Factors affecting power of tests for multiple binary outcomes

Edward J. Mascha\textsuperscript{a}∗† and Peter B. Imrey\textsuperscript{b}

Frequently in clinical studies a primary outcome is formulated from a vector of binary events. Several methods exist to assess treatment effects on multiple correlated binary outcomes, including comparing groups on the occurrence of at least one among the outcomes (‘collapsed composite’), on the count of outcomes observed per subject, on individual outcomes adjusting for multiplicity, or with multivariate tests postulating either common or distinct effects across outcomes. We focus on a 1-df distinct effects test in which the estimated outcome-specific treatment effects from a GEE model are simply averaged, and compare it with other methods on clinical and statistical grounds.

Using a flexible method to simulate multivariate binary data, we show that the relative efficiencies of the assessed tests depend in a complex way on the magnitudes and variabilities of component incidences and treatment effects, as well as correlations among component events. While other tests are easily ‘driven’ by high-frequency components, the average effect GEE test is not, since it averages the log odds ratios unweighted by the component frequencies. Thus, the average effect test is relatively more powerful when lower frequency components have stronger associations with a treatment or other predictor, but less powerful when higher frequency components are more strongly associated. In studies when relative effects are at least as important as absolute effects, or when lower frequency components are clinically most important, this test may be preferred. Two clinical trials are discussed and analyzed, and recommendations for practice are made. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: composite outcomes; generalized estimating equations; multivariate binary data; correlated binary outcomes; relative efficiency; clinical trials

1. Introduction

Clinical researchers frequently formulate a primary outcome from a vector of $K$ correlated binary outcomes anticipating that the study power will be increased relative to using a single binary outcome, and/or because the components collectively capture the disease or phenomenon of interest better than any single outcome event [1–6]. In some such studies it may also be important to draw conclusions about the individual components, while in others an overall assessment of the composite is sufficient. Existing methods to compare treatments on multiple correlated binary outcomes include comparing groups on the occurrence of at least one among the outcomes (the ‘collapsed composite’), on the count of positive components per subject, on the individual outcomes adjusting for multiple testing [7–9], or with multivariate tests, typically employing generalized estimating equations (GEE) [10, 11], random effect [12] or latent variable [13] models to adjust for the within-subject correlation. Although we discuss only binary outcomes, methods also exist to simultaneously assess treatment effects when the multiple outcomes per patient are of different types, such as binary and continuous, using latent variable models [14, 15] or GEE models [16], for example.

Lefkopoulou and Ryan [17] demonstrated that the relative power of the collapsed composite to a homogeneous ‘common’ treatment effect GEE test is a function of the correlations between the outcomes, the number of outcomes and the marginal proportions. The common effect test was found to be generally more powerful than the collapsed composite,

\textsuperscript{a}Departments of Quantitative Health Sciences and Outcomes Research, Cleveland Clinic, Cleveland, OH, U.S.A.
\textsuperscript{b}Department of Quantitative Health Sciences and Mellen Center for Multiple Sclerosis Treatment and Research, Cleveland Clinic, Cleveland, OH, U.S.A.
∗Correspondence to: Edward J. Mascha, Department of Quantitative Health Sciences (JNN-3), Cleveland Clinic, 9500 Euclid Ave, Cleveland, OH 44195, U.S.A.
†E-mail: maschae@ccf.org

except with highly correlated rare components. Legler et al. [18] showed that the common effect test is also more powerful than a heterogeneous effect test when effects are consistent, but not when effects vary greatly. Their assessment of relative efficacy also included the generalized least-squares test and rank sum test of O’Brien [19] and simultaneous individual comparisons using resampling to adjust for multiple comparisons [20]. Bull [21] expanded the common and distinct effects GEE tests for multiple binary outcomes to the multiple regression setting, allowing individual predictors to have either common or distinct effects on the outcome vector, and also discussed a semi-Bayesian shrinkage approach of which the common and distinct effects models are extreme special cases.

An interesting and sometimes problematic feature of several methods in common use for multiple binary outcomes is that the overall test of the outcome vector can be ‘driven’ by one or more components with high baseline (i.e. control group) incidence(s) and perhaps lower clinical importance compared with the others. For example, a primary outcome vector in perioperative medicine might include a low-incidence major cardiac event as well as serious infection, which has higher incidence but somewhat less clinical importance. Owing to varying incidences, similar relative reductions in each outcome gives the infection effect more weight than the major cardiac event effect in the overall treatment effect estimates for either the collapsed composite composite, count, or common treatment effect approaches. Overall treatment effects are difficult to interpret in such situations [4, 6, 22].

To address this problem we focus on a test of the average of the outcome-specific treatment effects in a GEE model. We use the GEE robust covariance matrix to construct the variance of the test statistic and thus account for varying incidences and correlations among components. By design, the average effect estimator and test are rather insensitive to differing baseline incidences across components, and much less likely than existing methods to be disproportionately weighted by high incidence outcomes. The average effect test also facilitates application of a priori clinical importance weights directly to the component treatment effect estimates.

We compare several methods on both clinical and statistical grounds. In Section 2.2 we review three non-GEE tests: the collapsed composite Wald test of any positive outcome versus none; the Mann–Whitney test of the counts of the number of positive responses; and simultaneous individual tests of each outcome component using bootstrap resampling to adjust for multiplicity. In Section 2.3 we discuss several GEE tests, including the common effect test, the distinct effects K-df test, the average effect 1-df test, and a treatment–outcome interaction test. In Section 2.4 we describe our simulation methods. In Section 3 relative power of the tests is assessed under various effect size scenarios, underlying correlations and ranges of marginal probabilities. In Section 4 we apply the methods to two studies. In Section 5 we discuss results and make recommendations for practice.

2. Methods

2.1. Initial model and data setup

We assume there are two groups, treatment (X = 1) and control (X = 0), N total subjects, n_i subjects in group i, and K outcomes for each subject. Y_{ijk} is the kth binary outcome for the jth subject in the ith group. Expansion to three or more groups is trivial for each method. Table I summarizes the tests described in Sections 2.2 and 2.3.

2.2. Non-GEE methods

2.2.1. Collapsed composite. The collapsed composite outcome is denoted for the jth subject in the ith group as Y_{ij} = \max_k(Y_{ijk}). We model the proportion of patients having this outcome as a function of treatment assignment in a logistic regression model: \logit(\pi) = \log[\pi/(1-\pi)] = z + \beta X, where \pi = P(Y_{ij} = 1|X), \beta is the log odds ratio of treatment (X = 1) versus control (X = 0), and z is the intercept, equal to the log odds of the collapsed composite outcome in the control group. The maximum likelihood estimates of the log odds ratio and its standard error are, respectively,

\hat{\beta} = \log \left[ \frac{n_{11}n_{00}}{n_{10}n_{01}} \right] \quad \text{and} \quad \hat{SE}_\beta = \sqrt{\frac{1}{\sum_{i=0}^{1} \sum_{r=0}^{1} \frac{1}{n_{ir}}}}.

where n_{ir} is the number of subjects in the ith treatment group with the rth response (1 = any, 0 = none). Treatment effect is assessed with a 1-df Wald test of \( H_0: \beta = 0 \) comparing \((\hat{\beta}/\hat{SE}_\beta)^2\) to a chi-square distribution with 1 df.

