10-17-2014

Applications of the Wei-Lachin multivariate one-sided test for multiple outcomes on possibly different scales

John M. Lachin
George Washington University

Applications of the Wei-Lachin Multivariate One-Sided Test for Multiple Outcomes on Possibly Different Scales

John M. Lachin*

The Biostatistics Center, The George Washington University, Rockville, Maryland, United States of America

Abstract

Many studies aim to assess whether a therapy has a beneficial effect on multiple outcomes simultaneously relative to a control. Often the joint null hypothesis of no difference for the set of outcomes is tested using separate tests with a correction for multiple tests, or using a multivariate T^2-like MANOVA or global test. However, a more powerful test in this case is a multivariate one-sided or one-directional test directed at detecting a simultaneous beneficial treatment effect on each outcome, though not necessarily of the same magnitude. The Wei-Lachin test is a simple 1 *df* test obtained from a simple sum of the component statistics that was originally described in the context of a multivariate rank analysis. Under mild conditions this test provides a maximin efficient test of the null hypothesis of no difference between treatment groups for all outcomes versus the alternative hypothesis that the experimental treatment is better than control for some or all of the component outcomes, and not worse for any. Herein applications are described to a simultaneous test for multiple differences in means, proportions or life-times, and combinations thereof, all on potentially different scales. The evaluation of sample size and power for such analyses is also described. For a test of means of two outcomes with a common unit variance and correlation 0.5, the sample size needed to provide 90% power for two separate one-sided tests at the 0.025 level is 64% greater than that needed for the single Wei-Lachin multivariate one-directional test at the 0.05 level. Thus, a Wei-Lachin test with these operating characteristics is 39% more efficient than two separate tests. Likewise, compared to a T^2-like omnibus test on 2 *df*, the Wei-Lachin test is 32% more efficient. An example is provided in which the Wei-Lachin test of multiple components has superior power to a test of a composite outcome.

Citation: Lachin JM (2014) Applications of the Wei-Lachin Multivariate One-Sided Test for Multiple Outcomes on Possibly Different Scales. PLoS ONE 9(10): e108784. doi:10.1371/journal.pone.0108784

Editor: Robert K. Hills, Cardiff University, United Kingdom

Received January 14, 2014; **Accepted** August 21, 2014; **Published** October 17, 2014

Funding: This work was supported with funding from NIDDK, NIH for the *Glycemia Reduction Approaches in Diabetes: A Comparative Effectiveness (GRADE) Study* through grant 1U01-DK-098246-01, John M. Lachin and David M. Nathan (Massachusetts General Hospital) co-Principal Investigators. This work was also partially supported by funding from NIDDK for the Diabetes Prevention Program Outcomes Study through grant U01-DK-48489, Sarah Fowler, Principal Investigator). The funders had no role in the conduct of this methodological research, the decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* Email: jml@bsc.gwu.edu

Introduction

In many studies an objective is to assess whether an experimental therapy (E) versus control (C) has beneficial effects on multiple component outcomes. This is becoming increasingly common in the evaluation of the comparative effectiveness of therapies. For example, the NIDDK-funded "**G**lycemia **R**eduction **A**pproaches in **D**iabetes: A Comparative **E**ffectiveness" (GRADE) Study will compare four agents commonly used to control glucose levels in type 2 (adult) diabetes [1], clinicaltrials.gov NCT01794143. The primary objective is to evaluate the durability of glucose control over 3–6 years of treatment, the primary outcome being the time to a confirmed rise of HbA1c (a measure of average glucose levels) ≥7% (the therapeutic target being a value <7%) using a logrank test. A secondary outcome is to compare each pair of treatments with respect to multiple components of effectiveness, specifically whether one treatment is superior to the other with respect to durability of control (event-times), absence of hypoglycemia over 3 years of treatment (proportions), and a lower mean body weight at 3 years. Herein we describe how such a test could be conducted and evaluate the power of the test or the required sample size.

For illustration, throughout we consider the case of two outcomes, say A and B, although all the procedures herein generalize to ≥2 outcomes. We wish to test the null hypothesis H_0: $(A_E \equiv A_C) \cap (B_E \equiv B_C)$ that the experimental therapy is equivalent to control for both outcomes versus the alternative H_1: $(A_E \succeq A_C) \cap (B_E \succeq B_C)$ with at least one strict superiority, where "$\equiv$" means equality for an outcome and where "$\succ$" means superiority. The test against such an alternative is called a multivariate one-directional (or one-sided) test.

Wei and Lachin [2] proposed a simple 1 *df* test for such a hypothesis that was described as a test against an ordered alternative, or a test of stochastic ordering. The test was later studied by Lachin [3] and Frick [4,5]. Herein the application of this test to multiple outcomes is described for a test of means, a test of proportions, a test of event times and a test with mixed components such as where one outcome is quantitative (using means) and another qualitative (using proportions). For each application, equations are also derived for evaluation of sample size and power of the test. Multiple model-based tests are also described. For an analysis of multiple mean differences we show that the Wei-Lachin test is more powerful than an analysis based on either separate tests for each outcome, multiplicity adjusted, or a multivariate T^2-like omnibus test. An example from a major clinical trial is presented.

Many other tests have also been proposed, principally in the setting of tests for differences in means. These are reviewed in the discussion section.

Wei-Lachin Multivariate One-Directional Test and Its Power

Three versions of the Wei-Lachin test are described. The first employs the measurements using the original scale of measurement. This test, however, is not invariant to scale transformations of the individual components. Two scale invariant tests are also described, one based on standardized values and another based on scale-independent Z-tests.

