intervention targeted at reducing surgery-induced inflammation, the composite might consist of several inflammation-related complications. Thoughtful and informed choices must be made, which include refraining from a composite that is too wide or from using the same composite across multiple studies unless truly appropriate.

Second, each component should be very plausibly affected by the intervention. Although true effects are of course unknown in advance, components should not generally be added to a composite "just in case" there is an effect. As shown later, including components that are only minimally (or not at all) affected by the exposure usually reduces power, and components with opposite effects certainly do.[4,12] Thus, a well-supported biological rationale should accompany each chosen component. Choosing components highly likely to be affected thus facilitates interpretation of the composite outcome and usually improves statistical power.

Third, the set of components should be parsimonious. For example, detailed disease categories, such as various types of serious infection, can often be combined into a single component. Such parsimonious categorization makes the composite endpoint easier to interpret. It also increases power because greater incidence will be available for analysis of the individual components, if performed. And finally, fewer individual component tests will be needed, thus reducing the adjustment required for multiple comparisons. However, caution must be taken not to combine categories that differ substantially in severity, such as superficial skin and deep sternal wound infections. As discussed below, components with very different severities should not usually be included in the same composite.

A related goal is that components should tend to be moderately correlated. Choosing components that are too similar (and thus very highly correlated) may mean that some of them are redundant with each other, whereas near zero correlation for most pairs of components would call into question the cohesiveness of the composite endpoint and may make interpretation difficult.

Finally and perhaps most importantly, components of a composite should ideally have the same severity, frequency, and treatment effect.[5,6,13–16] For example, a good composite might include only postoperative complications deemed to be serious in the average patient, expected to occur with similar frequency, and similarly affected by intervention. When these guidelines are not met, as is often the case, it is more difficult to interpret both the composite endpoint and overall treatment effect on it.[5,13,14] An important caveat is that if all components are equally important, differing frequencies and heterogeneous treatment effects do not really matter.[4] For example, a patient would likely choose treatment A over B regardless of which 2 complications were reduced by A, and regardless of their frequencies, as long as they were equally severe. In practice, however, equal severity across components is rare.

Differing severities and frequencies are also problematic because common statistical methods allow treatment effect estimates and tests to be driven by higher-frequency components. This is of special concern when higher-frequency components are less important than others and treatment effects also differ.[6,14,17] Heterogeneous treatment effects across the components of a composite, for example, reductions of 0%, 10%, and 50%, further complicate the interpretation of the overall treatment effect.

In the next section, we compare various statistical approaches for analyzing a binary-event composite endpoint, focusing on ways that modern methods can help with the challenges discussed above.

METHODS FOR ANALYZING A BINARY-EVENT COMPOSITE ENDPOINT

Choice of statistical method is an important part of designing a composite endpoint study because various methods provide different power, depending on the situation (see the section "Factors Affecting Powers of Tests for a Binary-Event Composite Endpoint" below). Available methods also differ in their interpretation and flexibility. Here, we describe and compare main features of "Standard Methods" and "Multivariate Methods" (both below). We show that multivariate methods are more flexible than standard methods, enabling them to help with problems of unequal severity and frequency, and to assess treatment effect heterogeneity across components. Throughout, we refer to a "global" or "overall" test as one that assesses the treatment effect across the entire composite endpoint, rather than evaluating individual components. We begin with an illustrative data example, the Smoking study.

Illustrative Example #1

In the "Smoking and Perioperative Outcomes" study,[18] Cleveland Clinic investigators assessed the association between smoking status and a composite of major postoperative complications using data from the American College of Surgeons National Surgical Quality Improvement Program[19] from 2006 to 2008. The 82,304 patients who reported smoking in the past year ("current smokers") were compared with 82,304 propensity matched[20,21] patients who reported having never smoked ("never smokers") on a vector of 13 major complications (Fig. 1). The composite was purposely wide because smoking affects many organ systems, both within and across patients. Results showed that current smokers were more likely to have major complications than never smokers. Throughout this article, methods for analyzing

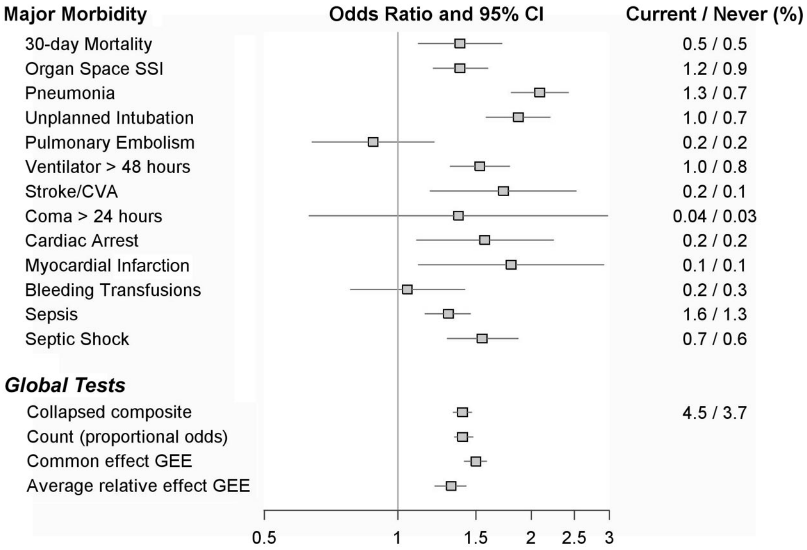

Figure 1. Smoking study results. Ten of the 13 individual components were significantly worse in smokers (confidence intervals [CIs] adjusted for multiple testing). Furthermore, all global test results show significantly higher odds of having the complication for current versus never smokers; that is, the odds ratio CIs all exceeded 1.0. SSI = surgical site infection; CVA = cerebrovascular accident; GEE = generalized estimating equation.

composite endpoints are illustrated and compared using the "Smoking study" (with data provided in Supplemental Digital Content 1, http://links.lww.com/AA/A258). All analyses were performed adjusting for age, which was still slightly imbalanced between current and never smokers after propensity score matching.