Whether or not the power for the collapsed composite increases with K is a function of the treatment effects, baseline incidences of the components and the correlations among components [2]. Although the collapsed composite is popular in practice, major limitations are that components with higher incidence can drive results, and that it does not facilitate either assessment of the consistency of effects across components or weighting of components by clinical importance [1, 5, 6, 22].
### Table I. Comparative tests.

<table>
<thead>
<tr>
<th>Test (Section)</th>
<th>Description</th>
<th>df</th>
<th>Details/notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Collapsed composite (2.2.1)</td>
<td>Compare groups on presence of any of $K$ outcomes, where $Y_{ij} = \max_k(Y_{ijk})$ using a logistic regression Wald test of $H_0: \beta = \text{log odds ratio} = 0$ versus $H_1: \beta \neq 0$</td>
<td>1</td>
<td>$Y_{ijk} =$ response for $i$th subj, $j$th group, $k$th outcome, $r = \text{response (0,1)}$, $\beta = \log[p_{ijr}/n_{ijr}]$, $\hat{SE}<em>\beta = \sqrt{\sum_r 1/n</em>{ir}}$</td>
</tr>
<tr>
<td>2. Count (2.1.2)</td>
<td>Compare groups on count of events $Y_{ij} = \sum Y_{ijk}$ using Wilcoxon rank sum test of $H_0: P(T &gt; S) = 0.5$ versus $H_1: P(T &gt; S) \neq 0.5$</td>
<td>1</td>
<td>Where $T$ and $S$ are randomly selected values from subjects on test and standard treatments, respectively</td>
</tr>
<tr>
<td>3. Minimum $P$-value bootstrap (2.1.3)</td>
<td>Simultaneously compare groups on each component with Pearson chi-square tests using bootstrap resampling to adjust for multiple testing and within-subject correlation $H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$ versus $H_1: \text{at least one} \beta_k \neq 0$</td>
<td>1</td>
<td>Reject $H_0$ if smallest bootstrap-adjusted $P$-value is $&lt; \alpha$. Decisions for remaining outcomes use the bootstrap stepdown multiple comparison procedure</td>
</tr>
<tr>
<td>4. Common effect (2.2.2)</td>
<td>Assuming an underlying common log odds ratio $\beta$ across the $K$ outcomes, test in a GEE model $H_0: \beta = \text{common odds ratio} = 0$ versus $H_1: \beta \neq 0$</td>
<td>1</td>
<td>Wald test of $\hat{\beta}/\hat{SE}<em>\beta$; $\hat{SE}</em>\beta$ is GEE robust standard error of $\hat{\beta}$</td>
</tr>
<tr>
<td>5–8. Distinct effects GEE tests (equation (4))</td>
<td>Tests 5–8 based on GEE models with separate log odds ratio $\beta_k$ estimated for each component. Hypotheses assessed with generalized Wald test, distinguished among tests by specification of $L$</td>
<td>1</td>
<td>$W = (L'\hat{\beta}/(L'\sum_{ij}L)\Sigma_{ij}L') \sim \chi^2_p$, $\hat{\beta}$ is vector of estimated log odds ratios, $\sum_{ij}L'$ is the GEE robust covariance matrix and $L$ a contrast vector or matrix</td>
</tr>
<tr>
<td>5. Distinct effects $K$-df (2.2.3)</td>
<td>Test whether the vector of treatment effects equals zero. $H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$ versus $H_1: \text{at least one} \beta_k \neq 0$</td>
<td>$K$</td>
<td>$L' = I_K$, the identity matrix. Test not sensitive to the direction of effects</td>
</tr>
<tr>
<td>6. Average effect (2.2.4)</td>
<td>Test whether the average (or weighted average) effect among the $K$ outcomes is equal to zero $H_0: L'\hat{\beta} = 0$ versus $H_1: L'\hat{\beta} \neq 0$</td>
<td>1</td>
<td>Using $L = (1/K)I_K$, $L'\hat{\beta}$ estimates the mean (log scale) relative treatment effect across $K$ components</td>
</tr>
<tr>
<td>7. Treatment–outcome interaction (2.2.6)</td>
<td>Test consistency of $K$ treatment effects across components with $K - 1$ df Wald test of $H_0: \beta_1 = \beta_2 = \cdots = \beta_K$ versus $H_1$: all $\beta$ not equal</td>
<td>$K - 1$</td>
<td>$L'<em>{(K-1)\times K}$ is matrix of $K - 1$ contrasts among the $K$ log-odds ratios, i.e. $L' = [I</em>{K-1}, -I_{K-1}]$</td>
</tr>
<tr>
<td>8. Var–Cov weighted avg (2.2.5)</td>
<td>Test whether $\hat{\beta}$ weighted by $W = \sum_{ij}^{-1}$ (inverse GEE robust cov matrix) = 0 Weighted average $\hat{\beta}<em>{e-w} = (I_K \sum</em>{ij}^{-1} \hat{\beta}/I_K \sum_{ij}^{-1} I_K$. $L = I_K$</td>
<td>1</td>
<td>Wald-like chi-square test of $(L' \sum_{ij}^{-1} I_K) \hat{\beta}_{e-w}$ yields results very similar to the common effect GEE test in most simulations</td>
</tr>
</tbody>
</table>

**2.2.2. Count.** For the count outcome, $Y_{ij} = \sum_{k=1}^{K} Y_{ijk}$ is the sum of the $K$ binary outcomes. Treatment effect on the count is often assessed using the two-tailed Mann–Whitney test, because the distribution of the count of a small number of correlated binary outcomes is non-Gaussian and dependent on the correlation structure.

Some assume that because the count outcome is ordinal, it necessarily yields more power than the collapsed composite. In fact, counterexamples are easy to construct, and systematic studies have not been done to assess the relative power of the count test to other methods of handling multiple binary events. The count test does not allow distinction between components or assessment of consistency of treatment effects. It also ignores the correlation among components, which can make interpretation difficult.

**2.2.3. Individual component tests adjusted for correlation and multiplicity.** When conclusions about treatment effects on individual outcomes are important, univariate analyses are needed, either alone or in conjunction with a global test [7, 23]. One approach is to modify results of univariate tests using Holm–Bonferroni or other standard multiplicity adjustments [24–27]. However, standard methods do not directly incorporate the correlations among components.
We therefore test the overall null hypothesis that all treatment effects are equal to zero against the alternative that at least one of them is non-zero using bootstrap resampling (20,000 samples with replacement), assessing the individual component treatment effects with Pearson chi-square tests. We use the resampling stepdown procedure to simultaneously adjust for multiple testing and inter-component correlation structure [8], and call this the minimum P-value bootstrap test.