Scale-Based Test For Multiple Outcomes

Let X_{ij} designate the jth outcome variable in the ith group with expectation $E(X_{ij})=\mu_{ij}$, $i=E,\ C;\ j=a,\ b$. The subscripts a, b are used through out to refer to the two outcomes. The jth outcome could be a quantitative measure or a binary variable (among others). Assume that a more favorable outcome is represented by a decreasing expectation for X. Let

$$\delta_a=\mu_{Ca}-\mu_{Ea} \tag{1}$$

$$\delta_b=\mu_{Cb}-\mu_{Eb}.$$

A positive value for each represents a beneficial effect of the experimental therapy over control for each outcome, and a negative value represents lack of benefit. The null and alternative hypotheses of interest are

$$H_0\ :\ \delta_a=0 \text{ and } \delta_b=0 \tag{2}$$

$$H_{1S}\ :\ \delta_a\geq 0 \text{ and } \delta_b\geq 0 \text{ and } sum(\delta_a,\delta_b)>0.$$

Thus, $H_{1\mathrm{S}}$ designates that the experimental therapy is at least as effective as control for both outcomes and is superior to control for either or both outcomes. This is called the multivariate one-directional hypothesis.

In the context of an analysis of repeated measures, or multivariate observations, Wei and Lachin [2] described a multivariate one-directional test, what they termed a test of stochastic ordering, i.e. a test of the null hypothesis that is directed towards an alternative hypothesis of the form $H_{1\mathrm{S}}$ in (2). Lachin [3,6] contrasts this test with other tests, such as the omnibus test.

Consider group-specific estimates $\hat{\mu}_{ij}$ with expectation μ_{ij}. Let $\hat{\delta}_a$ and $\hat{\delta}_b$ designate the estimates of the difference between the groups for each outcome as defined in (1), and $\hat{\Delta}=(\hat{\delta}_a\ \hat{\delta}_b)'$, where "′" designates the transpose. With large samples

$$\hat{\Delta}\sim\mathcal{N}(\Delta,\Sigma). \tag{3}$$

with expectation $\Delta=(\delta_a\ \delta_b)'$ and with a covariance matrix $\mathbf{\Sigma}$ that is consistently estimable with elements

$$\mathbf{\Sigma}=\begin{bmatrix}\sigma_a^2=V(\hat{\delta}_a) & \sigma_{ab}=Cov(\hat{\delta}_a,\hat{\delta}_b)\\ \sigma_{a\,b} & \sigma_b^2=V(\hat{\delta}_b)\end{bmatrix}. \tag{4}$$

The Wei-Lachin test is then provided by

$$Z_S=\frac{\mathbf{J}'\hat{\mathbf{\Delta}}}{\sqrt{\mathbf{J}'\hat{\mathbf{\Sigma}}\mathbf{J}}}=\frac{\hat{\delta}_a+\hat{\delta}_b}{\hat{\sigma}_S}, \tag{5}$$

$$\hat{\sigma}_S^2=\hat{V}(\hat{\delta}_a+\hat{\delta}_b)=\left[\hat{\sigma}_a^2+\hat{\sigma}_b^2+2\hat{\sigma}_{ab}\right]$$

using consistent estimates of the variances and covariance, where $\mathbf{J}=(1\ 1)'$. Asymptotically $Z_S\sim N(0{,}1)$ under H_0 from Slutsky's theorem. The test rejects H_0 in favor of $H_{1\mathrm{S}}$ when $Z_S\geq Z_{1-\alpha}$ at level α one-sided. The above generalizes to $K>2$ outcomes. Note that the test can also be obtained from the unweighted average of the group differences relative to its standard error that provides a convenient average measure of the group differences when all outcomes are measured on the same scale.

Specific applications include a large sample test of means [3] or proportions [7], a generalized linear regression model using quasi likelihoods with a covariance matrix estimated using the information sandwich, i.e. GEE [8]; or a normal errors model for the analysis of repeated measures [9]; or a proportional hazards model using the information sandwich [10]; or these estimates can be based on a distribution-free estimate such as the Mann-Whitney difference that provides a Wilcoxon test [3,11] with the Wei-Lachin [2] estimate of the covariance matrix. These and other methods allow for some observations for some outcomes in some subjects to be missing either completely at random or at random (conditionally).

Although often termed a multivariate one-directional (one-sided) test, it is possible to conduct a two-sided one-directional test that either E is superior to C for all components, or C is superior to E. In that case, the Wei-Lachin 1 df test statistic is referred to the two-sided critical value rather than the one-sided value. Herein we describe the one-sided test.

If beneficial values of X_a are lower, but those for X_b are higher, such as for a test of LDL and HDL, respectively, then the test would be constructed using the negative of the values for X_b such that $\delta_b=\mu_{Eb}-\mu_{Cb}$. If higher values of both measures demonstrate benefit for the treatment, then both δ_a and δ_b can be defined as the difference of treated minus control.

This test would be appropriate when all of the outcome measurements were on the same scale; for example, as for a test of a beneficial effect on both systolic and diastolic blood pressure (both mm Hg), or a test of a beneficial effect on both LDL and HDL (both mg/dl). Other variations described below would be appropriate for outcomes with different variances, or measures on different scales or mixtures of different types of measures, such as A being a quantitative variable and B being a binary variable.

An alternative approach commonly applied to test the superiority of an experimental therapy is to base the inference on the two separate one-sided tests. These tests would require a correction for multiple tests such as using the Holm [12] improved Bonferroni procedure which requires that the minimum of the two p-values be ≤ 0.025 (one-sided) and the other ≤ 0.05 in order to declare significance at the 0.05 level for the two tests. The corresponding alternative hypothesis is

$$H_{1P}: \ [\delta_a > 0 \text{ and/or } \delta_b > 0] \not\equiv H_{1S}. \tag{6}$$

However, the alternative H_{1P} includes the case where the experimental therapy is beneficial for one outcome but harmful for the other, such as where $\delta_a > 0$ and $\delta_b < 0$ or vice versa.

Yet another possible test would be the omnibus test using a T^2-like test of the null hypothesis H_0 versus

$$H_{1O}: \ [\delta_a \neq 0 \text{ and/or } \delta_b \neq 0] \not\equiv H_{1S}. \tag{7}$$

that is provided by

$$X_O^2 = \hat{\Delta}'\hat{\Sigma}^{-1}\hat{\Delta} \tag{8}$$

which is asymptotically distributed as chi-square on 2 (or more generally K) df. This is likewise inappropriate because the alternative includes cases where the experimental therapy is worse than control for either or both outcomes.