Standard Methods

Frequently used methods include comparing groups on the collapsed composite of "any event versus none," on the count of events per subject, or analyzing individual components separately. For the collapsed composite and count methods, component data are summarized into a single outcome for each subject before groups are compared. This simplification makes calculations easier, but at the cost of losing important details on the individual components.

Collapsed Composite

Perhaps the most frequently used composite outcome approach in perioperative medicine is to compare groups on the "collapsed composite" of any event versus none among a set of binary components, using either a χ^2 test or logistic regression. This was the primary analysis for the Smoking study, in which the collapsed composite incidence was 4.5% and 3.7% for current versus never smokers, respectively, with an odds ratio (95% confidence interval [CI]) of 1.40 (1.33, 1.47) ($P < 0.001$). Supplemental Digital Content 2, http://links.lww.com/AA/A259, contains the SAS code for this analysis and for all methods discussed below.

A major difficulty with collapsed composites is that higher-frequency components are overweighted, which is often problematic because those components tend to be less clinically important. Treatment effect estimates and test results are thus driven by a component(s) with the largest frequencies, overwhelming effects on smaller, and often more important, components. For example, a test comparing groups on the collapsed composite of higher-frequency infection and lower-frequency 30-day mortality would be driven by infections, potentially missing a clinically important difference in mortality caused by some unanticipated (noninfectious) mechanism. The any-versus-none structure also precludes importance weighting of components.

Largely inconsistent treatment effects are problematic for any global test, but particularly for the collapsed composite because differing individual effects can be hidden in the any-versus-none structure. For example, in the POISE I trial,[22] the effect of a β-blocker was assessed on a composite outcome of time to first event of nonfatal stroke, nonfatal myocardial infarction, or death. The β-blocker reduced the composite outcome, but individual component analyses showed a reduction only in nonfatal infarctions, whereas the other 2 increased. In such cases, the global test result lacks clean interpretation. Consequently, the investigators appropriately reported the overall composite as well as individual component results.

Count

The number of positive component events (i.e., "count") across the composite is sometimes chosen as the primary outcome. For example, in the Smoking study, the count would be the number of the distinct complications in the composite (Fig. 1) incurred by each patient. Groups are then compared on the count using a Mann-Whitney test, Poisson regression, or a proportional odds logistic regression model,[23] with the latter 2 allowing for covariable adjustment.

For the Smoking study, the components are all quite rare, with mean (SD) number of complications per patient of 0.06 (0.39) and 0.08 (0.45) for current and never smokers, respectively. The proportional odds logistic regression model gives an odds ratio (95% CI) of 1.40 (1.34, 1.48) ($P < 0.001$), meaning that current smokers are an estimated 1.4 times more likely to have more complications than never smokers.

Although attractive because it uses more information than the collapsed composite, the count can be difficult to interpret when components differ in severity. For example, a patient experiencing only the most severe outcome in a composite (e.g., death) may be worse off than a patient experiencing ≥2 less-severe events, but would be considered to have a better outcome. Furthermore, patients experiencing several highly correlated events may not be worse off than others experiencing fewer unrelated events. As with the collapsed composite, the count is also driven by higher-frequency components and does not directly facilitate clinical importance weighting of components.

Finally, because the count outcome is an ordinal variable, researchers might assume that a test based on it is generally more powerful than the binary any-versus-none collapsed composite test. Although often true, counterexamples are easy to construct, and determining which approach is more powerful in a specific situation is best done through simulations.[9]

Individual Component Analyses

When a composite primary outcome is chosen, individual components may be analyzed separately for several reasons. First, individual analyses are important in interpreting a global test result, that is, identifying which components are affected and describing the consistency of the treatment effect across components. Component analysis is particularly important if tests for treatment effect heterogeneity are statistically significant and/or strong heterogeneity across effects is observed (see "Multivariate Methods" section below).[24]

Second, investigators sometimes want to make conclusions about each component, regardless of the global test result or evidence for heterogeneity, as in the crystalloid-colloid trial discussed below in the "Data Application: Crystalloids Versus Colloids Trial" section. Alternatively, investigators may wish to demonstrate significant individual effects on some or all components of a composite in addition to a global effect.

Finally, sometimes no overall test is desired; individual analysis of components is the primary analysis. Here, investigators need to specify a priori the rule for claiming success of the intervention. Do all components need to be significant? At least one? A certain number? The decision on how many (or which specific) components are required needs to be made on clinical grounds. In general, power decreases as the number of components for which significance is required increases.

In each case above, adjustment to the significance criterion for multiple comparisons is needed to protect the overall type I error, except when all (or arguably, most) components are required to be significant before success is claimed.[1] Adjustment can be made using the traditional Bonferroni correction, or less conservatively using, for example, the Holm-Bonferroni method.[25]

Treatment effects on individual binary components can be assessed using Pearson χ^2 tests, using a correction for multiple comparisons. Such was the approach taken with the Smoking data, where 99.6% Bonferroni-corrected CIs are reported ($1-0.05/13 = 0.996$) (Fig. 1). For 10 of the 13 components, the odds of having the complication are significantly higher for current versus never smokers; only 1 of 13 odds ratio estimates is in the opposite direction (nonsignificant). When most individual effects are in the same direction, even if nonsignificant, any of the global tests are more powerful and easier to interpret.

A main limitation of individual component analysis compared with a global test is reduced power—one of the reasons composite outcomes are chosen in the first place. Power can be somewhat increased by using the Pearson χ^2 test in the context of resampling. Through resampling, we can simultaneously adjust for the discrete nature of the data, within-subject correlation and multiple testing. In our analysis of the crystalloid-colloid data (below in the "Data Application: Crystalloids Versus Colloids Trial" section), we used resampling and the Holm-Bonferroni multiple-comparison procedure to obtain adjusted *P* values for each component.[26] Basically, adjusted *P* values for each component are computed as the probability (over many thousands of resamples) that data sampled under the null hypothesis (i.e., no treatment effect) produces *P* values smaller than the observed *P* value for that component.

In the section "Factors Affecting Powers of Tests for a Binary-Event Composite Endpoint," we use the above resampling method to conduct a "minimum *P* value" test of the null hypothesis that at least 1 component has a nonzero treatment effect, and compare its power with other tests. Using a Bonferroni correction (here, same criterion as Holm-Bonferroni), the smallest observed *P* value is significant if it is smaller than α/k, where α is the significance level and k is the number of components. Of course, study hypotheses requiring significance on >1 component would be less powerful than this test.

