Published in final edited form as:
*Stat Med*. 2011 July 20; 30(16): 1917–1932. doi:10.1002/sim.4262.

# Alternative methods for testing treatment effects on the basis of multiple outcomes: simulation and case study

**Frank B. Yoon**[*,†], **Garrett M. Fitzmaurice**[‡,†], **Stuart R. Lipsitz**[†], **Nicholas J. Horton**[§], **Nan M. Laird**[‡], and **Sharon-Lise T. Normand**[†,‡]

[†]Harvard Medical School, Boston, MA

[‡]Harvard School of Public Health, Boston, MA

[§]Smith College, Northampton, MA

## Abstract

In clinical trials multiple outcomes are often used to assess treatment interventions. This paper presents an evaluation of likelihood-based methods for jointly testing treatment effects in clinical trials with multiple continuous outcomes. Specifically, we compare the power of joint tests of treatment effects obtained from joint models for the multiple outcomes with univariate tests based on modelling the outcomes separately. We also consider the power and bias of tests when data are missing, a common feature of many trials, especially in psychiatry. Our results suggest that joint tests capitalize on the correlation of multiple outcomes and are more powerful than standard univariate methods, especially when outcomes are missing completely at random. When outcomes are missing at random, test procedures based on correctly specified joint models are unbiased, while standard univariate procedures are not. Results of a simulation study are reported, and the methods are illustrated in an example from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) for schizophrenia.

### Keywords

joint tests; multiple outcomes; power; missing data; psychiatry

## 1 Analysis of multiple outcomes: review of methods and an example from psychiatry

### 1.1 Approaches to analysis of multiple outcomes

An increasingly common feature of many modern clinical studies is the inclusion of multiple outcomes that characterize the treatment effect, for example, by efficacy and safety measures. The desire to include more than one outcome arises for several reasons. Disease complexity may be such that a single outcome may not adequately characterize the disease, there may be lack of consensus on the most important clinical outcome, or there may be a desire to demonstrate clinical effectiveness on several outcomes. The most common approach to analysis of multiple outcomes involves separate testing of the outcomes with adjustment for multiplicity, such as Bonferroni adjustments, or combining the outcomes into a composite measure and performing a single univariate test. A less commonly used approach, especially in mental health research, is to jointly test the outcomes to evaluate the treatment effect.

[*]Correspondence to: yoon@hcp.med.harvard.edu, Dept of Health Care Policy, 180 Longwood Ave, Boston, MA 02115.

Joint testing of multiple outcomes in clinical studies has been discussed in the literature for some time. Previous work has employed the use of linear mixed models to specify the joint distribution of multiple continuous outcomes. Sammel, Lin, and Ryan [1] propose a multivariate linear mixed model which generalizes a latent variable approach by assuming a flexible correlation structure among the outcomes. Lin, Ryan, Sammel, et al. [2] propose a scaled linear mixed model to account for effect sizes that may differ across outcomes, which they estimate by maximum likelihood and the "working parameter" method. Roy, Lin, and Ryan [3] present scaled marginal models to test for a common effect on outcomes that are scaled by the marginal variances of the outcomes, rather than variances conditional on the random effects; their estimating equation approach is less efficient than maximum likelihood, but is robust to misspecification of the correlation structure. In recent work by Thurston, Ruppert, and Davidson, Bayesian methods have been employed to account for domain-specific effects on multiple outcomes measured on different scales [4].

In many settings multiple outcomes are correlated in the same direction. For example, in a study of antipsychotic medication for schizophrenics, a primary outcome is the Positive and Negative Syndrome Scale (PANSS), and secondary outcomes characterize side effects such as changes in weight and blood sugar levels; lower values indicate healthier patients, so we could expect the outcomes to be positively correlated. In this paper we focus on settings in which correlations of the multiple outcomes lie in the same direction. Joint testing of multiple outcomes capitalizes on the correlations among the outcomes and thereby has the advantage of yielding more powerful tests of the treatment effect. In contrast, standard univariate procedures, such as multiple testing with Bonferroni adjustments, tend to be overly conservative when the correlations among the outcomes are high; further limitations are discussed in §3.1. The use of a composite outcome also has important limitations, such as the need to rescale component outcomes, extreme sensitivity to missing data, and an inability to assess the treatment effect on individual components [5].

Despite the availability of methods and software to adopt a joint testing approach, the majority of practitioners continue to use a separate testing strategy. One reason for lack of adoption may be due, in part, to the lack of practical guidance regarding how to implement joint testing strategies. In this article we provide a practical approach to joint testing of continuous multiple outcomes capitalizing on recent advances in joint estimation. We compare power and type I error characteristics of the joint tests with conventional approaches, both in the settings of complete and incomplete data. Finally, to illustrate the use of joint tests we apply the practical approach to an example from a clinical trial in psychiatry.

The outline of this paper is as follows. In §1.2 we briefly introduce the example of the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) for schizophrenia. In §2 we propose joint models for multiple outcomes which are tested by the procedures discussed in §3. In §4 we describe our simulation study to assess the power and bias characteristics of the univariate and joint testing procedures, with complete and missing data, and present results from the study. In §5 we revisit the CATIE study and implement these testing procedures to data from Phase 1 in order to assess the safety and efficacy of atypical antipsychotics in the treatment of schizophrenia.

### 1.2 Example: the Clinical Antipsychotic Trials of Intervention Effectiveness for Schizophrenia

Antipsychotic drugs are widely used in the treatment of schizophrenia. First-generation ("conventional" or "typical") antipsychotics are known to be highly effective against psychotic symptoms, but have high rates of neurologic side effects; these side effects contribute significantly to non-compliance, which leads to relapse and rehospitalization. A

new class of second-generation ("atypical") antipsychotic drugs were promised to have enhanced safety with similar or better efficacy. However, the purported advantages in safety have been countered by other adverse effects, such as weight gain and changes in glucose and lipid metabolism. Of particular concern is that schizophrenia patients show high prevalence of metabolic syndrome, which is associated with increased risk for diabetes mellitus and coronary heart disease. Despite these potential risks atypical antipsychotics have been widely adopted in the treatment of schizophrenia. The Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project funded by the National Institute of Mental Health (NIMH) involved a clinical trial to study the effectiveness of atypical antipsychotics for the treatment of schizophrenia; this trial included numerous measures of safety and efficacy [6].

For our analysis in §5 we focus on the atypical antipsychotic quetiapine, one of the four used in CATIE, and the conventional antipsychotic perphenazine. Patients with tardive dyskinesia at baseline were ineligible for randomization to perphenazine; we do not consider this, although this baseline covariate could be included in the models we present in §2. The results from Lieberman et al. showed no difference between quetiapine and perphenazine based on the primary outcome, time to treatment discontinuation; however, in an analysis of the risk for metabolic syndrome in CATIE patients, Meyer, Davis, Goff et al. confirmed differential metabolic effects among the antipsychotic medications [7]. They showed, for example, that patients who were treated with quetiapine had the largest mean increase in waist circumference at the 3 month follow-up visit (0.7 inches, SE=0.2), while patients treated with perphanize had the smallest (−0.4 inches, SE=0.2). In our application in §5 we aim to evaluate *both* the efficacy and safety of the two drugs and consider five relevant and important outcomes: the Positive and Negative Syndrome Scale (PANSS), which measures the overall severity of the disease; and four that characterize the metabolic side effects, namely patient weight, levels of glucose, high-density lipoprotein (HDL) cholesterol, and triglycerides. Meyer et al. included these metabolic outcomes in their analysis and reported that clinicians should monitor these closely in treating schizophrenic patients with antipsychotic medications.