Limitations of using individual tests as primary analysis include the lack of an overall treatment effect estimator, difficulty in interpreting results when some components are significant and others are not, and that sensitivity to a broad range of alternatives, such as when only one component is affected, may be achieved at the cost of reduced power relative to other global tests when components are similarly affected [18, 28].

2.3. GEE methods

2.3.1. Model and data setup. We compare and contrast five GEE tests [10, 11] derived from GEE models in a common multivariate framework of K binary outcomes observed on each of N subjects. In each model, the logit [i.e. log \((\pi_i/(1−\pi_i))\)] of the probability of a positive response on each outcome, \(\pi_k\), is expressed as a function of treatment group, with a separate intercept, \(x_1, \ldots, x_K\) for each component. Let \(\pi_j\) be the \(K \times 1\) vector of outcome probabilities and \(Y_j\) the \(K \times 1\) vector of responses for the \(j\)th subject. Models contain either a single ‘common’ treatment effect parameter \(\beta\) (Section 2.3.2), or a vector of treatment effects \(\beta\), representing distinct treatment effects across components (Sections 2.3.3–2.3.5). Let \(\theta\) be the vector \([\alpha', \beta']\) of intercepts and treatment effect(s).

Our GEE estimators \(\hat{\beta}\) of \(\beta\) are solutions to the equations \(S(\theta) = \sum_{j=1}^{N} D_j \hat{V}_j^{-1}(Y_j−\pi_j(\hat{\theta})) = 0\), where \(\hat{V}_j\) is \(\hat{\phi}A_j^{1/2}Z_j^{1/2}R_j(\hat{\theta})Z_j^{-1/2}A_j^{1/2}\), the GEE working covariance matrix of \(Y_j\), \(D_j = \hat{c}_j(\pi_j(\theta))/\hat{\theta}\), \(R_j(\hat{\theta})\) is the estimated \(K \times K\) working correlation matrix, specified by the vector of correlation estimates \(\hat{\phi}\), \(\hat{A}_j\) is a \(K \times K\) diagonal matrix of \(\hat{\pi}_k(1−\hat{\pi}_k)\), \(\hat{\phi}\) is an estimated dispersion parameter, and the \(Z_j\) are the diagonal matrices of weights, which may be varied by component and/or subject but by default are constant.

We use the GEE robust ‘sandwich’ covariance estimator \(\sum R\) to adjust for within-subject correlations among components when making inference on \(\theta\). \(\sum R = \sum_{j=1}^{N} D_j \hat{V}_j^{-1}D_j\) is the model-based estimated covariance of \(\hat{\theta}\), \(I_1 = \sum_{j=1}^{N} D_j \hat{V}_j^{-1}Cov(Y_j \hat{V}_j^{-1}D_j\) and Cov\((Y_j) = (Y_j−\pi_j(\theta))(Y_j−\pi_j(\hat{\theta}))\). Although Wald tests are formulated, asymptotically equivalent generalized score tests could be constructed for each test [29, 30].

2.3.2. Common treatment effect. We first consider a GEE model with a common treatment effect \(\beta\) and logit link:

\[
\log\left[\frac{\pi_k}{1−\pi_k}\right] = x_k + \beta X
\]

where \(\pi_k = P[Y_{ijk}=1|X]\) is the probability that a subject in group \(X\) has outcome \(k\); \(x_k\) is the intercept for the \(k\)th outcome, the logit of \(P[Y_{ijk}=1|X=0]\); and \(\beta\) is the common log odds ratio of treatment \((X=1)\) versus control \((X=0)\).

In matrix notation, the model is

\[
\begin{align*}
\text{logit}(\pi_i) &= T_i'\theta \\
\end{align*}
\]

where \(\pi_i\) is the \(K \times 1\) vector of outcome probabilities for the \(i\)th group, on which the logit operates elementwise; \(\theta\) is a \((K+1) \times 1\) vector of parameters including the \(K\) intercepts and \(\beta\), the common log-odds ratio; and the design matrices \(T_{(K+1) \times (K+1)} = [I_K, i1_K]\) for each subject in group \(i\) each consist of an identity matrix corresponding to the \(K\) intercepts with an adjoining column \(i1_K\) indicating treatment assignment for the common effect parameter. The dimension of the overall design matrix \(T\) is \(NK \times (K+1)\).

We test \(H_0:\beta = 0\) with a 1-df GEE Wald chi-square test of \((\hat{\beta}/\hat{SE}_\beta)^2\), where \(\hat{SE}_\beta\) is the GEE robust standard error of \(\hat{\beta}\). Clinical importance or other weights could be applied to individual observations using the diagonal matrix \(Z_j\) in \(\hat{V}_j\).

This is the global odds ratio model discussed by Tilley [31] and the common effects global Wald statistic discussed by Bull [21]. Moreover, it is asymptotically equivalent to the homogeneous effect score test of Lefkopoulou and Ryan [17] which was derived under the independence working correlation [32].

Distinct effects GEE models.

Four distinct effects GEE tests are described (Sections 2.3.3–2.3.6) based on the following model, which differs from the common effect model in (1) in that a separate treatment effect \(\beta_k\) is estimated for each component:

\[
\log\left[\frac{\pi_k}{1−\pi_k}\right] = x_k + \beta_k X
\]

In matrix notation the model takes the form of (2) with \(\theta\) a \((2K) \times 1\) vector of parameters including the \(K\) intercepts plus \(K\) log-odds ratio parameters \(\beta=(\beta_1, \ldots, \beta_K)\) instead of a single \(\beta\), and design matrices \(T_{(2K) \times (2K)} = [I_K, i1_K]\) consisting
of an identity matrix for the $K$ intercepts with adjoining diagonal matrix $iI_K$ indicating treatment assignment for the outcome-specific effect parameters. The overall dimension of $T$ is $NK \times (2K)$.

We formulate hypothesis tests about the vector of treatment effects $\hat{\beta}$ using the generalized Wald statistic $[29, 30]$ as

$$ W = (L'\hat{\beta})'(L'\sum_{\beta}L)^{-1}(L'\hat{\beta}) \sim \chi^2_p $$

(4)

where $\sum_{\beta}$ is the variance–covariance matrix of $\hat{\beta}$, the lower right quadrant of $\sum_{R}$, and $L$ is a $K \times p$ contrast matrix or vector whose form will distinguish the tests below.

2.3.3. Distinct effects $K$-df test. We test $H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$ versus the alternative that at least one $\beta_k \neq 0$ using $L' = I_K$ in (4), resulting in a $K$-df Wald test. No estimate of overall treatment effect is available for this test. Since the treatment effects are squared in the test statistic, this is not expected to be sensitive to direction; it would just as likely reject for effects in opposite versus the same directions.