Multivariate Methods

A binary-event composite endpoint is often best analyzed using one of several multivariate (i.e., one record per component per patient) generalized estimating equation (GEE) methods.[27,28] These methods model the individual component data for each subject directly, instead of first summarizing results within each subject as do the collapsed composite and count methods. This flexibility enables them to address problems of unequal severity and frequency, and to assess treatment effect heterogeneity across components, while adjusting for within-subject correlation.

We first describe a "common effect" test in which a single common treatment effect across components is assumed and estimated. We then describe 2 "distinct effects" tests in which a different treatment effect for each component is estimated separately, and then hypotheses on the effects tested.

Common Effect Test

In the common effect test, a single "common" treatment effect odds ratio is estimated across the components of a composite.[7,8,29] This test is therefore most meaningful and powerful when components are similarly affected, although it remains useful in the face of moderate heterogeneity. Lefkopoulou and Ryan[7] showed that the common effect test is usually at least as powerful as the collapsed composite. Pogue et al.[24] concluded that this method was generally more powerful than other multivariate tests that also assume a common effect across components. Appendix 1 gives details on fitting the model.

For the Smoking study, the individual odds ratios appear fairly consistent across components, except for pulmonary embolism and bleeding (Fig. 1), so the common effect method seems reasonable. The estimated common effect odds ratio is 1.50 (95% CI: 1.41, 1.59), somewhat larger than the collapsed composite and count results (both 1.40), assuming an exchangeable or "equal" correlation between components (estimated correlation = 0.10).

Table 1. Smoking and Perioperative Complications Study (*n* = 82,304/Group)

Global tests[a]	Odds ratio (95% CI)	χ^2 (*df*)	*P* value
Collapsed composite (any-versus-none)	1.40 (1.33, 1.47)[b]	174 (1)	<0.001
Count of events	1.40 (1.34, 1.48)[c]	176 (1)	<0.001
Common effect[d] GEE	1.50 (1.41, 1.59)[e]	178 (1)	<0.001
Average relative effect[f] GEE	1.32 (1.21, 1.43)[e]	43 (1)	<0.001
Treatment-component interaction[g] GEE	—	123 (12)	<0.001

Results of all global tests for the smoking and perioperative complications study comparing *n* = *82,304 current smokers who were propensity matched to 82,304 never smokers.*

CI = confidence interval; GEE = generalized estimating equation models to adjust for within-subject correlation among components; unstructured pairwise correlations ranged from 0.55 to 0.89.

[a] Tests assessing the relationship between smoking status (current or never smoker) and a composite endpoint consisting of 13 postoperative complication events (Fig. 1).

[b,c,e] Odds ratio in current versus never smokers of: [b] at least 1 complication, [c] a higher complication count (proportional odds logistic regression), [e] overall odds of complications.

[d] Common effect: estimating a single treatment effect across all 13 components.

[f] Average relative effect: estimating, then averaging, the 13 distinct treatment effects.

[g] Test of whether the treatment effect differs across the 13 components.

When components have differing clinical severities, clinical importance weights (assigned a priori) can be applied to each component by weighting each observation. This is a distinct advantage over the unitary approach in the count and collapsed composite methods. Specific component weights are determined by investigators and reflect the clinical importance of the various component outcomes. They can be derived from expert judgment, a Delphi process,[30] or from previous work identifying patient perception, health-related quality of life, or cost.[31,32] Weights might also be derived based on previous associations between, for example, component complications and future health status (e.g., 1-year mortality).

However, the common effect test is driven by components with higher frequency, just as with the collapsed composite and count. Also, because the model assumes a common effect across components, it is especially important to report individual component results for accurate interpretation.

Distinct Effects Tests

Instead of assuming a common effect across all components, a more flexible option is to use a GEE distinct effects model in which a distinct treatment effect (and associated standard error) is estimated for each component.[9,10] This "distinct effects" model may also be more appropriate for some studies than assuming a common effect. We show below that with this model one can test whether the average relative effect equals zero, test whether the treatment effects are consistent across components, and apply clinical importance weights directly to each component effect. More details on these and other distinct effects tests are given in Reference 9, and briefly in Appendix 2.

Average relative effect test. As noted previously, the collapsed composite, count, and GEE common effect tests are easily driven by high-frequency components. This is because they are designed to be sensitive to reductions in the actual number of events between groups, and as such are much more sensitive to the absolute differences between proportions than the relative differences (e.g., relative risk). For example, in the above-mentioned tests, an absolute reduction from 0.50 to 0.40 for component A, a 20% relative reduction, would receive much more weight than a reduction from 0.10 to 0.05 for component B, even though B has a much larger relative reduction (50%).

The average relative effect test avoids this problem by simply averaging the component-specific treatment effects (i.e., log-odds ratios) from the distinct effects model, and testing whether the average is equal to zero.[9,10] Because each treatment effect receives equal weight in calculating the average, this test is not driven by higher-frequency components as are the common effect and other tests. Thus, the test is especially appropriate when composite components have differing incidences and the relative effects are at least as important as the absolute effects. Also, clinical importance weights can be applied directly to the treatment effects, and a priori designated subsets of the components can also be tested.

For the Smoking study, the average relative effect odds ratio (95% CI) is 1.32 (1.21, 1.43), considerably smaller than the common effect odds ratio of 1.50. This is because more frequent complications (e.g., pneumonia) also tended to have larger treatment effects, thus weighting the common effect odds ratio toward them. But the average relative effect odds ratio, as expected, is not dependent on the incidences.

Test for heterogeneity of treatment effects. Heterogeneity of treatment effects across components can be assessed by a treatment-by-component interaction test in the distinct effects GEE model (Appendix 1, code 5.1),[9,10] and should accompany global test results.[24] A *significant* test for heterogeneity is usually convincing evidence of underlying differences in treatment effect across components, and would indicate that individual assessment of components is needed, especially if observed effects are in opposite directions! However, a *nonsignificant* result does not assure that the effects are homogeneous, because a study powered for a global test may well be underpowered for the heterogeneity test. Nonetheless, this test can facilitate interpretation of global treatment effect estimates and guide decisions on whether a common or distinct effect method (if either) is most appropriate for the global analysis.