In our analysis we use the mean changes in these outcomes over a 3-month period to evaluate the effect of perphenazine versus quetiapine; this timeframe is similar to the 3-month period used in Meyer et al. A summary of these mean changes for the five outcomes is shown in Table 1. Perphenazine and quetiapine both reduced the severity of disease through the PANSS score; however, patients who took quetiapine exhibited more weight gain and increased glucose and triglyceride levels. The analysis in Lieberman et al. focused primarily on time to discontinuation of treatment; the secondary outcomes, such as metabolic changes, were also presented, but no correction for multiple testing on the outcomes was performed [8]. Here, we propose to jointly test both treatment efficacy and safety through the joint models and tests discussed in §2 and §3.

## 2 Models and estimation

### 2.1 Overview

We focus mainly on the characteristics of joint tests for treatment effects on the basis of multiple outcomes; specifically, we compare the power of tests of treatment effects obtained from joint models for the outcomes with tests based on modelling the outcomes separately. In order to formulate joint tests for multiple outcomes, we must first specify models for the outcomes which characterize their joint distributions. It is expected that tests based on joint models will capitalize on the correlation among the outcomes in order to provide more powerful tests of treatment effects; when all the outcomes are considered to be

manifestations of a common underlying treatment effect, a combined joint analysis has the potential to increase statistical power.

We consider multivariate models to characterize the joint distribution of $K$ outcomes in a clinical study of a single treatment versus comparator. Denote by $Y_i^{(k)}$ the $k$th outcome measured for the $i$th subject, where $k = 1, \ldots, K$, $i = 1, \ldots, 2n$, and $n$ represents the number of subjects in each of the treated and comparison groups. We denote by $\beta^{(k)}$ the treatment effect on the $k$th outcome; $\widehat{\beta}^{(k)}$ is estimated from the observed data. In the univariate setting $\widehat{\beta}^{(k)}$ is a difference in sample means of the $k$th outcome between the treated and comparison groups; in the joint setting, $\widehat{\beta}^{(k)}$ can be estimated from a model for the joint distribution of the multiple outcomes, for example, by maximum likelihood.

### 2.2 Linear mixed effects models for multivariate normal outcomes

We consider an example of multiple outcomes that are generated from linear mixed effects models (LMM) with a subject-specific common random effect [9]. The random effect induces a correlation among the outcomes of interest; for example, subjects with higher values of the random effect will have higher values of his or her outcomes. Throughout this section treatment assignment is assumed to be random so that it is sufficient to model the outcomes and treatment assignment only; in an extension of these LMM models, additional covariates may be included in order to control for possible confounding, particularly in the non-randomized setting.

For the $i$th subject, with treatment assignment $X_i = 1$ if treated and $X_i = 0$ for the comparison group, the $k$th outcome $Y_i^{(k)}$, is modelled as:

$$Y_i^{(k)} = \alpha^{(k)} + \beta^{(k)} X_i + \gamma^{(k)} b_i + e_i^{(k)} \tag{1}$$

where $b_i \sim N\left(0, \sigma_b^2\right)$ is the subject-specific random effect, and the $e_i^{(k)} \sim N\left(0, \sigma_e^2\right)$ are independent noise components. We place the restriction that $\gamma^{(K)} = 1$ so that the model is identifiable. The marginal variance of $Y_i^{(k)}$ is $\left(\gamma^{(k)}\sigma_b\right)^2 + \sigma_e^2$, and the covariance between $Y_i^{(k)}$ and $Y_i^{(j)}$, $k \neq j$, is $\gamma^{(k)}\gamma^{(j)}\sigma_b^2$; thus, the correlation between them is

$$Corr\left(Y_i^{(k)}, Y_i^{(j)}\right) = \frac{\gamma^{(k)}\gamma^{(j)}\sigma_b^2}{\sqrt{\left(\gamma^{(k)}\sigma_b\right)^2 + \sigma_e^2}\sqrt{\left(\gamma^{(j)}\sigma_b\right)^2 + \sigma_e^2}}$$

The scale parameters $\gamma = (\gamma^{(1)}, \ldots, \gamma^{(K)})'$ characterize the variance-covariance structure of the multiple outcomes; of particular interest in our paper are positive pairwise correlations, which we explore in the simulation study of §4. If the multiple outcomes for the $i$th subject $\left(Y_i^{(1)}, \ldots, Y_i^{(K)}\right)'$ are assumed to be conditionally independent given $b_i$, then likelihood-based inferences about the fixed effects (and random effects variance component) can be based on the marginal likelihood obtained by integrating over the distribution of $b_i$. Although, in general, this is a non-standard LMM because of the inclusion of the scale parameters $\gamma$, the model can be fit, with relatively minor modifications, using existing statistical software for LMMs (for example, the `nlme' package in R or PROC NLMIXED in SAS).

Finally, we note that model (1) yields a parsimonious model for the variance-covariance among the K outcomes that is characterized by K+1 parameters instead of $\frac{1}{2}K(K+1)$. This model places restrictions on the patterns of correlations that make it more suitable for the analysis of outcomes that are correlated in the same direction. In settings where these restrictions might not hold, we note that a more general model for the variance-covariance among the K outcomes can be obtained by allowing an unstructured covariance for the errors instead of inducing correlation via the inclusion of random effects. The latter model can be fit using standard statistical software for general linear models for correlated data (e.g., PROC MIXED in SAS); in §5 we illustrate the applications of this model and model (1) to the data from the CATIE trial.

## 3 Testing multiple outcomes

### 3.1 Univariate testing

On the *K* outcomes, we say that a treatment is clinically effective if clinically meaningful treatment effects are observed and if statistical significance is demonstrated at a pre-specified significance level α for one or more of the outcomes, while controlling the overall type I error rate at, say, α = 0.05. An important consideration in testing multiple outcomes is the degree to which the outcomes are correlated. In order to facilitate and focus the discussion on the relationship between correlation and testing characteristics, we consider the equi-correlated case in which all pairs of outcomes $(Y^{(k_1)}, Y^{(k_2)})'$, $k_1 \neq k_2$ have constant positive correlation ρ, by setting $\gamma^{(k)} = 1\ \forall k$, in the LMM (1); this is the setting we explore in the simulation study in §4. If the outcomes are small in number and weakly correlated ($0 \leq \rho \leq 0.1$) then inference can be safely made using a univariate adjustment method, such as the standard Bonferroni procedure which partitions α evenly among the *K* tests [10]. In the standard Bonferroni procedure, each of the hypotheses $H_{0k} : \beta^{(k)} = 0$ is tested with significance level $\frac{\alpha}{K}$; if the treatment is shown to have a significant effect on any of the *K* outcomes, then it is considered to be clinically effective. The standard Bonferroni procedure in this setting of the LMM (1) is based on the univariate *t*-test for each of the mean differences in outcomes between the treated and comparison groups, that is, when the treatment effect $\beta^{(k)}$ on the *k*th outcome is estimated separately from observations of $Y^{(k)}$, rather than by full specification of the likelihood for all the outcomes by the LMM (1).