Maximin Efficiency of the Wei-Lachin Test

For the case of two measures as herein, the restricted alternative multivariate one-dimensional hypothesis $H_{1\mathrm{S}}$ in (2) corresponds to all points in the positive orthant of the two-dimensional parameter space for (δ_a,δ_b). Since the test is a sum of the two estimates, the rejection region is defined by the line of values $(\hat{\delta}_a,\hat{\delta}_b)$ satisfying $Z_S = Z_{1-\alpha}$ that simply connects the points $(\delta_\alpha,0)$ and $(0,\delta_\alpha)$ where $\delta_\alpha = Z_{1-\alpha}\hat{\sigma}_S$. Thus the rejection region principally includes an area of the positive orthant away from the origin, but also includes elements of the sample space where either $\hat{\delta}_a < 0$ or $\hat{\delta}_b < 0$, but not both. With large sample sizes, the probability of such points is negligible for true values (δ_a,δ_b) away from zero, i.e towards the central projection (the 45° line) of the positive orthant. Lachin [6] provides figures to illustrate these relationships.

For a given pair of values $\Delta_1 = (\delta_{a1} \ \delta_{b1})'$ specifying a point in the positive orthant $(\delta_{a1},\delta_{b1})$, it is readily shown [13] that the optimal likelihood ratio test of H_0: $\Delta = (0 \ 0)'$ versus the point alternative H_{Δ_1}: $\Delta = \Delta_1$ based on (3) is

$$\chi^2_{LR} = \frac{(\Delta'_1\Sigma^{-1}\hat{\Delta})^2}{\Delta'_1\Sigma^{-1}\Delta_1} \tag{9}$$

where χ^2_{LR} is distributed as chi-square on 1 df under H_0. Note that χ^2_{LR} is based on a weighted sum of the estimated differences $(\hat{\Delta}) = (\hat{\delta}_a \ \hat{\delta}_b)'$. Thus, for a given Σ, every point $\Delta_1 = (\delta_{a1},\delta_{b1})$ that defines a unique alternative hypothesis value in the two dimensional parameter space entails a different optimal linear combination of the observed $\hat{\Delta}$. Further, the same weights are optimal for any alternative hypothesis defined by points proportional to $(\delta_{a1}/\sigma_a, \ \delta_{b1}/\sigma_b)$ with the same correlation, such as the point $(c\delta_{a1}/\sigma_a, \ c\delta_{b1}/\sigma_b)$ for any $c>0$. This implies that the same weights would be optimal for all points in the parameter space falling on the vector projection defined by the specified $(\delta_{a1}/\sigma_a, \ \delta_{b1}/\sigma_b)$. Thus, there are an infinite number of alternative hypotheses corresponding to all possible projections in the positive orthant, each with a different optimal test.

Unfortunately it is not known which projection is optimal since the actual parameter values (δ_a,δ_b) are unknown. However, Frick [4,5] showed that the Wei-Lachin test is maximin efficient with respect to whichever weighted test is in fact optimal under the condition that $\hat{\Sigma}\mathbf{J} \geq 0$. That is, among the family of linear combinations of the estimates, the Wei-Lachin test minimizes the loss in efficiency (power) relative to the unknown optimal linear combination when this condition applies, in which case it is the optimal robust linear test of $H_{0\mathrm{S}}$ versus $H_{1\mathrm{S}}$. For two or more measures with positive correlations, as would be the case under the alternative hypothesis, Frick's condition $\Sigma\mathbf{J} \geq 0$ is satisfied.

When this simple condition does not apply, Frick [4] shows that a simple weighted test is provided by

$$Z_{S,L} = \frac{\mathbf{L}'\hat{\Delta}}{\sqrt{\mathbf{L}'\hat{\Sigma}\mathbf{L}}} \tag{10}$$

that is also maximin efficient where $\mathbf{L}$ satisfies the restriction $\mathbf{L}'\hat{\Sigma}\mathbf{J} = 1$. For a given $\hat{\Sigma}$, the vector $\mathbf{L}$ is obtained as $\mathbf{L} = \mathbf{B}'\hat{\Sigma}$ where $\mathbf{B}$ is the quadratic program solution to $\min_{\mathbf{y}}[\mathbf{y}'\hat{\Sigma}^{-1}\mathbf{y}]$ under the constraints that $y_i \geq 0 \ \forall i$ and $\mathbf{y}'\mathbf{J} = 1$. This test will principally be required in cases where the null hypothesis applies, or the treatment is inferior for some of the component outcome measures. A SAS program for this computation is available from the author (see Discussion).

Scale-based Test for Multiple Means

To illustrate the construction of the Wei-Lachin test, consider a large sample test for a difference between groups in the means of two outcomes where it is assumed that $X_{ij} \sim f(\mu_{ij},\psi^2_{ij})$ with some distribution f where $\psi^2_{ij} = V(X_{ij})$ is the variance of the observations for the jth outcome in the ith group, or the residual variance after adjusting for other covariates, and $\psi_{iab} = Cov(X_{ia},X_{ib})$, $i = E, \ C$; $j = a, \ b$. To simplify, assume that there is a common covariance matrix in the two groups (homoscedasticity) with correlation $\rho_{ab} = \psi_{ab}/(\psi_a\psi_b)$. Then asymptotically

$$\begin{pmatrix} \overline{X}_{ia} \\ \overline{X}_{ib} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_{ia} \\ \mu_{ib} \end{pmatrix}, \begin{pmatrix} \psi^2_a/n_{ia} & \psi_{ab}\dfrac{n_{iab}}{n_{ia}n_{ib}} \\ \psi_{ab}\dfrac{n_{iab}}{n_{ia}n_{ib}} & \psi^2_b/n_{ib} \end{pmatrix}\right] \tag{11}$$

where $(n_{ia}, \ n_{ib}, \ n_{iab})$ are the numbers in the ith group with observed values for outcome A and B separately and jointly, $i = E$, C [3].