For the Smoking study, the test for heterogeneity of treatment effects is highly significant ($P < 0.001$, Table 1),

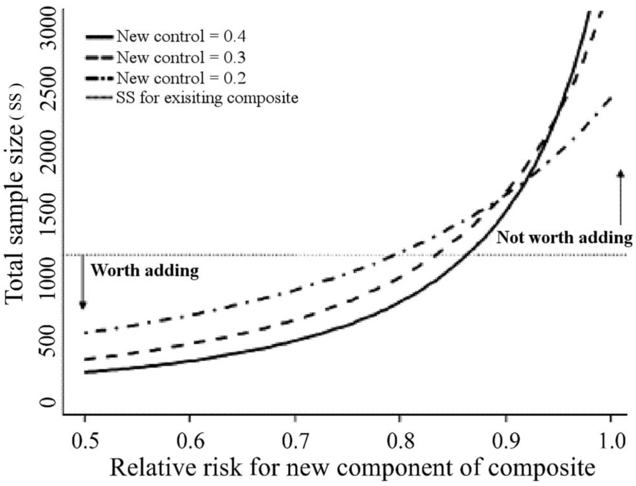

Figure 2. Effect on sample size (vertical axis) of adding a new component to a collapsed composite. Original composite requires 1164 patients to detect a 50% reduction from a control group incidence of 0.10 with 90% power at the 0.05 significance level. Plotted is required sample size to detect the resulting difference after adding a new component versus the relative risk of new component (horizontal axis), stratified by new control group incidence (3 curved lines) of 0.2, 0.3, or 0.4. Portions of the curves below the horizontal line indicate components worth adding. Least possible overlap between components is assumed. (Reprinted with modifications from Neaton et al.,[4] with permission.)

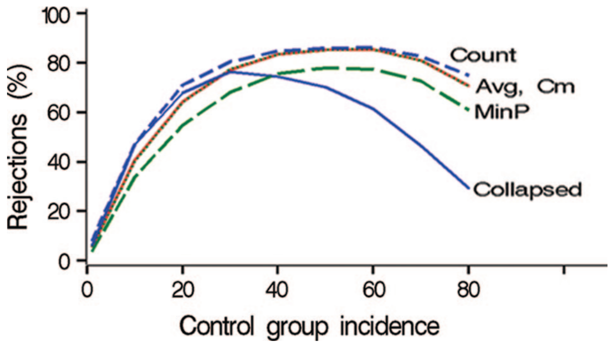

Figure 3. Power as function of baseline incidence, with consistent treatment effects (odds ratio = 0.74) across 4 outcomes. n = 500/group, 5000 simulations, within-subject correlation = 0.50. Tests: Avg = average relative effect; Cm = common effect; Ct = count; MinP = minimum P value; collapsed = collapsed composite. (Reprinted with modifications from Mascha and Imrey,[9] with permission.)

consistent with the variation in odds ratios across components seen in Figure 1, even though most are in the same direction. The significant heterogeneity test led us to test the individual treatment effects in addition to the global effect. However, when most of the effects are in the same direction (here, 12 of 13), a significant test for heterogeneity does not necessarily indicate that results for global tests such as the average relative effect or common effect are not meaningful. This is especially true for "quantitative" interactions, i.e., when effects are different but in the same direction. Table 1 summarizes results for all of the discussed tests for the Smoking study.

FACTORS AFFECTING POWERS OF TESTS FOR A BINARY-EVENT COMPOSITE ENDPOINT

Statistical power is an important consideration when designing any trial, but is especially challenging for composite outcome studies because many factors influence power, often in nonintuitive ways. Using simulations, we thus highlight factors that most affect power, including treatment effect, number of components, incidence, correlation, and consistency of treatment effects.[9] We also show which statistical tests are most powerful in particular situations.

Treatment Effect and Number of Components

Whether adding a new component to a composite will increase power depends mainly on the treatment effect and control group incidence of the new component in relation to existing components, and the correlation between new and existing components. For any of the discussed methods, adding a component with too small a frequency or too small a relative risk reduces power.

Figure 2 from Neaton et al.[4] shows the relationship (for the collapsed composite) between required sample size and the relative risk of a new component, for 3 scenarios of new control group incidence. For the original composite (i.e., before adding a new component), 1164 patients are needed to detect a relative risk of 0.5 from a control group incidence of 0.1 with 90% power at the 0.05 significance level (horizontal line).

The curved lines represent 3 possible scenarios of control group incidence for the collapsed composite (0.2, 0.3, 0.4) after adding a new component. Portions of these curves below the horizontal line indicate combinations of control group incidence and new component relative risk that make the new component worth adding to the composite, because the required sample size to detect the new difference between treatment and control is reduced.

As seen, if adding a new component increases the control group incidence to 0.2 (top dotted line), the relative reduction from control for the new component needs to be at least 20% (curve crosses horizontal line at relative risk of 0.80) to reduce the sample size needed for detecting the resulting population difference with 90% power at the 0.05 significance level. However, if the new component had control and treatment incidences of 0.10 and 0.06, respectively, for a relative risk of 0.6, the new collapsed composite would require only n = 676. New components resulting in control group incidences of 0.3 or 0.4 would need relative risks of approximately 0.83 and 0.87 or stronger, respectively, to be worth adding. Scenarios in Figure 2 assume the least possible overlap between the original and new components (e.g., control incidences of 0.1 for both components results in 0.2 for new composite). Higher correlation among components decreases power for each of the global tests (Fig. 4).

Incidence of the Components

Figure 3 depicts power (vertical axis) as a function of a common control group incidence (horizontal axis) when treatment effects are consistent (each of 4 components has an odds ratio of 0.74), with moderate/high within-subject correlation (0.50).