Stepwise procedures have also been developed that are more powerful than the standard Bonferroni method. The Holm "step-down" procedure sequentially tests ranked p-values as does the Hochberg "step-up" procedure [11, 12, 13]. Table 2 shows some situations for *K* = 3 outcomes in which the Bonferroni procedure and the Holm and Hochberg variants reach different conclusions. When a treatment is defined as clinically effective if at least one outcome is improved in the treatment group, then Holm's procedure is identical to the standard Bonferroni procedure in that one will reject the global null hypothesis if the other rejects. Specifically, at the first step of the Holm's procedure, the smallest p-value from the *K* tests is compared to level $\frac{\alpha}{K}$, and if the result is significant, no further tests are conducted; the standard Bonferroni procedure follows the same algorithm under our definition of a clinically effective treatment. The standard Bonferroni and Holm's procedures have the advantage of not requiring any assumption of the joint distribution of the test statistics. In general, if the Bonferroni procedure rejects a hypothesis at level α, then both the Holm and Hochberg procedures will also reject. However, the converse is not always true for the Hochberg adjustment, which requires additional assumptions on the distribution of the test statistics [13]. (In the setting of our simulation study in §4, the Hochberg procedure was

almost identical to the standard Bonferroni procedure in terms of statistical power, so we removed these variants of the Bonferroni adjustment from our discussion.)

Sankoh et al. demonstrated that with uniform treatment effects, separate testing using Bonferroni-based adjustments provides adequate type I error rate and yields high power when $\rho \leq 0.1$, reasonable power for $0.2 \leq \rho \leq 0.6$, and deflated type I error rates and loss of power when the outcomes are highly correlated, $\rho > 0.6$ [10]. In summary, separate testing of the constituent outcomes, with adjustment for multiplicity of testing to control the experimentwise error rate $\alpha$, is a common and straightforward approach to the analysis of multiple outcomes. When the outcomes are moderately correlated, Bonferroni-based procedures tend to have adequate power, but are overly conservative when the outcomes are highly correlated. Indeed, as our simulation results in §4 show, Bonferroni procedures tend to perform adequately in most settings, although joint tests based on the LMM (1), for example, are generally more powerful than univariate procedures.

Multiplicity can also affect the power of univariate, Bonferroni-based procedures in surprising ways. Intuitively, one might expect that an increase in the number of outcomes to be tested would lead to a less powerful test of the treatment effect, simply because the level $\alpha$ would be partitioned over more univariate tests; however, this is not a conventional truth. The results of a small simulation of 1,000 datasets are presented in Table 3, in which data were generated according to the LMM (1). Forty treated ($X = 1$, $n_1 = 40$) and forty comparison ($X = 0$, $n_0 = 40$) observations were generated, each containing $K = \{5, 10, 20\}$ outcomes (other sample size configurations were simulated with similar results). Two settings of the treatment effect $\boldsymbol{\beta} = (\beta^{(1)}, \ldots, \beta^{(K)})'$ were considered: (a) $\boldsymbol{\beta} = (0.6, 0, \ldots, 0)'$ defined as a "treatment effect on one outcome"; and (b) $\boldsymbol{\beta} = (0.3, \ldots, 0.3)'$ defined as a "smaller uniform effect on all outcomes." Separate *t*-tests with Bonferroni adjustments were implemented to test for a treatment effect on any outcome. Table 3 shows that the power of the standard Bonferroni procedure does not always decrease with increasing $K$ for fixed correlation, contrary to basic intuition. For example, under zero correlation and a uniform treatment effect, the power actually increases with more outcomes. In part (a) with a treatment effect on just one outcome, the power understandably decreases as $K$ increases, because the significance level $\alpha$ is partitioned over more tests on outcomes with no treatment effect. In part (b), although $\alpha$ is more highly partitioned with increasing $K$, underlying each outcome is a true treatment effect so that increasing $K$ yields more power, specifically when correlation is zero; on the other hand, under high correlation, increasing $K$ yields lower power of the Bonferroni procedure, because the multiple univariate tests become redundant. Table 3 highlights that the conservativeness of the Bonferroni procedure with increasing correlation among outcomes is dependent upon specification of the alternative hypothesis. For certain alternatives, it is less sensitive to the strength of the correlation. The lesson here is that while Bonferroni procedures are well-suited to controlling false rejections due to multiplicity, it is not often clear in which situations they are optimal in the sense of statistical power.

### 3.2 Joint testing

We are interested in two types of joint tests:

**(i)** a *global test* (*K-df*) to evaluate the null hypothesis $H_0 : \beta^{(1)} = \ldots = \beta^{(K)} = 0$; and

**(ii)** a *single degree of freedom* (*1-df*) *test* against the alternative $H_A : \Sigma_{k=1}^{K} w_k \beta^{(k)} \neq 0$,

for pre-specified weights $w_k$. For example, if there are $K = 5$ outcomes, with one primary and the rest secondary, then the primary outcome could be given weight $w_1 = 0.6$ and the other four weights $w_2 = w_3 = w_4 = w_5 = 0.1$. In another scenario, if all the outcomes are of primary interest, then all weights could equal $w_k = 1$, $\forall k$. Test statistics are based on

estimates from a joint model, such as the LMM (1). Once specified, from the model we can obtain joint estimates of the $\beta^{(k)}$'s and their variance-covariance matrix. Using these estimates and their variance-covariance matrix, Wald statistics can be formed to test either the global or 1-*df* null hypotheses outlined above. For example, the Wald statistic for the 1-*df* test is given by $\frac{\left(\Sigma_k w_k \widehat{\beta}^{(k)}\right)^2}{Var\left(\Sigma_k w_k \widehat{\beta}^{(k)}\right)}$ which has a $\chi^2(1)$ distribution.

As noted earlier, joint testing of multiple outcomes is a less common approach because of the difficulty in specifying their joint distribution. In the setting of multiple outcomes assumed to have normal distributions, the standard comparison of two treatment groups can be based on Hotelling's $T^2$ statistic, which provides a measure of the distance between two population mean outcome vectors [14]. The statistic is based on the Mahalanobis distance between the two sample mean vectors of the treated and comparison groups; the test is useful for evaluating an overall difference between the mean outcomes of the treated and comparison groups.

As with Hotelling's $T^2$ test, the 1-*df* test based on the LMM (1) does not generally determine which specific outcomes may be improved by treatment; however, the weights $\{w_k\}$ may be defined *a priori* so that some outcomes are considered more important in evaluating the hypothesis. One limitation of the 1-*df* test is that it cannot detect the direction of the treatment effect; in some cases, one may wish to transform the outcome(s) so that the estimated effects are in the desired direction. In particular, because the 1-*df* test statistic is a weighted sum of multiple statistics, opposing estimates (that is, positive and negative effects) might cancel out so that the treatment effect is deemed insignificant when in fact there is a true effect. As noted by Roy, Lin, and Ryan, 1-*df* global tests will have the most power in the presence of a common effect size on the multiple outcomes [3].