Bull [21] also discussed the test presented here, which is asymptotically equivalent to the $K$-df GEE score test proposed by Legler and Ryan [18] in which an independence working correlation was assumed [32]. Legler and Ryan [18] showed that the homogeneous score test of Lefkopoulou and Ryan [17] (and by implication, the common effects test in 2.3.2) was in general more powerful than their distinct effects $K$-df test for alternatives with similar effects, and less powerful when effects are quite different. Note that alternative specifications of $L$ may also be used to generate $p < K$ df tests of subsets of the log-odds ratios, and/or tests incorporating clinical importance weights for the components.

2.3.4. Average treatment effect 1-df test. We test whether the average or weighted average effect among the $K$ outcomes is equal to zero by testing $H_0: L'\hat{\beta} = 0$ with a 1-df Wald test based on (4), with $L = (1/K)I_K$ or other $L = (l_1, \ldots, l_K)$, with all $l_k \geq 0$ and $L'L = 1$.

When using weights $l_k = 1/K$, $L'\hat{\beta}$ estimates the average (log scale) relative treatment effect across the $K$ components, regardless of their underlying incidences or variances. Component frequencies (via their variances) and correlations among components are accounted for only in the denominator of the test statistic. As such, the average effect test is useful for situations in which relative effects are at least as important as absolute effects. Unequal weights summing to 1.0 can be used to reflect a priori differential clinical importance across components. The average effect test thus has an advantage over the collapsed composite, count, and common effect tests, for which direct weighting of treatment effects is not possible (although individual observations can be weighted in the common effect test).

The estimated variance $L'\sum_{\beta}L$ of $L'\hat{\beta}$ is a weighted sum of the variances and covariances of the $K$ treatment effects, with variances of individual betas equal to those of the univariate log-odds ratios, such that

$$ \hat{V}(AE) = \hat{V} \left( \sum_{k=1}^K l_k\hat{\beta}_k \right) = \sum_{k=1}^K l_k^2 \hat{V}(\hat{\beta}_k) + 2 \sum_{k=1}^K \sum_{k' > k} l_k l_{k'} Cov(\hat{\beta}_k, \hat{\beta}_{k'}) $$

(5)

The variance of the average effect estimator thus accounts for within-subject correlation across components (smaller correlation yielding lower variance) and differing variances of $\hat{\beta}_k$ due to differing baselines (variances smaller when proportions closer to 0.5, as seen in formula for $SE_{\hat{\beta}}$ in 2.2.1).

Bull [21] described a ‘global Wald statistic’ and test equivalent to the one described above, and suggested a vector of equal weights or weights inversely proportional to functions of the robust variance and covariance estimates, giving tests analogous to those of O’Brien [19] and Stram et al. [33] as special cases.

2.3.5. Variance–covariance weighted average test. We also constructed a test weighting $\hat{\beta}$ by the inverse of the estimated robust GEE variance–covariance matrix. Letting $W_{K \times K} = \sum_{\beta}^{-1}$ and $L = I_K$, the inverse variance–covariance weighted average of $\hat{\beta}$ is thus $\hat{A}_{-w} = I_K \sum_{\beta}^{-1} \hat{\beta}/I_K \sum_{\beta}^{-1} I_K$, with estimated variance $\hat{V}(A_{-w}) = (L'W)\sum_{\beta}(L'W)'/(L'WL)(L'WL)' = (I_K \sum_{\beta}^{-1} I_K)^{-1}$. The resulting 1-df chi-square test of $(I_K \sum_{\beta}^{-1} I_K)\hat{A}_{-w}^2$ might be expected to have properties similar to the common effect test in Section 2.3.2 because the numerators for both tests give higher weight to effects with lower variance and less-correlated components. In simulations, this weighted average test yielded almost identical power to the common effect test for a wide array of scenarios. Therefore, we do not present simulation results for it in Section 3. However, these results give intuition into the weighting inherent in the common effect estimator and test.

2.3.6. Treatment–outcome interaction. We test the consistency of treatment effects across components with a $(K - 1)$ df Wald test of $H_0: \beta_1 = \beta_2 = \cdots = \beta_K$, based on (4) and specifying $L'_{(K-1) \times K}$ to be a matrix of $K - 1$ contrasts among the $K$ log-odds ratios, i.e. $L' = [I_{K-1}, -I_{K-1}]$. 

A significant interaction test would imply that the effects differed across components. However, the test is sensitive to opposite effects as well as same-direction effects, its power in clinical studies is often low, and equal effects would be most cases not even be expected. Nevertheless, this test may help interpret results of overall treatment effect or individual effect tests in certain situations.

2.4. Simulation methods to assess relative power of tests

We simulated multivariate binary data using the method of Oman and Zucker [34] and Oman [35], which ensures that all pairwise correlations meet the constraints imposed by binary responses [36, 37]. For binary outcomes the joint probabilities \( \pi_{kk'} = P(Y_k = 1, Y_{k'} = 1) \) for outcomes \( k \) and \( k' \) must satisfy \( \max(0, \pi_k + \pi_{k'} - 1) \leq \pi_{kk'} \leq \min(\pi_k, \pi_{k'}) \). This leads to the Frechet bounds [38] which constrain the pairwise correlations \( r_{kk'} \) as:

\[
-\min\{f(\pi_k, 1-\pi_{k'}), f(1-\pi_k, \pi_{k'})\} \leq r_{kk'} \leq \min\{f(\pi_k, \pi_{k'}), f(\pi_{k'}, \pi_k)\}
\]

where \( f(u, v) = [u(1-v)/(v(1-u))]^{1/2} \).

In our simulations we vary within-subject correlation from 0.0 to 0.9, depending on the constraints for the particular scenario. Simulation correlation structures are either exchangeable \( (r_{kk'} = \rho) \) or AR(1) \( (r_{kk'} = \rho^{k-k'}) \). We use the AR(1) structure as a tool to create data with differing correlations across pairs of components. Data were analyzed assuming exchangeable or unstructured working correlation matrices. We also vary effect size and the range and consistency of baseline incidences.

Results for the assessed scenarios are presented in Section 3. Section 3.1: size of tests; Section 3.2: power when consistent odds ratios—including assessment of effects of baseline incidence, magnitude of the correlation among components, choice of correlation structure, and magnitude of treatment effect; Section 3.3: power when some components unaffected—varying whether larger or smaller baseline affected; Section 3.4: power when inconsistent non-zero odds ratios; Section 3.5: power when opposite effects. Scenarios in Sections 3.1–3.4 include both equal and unequal baseline incidences. Chosen sample size depends on effect size. All methods are compared on each simulated data set.

Simulations include 10 000 runs for assessment of size, and 5000 otherwise. Thus, for assessment of size, simulation standard error for reported rejection percents is approximately 0.0022 at the employed 0.05 significance level. For other scenarios the error is 0.007, 0.0065, and 0.004, respectively, for rejection fractions of 0.50, 0.70, and 0.90.