In this scenario, power for each test increases considerably as control group incidence increases, and then peaks

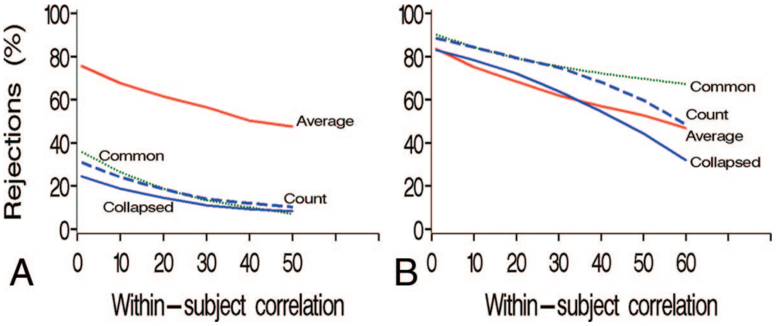

Figure 4. *Unequal* baselines with *some* components affected; power as function of within-subject correlation. Four outcomes, n = 1000 per group, 5000 simulations. Affected components in bold type. A, *Smaller* baseline reduced 50% (treatment: **0.05**, 0.10, 0.20, 0.20; control: **0.10**, 0.10, 0.20, 0.20). B, *Larger* baseline reduced 50% (treatment: 0.10, 0.10, **0.10**, 0.20; control: 0.10, 0.10, **0.20**, 0.20). Tests: common = common effect; average = average relative effect; collapsed = collapsed composite. (Reprinted with modifications from Mascha and Imrey,[9] with permission.)

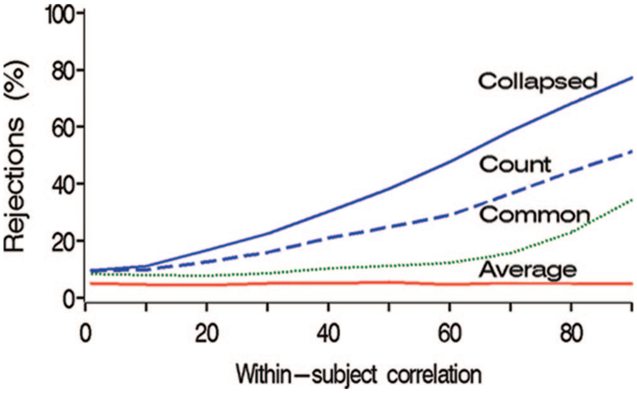

Figure 5. *Opposite* effects (on log-odds ratio scale) in same composite. Power appropriately stays at nominal level for average relative effect test, but increases with within-subject correlation for the other tests. Two outcomes; n = 1000 per group; 5000 simulations; incidences are 0.20 and 0.11 for treatment and 0.15 and 0.15 for control for the 2 outcomes, giving respective log-odds ratios of +0.36 and −0.36. Tests: common = common effect; average = average relative effect; collapsed = collapsed composite. (Reprinted with modifications from Mascha and Imrey,[9] with permission.)

and decreases slightly. The average relative effect, common effect, and count tests all have similar powers regardless of control group incidence. However, the collapsed composite has considerably less power than other tests for control group incidence more than approximately 0.3. The minimum *P* value test is designed to be more powerful with heterogeneous effects, so naturally has somewhat lower power here where the odds ratio is consistent across components.

Within-Subject Correlation

Lower within-subject correlation increases power for all tests, regardless of control group incidence. For example, if Figure 3 were drawn using a correlation of 0.10 instead of 0.50, each of the curves would be shifted upward. Lower correlation intuitively increases power because as component responses are less similar within subject, more information is represented by the multiple components. Power increasing with decreasing correlation is also depicted for all tests in Figures 4 and 5, with inconsistent treatment effects and baseline incidences.

Inconsistent Treatment Effects

As seen, when treatment effects and incidences are similar across components (Fig. 3), the multivariate tests and the count test have similar power, and are better than the collapsed composite at higher incidences. But in practice, it is likely that both treatment effects and incidences will differ across components, at least to some degree. When only some components are affected by treatment (and the remainder have zero effect), power is naturally lower for any particular global test than when all components are affected by treatment. Power is especially reduced when 1 or more components move in opposite directions. We explore here the relative powers of the tests when both treatment effects and control group incidences differ across components.

Smaller versus larger component affected. Figure 4 depicts statistical power for a scenario with 4 components and unequal control group incidences (0.10, 0.10, 0.20, 0.20), where either a larger incidence (Fig. 4A) or smaller incidence (Fig. 4B) control group component is reduced by treatment. In Figure 4A, where the only effect is a 50% reduction in a *smaller* component (0.10 reduced to 0.05), the average relative effect test is much more powerful than the common effect, count, and collapsed composite tests. In Figure 4B, where the only effect is a 50% reduction in a *larger* baseline component (0.20 reduced to 0.10), power of the average relative effect test is very similar to Figure 4A because the relative treatment effects (50% reduction) are the same. However, power for the common, count, and collapsed composite tests have increased substantially because the absolute reduction induced by treatment is doubled compared with Figure 4A (difference of 0.10 vs 0.05), and these tests are sensitive to absolute rather than relative effects.

Compared with the collapsed composite, count, and common effect tests, an advantage of the average relative effect test is that its power remains unchanged whether an intervention affects components with smaller or larger baseline incidences. Compared with the other tests, the minimum *P* value and interaction tests have higher power in both figures as is common when effects across components are quite disparate (not shown).

Reductions versus increases. A similar conclusion holds for scenarios with equal baseline incidences, where 2 of the 4 components are affected, and either both are reduced or increased by the same *relative* magnitude (say, log-odds ratios of either +0.35 or −0.35). Power of the average relative effect test does not change, but the other tests have considerably less power for the relative reductions than they do for relative increases (not shown).

Opposite effects in the same composite. Finally, when all components are affected, and of equal magnitude, but the effects are split in opposite directions on the log-odds ratio scale within a composite, a test that cancels the opposite effects and tends toward nominal (say 5%) power is desired. However, we see from Figure 5 that only the average relative effect test has consistently very low power, whereas the common effect, collapsed, and count tests gain power as within-subject correlation increases. As expected, the minimum P value and interaction tests have extremely high power for effects in opposite directions (>95%, not shown).