Models for the joint distribution of the *K* multiple outcomes such as the LMM (1) can also be used to construct a *K*-*df* test. LMM-based *K*-*df* tests allow for specification of the covariance structure, in contrast to the Hotelling $T^2$ statistic which utilizes a general unstructured covariance matrix. As a result, the LMM-based *K*-*df* test can be more powerful than Hotelling's $T^2$ test when the covariance is correctly specified, for example, under compound symmetry. Both the LMM-based *K*-*df* test and Hotelling's $T^2$ have the advantage of being able to detect the direction of treatment effects, as compared to the 1-*df* test; however, they have less power than the 1-*df* test when the treatment has a common effect on all the outcomes in the study.

## 4 Simulation study

We investigated the performance of the univariate and joint testing procedures with varying treatment effects and correlations among the outcomes. A necessary emphasis here is that our simulation study was designed for the setting of a constant positive correlation between all pairs of outcomes by setting $\gamma^{(k)} = 1, \forall k$ in the LMM (1). Monte Carlo simulations of 10,000 repetitions were used to assess the type I error and power properties of the global *K*-*df* and 1-*df* tests based on the LMM (1). Specifically, we examine global and 1-*df* tests based on maximum likelihood estimation of the treatment effects $(\beta^{(1)}, \ldots, \beta^{(K)})'$, and compare their properties against Hotelling's $T^2$ test and a univariate test with Bonferroni adjustments for multiplicity. The joint tests based on estimates from the LMM that specifies the (correct) joint distribution of the multivariate outcomes should capitalize on the correlation among the outcomes in order to yield more powerful (and unbiased) tests of treatment effects.

The performance of the univariate and joint testing procedures is also evaluated under the situation of missing data. When data are missing, the effective sample size is reduced so that a loss of power is expected for any given test. The test procedures, however, use the available information in different ways. The univariate Bonferroni procedure estimates the treatment effect on an outcome with all the available data on that outcome; on the other hand, Hotelling's $T^2$ discards the ith subject in computing the test statistic if any of subject *i*'s outcomes are missing. Joint testing makes the most efficient use of the available information by capitalizing on the association among the outcomes through a joint model. For our simulation study, two forms of missingness are considered and described in §4.1.2 and §4.1.3.

## 4.1 Design of simulation

**4.1.1 Parameter configuration—**Data for our simulation study were generated by the LMM (1). We assume that the intercept terms $\alpha^{(k)} = 0\ \forall k$, an assumption which does not affect the general results presented here. We considered three specifications of the treatment effect $\boldsymbol{\beta} = (\beta^{(1)}, \ldots, \beta^{(K)})'$, following the LMM (1). The first specification of the treatment effect places an effect $\beta^{(1)} > 0$ on the first outcome, while the other effects are zero, $\beta^{(k)} = 0$ for $k \neq 1$; for example, in the simulation presented here, $K = 5$ and $\boldsymbol{\beta} = (0.6, 0, 0, 0, 0)'$. A second specification places a uniform treatment effect (also known as "common effect") on all the outcomes $\beta^{(j)} = \beta^{(k)} > 0\ \forall j \neq k$; for example, $\beta^{(k)} = 0.3\ \forall k$. The last specification places a non-constant, positive effect on all but one of the outcomes; for example, $\boldsymbol{\beta} = (0.6, 0.45, 0.3, 0.15, 0)'$ for the third specification of the treatment effect $\boldsymbol{\beta}$ in the simulation study.

The covariance-variance structure for the multiple outcomes is determined by the random effect $b_i$ and independent error terms $\boldsymbol{\varepsilon}_i = \left(\varepsilon_i^{(1)}, \ldots, \varepsilon_i^{(K)}\right)'$, for which $b_i \sim N(0, \rho$ and $\varepsilon_i^{(k)} \sim N\{0, (1-\rho)\}\ \forall k$ To facilitate discussion here, we consider the setting in which $\gamma^{(k)} = 1\ \forall k$, in order to induce a constant pairwise correlation for all pairs of outcomes (compound symmetry); more complex correlation structures can be defined by changing the values of the scale parameters $\gamma^{(k)}$ for $k = 1, \ldots, K-1$, with $\gamma^{(K)} = 1$. With the $\gamma^{(k)}$'s set to unity, $Var(Y^{(k)}|X, \beta^{(k)}) = 1\ \forall k$, so that the outcomes are all measured on the same scale, and the outcomes all have a constant pairwise correlation, ρ. The simulation parameter of most interest is ρ, which affects the power of the univariate and joint tests in different ways. While the assumption of compound symmetry for the correlation structure might not be tenable in real examples; we make this assumption here in the simulation in order to illuminate the meaningful influence of correlation on the characteristics of the univariate and joint tests, particularly under the different specifications of the treatment effect and under different patterns of missing data.

With the foregoing specification, 40 treated ($X = 1$) and 40 comparison ($X = 0$) subjects with $K = 5$ outcomes each were generated by the LMM (1). Other settings of sample size and number of outcomes were considered, but not shown in the results in §4.2. The relationship between correlation and testing characteristics were similar for different settings of these parameters. For example, an increase in sample size yielded greater power for the univariate and joint tests, all other things held equal, but the general trends and relationships did not change.

**4.1.2 Missing completely at random—**When data are missing completely at random (MCAR), the probability that observation $Y_i^{(k)}$ is missing is a constant π, for all subjects *i* and outcomes *k*, and the test procedures will generally suffer a loss of power, although to differing degrees. For example, when data are MCAR with probability $\pi = 0.2$ in a sample of

100 subjects, the Bonferroni procedure estimates the treatment effect on each of the $K$ outcomes with an expected $(1-\pi)100 = 80$ observations. On the other hand, for $K = 5$ outcomes, say, Hotelling's $T^2$ statistic uses only an expected $(1-\pi)^5 100 = 33$ complete observations, that is, 33 individuals whose 5 outcomes are all observed. Joint tests based on a joint model, such as the LMM (1) use all the available information, like the Bonferroni procedure, but also capitalize on the correlation of the outcomes so that they recover the loss of power when the outcomes are highly correlated.

**4.1.3 Missing at random—**When data are missing at random (MAR), the probability that an observation is missing can depend on observed outcomes and treatment assignment. This dependency can induce bias in test procedures when it is ignored. Patterns of MAR missingness will lead to biased estimates of the treatment effect, thereby inflating type I error if it is not properly accounted for in the estimation.