3. Results

In general, the relative powers of the assessed tests depended on the magnitudes and similarities of their baseline incidences, consistency of treatment effects, and correlations among components.

3.1. Size

Size assessment (proportion rejections when null hypothesis is true) included simulation scenarios with either all baseline incidences the same (0.05 or 0.10) or varied (0.10 to 0.25 by 0.05), exchangeable or AR(1) correlation structures (with range of \( \rho \) depending on Frechet bounds), from 2 to 6 components, and \( N \) of 100 or 300 per group.

All tests had adequate size, between 4 and 6 per cent for alpha of 5 per cent. For scenarios with \( K = 4 \), \( N = 100 \) per group, exchangeable correlations, and either equal 0.10 incidences (Figure 1(a)) or incidences varying from 0.10 to 0.25 (Figure 1(b)), sizes of all tests were well controlled for all correlations assessed. Unequal correlations among components did not noticeably change these test results (not shown). Size of the treatment–outcome interaction test, also within 4–6 per cent, is omitted from Figure 1 to ease comparisons among the remaining tests.

3.2. Consistent odds ratios

3.2.1. Equal baselines. With consistent treatment effects and equal baselines across components, the relative powers of the tests depended on the combination of component incidences and correlations.

Effect of baseline incidence. In Figure 2, power is plotted versus common baseline incidence, with consistent odds ratio of 0.74 and \( \rho \) of 0.10 (Figure 2(a)) or 0.50 (Figure 2(b)). Except for the interaction test, power increases for all tests from a zero baseline, then decreases after baseline of 0.30 for the collapsed composite and about 0.50, less markedly, for other tests. Power is similar for the average effect, common effect and count tests at low correlation (Figure 2(a)). These three are considerably more efficient than the collapsed composite when baseline incidences higher than 0.10 or 0.30 are combined with exchangeable \( \rho \) of 0.10 (Figure 2(a)) or 0.50 (Figure 2(b)), respectively. At \( \rho = 0.50 \) the count and collapsed composite are best among tests at very low baselines, while the count test is best at very high baselines.
E. J. MASCHA AND P. B. IMREY

Figure 1. Size of tests as function of within-subject correlation. $K = 4$ outcomes, $N = 100$ for treatment and control groups, 10,000 simulations. Left panel (a): same frequency of 0.10 for each outcome in each group. Right panel (b): differing baseline frequencies across components (same for each group): 0.10, 0.15, 0.20, 0.25.

Figure 2. Consistent effects (odds ratio $= 0.74$). Power as function of common baseline incidence. $K = 4$ outcomes, $N = 500$ per group, 5000 simulations. Symbols: $A =$ Average effect, $CM =$ Common effect, $CT =$ Count. Left panel (a): $\rho = 0.10$. Right panel (b): $\rho = 0.50$. (Figure 2(b)). The $K$-df and minimum $P$-value test are both weaker than other tests, except the collapsed composite, at all baseline incidences. The treatment–outcome interaction test has size close to the nominal level (5 per cent), as expected with consistent effects.

Effect of correlations. In general, standard errors increased and thus power decreased with increasing correlation among components (Figure 3). Magnitude of correlation also affected the relative powers of the tests. With low baseline incidence (0.10) and consistent odds ratio (0.67), the common and average effect tests had similar powers, and were slightly more efficient than the collapsed composite and count tests at very low correlations, but slightly worse at correlations above 0.20 (Figure 3(a)). With even lower baselines (0.05) combined with moderate-to-high correlations, the collapsed composite and count tests were markedly more efficient than the common and average effect tests, but no different when $\rho$ approached zero (not shown). At high baseline incidence (0.50), the common, average, and count tests had almost identical power, but the collapsed composite was substantially lower, although similar at very high correlation (not shown).

The minimum $P$-value and $K$-df tests were weaker than other tests, since they favor inconsistent effect alternatives. The $K$-df test becomes relatively less powerful than the minimum $P$-value test between $\rho = 0.0$ and $\rho = 0.5$ (Figure 2(a) versus 2(b), and Figure 3); both are weaker than other tests, except the interaction test, at all correlations. The relative advantage of the common effect test over the $K$-df test for consistent effects agrees with Legler et al. [18].

Varying treatment effect. The magnitude of a consistent treatment effect did not substantially change the relative efficiencies of the tests, given a common baseline incidence (not shown).

Correlation structure. An underlying AR(1) structure gave slightly more power than an exchangeable structure when the same value of $\rho$ was used for each method, regardless of whether the data were analyzed using exchangeable or unstructured correlation (not shown). This is intuitive since more power is expected when the average correlation among components is decreased [28].
3.2.2. Unequal baselines. Patterns similar to those for equal baselines were observed when baselines differed but effects remained consistent. An exception is that the common, count and collapsed composite tests (herein called ‘competitor tests’) gain power over the average effect test as the difference in baseline frequency increases, particularly when some baselines are very small. This is because the variance of the average effect test gives equal weight to variances of effects for each component, thus reflecting the increased variance of log-odds ratios as incidence declines, but competitor test variances for the single treatment effect give more weight to components with higher incidence.

For example, with 4 baselines ranging from 0.08 to 0.20, and a true log-odds ratio of $-0.33$ for each component, the average effect and common effect tests had the same mean estimated log-odds ratio across simulations, but standard errors for the average effect test were about 6 per cent higher and power thus 4 to 8 per cent lower (Figure 3(b)). With baseline frequencies closer to zero, ranging from 0.02 to 0.12, and consistent log-odds ratio of $-0.30$, the average effect test was 15–20 per cent less powerful than the common effect test, with mean standard errors 19 per cent higher and log-odds ratios 2 per cent higher (not shown).

Correlation and baseline incidence. Scenarios were simulated with 4 components and baselines of 0.08–0.20 with consistent relative risk of 0.25 and varying magnitudes of correlation among the components. With consistent effects, neither the actual nor relative powers depended heavily on whether the most correlated pairs of components were those with higher or lower baseline incidence (not shown).

3.3. Some unaffected components

Scenarios were simulated in which only some components were affected by treatment. Both simulated and analyzed correlation structures were exchangeable.

3.3.1. Equal baselines. The average effect test was more powerful than the common, count and collapsed tests, respectively, when baseline incidences were the same and some components reduced, but relatively less powerful when some components increased. The $K$-df and minimum $P$-value tests were typically the most powerful when most components were unaffected, especially at high correlations, given their sensitivity to inconsistent effects.

When 2 of 4 components were affected, each at a 25 per cent relative reduction from a 0.20 baseline (log-odds ratio of $-0.35$), the average effect test was more powerful than the competitor tests at all correlations, although consistently less powerful than the $K$-df and minimum $P$-value tests (Figure 4(a)). At $\rho=0$ and $\rho=0.50$, respectively, the average effect test was 6 and 29 per cent more powerful than the common effect test due to larger mean treatment effect estimates (7–17 per cent higher) but the same mean standard errors. With only 1 of 4 components affected, the advantage of the average effect test over the competitor tests was much larger (not shown).