The average relative effect test may thus be preferred when incidences differ and it is not known which components will be affected more by treatment, and/or when direction of effects is not known. But when absolute effects are a priori deemed more important than relative effects, and weighting results by components with highest frequency is desired, then the common effect test is preferable.

Data Application: Crystalloids Versus Colloids Trial

In the trial of Crystalloids Versus Colloids During Surgery (ClinicalTrials.gov identifier: NCT00517127), researchers at Medical University of Vienna and their Cleveland Clinic collaborators are investigating whether intraoperative fluid management using colloids improves major perioperative complications compared with crystalloids. The primary outcome is a composite of binary complications in 6 organ systems (Table 2). Particular components were chosen because each is considered clinically serious and likely to be affected by intervention. Some of them, however, such as infection, are expected to occur more frequently than others. But because intervention could well improve some components while worsening or not affecting others, individual components analysis was planned as well. As is often the case, detecting an overall effect is insufficient; the investigators also want to know which complications, if any, are affected.

Table 2. Major Complications Defining the Composite Endpoint in the Crystalloid-Colloid Trial

Organ system[a]	Complication definition
1. Cardiac	Acute heart failure, myocardial infarction, ventricular arrhythmia
2. Pulmonary	Pulmonary embolism, pulmonary edema, respiratory failure, pneumonia
3. Renal	Dialysis
4. Coagulation	Bleeding
5. Gastrointestinal	Bowel and surgical anastomosis stricture/obstruction or anastomotic leak, fistulas, peritoneal effusions
6. Infectious	Deep or organ-space surgical site infection, sepsis

[a] Each system was considered as a binary event, such that component complications 1, 2, 5, and 6 are themselves collapsed composites.

Because the trial is ongoing, an illustrative dataset was simulated and then analyzed using each of the discussed methods; the planned analysis for the ongoing trial is the common effect GEE method. The underlying scenario had $n = 800$ patients per group and baseline event proportions (crystalloid group) of 0.02, 0.04, 0.06, 0.08, 0.10, and 0.12, corresponding to components 1 through 6 in Table 2, respectively. Underlying treatment effects were made to differ across components, with zero effect on the largest 2 baselines (gastrointestinal and infection) and a consistent 30% relative reduction on the remainder. The underlying scenario had a common correlation of 0.20 between all pairs of components, whereas in the GEE analyses, an unstructured correlation was used (because making no assumptions is best, unless not feasible).

Global treatment effect estimates for the collapsed composite, count, and common effect methods are similar (with odds ratios between 0.82 and 0.85), each nonsignificant, and as expected, each close to the effects of the 2 components with largest frequency (gastrointestinal and infection; Table 3). The average relative effect odds ratio of 0.74 is a more representative summary of the 6 relative effects (i.e., log-odds

Table 3. Crystalloid-Colloid Trial Simulated Data Example (n = 800 per Group)

	P_{T/P_C}[a]	Odds ratio (95% CI)[b]	χ^2 (df)	P value	Adjusted P value[c]
Individual analyses					
1. Cardiac	0.016/0.029	0.56 (0.28, 1.1)[d]	2.8 (1)	0.096	0.314
2. Pulmonary	0.038/0.048	0.78 (0.48, 1.3)[d]	1.0 (1)	0.323	0.679
3. Renal, dialysis	0.044/0.064	0.67 (0.43, 1.05)[d]	3.1 (1)	0.078	0.314
4. Coagulation	0.060/0.091	0.64 (0.44, 0.93)[d]	5.5 (1)	0.019	0.097
5. Gastrointestinal	0.090/0.100	0.89 (0.64, 1.2)[d]	0.45 (1)	0.503	0.746
6. Infection	0.120/0.120	0.98 (0.72, 1.3)[d]	0.02 (1)	0.878	0.878
Global methods					
Collapsed composite (any)	0.24/0.27	0.85 (0.68, 1.07)	1.9 (1)	0.169	—
Count of events	0.37 (0.80)/0.46 (0.95)[e]	0.84 (0.67, 1.05)	2.3 (1)	0.132	—
Common effect[f] GEE	—	0.82 (0.66, 1.03)	3.0 (1)	0.086	—
Average relative effect[g] GEE	—	0.74 (0.57, 0.96)	5.2 (1)	0.023	—
Treatment-outcome interaction[h] GEE	—	—	5.8 (5)	0.322	—

Comparing methods of analyzing a composite endpoint consisting of binary events using simulated data based on the crystalloids versus colloids randomized trial.

CI = confidence interval; GEE = generalized estimating equation model; pairwise correlations ranged from 0.13 to 0.34.

[a] Proportion with outcome in T (colloid) and C (crystalloid) groups. Random sample from underlying scenario in which components 1–4 had 30% reduction and components 5 and 6 had zero effect.

[b] Estimated odds of outcome in colloid versus crystalloid patients.

[c] χ^2 test with resampling and stepdown multiple comparison procedure (50,000 resamples).

[d] Univariate logistic regression.

[e] Mean (SD) of count of complications across components per subject (proportional odds model).

[f] Common effect: estimating a single treatment effect across all 13 components.

[g] Average relative effect: estimating, then averaging, the 13 distinct treatment effects.

[h] Test of whether the treatment effect differs across the 13 components.

ratios), and as such is farther from zero and the only significant test ($P = 0.024$). None of the individual effects was significant using a univariate test and correcting for multiple testing (Table 3, last column). Thus, only the average relative effect test detected the 4 underlying 30% treatment effects. As shown in Figure 4A, when only lower-frequency components are affected, this test is more powerful than comparator tests, especially with small baselines.

The treatment-outcome interaction test is nonsignificant ($P = 0.32$), even though there is a strong underlying interaction and large sample size. This supports findings that the GEE interaction test may often be underpowered, especially with low-frequency components.[9]

A practical advantage of the multivariate tests we have discussed is that clinical importance weights can be directly applied, either to the estimated treatment effects (average relative effect method) or the individual data points (common effect method). For example, suppose investigators had agreed a priori that coagulation, cardiac, pulmonary, and renal complications were twice as important as infection and gastrointestinal complications. Assigning double weight for the treatment effects on these outcomes, we obtain a weighted average relative effect odds ratio (95% CI) of 0.70 (0.53, 0.94) ($P = 0.016$), somewhat stronger than the unweighted result because the 4 affected components were emphasized by the weighting.