Then $\hat{\delta}_a = (\bar{X}_{Ca} - \bar{X}_{Ea})$ and $\hat{\delta}_b = (\bar{X}_{Cb} - \bar{X}_{Eb})$ and $\hat{\Delta} = (\hat{\delta}_a \ \hat{\delta}_b)'$ is asymptotically distributed as in (3) with covariance matrix

$$\Sigma = \begin{bmatrix} \psi^2_a\left(\dfrac{1}{n_{Ea}} + \dfrac{1}{n_{Ca}}\right) & \psi_{ab}\left(\dfrac{n_{Eab}}{n_{Ea}n_{Eb}} + \dfrac{n_{Cab}}{n_{Ca}n_{Cb}}\right) \\ \psi_{ab}\left(\dfrac{n_{Eab}}{n_{Ea}n_{Eb}} + \dfrac{n_{Cab}}{n_{Ca}n_{Cb}}\right) & \psi^2_b\left(\dfrac{1}{n_{Eb}} + \dfrac{1}{n_{Cb}}\right) \end{bmatrix} \tag{12}$$
$$= \begin{bmatrix} \sigma^2_a & \sigma_{ab} \\ \sigma_{ab} & \sigma^2_b \end{bmatrix}.$$

where the variances ψ^2_a, ψ^2_b and covariance ψ_{ab} can be estimated directly from the available observations [3] under the homoscedasticity assumption. The estimated variance of the sum of mean differences is

$$\hat{\sigma}^2_S = \hat{V}(\hat{\delta}_a + \hat{\delta}_b) = \hat{\sigma}^2_a + \hat{\sigma}^2_b + 2\hat{\sigma}_{ab}. \tag{13}$$

These then provide the test statistic Z_S in (5), or $Z_{S,L}$ in (10) if Frick's condition is not satisfied.

Standardized Score Test for Multiple Means

For an analysis of the means of quantitative variables, the Wei-Lachin test Z_S is not invariant to a change of scale for either of the two measures. In cases where there is a mixture of quantitative variables with different dispersions or units, such as LDL measured in mg/dl and systolic blood pressure measured in mm Hg, it is more meaningful to compute a scale-invariant test using the average of the corresponding standardized differences. This might also be preferred when the variances of the measures differ substantially, even though measured on the same scale.

Let Y_{ij} denote the standardized value $Y_{ij}=X_{ij}/\psi_j$ with $V(Y_{ij})=1$. Then the standardized difference between groups for the jth outcome is

$$\hat{\delta}_{Yj}=\bar{Y}_{Cj}-\bar{Y}_{Ej}=(\bar{X}_{Cj}-\bar{X}_{Ej})/\hat{\psi}_j=\hat{\delta}_j/\hat{\psi}_j \qquad (14)$$

where $\hat{\Delta}_Y=(\hat{\delta}_{Ya}\ \hat{\delta}_{Yb})'$ is asymptotically normally distributed with expectation $(\delta_a/\psi_a\ \delta_b/\psi_b)'$ and covariance matrix

$$\Sigma_Y=\begin{bmatrix} \dfrac{1}{n_{Ea}}+\dfrac{1}{n_{Ca}} & \rho_{ab}\left(\dfrac{n_{Eab}}{n_{Ea}n_{Eb}}+\dfrac{n_{Cab}}{n_{Ca}n_{Cb}}\right) \\ \rho_{ab}\left(\dfrac{n_{Eab}}{n_{Ea}n_{Eb}}+\dfrac{n_{Cab}}{n_{Ca}n_{Cb}}\right) & \dfrac{1}{n_{Eb}}+\dfrac{1}{n_{Cb}} \end{bmatrix}. \qquad (15)$$

The resulting standardized Wei-Lachin test is then provided by

$$Z_{S,Y}=\frac{\mathbf{J}'\hat{\Delta}_Y}{\sqrt{\mathbf{J}'\hat{\Sigma}_Y\mathbf{J}}}=\frac{\hat{\delta}_a/\hat{\psi}_a+\hat{\delta}_b/\hat{\psi}_b}{\hat{\sigma}_{S,Y}} \qquad (16)$$

where

$$\sigma^2_{S,Y}=\left(\frac{1}{n_{Ea}}+\frac{1}{n_{Ca}}\right)+\left(\frac{1}{n_{Eb}}+\frac{1}{n_{Cb}}\right)+ 2\rho_{ab}\left(\frac{n_{Eab}}{n_{Ea}n_{Eb}}+\frac{n_{Cab}}{n_{Ca}n_{Cb}}\right) \qquad (17)$$

that is consistently estimated from the estimate of the correlation $\hat{\rho}_{ab}$. When the variances of the outcomes are equal $(\hat{\psi}_a=\hat{\psi}_b)$, then $Z_{S,Y}=Z_S$. With equal sample sizes and no missing values, $n_{ia}=n_{ib}=n_{iab}=n=N/2$, $(i=E,C)$, then

$$Z_{S,Y}=\frac{\sqrt{N}\left[\hat{\delta}_a/\hat{\psi}_a+\hat{\delta}_b/\hat{\psi}_b\right]}{\sqrt{8(1+\hat{\rho}_{ab})}}. \qquad (18)$$

As above, with positive correlations, Frick's condition $\Sigma_Y\mathbf{J}\geq 0$ is satisfied. If not, then the weighted test is provided by $Z_{S,L}$ using $\hat{\Delta}_Y=(\hat{\delta}_a/\hat{\psi}_a\ \ \hat{\delta}_b/\hat{\psi}_b)'$ and $\hat{\Sigma}_Y$ in (10).

Z-Based Test

In some cases, it may be desired to conduct a test with mixtures of quantitative and qualitative outcomes (or other types), e.g. combining tests for means, proportions and/or life-times. In such cases a multivariate one-directional test with respect to the multiple outcomes can be obtained from a combination of the individual Z-test values of the form

$$Z_{S,z}=\frac{z_a+z_b}{\sqrt{2+2Cov(z_a,z_b)}} \qquad (19)$$

where $z_j=\hat{\delta}_j/\hat{\sigma}_j$ and the covariance matrix of the Z-tests (Σ_z) has variances $V(z_j)=1$ $(j=a,b)$ and $Cov(z_a,z_b)=Corr(\hat{\delta}_a,\hat{\delta}_b)=\dfrac{\sigma_{ab}}{\sigma_a\sigma_b}$ with elements from (12). If $n_{ia}=n_{ib}=n_{iab}$ for $i=E,C$ then $Cov(z_a,z_b)=\rho_{ab}$.