Our MAR mechanism for the simulation study is specified by the following model. Let $\boldsymbol{Y}_i = \left(Y_i^{(1)}, \ldots, Y_i^{(K)}\right)'$ be the vector of $K$ outcomes for subject $i$ and $X_i$ be the treatment assignment. Assume that the first outcome $Y_i^{(1)}$ is always observed, and the pattern of missingness on the other outcomes $\left(Y_i^{(2)}, \ldots, Y_i^{(K)}\right)'$, depends on $Y_i^{(1)}$. Let $\pi_i\left(Y_i^{(1)}, X_i\right)$ be the probability of missingness for each of these $K-1$ outcomes for subject $i$, where

$$\log\left(\frac{\pi_i}{1-\pi_i} | Y_i^{(1)}, X_i\right) = \delta_0 + \delta_1 Y_i^{(1)} + \delta_2 X_i + \delta_{12} Y_i^{(1)} X_i \tag{2}$$

By this missingness model, the baseline probability of missingness is given by $\delta_0$; for ease of discussion, we assume that $\delta_1 = \delta_2 = 0$, so that missingness depends on the interaction between treatment and outcome $Y^{(1)}$ through $\delta_{12} > 0$. By this construction, the outcomes in the comparison group ($X = 0$) are missing completely at random, while the outcomes in the treated group ($X = 1$) are missing at random with probabilities that depend on the values of $Y^{(1)}$. (When $\delta_{12} = 0$, then this setting reduces to the MCAR mechanism for both comparison and treated observations.) Without a correctly specified model for the multiple outcomes, such as the LMM (1), it is expected that tests based on data with this pattern of missingness will be biased, namely the Bonferroni and Hotelling's $T^2$ test procedures. In the missingness model as defined, treated subjects with larger values of $Y^{(1)}$ will have more outcomes missing when $\delta_{12} > 0$. When the treatment has effect only on $Y^{(1)}$, this might pose little problem for univariate tests; however, when the correlation among outcomes is large, then the missingness will contribute to biased tests.

When the outcomes are jointly modelled, for example, by the LMM (1), then maximum likelihood estimates of the treatment effect can be obtained. In particular, when the model is correctly specified, then resulting estimates of the treatment effects $\beta$ will be unbiased, and test procedures based on these estimates will maintain the nominal type I error rate.

## 4.2 Results

The power of the joint and univariate tests, in the presence of complete and MCAR data, are shown in Figure 1 for $K = 5$ outcomes. Simulation results for larger number of outcomes, for example, $K = 10$, are not presented here; such a presentation would have involved results that were not directly comparable across varying values of $K$, because the specifications of the treatment effect would not have had an obvious interpretation across values of $K$. However, it was noted that the results and conclusions from those simulations are similar to

these presented here; namely, the power characteristics of the test procedures in relation to each other were the same for $K = 5$ versus $K = 10$ outcomes.

For two of the three configurations of the treatment effect, the 5-*df* joint test based on the LMM (1) outperforms the other testing procedures, especially in the presence of highly correlated outcomes; in fact, the power of the 5-*df* test approaches unity when the correlation is large, even when data are MCAR. Hotelling's $T^2$ closely tracks the power of the 5-*df* test as well, when the data are completely observed. The small discrepancy in power between Hotelling's $T^2$ and the 5-*df* joint test in the case of complete data is attributable to the correctly specified correlation model (compound symmetry) for the latter. In contrast to the 5-*df* test, the 1-*df* joint test has deflated power for the two settings in which the 5-*df* performs best; however, as expected, when the treatment effect is uniform on all the outcomes then the 1-*df* test produces the greatest power among the testing procedures considered here.

When the treatment has effect on one outcome (for example, $\beta^{(1)} = 0.6$ in the first row of Figure 1), the power of the Bonferroni procedure does not change with the strength of the correlation of the outcomes; that is, there appears to be no loss of power for the Bonferroni procedure when correlation increases. A basic explanation of this result is that the Bonferroni procedure likely depends on the test for $\beta^{(1)}$ and not on the tests for the other outcomes, which do not exhibit a treatment effect under the alternative. However, when the treatment has an effect on more than one outcome, the Bonferroni procedure loses power as the correlation among the outcomes increases. This reinforces the earlier point made in §3.1 about the power characteristics of the Bonferroni procedure that depend on specification of the alternative hypothesis.

In the presence of data that are MCAR, similar patterns emerge for the relative power of the joint and univariate test procedures, with the exception of Hotelling's $T^2$. A subject who is missing one or more outcomes is not included in calculating the test statistic of Hotelling's $T^2$; thus, the Hotelling's $T^2$, based on a much smaller effective sample size, suffers a considerable loss of power when outcomes are missing. On the other hand, the univariate Bonferroni procedure uses all the observed information in constructing the separate *t*-tests; the joint tests based on maximum likelihood estimates from the LMM also maintain good power when data are MCAR (in fact, the power of the 5-*df* test approaches unity with increasing correlation). These joint tests capitalize on the correct specification of the model that generate the multivariate outcomes, and thus should yield the most powerful and also unbiased tests.

Further analysis (not shown here) of the relative power between situations in which data are missing and completely observed shows that relative power increases with correlation ρ; that is, if $Power_m$ is the power of a given test when data are missing and $Power_c$ is the power of the same test when data are completely observed, then the plot of $\frac{Power_m}{Power_e}$ versus ρ has a positive slope. In other words, when data are missing, the test procedures suffer the most loss of power when the outcomes are weakly correlated, but lose little power under high correlation. For example, although the 5-*df* test loses considerable power when data are missing under low correlation, it loses less power under high correlation; in fact, under MCAR data the power of the 5-*df* test approaches unity with increasing ρ in two of the scenarios in Figure 1. We provide a basic explanation for this in the setting of two outcomes: let $Y_i^{(k)}$ be the observed *k*th outcome, for $k = 1, 2$, for the *i*th subject, and let the indicator $R_i^{(k)}=1$ if $Y_i^{(k)}$ is observed and $R_i^{(k)}=0$ if missing, with $P\left(R_i^{(k)}=1\right)=1-\pi$. Let $\bar{Y}^{(k)}$ be the sample

mean of the complete observations for the *k*th outcome and $\bar{Y}_{miss}^{(k)} = \frac{\Sigma_i R_i^{(k)} Y_i^{(k)}}{\Sigma_i R_i^{(k)}}$ for the incomplete observations. The covariance between sample means $Cov\left(\bar{Y}^{(1)}, \bar{Y}^{(2)}\right)$ is a function of the covariances $Cov\left(Y_i^{(1)}, Y_i^{(2)}\right)$; likewise, the covariance between sample means from incomplete data $Cov\left(\bar{Y}_{miss}^{(1)}, \bar{Y}_{miss}^{(2)}\right)$ is a function of the covariance given by

$$Cov\left(R_i^{(1)}Y_i^{(1)}, R_i^{(2)}Y_i^{(2)}\right) = E\left\{\left(R_i^{(1)}Y_i^{(1)}\right)\left(R_i^{(2)}Y_i^{(2)}\right)\right\} - E\left(R_i^{(1)}Y_i^{(1)}\right)E\left(R_i^{(2)}Y_i^{(2)}\right) = (1-\pi)^2 Cov\left(Y_i^{(1)}Y_i^{(2)}\right).$$

The covariance of $\bar{Y}_{miss}^{(1)}$ and $\bar{Y}_{miss}^{(2)}$ is proportional to the covariance of the $\bar{Y}^{(1)}$ and $\bar{Y}^{(2)}$ by the factor $(1 - \pi)^2$; in the side-by-side plots of Figure 1, this means that the plots in the right column (missing data) can be viewed as rescaled, along the horizontal axis (ρ), versions of the plots on the left (complete data). Under no missing data and non-uniform treatment effects, correlation increases the power of some tests. Under data that are missing completely at random, an increase in correlation accounts for missing information; the benefit is particularly noticeable for the joint tests. However, the Bonferroni and Hotelling's procedures do not generally capitalize on the correlation when data are missing.