DISCUSSION

A key first step in designing a nonrandomized or randomized study in which the primary outcome is a binary-event composite endpoint is to carefully choose the components of the composite. Components should capture the underlying disease or outcome of interest and also be very plausibly affected by the intervention. The chosen set should be parsimonious and nonredundant, qualities that help interpretation of the composite and improve power of individual component analyses. A corollary is that investigators should exercise considerable caution in adopting an unmodified composite endpoint from a previous study unless it truly represents the current outcome of interest. Finally, although challenging, investigators should strive to choose components with similar baseline frequency, treatment effects, and clinical severity.[5,6,13–16] In practice, however, strong similarity across all components in each of these 3 dimensions is rarely possible. Fortunately, newer statistical methods can help ameliorate shortcomings of composite endpoints while simultaneously improving power.[8–10] The second key planning step is thus choosing the appropriate statistical approach.

We discussed 2 multivariate (i.e., multiple outcome per subject) GEE methods for comparing groups on a binary-event composite endpoint: the common effect test, which assumes and estimates a common treatment effect across components,[7,8] and the average relative effect test, which estimates and then averages effects across components.[9,10] Compared with standard methods, collapsed any-versus-none, count of events, and individual component analyses, advantages of the multivariate approaches include use of more information per subject, ability to apply clinical importance weights, and in most cases greater statistical power. The average relative effect test has the important additional advantage of preventing results from being driven by higher-frequency components. It is thus preferred when components differ in both frequency and treatment effect, and the relative treatment effects (e.g., odds ratios) are at least as important as absolute effects (e.g., differences in proportions).

An additional distinct advantage of multivariate over conventional methods is that clinical importance weights can be applied. Although some investigators might be reticent to apply inherently subjective clinical importance weights, failing to incorporate them assumes equal weights and is thus often worse.

Relative powers of methods for analyzing a binary-event composite depend on the frequencies, treatment effects, and correlations among components.[7–9] For most tests, power decreases as correlation among components increases, which is worth remembering when choosing components. As incidence increases, power initially increases for all tests, but then peaks and decreases. Notably, the frequently used collapsed composite is considerably *less* powerful than other methods when incidences are moderate to high (say >0.15) and effects are consistent. For all global tests, power decreases as treatment effect heterogeneity increases.

Although primary analysis of a composite endpoint is usually a global assessment of the treatment effect across components, it is vital to assess and report the heterogeneity of treatment effects.[24] When heterogeneity is detected, and/or when inference on individual components is desired, as is often the case, individual component analyses are clearly needed, independent of the global test results.

Design of studies using composite endpoints thus requires careful planning of the components and the analytic method for best interpretation and power. Newer statistical methods often help meet the practical challenges of using composite endpoints in clinical trials.

APPENDIX 1: COMMON EFFECT GENERALIZED ESTIMATING EQUATION MODEL DETAILS

ID	Smoker	Component	Outcome
1	0	1	1
1	0	2	0
1	0	3	1
1	0	13	1
2	1	1	0
2	1	2	0
2	1	3	1
2	1	13	1

The Smoking study data are set up (as for each multivariate method discussed) with 1 row per component per patient, such as:

and so on, where "Smoker" is 1 for current and 0 for never smokers, and "Outcome" is 1 if a complication was observed and 0 otherwise, for each of components (i.e., complications) 1 to 13.

The common effect log-odds ratio is estimated in a multivariate logistic model such as:

$$\log\left[\frac{\pi_k}{1-\pi_k}\right] = \alpha_k + \beta X, \tag{1}$$

in which the incidence (π_k) for each outcome is modeled as a function of treatment group X (here, smoker status). A within-subject correlation structure must be specified; logical options are an unstructured (distinct correlation for each pair of components) or exchangeable (same correlation for all pairs) correlation. A single treatment effect ($\hat{\beta}$) is then estimated, along with a separate control group incidence ($\hat{\alpha}_k$) per component. The common effect odds ratio (i.e., odds of outcome in treated versus control patients) is estimated by exponentiating $\hat{\beta}$, as in logistic regression. The null hypothesis that the common effect log-odds ratio equals 0, or equivalently, that the odds ratio = 1, is assessed using a 1-*df* Wald χ^2 test of $\hat{\beta}/\hat{S}E_{\hat{\beta}}$, where $\hat{S}E_{\hat{\beta}}$ is the standard error of the estimated treatment effect, $\hat{\beta}$. Supplemental Digital Content 2 (http://links.lww.com/AA/A259) contains the SAS code to conduct this test.

APPENDIX 2: DISTINCT EFFECT GENERALIZED ESTIMATING EQUATION MODEL DETAILS

For the distinct effects model, a multivariate generalized estimating equation (GEE) model can be used to estimate a distinct treatment effect, β_k, and standard error for each component, as in the equation

$$\log\left[\frac{\pi_k}{1-\pi_k}\right] = \alpha_k + \beta_k X, \qquad (2)$$

where the only difference from model (1) is the component-specific log-odds ratio, β_k, instead of the common effect, β. As shown in more detail in Mascha and Imrey (2010)[9] and Bull (1998),[10] a generalized Wald test can be used to conduct several distinct effects tests (beyond what is discussed here), simply by changing the contrast matrix or vector $\mathbf{L}'$ in the following test statistic

$$W = (\mathbf{L}'\hat{\boldsymbol{\beta}})'(\mathbf{L}'\hat{\boldsymbol{\Sigma}}_{\beta}\mathbf{L})^{-1}(\mathbf{L}'\hat{\boldsymbol{\beta}}) \sim \chi^2_{p}, \qquad (3)$$

where $\hat{\boldsymbol{\beta}}$ is the vector of estimated treatment effects from the GEE model, $\hat{\Sigma}_{\beta}$ is the robust variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, $\mathbf{L}'$ is a $p \times K$ contrast matrix or vector and p is the number of rows in the contrast $\mathbf{L}'$. We thus test the null hypothesis

$$\begin{aligned}&\text{H0: } \mathbf{L}'\boldsymbol{\beta} = 0 \textit{ versus the alternative}\\ &\text{H1: } \mathbf{L}'\boldsymbol{\beta} \neq 0.\end{aligned}$$

The tests below are based on Equation (3) and are distinguished by the particular form of the contrast $\mathbf{L}'$ that is used. Supplemental Digital Content 2 (http://links.lww.com/AA/A259) contains the practical SAS code to conduct the tests below.