Under the alternative hypothesis where the components $\{\hat{\delta}_j\}$ or $\{z_j\}$ are expected to be positive, then the covariance will likewise be expected to be positive and Frick's condition $\Sigma_z\mathbf{J}\geq 0$ is readily satisfied. If this condition is not be satisfied, we would use the test $Z_{S,L}$ using $\mathbf{Z}=(z_a z_b)'$ and $\hat{\Sigma}_z$ in lieu of $\hat{\Delta}$ and $\hat{\Sigma}$ in (10).

It should be noted that this Z-based test is analogous to the Gastwirth [14] miximin efficient robust test (MERT) that is a obtained using the sum of the extreme Z-tests from a set of tests against a closed family of alternatives. For a family with only 2 alternatives (or tests), the MERT is equivalent to the above Z-based test.

Comparison of the Tests for Means

When the variances are equal $(\hat{\psi}_a=\hat{\psi}_b)$, it can readily be shown that the standardized scores test equals the scale-based test $(Z_S=Z_{S,Y})$ regardless of the sample sizes or sample fractions. When the group sample sizes are equal with no missing values, it can also be shown that the standardized scores test equals the Z-based test $(Z_{S,Y}=Z_{S,Z})$. When both the variances and sample sizes are equal, then all three tests are equal.

Direct computation of the three tests $(Z_S,\ Z_{S,Y},\ Z_{S,z})$ over a range of sample sizes, variances and group differences shows that $\bar{Z}_{S,z}\overset{1.009}{>}\bar{Z}_{S,Y}\overset{1.032}{>}\bar{Z}_S$, i.e. with given proportionalities. Thus, $Z_{S,Y}$ and $Z_{S,z}$ are virtually equivalent with $corr(Z_{S,Y},Z_{S,z})=0.988$ over the range of alternatives considered. These two tests are about 3% greater than the scale-based test with respective correlations of 0.977 and 0.953. Thus, on this basis the standardized scores or Z-based test would appear to be preferable.

General Expressions for Power and Sample Size for the Tests

For each variation of the test, expressions for the evaluation of sample size and power are readily obtained. Under H_{1S} with specified values (δ_a,δ_b), let $\sigma^2_S=V(\hat{\delta}_a+\hat{\delta}_b)$ that may be a function of (δ_a,δ_b) depending on the underlying model. Also, let $\sigma^2_S=\phi^2_S/N$ represent the factorization of this variance into a term ϕ^2_S and N. Therefore, from standard equations [15], the power of the test to reject H_{1S} for specified values (δ_a,δ_b) is provided by $1-\beta=\Phi(Z_{1-\beta})$ where

$$Z_{1-\beta}=\frac{\sqrt{N}(\delta_a+\delta_b)}{\phi_S}-Z_{1-\alpha} \qquad (20)$$

and where the variance σ^2_S is factored as

$$\sigma^2_S=\frac{\phi^2_S}{N}=\left[\frac{\phi^2_a+\phi^2_b+2\phi_{ab}}{N}\right] \qquad (21)$$

and the individual variances and the covariance are factored as $\sigma^2_a=\phi^2_a/N$, $\sigma^2_b=\phi^2_b/N$, and $\sigma_{ab}=\phi_{ab}/N$. Specific expressions are presented below. Conversely, the sample size required to provide power $1-\beta$ to detect specified values (δ_a,δ_b) is provided by

$$N=\left[\frac{(Z_{1-\alpha}+Z_{1-\beta})\phi_S}{\delta_a+\delta_b}\right]^2. \tag{22}$$

To evaluate these equations, is it necessary to provide the components of ϕ_S^2, i.e. $(\phi_a^2,\ \phi_b^2,\ \phi_{ab})$, and to specify the values (δ_a,δ_b) representing the minimal degree of superiority of treatment both outcomes of clinical interest.

For the standardized scores test in (16) the variance is likewise factored as $\sigma_{S,Y}^2=\phi_{S,Y}^2/N$. Then power is obtained from

$$Z_{1-\beta}=\frac{\sqrt{N}(\delta_a/\psi_a+\delta_b/\psi_b)}{\phi_{S,Y}}-Z_{1-\alpha} \tag{23}$$

and the required sample size from

$$N=\left[\frac{(Z_{1-\alpha}+Z_{1-\beta})\phi_{S,Y}}{\delta_a/\psi_a+\delta_b/\psi_b}\right]^2. \tag{24}$$

Likewise, for the Z-based test in (19), power is obtained from

$$Z_{1-\beta}=\frac{\sqrt{N}(\delta_a/\phi_a+\delta_b/\phi_b)}{\sqrt{2+2Corr(\hat{\delta}_a,\hat{\delta}_b)}}-Z_{1-\alpha} \tag{25}$$

and the required sample size from

$$N=\left[\frac{(Z_{1-\alpha}+Z_{1-\beta})\sqrt{2+2Corr(\hat{\delta}_a,\hat{\delta}_b)}}{\delta_a/\phi_a+\delta_b/\phi_b}\right]^2 \tag{26}$$

where $Corr(\hat{\delta}_a,\hat{\delta}_b)=\phi_{ab}/(\phi_a\phi_b)$. Expressions for the correlation are provided below for specific cases.

Also, each of the above expressions for power can be expressed as $E(Z)=Z_{1-\alpha}+Z_{1-\beta}$ where $E(Z)$ is also termed the non-centrality parameter of the test. Thus, the first term on the right hand side of (20), (23) and (25) is the respective expression for $E(Z)$.

Sample Size and Power for Tests for Means

To assess sample size and power for a test, let $E(n_{ia}\ n_{ib}\ n_{iab})=N(\xi_{ia}\ \xi_{ib}\ \xi_{iab})$ denote the expected numbers observed in the ith group, where N is the total sample size in the two groups with at least one observed measurement (not including any subject missing both A and B measurements).