To summarize the results, the average power of the test procedures over the three configurations of the treatment effect is presented in the fourth row of Figure 1; in a sense, this plot of the average power is itself a scenario in which the treatment has a non-uniform effect on all the outcomes. It can be argued that this is the most likely situation in a real study, as opposed to a study which might believe the treatment has an effect on just one outcome or a uniform effect on all outcomes. The plots of the average power show that under mild correlation, the Bonferroni procedure maintains adequate power (as claimed in Sankoh et al. [10]), compared to joint tests that were expected to capitalize on the correctly specified joint distribution of the multiple outcomes. It is rather clear, however, that when correlation is moderately large ($\rho \geq 0.4$), one should consider the use of joint tests in order to maximize power. Notably, when data are missing, Hotelling's $T^2$ test should be avoided, because of the drastic loss of power.

When data are MAR according to the model (2), our attention turns to the type I error characteristics of the test procedures. The plot of the type I error rate versus correlation among the outcomes is shown in Figure 2; in these plots, the parameters of the MAR model (2) are set such that in the comparison group ($X = 0$) the data are MCAR with $\pi \approx 0.2$, while in the treated group ($X = 1$) the odds ratio of missingness for outcomes $Y^{(k)}$, $k \neq 1$, under a unit increase in $Y^{(1)}$ is $\exp(\delta_{12})$. In general the joint tests based on the LMM (1) are unbiased for all levels of correlation, while Hotelling's $T^2$ is the most sensitive to MAR conditions, particularly when the odds ratio of missingness $\exp(\delta_{12})$ is large. The Bonferroni procedure is relatively unbiased for small $\exp(\delta_{12})$; however, when $\exp(\delta_{12})$ is large, the type I error of the Bonferroni procedure becomes inflated with increasing correlation. Both the Hotelling's $T^2$ and Bonferroni procedures tend to exclude observations with large values of $Y^{(1)}$, such that the resulting estimates of the treatment effect are biased. While bias is of general concern for Hotelling's $T^2$ procedure, the Bonferroni procedure is relatively unbiased under weakly correlated outcomes ($\rho \leq 0.2$); however, high correlation induces greater bias for the Bonferroni procedure when the degree of missingness due to $Y^{(1)}$ is large, that is, for large $\exp(\delta_{12})$.

## 5 Application

The discussion in §1 introduced the CATIE study for schizophrenia and briefly outlined our plan for analysis of five clinical outcomes to evaluate the efficacy and safety of the conventional antipsychotic, perphenazine, and one of the atypicals, quetiapine. The five outcomes included PANSS, which measures the overall severity of schizophrenic symptoms, and four outcomes that characterize the metabolic side effects of the medications. With the exception of HDL cholesterol, an increase from baseline in an outcome reflects an adverse event; for example, weight gain is an undesirable symptom of metabolic syndrome, as discussed in Meyer et al. [7]. In the ensuing discussion in §2 and §3, we proposed joint models and tests of treatment effects that would be able to capitalize on the correlations among the outcomes and provide more powerful tests of the effects. In the simulation study of §4 we showed that joint tests are able to capitalize on the correlation between outcomes and yield more powerful tests of treatment effects. In this section we implement joint tests to evaluate the efficacy and safety effects of perphenazine and quetiapine from the CATIE Schizophrenia Trial. Boxplots of the individual 3-month changes in the five outcomes are shown in Figure 3, and correlations are shown in Figure 4.

The linear mixed model (1) for multiple outcomes was fit using PROC NLMIXED in SAS, so that there were $K + 1$ variance-covariance parameters, as opposed to $\frac{1}{2}K(K+1)$ for joint model with a general correlation structure. To compare inferences with different assumptions about the correlation matrix, the latter model was also fit with PROC MIXED in SAS with an unstructured correlation. With these models, the treatment effect was evaluated using both a global 5-*df* test and 1-*df* test as specified in §2.1; for comparison to standard univariate procedures, we also tested the individual treatment effects with maximum likelihood estimates from the model. In addition to the estimates produced by the model, standard univariate *t*-tests were conducted which formed the basis of the Bonferroni procedure.

The estimates from univariate models and the joint LME and corresponding test results are shown in Table 4. In the first set on the left under “Univariate *t*-tests” the crude estimates and corresponding tests are shown; the Bonferroni procedure finds a treatment effect through the outcome “Weight” which has p-value less than $\frac{\alpha}{K}=0.01$, the Bonferroni-adjusted p-value. The estimates for the joint model do not differ significantly from the crude estimates of differences in sample means, and likewise, the Bonferroni procedure based on these estimates detects a treatment effect through “Weight,” but barely with p-value = 0.008 < 0.01 Both Hotelling's $T^2$ and the 5-*df* test from the joint model find at the $\alpha = 0.05$ level a significant effect of the treatment through the 5 outcomes, while the 1-*df* test does not. The latter result is not surprising, given the lack of a common effect on the outcomes; in the 1-*df* test the outcomes were equally weighted, so the four outcomes that yield insignificant results have influence on the overall 1-*df* test result. However, if the effect on the outcome “Weight” is weighted, say, by $w_k = 0.6$, and the rest by $w_j = 0.1$, $j \neq k$, then the resulting p-value is 0.027, yielding a significant result. Of course, one could use different weights to yield other significant results, so long as these weights (and corresponding assumptions) are set in advance as part of study protocol. As an alternative for a more powerful 1-*df* test of the treatment effect, Roy, Lin, and Ryan provide a scaled marginal model and relevant tests for a common effect [3]. In one related approach (not shown here), we scaled the outcomes by their estimated standard deviations, changed the sign of the cholesterol outcomes, and estimated the effects through the joint model based on the scaled data; the 1-*df* test did not yield a significant result.

Table 4 also shows that, across the estimates, the standard errors from the LMM (1) are very similar to each other, except for the estimated treatment effect on triglyceride levels. This is likely a feature of our model which restricts the number of parameters in the variance-covariance matrix at $K + 1$, rather than a full $\frac{1}{2}K(K+1)$ parameters for a general structure. We emphasize, as in §2.2, that the LMM is suited for the analysis of outcomes that are correlated in the same direction; see Figure 4. As a partial check on this assumption, we also fit a joint model with an unstructured variance-covariance matrix, results for which are presented in Table 5. Qualitatively, the results from this model do not differ from the conclusions made by the LMM. The corresponding 5-*df* test concludes with a significant effect of quetiapine over perphenazine on the five outcomes, while the 1-*df* test does not. One difference is that the corresponding standard errors from the general covariance model in Table 5 are more similar to those of the univariate estimates in Table 4. This suggests that a general covariance structure may be more appropriate for the analysis of the CATIE trial here, although the conclusions from the tests based on the more restrictive LMM do not differ.