Average Relative Effect Test

The average effect GEE test is constructed using Equation (3) and inserting the K-length vector ($1/K\ 1/K\ 1/K \ldots$) as the contrast $\mathbf{L}'$, so that $\mathbf{L}'\boldsymbol{\beta} = 1/\mathrm{K}\boldsymbol{\Sigma}_{k=1}^{K}\boldsymbol{\beta}_k$, the average log-odds ratio across components. We thus test the null hypothesis that the average log-odds ratio = 0 with a 1-*df* χ^2 test.

Treatment-Outcome Interaction "Heterogeneity" Test

Homogeneity of the treatment effects across components can be assessed using a *K*-1 *df* treatment-component interaction test based on Equation (3) and using a *K*-1 by *K* matrix of contrasts for $\mathbf{L}$, where $\mathbf{L}'\boldsymbol{\beta}$ based on $K = 4$ components would be as follows:

$$\mathbf{L}'\boldsymbol{\beta} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_1 - \beta_3 \\ \beta_1 - \beta_4 \end{bmatrix}. \qquad (4)$$

DISCLOSURES

Name: Edward J. Mascha, PhD.
Contribution: This author helped design the study, conduct the study, analyze the data, and write the manuscript.
Attestation: Edward J. Mascha approved the final manuscript.
Name: Daniel I. Sessler, MD.
Contribution: This author helped write the manuscript.
Attestation: Daniel I. Sessler approved the final manuscript.

REFERENCES

1. Sankoh A, D'Agostino R, Huque M. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. Stat Med 2003;22:3133–50
2. Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. BMJ 2010;341:c3920
3. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA 2003;289:2554–9
4. Neaton J, Gray G, Zuckerman B, Konstam M. Key issues in end point selection for heart failure trials: composite end points. J Cardiac Fail 2005;11:567–75
5. Chi G. Some issues with composite endpoints in clinical trials. Fundam Clin Pharmacol 2005;19:609–19
6. Tomlinson G, Detsky A. Composite end points in randomized trials: there is no free lunch. JAMA 2010;303:267–8
7. Lefkopoulou M, Ryan L. Global tests for multiple binary outcomes. Biometrics 1993;49:975–88
8. Legler JM, Lefkopoulou M, Ryan LM. Efficiency and power of tests for multiple binary outcomes. J Am Stat Assoc 1995;90:680–93
9. Mascha EJ, Imrey PB. Factors affecting power of tests for multiple binary outcomes. Stat Med 2010;29:2890–904
10. Bull SB. Regression models for multiple outcomes in large epidemiologic studies. Stat Med 1998;17:2179–97
11. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. N Engl J Med 1995;333:1581–8
12. Bethel MA, Holman R, Haffner SM, Califf RM, Huntsman-Labed A, Hua TA, McMurray J. Determining the most appropriate components for a composite clinical trial outcome. Am Heart J 2008;156:633–40
13. Pocock S. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis and interpretation. Control Clin Trials 1997;18:530–45
14. Ferreira-Gonzalez I, Permanyer-Miralda G, Busse JW, Devereaux PJ, Guyatt GH, Alonso-Coello P, Montori VM. Composite outcomes can distort the nature and magnitude of treatment benefits in clinical trials. Ann Intern Med 2009;150:566–7
15. Myles PS, Devereaux PJ. Pros and cons of composite endpoints in anesthesia trials. Anesthesiology 2010;113:776–8
16. Montori V, Permanyer G, Ferreira I, Busse J, Pacheco-Huergo V, Bryant D, Alonso J, Akl E, Domingo-Salvany A, Mills E, Wu P, Schünemann HJ, Jaeschke R, Guyatt G. Validity of composite end points in clinical trials. BMJ 2009;330:3

17. Ferreira-Gonzalez I, Permanyer-Miralda G, Busse J, Bryant D, Montori V, Alonso-Coello P, Walter S, Guyatt G. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. J Clin Epidemiol 2007;60:651–7
18. Turan A, Mascha EJ, Roberman D, Turner P, You J, Kurz A, Sessler DI, Saager L. Smoking and perioperative outcomes. Anesthesiology 2010;114:837–46
19. Khuri SF. The NSQIP: a new frontier in surgery. Surgery 2005;138:837–43
20. Blackstone E. Comparing apples and oranges. J Thorac Cardiovasc Surg 2002;123:8–15
21. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55
22. Poise Study Group. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. Lancet 2008;371:1839–47
23. Hosmer D, Lemeshow S. Applied Logistic Regression. 2nd ed. New York: John Wiley & Sons, 2000
24. Pogue J, Thabane L, Devereaux PJ, Yusuf S. Testing for heterogeneity among the components of a binary composite outcome in a clinical trial. BMC Med Res Methodol 2010;10:49
25. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65–70
26. Westfall P, Young S. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustments. New York: John Wiley & Sons, 1993
27. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22
28. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986;42:121–30
29. Tilley B, Marler J, Geller N, Lu M, Legler J, Brott T, Lyden P, Grotta J, for the National Institute of Neurological Disorders and Stroke (NINDS) rt-PA Stroke Trial Study Group. Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA Stroke Trial. Stroke 1996;27:2136–42
30. Linstone HA, Turoff M. The Delphi Method: Techniques and Applications. Reading, MA: Addison-Wesley, 1975
31. Follmann D, Wittes J, Cutler JA. The use of subjective rankings in clinical trials with an application to cardiovascular disease. Stat Med 1992;11:427–37
32. Eurich DT, Majumdar SR, McAlister F, Tsuyuki R, Yasui Y, Johnson J. Analyzing composite outcomes in cardiovascular studies: traditional Cox proportional hazards versus quality-of-life-adjusted survival approaches. Open Med 2010;4:40–8