The Scale-Based Test

From (12), the covariance matrix $Cov(\hat{\delta}_a\ \hat{\delta}_b)$ can be factored as $\Sigma=\Omega/N$ where

$$\Omega=\begin{bmatrix}\psi_a^2\left(\frac{1}{\xi_{Ea}}+\frac{1}{\xi_{Ca}}\right) & \psi_{ab}\left(\frac{\xi_{Eab}}{\xi_{Ea}\xi_{Eb}}+\frac{\xi_{Cab}}{\xi_{Ca}\xi_{Cb}}\right)\\ \psi_{ab}\left(\frac{\xi_{Eab}}{\xi_{Ea}\xi_{Eb}}+\frac{\xi_{Cab}}{\xi_{Ca}\xi_{Cb}}\right) & \psi_b^2\left(\frac{1}{\xi_{Eb}}+\frac{1}{\xi_{Cb}}\right)\end{bmatrix} \tag{27}$$
$$=\begin{bmatrix}\phi_a^2 & \phi_{ab}\\ \phi_{ab} & \phi_b^2\end{bmatrix}$$

and $\sigma_S^2=\phi_S^2/N$ where

$$\phi_S^2=\left[\frac{\psi_a^2(\xi_{Ea}+\xi_{Ca})}{\xi_{Ea}\xi_{Ca}}+\frac{\psi_b^2(\xi_{Eb}+\xi_{Cb})}{\xi_{Eb}\xi_{Cb}}+\right.$$
$$\left.2\psi_{ab}\left(\frac{\xi_{Eab}}{\xi_{Ea}\xi_{Eb}}+\frac{\xi_{Cab}}{\xi_{Ca}\xi_{Cb}}\right)\right]. \tag{28}$$

When the groups are of equal size with the same fractions observed $(\xi_{ia}\ \xi_{ib}\ \xi_{iab})=(\xi_a\ \xi_b\ \xi_{ab})$ for $i=E,C$, then

$$\phi_S^2=2\left[\frac{\psi_a^2}{\xi_a}+\frac{\psi_b^2}{\xi_b}+\left(\frac{2\psi_{ab}\xi_{ab}}{\xi_a\xi_b}\right)\right] \tag{29}$$

When there are equal-sized groups with no missing observations then $\xi_a=\xi_b=\xi_{ab}=0.5$ and

$$\phi_S^2=4\left[\psi_a^2+\psi_b^2+2\psi_{ab}\right]. \tag{30}$$

Then the power or sample size required to detect specified values δ_a and δ_b are provided by (20) or (22), respectively.

For example, suppose we desire to test the treatment group differences in both systolic (A) and diastolic (B) blood pressures, lower values of each being better. From existing data the respective SDs are $\psi_a=13$ mm Hg and $\psi_b=7$ mm Hg. The correlation of the two is $\rho_{ab}=0.6$ which yields $\psi_{ab}=(0.6)(13)(7)=54.6$. Assume that we wish to detect a treatment group difference equal to 0.25 SD for each measure, so that $\delta_a=(0.25)(13)=3.25$and $\delta_b=(0.25)(7)=1.75$. For equal-sized groups with no missing observations then $\xi_a=\xi_b=\xi_{ab}=0.5$ and $\phi_S^2=4[13^2+7^2+2(54.6)]=1308.8$. For a one-sided test at the 0.05 level, the sample size required to provide power of at least 0.9 is provided by

$$N=\left[\frac{(1.645+1.282)\sqrt{1308.8}}{3.25+1.75}\right]^2=448.52 \tag{31}$$

or 225 subjects per group (rounded up). From equation (20), with this sample size the power to detect smaller differences of 0.2 SD with $\delta_a=2.6$ and $\delta_b=1.4$, then the power using $N=450$ is provided by

$$Z_{1-\beta}=\left[\frac{\sqrt{450}(2.6+1.4)}{\sqrt{1308.8}}-1.645\right]=0.700 \tag{32}$$

with power $\Phi(0.7)=0.758$. Below we also examine the power for this example using the other tests.

The Test Using Standardized Means or Z-Scores

When the component measurements have different units or scales of measurement, then either the test based on the standardized values or the individual Z-tests is invariant to scale transformations, and, therefore, preferred. This test may also be preferred when the component measures have different variances, even when measured on the same scale.

For the standardized-scores test, from (16),

$$\phi^2_{S,y} = \frac{(\xi_{Ea}+\xi_{Ca})}{\xi_{Ea}\xi_{Ca}} + \frac{(\xi_{Eb}+\xi_{Cb})}{\xi_{Eb}\xi_{Cb}} + 2\rho_{ab}\left(\frac{\xi_{Eab}}{\xi_{Ea}\xi_{Eb}} + \frac{\xi_{Cab}}{\xi_{Ca}\xi_{Cb}}\right). \quad (33)$$

When there are equal-sized groups with no missing observations (all $\{\xi\}=0.5$) then $\phi^2_{S,Y}=8(1+\rho_{ab})$. Power and sample size are then obtained from (23) and (24).

For the above example, with equal sample sizes and no missing data, then $corr(\hat{\delta}_a,\hat{\delta}_b)=corr(X_a,X_b)=\rho_{ab}=0.6$. Since the difference is specified as a fraction of the standard deviation, $\delta_a=(0.25)\psi_a$ and $\delta_b=(0.25)\psi_b$, then $\delta_a/\psi_a=\delta_b/\psi_b=0.25$ and the required sample size is

$$N = \left[\frac{(1.645+1.282)\sqrt{8(1+0.6)}}{2(0.25)}\right]^2 = 438.65 \quad (34)$$

that is slightly less than the N required for the scale-based test. This indicates that for this example, the test based on standardized scores would have greater power for a given N.

The same numerical result also is obtained using the Z-based test since in this case the two tests are equal.

Relative Efficiency Versus Other Tests

It is also instructive to compare the efficiency of the Wei-Lachin test versus two one-sided tests or an omnibus test. We do so here in the context of a test for means, and these results apply in general to other tests as well. Standard methods for the evaluation of the asymptotic relative (Pitman) efficiency (ARE) of two tests under a local alternative would not account for the necessary adjustment to the significance level for two tests. However, the ARE can be interpreted as the ratio of sample sizes needed to provide the same level of power for a specific alternative. This ratio of sample sizes can be derived directly from (22) relative to the like expression for either two separate tests or the omnibus test.