## 6 Discussion

In this paper we proposed the use of joint tests to evaluate treatment effects that are characterized by multiple outcomes; such settings occur frequently in mental health research. Standard approaches for multiplicity in testing, such as Bonferroni-based procedures, are effective when the outcomes are small in number and weakly correlated; however, as our simulated results show, the Bonferroni procedure becomes overly conservative under moderate to high correlation. We note that our simulation results depend on the correct specification of the linear mixed effects model (1), specifically under compound symmetry for the correlation structure, $\gamma^{(k)} = 1, \forall k$. In the application to the CATIE trial we fit both the linear mixed model (1) and the model with a general correlation structure and find that the qualitative results do not change. As a general check of our results, the standard Hotelling's $T^2$ test does not make any restrictive assumption about the correlation structure and closely tracked the joint *K*-*df* tests in our simulation study. Nonetheless, further investigation about the impact of an incorrectly specified correlation structure may be warranted.

An alternative approach is to use composite outcomes, which summarize multiple outcomes into a single measure. The advantage is that standard univariate procedures can be implemented without the need to adjust for multiplicity. However, when the outcomes are measured on different scales or represent different underlying mechanisms of a treatment effect, a composite outcome might not capture the true underlying effect. The use of joint models offers a flexible way to test multiple outcomes on equal footing; in particular, joint tests based on such models can capture the correlation of the outcomes and thus are more powerful in settings with moderate to large correlation, compared to standard Bonferroni procedures. In particular, when a model for the outcomes is correctly specified, the joint tests are very robust when data are missing at random, as shown in our simulation results.

Despite the availability of methods and software to fit joint models for multiple continuous outcomes, this approach is infrequently used in practice. The primary aim of our work was to show that joint tests are able to capitalize on the associations among outcomes, yielding more powerful tests under high correlation and unbiased tests when data are missing at random. Specification of a joint model to estimate effects of treatment on multiple outcomes is straightforward, namely through the linear mixed model (1), and available software can easily fit these models.

While the simulation study presented in this paper assumes multivariate normal outcomes, this assumption is less tenable in many studies. In particular, multiple outcomes are typically measured on difference scales or are non-commensurate, meaning that the multiple outcomes are a mixture of discrete and continuous outcomes. Further work needs to be done in this setting; the major challenge is the specification of a joint model for the non-commensurate outcomes.

## Acknowledgments

The authors were supported in part by NIH grant R01-MH 054693. The authors are grateful for the programming assistance of Rita Volya, MS. Data from the CATIE study were provided by Richard Frank (R01-MH 06972 and MacArthur Foundation Grant Number 89045-0).

## References

[1]. Sammel M, Lin X, Ryan L. Multivariate linear mixed models for multiple outcomes. Statistics in Medicine. 1999; 18:2479–2492. DOI: 10.1002/(SICI)1097-0258(19990915/30)18:17/18¡2479::AID-SIM270¿3.0.CO;2-F. [PubMed: 10474154]

[2]. Lin X, Ryan L, Sammel M, Zhang D, Padungtod C, Xu X. A scaled linear mixed model for multiple outcomes. Biometrics. 2000; 56:593–601. DOI: 10.1111/j.0006-341X.2000.00593.x. [PubMed: 10877322]

[3]. Roy J, Lin X, Ryan L. Scaled marginal models for multiple continuous outcomes. Biostatistics. 2003; 4:371–383. [PubMed: 12925505]

[4]. Thurston SW, Ruppert D, Davidson PW. Bayesian models for multiple outcomes nested in domains. Biometrics. 2009; 65:1078–1086. DOI: 10.1111/j.1541-0420.2009.01224.x. [PubMed: 19397588]

[5]. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: Greater precision but with greater uncertainty. New England Journal of Medicine. 2003; 289:2554–2559.

[6]. Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO, et al. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. New England Journal of Medicine. 2005; 353:1209–1223. [PubMed: 16172203]

[7]. Meyer JM, Davis VG, Goff DC, McEvoy JP, Nasrallah HA, Davis SM, et al. Change in metabolic syndrome parameters with antipsychotic treatment in the CATIE Schizophrenia Trial: Prospective data from phase 1. Schizophrenia Research. 2008; 101:273–286. [PubMed: 18258416]

[8]. Kraemer HC, Glick ID, Klein DF. Clinical trials design lessons from the CATIE study. American Journal of Psychiatry. 2009; 166:1222–1228. DOI: 10.1176/appi.ajp.2009.08121809. [PubMed: 19797435]

[9]. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982; 38:963–974. [PubMed: 7168798]

[10]. Sankoh AJ, D'Agostino RB Sr. Huque MF. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. Statistics in Medicine. 2003; 22:3133–3150. DOI: 10.1002/sim.1557. [PubMed: 14518019]

[11]. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979; 6:65–70.

[12]. Hochberg YA. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988; 75:800–802. DOI: 10.1093/biomet/75.4.800.

[13]. Huang Y, Hsu JC. Hochberg's step-up method: cutting corners off Holm's step down method. Biometrika. 2007; 94:965–975. DOI: 10.1093/biomet/asm067.

[14]. Hotelling H. The generalization of Student's ratio. Annals of Mathematical Statistics. 1931; 2:360–378.

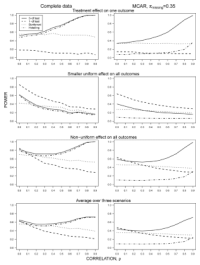

**Figure 1.**
Power of univariate and joint tests with data that are complete and missing completely at random (MCAR). Outcomes were generated by the LMM (1) and estimated by the same model to construct joint 5-df and 1-*df* tests of the treatment effect. The first configuration of the treatment effect (first row) is given by $\beta = (0.6, 0, 0, 0, 0)'$. In the second row, a smaller uniform effect is placed on all the outcomes by setting $\beta = (0.3, 0.3, 0.3, 0.3, 0.3)'$. In the third row, a non-uniform treatment effect is placed on the $K = 5$ outcomes by setting $\beta = (0.6, 0.45, 0.3, 0.15, 0)'$. The fourth row shows the average power over the three scenarios.

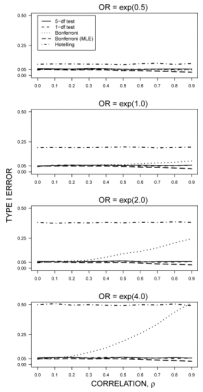


**Figure 2.**

Type I error when data are missing at random, at different odds of missingness. Data for the comparison ($X = 0$) group are missing completely at random with constant probability ≈ 0.2. The odds ratio (OR) of missingness between the treated ($X = 1$) and comparison group is given by $\exp(\delta_{12})$ according to the model (2). Univariate Bonferroni tests were conducted by estimating the treatment effects separately by outcome (dotted line) and jointly by the LMM (1) (long-dashed line).

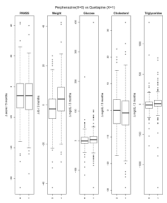

**Figure 3.**
Distributions of individual outcomes between treated and comparison groups. Each individual outcome represents the change from baseline to follow-up at 3 months.