With some components reduced, regardless of the common baseline incidence, the average test is more powerful than the competitor tests, especially with moderate to high correlations. Whereas the average effect test averages the log-odds ratios across components, the competitor tests give less weight to the effect(s) for the reduced components because of the relatively lower pooled incidence for those components.

In contrast, when some baselines are increased by treatment and the others unaffected, the relative (not actual) power of the average effect test is reduced. In Figure 4(b) the baseline incidence of 0.20 is increased to 0.26 for 2 of 4
Figure 4. Some unaffected components, equal baselines. Power as function of within-subject correlation. $K=4$ outcomes, $N=800$ per group, 5000 simulations. Left panel (a): 2 of 4 components reduced, log-odds ratio $= -0.35$. Incidences are 0.15, 0.15, 0.20, 0.20 for treatment and 0.20 for all four outcomes for control group. Right panel (b): 2 of 4 increased, log-odds ratio $= +0.35$. Incidences are 0.26, 0.26, 0.20, 0.20 for treatment and 0.20 for all 4 outcomes for control group.

Figure 5. Some unaffected components, unequal baselines. Power as function of within-subject correlation. $K=4$ outcomes, $N=1000$ per group, 5000 simulations. Symbols: $K =$ df distinct effect, $M =$ Minimum $P$-value, $I =$ Interaction. Left panel (a): smaller baseline reduced 50 per cent. Incidences are 0.05, 0.10, 0.20, 0.20 for treatment and 0.10, 0.10, 0.20 and 0.20 for control for the 4 respective outcomes. Right panel (b): larger baseline reduced 50 per cent. Incidences are 0.10, 0.10, 0.10, 0.20 for treatment and 0.10, 0.10, 0.20 and 0.20 for control for the 4 respective outcomes.

components, corresponding to a log-odds ratio of 0.35 for each, the opposite of the effects in Figure 4(a). While the power of the average effect test does not change, the competitor tests gain power (to different degrees), since the affected components have higher pooled incidence than the unaffected ones.

The consistency of the actual power of the average effect test under varying incidences but the same relative effects is its main distinction from, and potential advantage over, the collapsed composite, count and common effect tests.

3.3.2. Unequal baselines. Simulations included scenarios with some unaffected components and differing baselines. Power of the average effect test was consistent regardless of whether a component with larger or smaller baseline was affected, given the same relative effect, whereas power for other methods depended on which baseline was affected. The $K$-df and minimum $P$-value tests were often most powerful.

When the smaller of several baseline incidences is affected, the average effect test is more powerful than the competitor tests. An example is Figure 5(a) with four components, baselines of 0.2, 0.2, 0.1, 0.1, and where the only effect is a 50 per cent reduction in the smaller baseline. For the average effect estimator the 50 per cent reduction in the first component is weighted the same as the three null effects; for the competitors, the null effects receive more weight due to higher incidences. The $K$-df and minimum $P$-value tests are most powerful.

In contrast, when the largest of several baselines is affected, the average effect test has among the lowest powers among the tests. Figure 5(b) shows 4 components with the same baselines as Figure 5(a), but where the only effect is a
3.4. Inconsistent odds ratios—quantitative interactions

Scenarios consisting of differing relative effects in the same direction, all non-zero, were simulated with either equal or unequal baseline incidences. Patterns are quite similar to those when only some components are unaffected (Section 3.3). Regardless of whether the treatment effect represents an increase or decrease from baseline, the average effect test tends to be more (less) powerful than the competitor tests when the relatively more affected components have smaller (larger) pooled incidence. The actual power of the average effect test depends little on whether a large or small baseline is affected.

3.4.1. Equal baselines. Scenarios had 2 components, with a common control group incidence of either 0.15 or 0.20. Treatment effects for the 0.15 baselines were relative increases or decreases of 25 and 50 per cent, and for the 0.20 baselines were log-odds ratios (positive and negative) of 0.25 and either 0.75 or 1.25. The average test was more powerful than the competitors with both components decreased, but less powerful with both increased (not shown).

3.4.2. Unequal baselines. Two scenarios with the same unequal baselines (0.20 and 0.30) are presented, one with the smaller baseline more affected (Figure 6(a)) and another with the larger baseline more affected (Figure 6(b)). Log-odds ratios (odds ratios) are $-0.25 (0.78)$ and $-0.75 (0.47)$ in each scenarios. Sample size is 200 per group, with exchangeable correlation for both simulation and analysis.

When the component with the smaller pooled incidence has the bigger effect, the average test is slightly more powerful than the $K$-df test and minimum $P$-value test, and considerably more powerful than the common effect, count and collapsed composite tests (Figure 6(a)). However, when the component with the larger baseline and pooled incidence has the bigger effect, the $K$-df and minimum $P$-value tests and then the competitor tests are more powerful than the average effect test (Figure 6(b)). Lower power is observed for the average effect test compared with the common effect test due to slightly smaller observed betas and slightly larger standard errors.

3.5. Opposite effects—qualitative interactions

The average effect, common effect, collapsed composite, and count tests are designed for alternatives with effects in the same direction, although each has good power for inconsistent effects. In scenarios with both positively and negatively affected components, these tests would be expected to reject the null hypothesis less frequently, with perhaps the most appropriate analysis being to assess components individually. We assessed relative powers of the tests against equal but
opposite log-odds ratios of ± 0.36 (odds ratios of 0.70 and 1.43). Simulation scenarios include either 2 or 4 components, and equal baseline incidences of 0.10, 0.15, or 0.20.

When effects are opposite on the log-odds scale and baselines the same, the average effect test has power near 5 per cent, as expected since the underlying effects average to zero (Figure 7). The collapsed composite, then the count and common effect tests rejected more often than the average effect test, and increasingly so with higher correlation. The K-df and minimum P-value tests have highest power when effects are strong but opposite (95 to 100 per cent in Figure 7) since neither is sensitive to the direction of effects. Although they have strong power for quantitatively ‘friendly’ interactions (effects in same direction), their strong power for qualitative interactions makes them less attractive than the other tests when the goal of the test is to specifically target an overall effect (as in the example of Section 4.1 below).