Pairwise Tests. Consider the power of the test for means with equal group sample sizes and residual variance ψ^2_j for the jth outcome where each is measured on the same scale so that the original scale-based test is appropriate. For a given alternative $(\delta_a>0,\delta_b>0)$. For two tests with equal-sized groups, each being of size $N/2$, with no missing data $(\xi_{ia}=\xi_{ib}=\xi_{iab}=1/2)$, the variance of the difference for the jth outcome is

$$V(\hat{\delta}_j)=4\psi^2_j/N \quad (35)$$

assuming homoscedasticity. Then the equivalent expression for the total sample size required based on the separate tests is provided by

$$N_P = \max\left\{\left[\frac{(Z_{1-\alpha/2}+Z_{1-\beta})2\psi_a}{\delta_a}\right]^2, \left[\frac{(Z_{1-\alpha/2}+Z_{1-\beta})2\psi_b}{\delta_b}\right]^2\right\} \quad (36)$$

using the Bonferroni correction for 2 one-sided tests. To simplify, assume that the differences of interest are a common fraction v of the standard deviations, i.e. $\delta_a=v\psi_a$ and $\delta_b=v\psi_b$ in which case

$$N_P = \left[\frac{2(Z_{1-\alpha/2}+Z_{1-\beta})}{v}\right]^2 \quad (37)$$

Let N_S denote the total sample size required for the Wei-Lachin test as obtained from (22) with the value ϕ^2_S that is obtained from (30) to yield

$$\begin{aligned} N_S &= \left[\frac{(Z_{1-\alpha}+Z_{1-\beta})}{\delta_a+\delta_b}\right]^2 4\left[\psi^2_a+\psi^2_b+2\psi_{ab}\right] \\ &= \left[\frac{(Z_{1-\alpha}+Z_{1-\beta})}{v[\psi_a+\psi_b]}\right]^2 4\left[\psi^2_a+\psi^2_b+2\psi_{ab}\right]. \end{aligned} \quad (38)$$

Thus, the ratio of sample sizes needed with the two-pairwise one-sided tests versus the Wei-Lachin test is

$$\begin{aligned} \frac{N_P}{N_S} &= \left[\frac{2\left(Z_{1-\alpha/2}+Z_{1-\beta}\right)}{Z_{1-\alpha}+Z_{1-\beta}}\right]^2 \left[\frac{(v[\psi_a+\psi_b])^2}{4v^2\left[\psi^2_a+\psi^2_b+2\psi_{ab}\right]}\right] \\ &= \left[\frac{\left(Z_{1-\alpha/2}+Z_{1-\beta}\right)}{Z_{1-\alpha}+Z_{1-\beta}}\right]^2 \left[\frac{\psi^2_a+\psi^2_b+2\psi_a\psi_b}{\psi^2_a+\psi^2_b+2\psi_{ab}}\right] \end{aligned} \quad (39)$$

Since $Z_{1-\alpha/2}>Z_{1-\alpha}$ and $\psi_a\psi_b\geq\psi_{ab}$, then $N_P>N_S$.

For example, consider a one-sided test at the 0.05 level (0.025 adjusted for two tests) with 90% power to detect an improvement E versus C at any level v. Assume a correlation among the A and B measures of 0.5 and variances $\psi^2_a=\psi^2_b = 1$. Then the ratio of sample sizes is

$$\frac{N_P}{N_S} = \left[\frac{1.96+1.282}{1.645+1.282}\right]^2 \left[\frac{4}{2+2(0.5)}\right] = 1.64 \quad (40)$$

which indicates that two separate tests requires a 64% greater sample size than does the Wei-Lachin test for this α and β, or that the Wei-Lachin test is 39% more efficient. These results apply approximately to other tests such as the test for proportions or the test of life-times.

The Omnibus MANOVA Test. Similarly, the omnibus multivariate analysis of variance (MANOVA) T^2-like test of H_0 versus the general alternative H_{1O} in (7) is provided by $X^2=(\hat{\delta}_a\ \hat{\delta}_b)\hat{\Sigma}^{-1}(\hat{\delta}_a\ \hat{\delta}_b)'$ that is asymptotically distributed as chi-square on 2 df. The corresponding non-centrality parameter is

$$\begin{aligned} \theta^2 &= (\delta_a\delta_b)\Sigma^{-1}(\delta_a\delta_b)' \\ &= (v\psi_a v\psi_b)\Sigma^{-1}(v\psi_a v\psi_b)' \end{aligned} \quad (41)$$

where the inverse covariance matrix is

$$\Sigma^{-1} = \frac{N}{4\left(\psi^2_a\psi^2_b-\psi^2_{ab}\right)}\begin{bmatrix} \psi^2_b & -\psi_{ab} \\ -\psi_{ab} & \psi^2_a \end{bmatrix}. \quad (42)$$

Thus

$$\theta^2=\frac{N_O}{2}\left[\frac{v^2\left(\psi_a^2\psi_b^2-\psi_a\psi_b\psi_{ab}\right)}{\psi_a^2\psi_b^2-\psi_{ab}^2}\right]=N_O\phi^2. \tag{43}$$

The non-centrality parameter for a test at level α on K df that provides power $1-\beta$, designated as $\theta^2(\alpha,\beta,K)$, is readily obtained, such as from the SAS function CNONCT. Then the required sample size is provided by

$$N_O=\theta^2(\alpha,\beta,df)/\phi^2 \tag{44}$$

For the above example, $\theta^2(0.05{,}0.10{,}2)=12.654$ and

$$N_O=\frac{(12.654)(2)(\psi_a^2\psi_b^2-\psi_{ab}^2)}{v^2\left(\psi_a^2\psi_b^2-\psi_a\psi_b\psi_{ab}\right)}. \tag{45}$$

Then, for the above example, the inverse efficiency relative to the Wei-Lachin test is provided by the ratio of N_O to N_S in (38) to yield

$$\frac{N_O}{N_S}=\frac{\dfrac{(12.654)(2)(1-0.25)}{(1-0.5)}}{\left[\dfrac{1.645+1.282}{2}\right]^2(12)}=1.477 \tag{46}$$

and the Wei-Lachin test is 32% more efficient for these operating characteristics. If the computation is conducted for a two-sized Wei-Lachin test, then N$_O/N_S=1.204$ and the Wei-Lachin test is 17% more efficient.