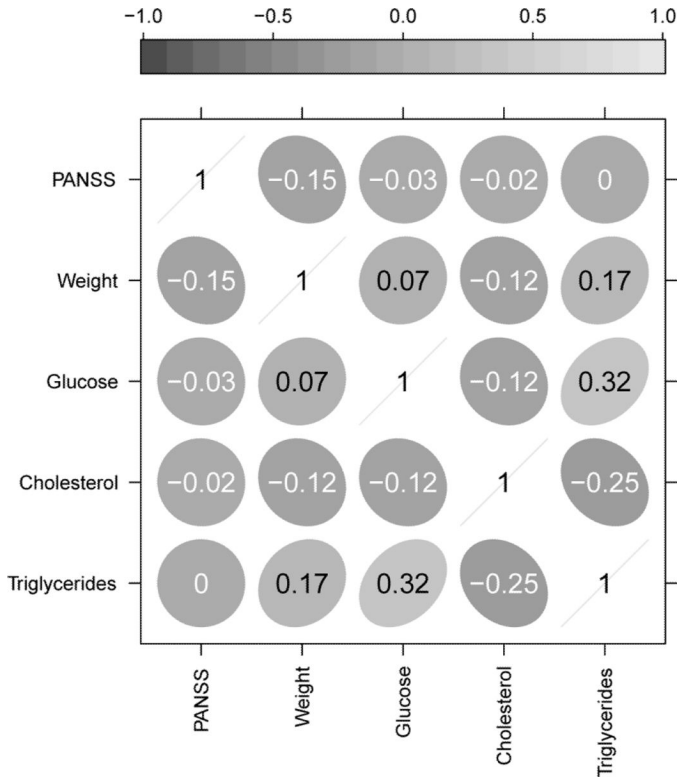


**Figure 4.**
Correlations among the CATIE outcomes.

## Table 1

CATIE outcomes: change from baseline to follow-up visit at 3 months. Sample means and standard deviations were computed with the available data, that is, missing values were ignored. With the exception of HDL cholesterol, an increase from baseline indicates an adverse effect for the outcomes.

| | Perphenazine $n = 261$ | | Quetiapine $n = 337$ | |
|---|---|---|---|---|
| **Outcome** | $\bar{x}$ **(s)** | **%miss.** | $\bar{x}$ **(s)** | **%miss.** |
| PANSS | −6.2 (14.9) | 34 | −6.7 (14.4) | 39 |
| Weight, *lb* | −2.1 (10.3) | 33 | 3.3 (10.8) | 40 |
| Glucose, *mg/dL* | 0.8 (28.8) | 37 | 6.0 (41.5) | 42 |
| HDL Cholesterol, *mg/dL* | 0.2 (7.5) | 37 | −1.0 (9.0) | 42 |
| Triglycerides, *mg/dL* | −3.9 (177) | 37 | 15.3 (199) | 42 |

## Table 2

Examples of the standard Bonferroni procedure and its variants, Hochberg's and Holm's procedures for testing of $K = 3$ outcomes at level $\alpha = 0.05$. Hypotheses are sorted $H_{01},H_{02},H_{03}$ according to the sorted p-values.

| Sorted p-values | | Bonferroni | Hochberg's step-up | Holm's step-down |
|---|---|---|---|---|
| (0.01,0.02,0.05) | **Step 1** | Reject $H_{01}$ | Accept $H_{03}$ | Reject $H_{01}$ |
| | **Step 2** | | Reject $H_{01},H_{02}$, stop | Reject $H_{02}$ |
| | **Step 3** | | | Stop |
| (0.02,0.02,0.04) | **Step 1** | No rejection | Reject $H_{01},H_{02}$, $H_{03}$ | No rejection |
| | **Step 2** | | | |
| | **Step 3** | | | |
| (0.03,0.03,0.03) | **Step 1** | No rejection | Reject $H_{01},H_{02},H_{03}$ | No rejection |
| | **Step 2** | | | |
| | **Step 3** | | | |

## Table 3

Power (%, based on 1,000 repetitions) of univariate tests with Bonferroni adjustment compared with varying number *K* of outcomes and treatment effects: (a) for a treatment effect on one outcome, $\boldsymbol{\beta} = (0.6,0,\ldots, 0)'$; and (b) for a smaller uniform effect on all outcomes, $\boldsymbol{\beta} = (0.3,\ldots, 0.3)'$.

| | K | Correlation, ρ | | |
|---|---|---|---|---|
| | | **0.0** | **0.3** | **0.9** |
| (a) Treatment effect on one outcome | **5** | 56 | 53 | 58 |
| | **10** | 46 | 44 | 44 |
| | **20** | 37 | 39 | 36 |
| (b) Smaller uniform effect on all outcomes | **5** | 44 | 33 | 18 |
| | **10** | 49 | 42 | 18 |
| | **20** | 56 | 39 | 12 |

**Table 4**

Estimates and tests of the treatment effect on $K = 5$ outcomes. Two sets of estimates are presented, one from standard t-tests based on differences in sample means of each outcome between treated and comparison groups and the other from the linear mixed model (1). *p*-values from joint tests are also presented.

| | Univariate *t*-tests | | | Linear mixed model | | | Joint tests: p-values | | |
|---|---|---|---|---|---|---|---|---|---|
| **Outcome** | $\hat{\beta}$ | se($\hat{\beta}$) | p-value | $\hat{\beta}$ | se($\hat{\beta}$) | p-value | Hotel. $T^2$ | 5-df | 1-*df* |
| PANSS | −0.5 | 1.5 | 0.74 | −0.5 | 2.1 | 0.80 | | | |
| Weight | 5.5 | 1.1 | $< 10^{-4}$ | 5.5 | 2.1 | 0.008 | | | |
| Glucose | 5.1 | 3.7 | 0.17 | 5.1 | 2.3 | 0.026 | $< 10^{-4}$ | 0.036 | 0.18 |
| Cholesterol* | −1.2 | 0.9 | 0.16 | −1.2 | 2.1 | 0.57 | | | |
| Trigylcerides | 19.2 | 19.8 | 0.33 | 19.5 | 20.0 | 0.33 | | | |

*Cholesterol represents high-density lipoproteins; a decrease from baseline is considered to be an adverse event.

**Table 5**

Estimates and tests of the treatment effect on K = 5 outcomes, under the assumption of a general covariance structure for the outcomes.

| | General correlation | | | Joint tests: p-values | |
|---|---|---|---|---|---|
| **Outcome** | $\hat{\beta}$ | se($\hat{\beta}$) | **p-value** | **5-df** | **1-*df*** |
| PANSS | −0.5 | 1.5 | 0.76 | | |
| Weight | 5.5 | 1.1 | $< 10^{-4}$ | | |
| Glucose | 5.2 | 3.8 | 0.22 | $< 10^{-4}$ | 0.17 |
| Cholesterol* | −1.2 | 0.9 | 0.20 | | |
| Trigylcerides | 20.1 | 20.0 | 0.33 | | |