4. Applications

4.1. NINDS t-PA stroke trial

Investigators in the NINDS t-PA stroke multi-center randomized controlled trial [39] assessed the effect of tissue plasminogen activator (t-PA) versus placebo on neurological function after ischemic stroke on N = 333 patients. A vector of four dichotomized (success/fail) neurological assessments, including the dichotomized NIH stroke scale, Glasgow Outcome Scale (GOS), Barthel Index, and Modified Rankin Scale (MRS), was chosen as the primary outcome for Part II of the trial because none of the components by themselves adequately describe post-stroke neurological function. In planning the trial, efficacy was defined as a consistent and persuasive t-PA benefit across the 4 components. The common effect GEE method was used to assess the treatment effect on the primary outcome vector at 90 days post randomization [31]. Individual analyses were planned only if the global test was significant at the 0.05 level, with no adjustment for multiple comparisons. We re-analyzed the t-PA primary outcome data using the assessed methods. The data set is in the public domain and distributed through the National Technical Information Service (www.ntis.gov).

Global methods. Treatment effects were quite consistent and the control group incidences of the outcome ranged from 20 to 38 per cent across the components (Table II). A GEE model with unstructured correlation was used to perform the common effect, average effect, K-df distinct effect and treatment–outcome interaction tests; pairwise correlations among the components ranged from 0.55 to 0.89 (see full working correlation matrix in Table II). Estimates and P-values for the average effect and common effect GEE methods were almost identical (Table II). The 4-df distinct effects test had similar chi-square value to the other tests, but was not significant. There was no evidence of treatment–outcome interaction (P = 0.96) which, coupled with the significant average treatment effect result and similar observed effects on each component, is compatible with an underlying consistent treatment effect.

This example is consistent with our simulation results where the average effect test had similar power to the common effect test when the treatment effect was consistent across components and baseline incidences were somewhat different.
Table II. NINDS t-PA study randomized trial (Part II, N=333): comparing methods.

<table>
<thead>
<tr>
<th>Individual components</th>
<th>$P_t / P_c$</th>
<th>LOR (SE)$^\dagger$</th>
<th>Odds ratio$^\ddagger$ (95 per cent CI)</th>
<th>Chi-square</th>
<th>$P$-value</th>
<th>Adjusted $P$-values$^\ddagger$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. MRS=0</td>
<td>0.39/0.26</td>
<td>0.58 (0.24)$^\dagger$</td>
<td>1.79 (1.12, 2.85)</td>
<td>6.0 (1)</td>
<td>0.014$^\ddagger$</td>
<td>0.039</td>
</tr>
<tr>
<td>2. Barthel Index $\geq$ 95</td>
<td>0.50/0.38</td>
<td>0.51 (0.22)$^\dagger$</td>
<td>1.66 (1.07, 2.57)</td>
<td>5.2 (1)</td>
<td>0.023$^\ddagger$</td>
<td>0.043</td>
</tr>
<tr>
<td>3. GOS=1</td>
<td>0.44/0.32</td>
<td>0.54 (0.23)$^\dagger$</td>
<td>1.71 (1.09, 2.68)</td>
<td>5.5 (1)</td>
<td>0.019$^\ddagger$</td>
<td>0.043</td>
</tr>
<tr>
<td>4. NIHSS $=0$ or 1</td>
<td>0.31/0.20</td>
<td>0.58 (0.26)$^\dagger$</td>
<td>1.79 (1.08, 2.96)</td>
<td>5.2 (1)</td>
<td>0.023$^\ddagger$</td>
<td>0.043</td>
</tr>
</tbody>
</table>

Global tests (see details in Table I)

| Collapsed composite     | —          | 0.52 (0.22)         | 1.69 (1.09, 2.60)                     | 5.6 (1)    | 0.018     | —                             |
| Count                  | —          | —                   | —                                      | 7.2 (1)    | 0.007     | —                             |
| Common effect GEE$^\ast$ | —          | 0.54 (0.21)         | 1.71 (1.15, 2.56)                     | 6.9 (1)    | 0.009     | —                             |
| Distinct effects       | —          | —                   | —                                      | 7.2 (4)    | 0.128     | —                             |
| K-df GEE$^\ast$        | —          | 0.55 (0.21)         | 1.74 (1.15, 2.62)                     | 7.0 (1)    | 0.008     | —                             |
| Average effect GEE$^\ast$ | 0.54 (0.21) | 1.71 (1.14, 2.56) | 6.9 (1)                              | 0.009     | —         | —                             |
| Var-Cov weighted average GEE$^\ast,\parallel$ | 0.54 (0.21) | 1.71 (1.14, 2.56) | 6.9 (1)                              | 0.009     | —         | —                             |
| Treatment-outcome interaction GEE$^\parallel$ | —          | —                   | —                                      | 0.3 (3)    | 0.964     | —                             |

MRS= Modified Rankin Scale; GOS = Glasgow Outcome Scale; NIHSS = National Institutes of Health Stroke Scale. Individual component outcomes are binary indicators with the given criterion for having the event.

$^*$Proportion of patients with observed success in treatment and control groups.

$^\dagger$LOR: log odds ratio of t-PA versus placebo; SE: standard error.

$^\ddagger$Odds of successful outcome in t-PA group versus placebo.

$^\ast$Chi-square test with resampling step-down multiple comparison procedure (50,000 resamples).

$^\parallel$Univariate logistic regression.

$^\parallel$Using the test in 2.3.5 with estimated inverse robust covariance matrix as weight matrix.

$^\ast\ast$Estimated unstructured GEE working correlation matrix for distinct effects models (nearly identical results obtained for common effect GEE model).

from each other but not very close to zero (Figure 3(b)). The count test is slightly more significant than either the common effect or average effect tests, while the K-df and minimum $P$-value tests are less significant, all consistent with the results in Figure 3(b) at a within subject correlation of 0.50.

Individual comparisons. We also analyzed the components individually to assess whether the treatment effect was consistent. Using a 0.05 criterion for each comparison, as planned in the trial, all 4 components are significant (Table II, individual logistic regressions). However, with standard Holm–Bonferroni multiple comparison correction, none are significant. Using the bootstrap stepdown procedure as in our simulations to simultaneously adjust for multiple comparisons and correlation among components, all components are significant at the overall 0.05 level (Table II, last column).

4.2. Crystalloid versus Colloid fluid management randomized trial

In the ongoing randomized controlled trial Crystalloids versus Colloids During Surgery (ClinicalTrials.gov identifier: NCT00517127), researchers at the 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 collapsed composite of binary complication events in six organ systems, four of which are collapsed composites themselves (Table III). However, because the intervention could well improve some components while worsening or not affecting others, secondary analyses will analyze the individual components as well. As is sometimes the case, detecting an overall effect is not sufficient—the investigators want to know which complications, if any, are affected. The particular components were chosen because each is considered very serious and likely to be affected by intervention. Some of them, however, such as infection, are expected to occur more frequently than others, such as renal/dialysis or cardiac problems.
Data for this ongoing trial are not available for analysis. However, the expected underlying scenario would be consistent with, for example, 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 III, respectively. It is also plausible that the treatment effects would be inconsistent across components, with no effect on the largest two baselines (GI and infection), for example, and some consistent relative reduction on the rest. In such a scenario, the relative powers of the comparative tests would be ordered as those seen in Figure 5(a), where the average effect test was more powerful than the common effect, count, and collapsed composite, but less powerful than the $K$-df and minimum $P$-value tests. In such situations, where the least frequent components are the most serious, the average effect test can prove advantageous to detect alternatives where most effects are non-zero. On the other hand, if the treatment instead reduced the larger baselines but had no effect on the smaller ones, the average effect test would not change in absolute power, as in Figure 5(b) compared with Figure 5(a), but the comparator tests would gain power to a degree dependent on the size of the treatment effects and size and range of baseline incidences.

**Clinical Importance Weights.** The multivariate methods under consideration allow weighting of components by clinical importance. Suppose, for example, that investigators in the study had agreed a priori that coagulation, cardiac, pulmonary, and renal complications were twice as important as infection and gastrointestinal complications. We would then give double weight to each individual observation on the former outcomes in the covariance matrix for the common effect test, and double weight for the treatment effects on these outcomes in the contrast for the average effect test. Results for the common effect and average effect tests would be expected to be stronger than the unweighted results because the four affected components are emphasized by the reweighting.

Flexibility of the average effect test also permits straightforward component subset analyses by specifying the appropriate contrast. For example, one could use $L = \{0 \ 0 \ 0 \ 0.5 \ 0.5\}$ to isolate the combined treatment effect for the pulmonary and cardiac components.

### 5. Discussion

Composite primary outcomes consisting of several binary events are common in both randomized and non-randomized studies because they are more clinically relevant and/or expected to be more powerful than a single binary outcome. Deciding which outcomes to include in a composite and the best statistical approach to assess treatment effects are challenges for biomedical researchers [6].

Ideally, for a composite to be easily interpretable, its elements should have similar importance, or failing that, at least similar frequencies and effects. Nevertheless, considerable deviations from each guideline are often observed [3, 7, 22, 40, 41]. For studies with varying baseline frequencies, the collapsed composite, count, and GEE common effect tests give more weight to treatment effects on components with higher incidence than to those with lower incidence. In some situations, one may prefer a method which is not easily ‘driven’ by higher frequency components, especially those of lesser clinical importance, but instead focuses on the relative effects. For example, in choosing treatments, individual patients might be more interested in reducing their chance of an infrequent serious event than a somewhat less serious but more common event.

Therefore, for situations in which relative treatment effects are at least as important as absolute effects, we highlight properties of a test of the average log-odds ratio across components. The average effect estimator gives equal weight to observed treatment effects regardless of the frequency of the components, while varying incidences and correlations among components are accounted for in the standard error. The average effect test assesses whether the average relative treatment effect across the components is equal to zero. As such, it generally gains power as the majority of relative effects are further from zero and loses power as the majority is either close to zero or when the effects are in opposite directions. We do not advocate using the average effect test or any of the global tests in isolation. Rather, interpretation should also consider the amount of treatment effect heterogeneity exhibited across the individual components [2].

**Table III. Major complications for crystalloid-colloid trial.**

<table>
<thead>
<tr>
<th>Organ system*</th>
<th>Complication definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Cardiac</td>
<td>Acute heart failure, myocardial infarction, ventricular arrhythmia</td>
</tr>
<tr>
<td>2. Pulmonary</td>
<td>Pulmonary embolism, pulmonary edema, respiratory failure, pneumonia</td>
</tr>
<tr>
<td>3. Renal</td>
<td>Dialysis</td>
</tr>
<tr>
<td>4. Coagulation</td>
<td>Bleeding</td>
</tr>
<tr>
<td>5. Gastrointestinal</td>
<td>Bowel and surgical anastomosis, stricture/obstruction</td>
</tr>
<tr>
<td>6. Infectious</td>
<td>Deep or organ-space surgical site infection, sepsis</td>
</tr>
</tbody>
</table>

*Each system considered as a binary event, such that 1, 2, 5 and 6 are collapsed composites in themselves.
We compared the power of the average effect test to competing methods under various scenarios of baseline proportions, treatment effects, and within-subject correlations. Our simulations verify that when treatment effects differ across components, the power of the average effect test does not depend on the relative incidence(s) of the more or less affected components to the extent of the collapsed composite, count and common effect tests. When baselines differ and effects are inconsistent, treatment effect estimators of these tests are weighted towards components with higher pooled incidence, whereas the average effect is not a function of the pooled incidences. Therefore, the average effect test is relatively more powerful than competitor tests when lower pooled frequency components are more affected than higher frequency components, and relatively less powerful when higher baselines more affected. Even with common baseline incidences and consistent effects, the power of competitor tests depends heavily on whether baselines are increased versus decreased by treatment, since higher pooled incidences increase their power. In contrast, the average effect test focuses on the magnitude of the relative effects themselves.

In general, we found that relative efficiencies of the assessed tests are a complex function of incidences, treatment effects, and component correlations, consistent with the findings of Legler et al. [18], who studied several of these methods. Depending on the situation, any of the methods could be most powerful. For consistent treatment effects and baseline incidences, power increases and then decreases for all tests as baseline increases, but peaks noticeably earlier for the collapsed composite. The collapsed composite and count tests are most powerful only with the combination of consistent effects, low baseline incidence and high correlation. Power typically decreases as correlation increases, although not consistently for the K-df and treatment–outcome interaction tests.

Study design is challenging with a composite of multiple binary events as primary outcome since baseline incidences, treatment effects and correlations are difficult to estimate beforehand. Availability of comprehensive registries may facilitate estimation of historical baseline incidences and correlations among outcomes. Conjecturing the treatment effect is more challenging, and a strength of the average effect test over other methods is that its power depends minimally on which baseline frequencies are most affected. Since such information is usually unavailable, the average effect test could help prevent trials from being under- or over-powered. Although formulae and programs for assessing sample size for GEE-based designs exist [42, 43], our available simulation programs (available upon request from the first author) may assist trial designers in choosing among the specific methods presented here.

Applying differential clinical importance weights to components is intuitive [2] but controversial. Such weights are inherently subjective, and therefore difficult to agree on or generalize across studies [44]. However, sometimes assuming equal weights may be unsatisfying and results difficult to interpret [4]. Where weights are agreed upon, the average effect estimator facilitates the direct weighting of the treatment effect log-odds ratios. Importance weighting is intrinsic to the average effect and other distinct effect methods with less than K-df, since an underlying common effect across outcome components is not being assumed.

In conclusion, we investigated many scenarios while varying key parameters that could affect the relative power of tests for multiple binary outcomes. Our results are intuitive, and where comparable, agree well with previous research. The GEE average effect test is less sensitive to baseline frequencies of components than other tests. As such, it is a potentially attractive option when the goal is to estimate the average relative effect of an intervention or treatment, particularly when the effects are expected to differ across components and may be most manifest in those less common.

Acknowledgements

The authors thank the associate editor and the reviewers for constructive comments and suggestions. These have led to considerable improvements in this paper.

References