Power of Tests for Multiple Proportions, and Mixtures of Proportions and Means

Test for Multiple Proportions

Now consider a large sample test for a difference between groups in the probabilities (π_{ij}) of two Bernoulli variables X_a and X_b where the corresponding sample proportions are distributed as $p_{ij}\sim N(\pi_{ij},\psi_{ij}^2/n_{ij})$ with Bernoulli variance $\psi_{ij}^2=\pi_{ij}(1-\pi_{ij})$ for the jth outcome within the ith group and sample sizes $n_{ij}=N\xi_{ij}$, $i=E,C$; $j=a,b$. The covariance of the Bernoulli variables within the ith group, $Cov(X_{ia},X_{ib})$, is simply

$$\psi_{iab}=E(X_{ia}X_{ib})-E(X_{ia})E(X_{ib})=\pi_{iab}-\pi_{ia}\pi_{ib} \tag{47}$$

where π_{iab} is the probability that both variables are positive [7]. Again we assume that a lower probability is better. If not, the (0, 1) categories should be reversed.

Then $\hat{\delta}_a=(p_{Ca}-p_{Ea})$ and $\hat{\delta}_b=(p_{Cb}-p_{Eb})$ and $\hat{\Delta}=(\hat{\delta}_a\ \hat{\delta}_b)'$ is asymptotically distributed as in (3) with expectation $\Delta=(\delta_a\ \delta_b)'$ where $\delta_j=(\pi_{Cj}-\pi_{Ej})$ and with covariance matrix

$$\Sigma=\begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}=$$
$$\begin{bmatrix}\dfrac{\psi_{Ea}^2}{n_{Ea}}+\dfrac{\psi_{Ca}^2}{n_{Ca}} & \dfrac{\psi_{Eab}n_{Eab}}{n_{Ea}n_{Eb}}+\dfrac{\psi_{Cab}n_{Cab}}{n_{Ca}n_{Cb}}\\ \dfrac{\psi_{Eab}n_{Eab}}{n_{Ea}n_{Eb}}+\dfrac{\psi_{Cab}n_{Cab}}{n_{Ca}n_{Cb}} & \dfrac{\psi_{Eb}^2}{n_{Eb}}+\dfrac{\psi_{Cb}^2}{n_{Cb}}\end{bmatrix} \tag{48}$$

that is consistently estimable from the sample quantities [7]. Then the statistic Z_S is constructed as in (5) based on the sample estimate of the variance σ_S^2 as in (13). Note that in this case, since all measures are based on Bernoulli variables, there is no advantage to using the test based on standardized scores. Alternately, the Z-based test would be constructed as in (19) with $\widehat{Corr}(\hat{\delta}_a,\hat{\delta}_b)=\hat{\sigma}_{ab}/(\hat{\sigma}_a\hat{\sigma}_b)$.

For the assessment of sample size or power the covariance would be factored as $\Sigma=\Omega/N$ with terms $(\phi_a^2,\phi_b^2,\phi_{ab}^2)$ and where $Corr(\hat{\delta}_a,\hat{\delta}_b)=\phi_{ab}/(\phi_a\phi_b)$.

For example, assume that the outcomes in the control group are expected to have probabilities $\pi_{Ca}=\pi_{Cb}=0.4$ with joint probability $\pi_{Cab}=0.2$ and that the respective probabilities in the experimental group are $\pi_{Ea}=\pi_{Eb}=0.3$ with joint probability $\pi_{Eab}=0.15$. Then $\psi_{Ea}^2=\psi_{Eb}^2=(0.3)(0.7)$, $\psi_{Ca}^2=\psi_{Cb}^2=(0.4)(0.6)$, $\psi_{Eab}=(0.15-0.3^2)$, and $\psi_{Cab}=(0.20-0.4^2)$. With equal sized groups and no missing observations, then $\xi_{ia}=\xi_{ib}=\xi_{iab}=1/2$ $(i=E,C)$ and

$$\phi_s^2=$$
$$4\left[(0.3)(0.7)+(0.4)(0.6)+(0.15-0.3^2)+(0.20-0.4^2)\right]=2.20 \tag{49}$$

with the resulting computation

$$N=\left[\frac{(1.645+1.282)\sqrt{2.2}}{2(0.1)}\right]^2=471.2. \tag{50}$$

The correlation of the estimates is

$$\begin{aligned}Corr(\hat{\delta}_a,\hat{\delta}_b)&=\frac{\sqrt{2}(\psi_{Eab}+\psi_{Cab})}{\sqrt{\psi_{Ea}^2+\psi_{Ca}^2}+\sqrt{\psi_{Eb}^2+\psi_{Cb}^2}}\\ &=\frac{\sqrt{2}\left[(0.15-0.3^2)+(0.20-0.4^2)\right]}{2\sqrt{(0.3)(0.7)+(0.4)(0.6)}}=0.10541.\end{aligned} \tag{51}$$

Then the test based on Z-test values would require

$$N=\left[\frac{\dfrac{(1.645+1.282)\sqrt{2+0.10541}}{0.1}}{2\dfrac{}{\sqrt{2}\sqrt{(0.3)(0.7)+(0.4)(0.6)}}}\right]^2=405.85. \tag{52}$$

Thus, the Z-based test is again more efficient than the scale-based test.

Tests for Means and Proportions

Scale-Based Test. It is also possible to determine the joint distribution of a test for means of one outcome and a test for proportions of another. Let X_A denote a quantitative measurement with means μ_{ia} and variance ψ_a^2, assuming homoscedasticity, and X_{ib} denote a binary variable with probability π_{ib} and variance $\psi_{ib}^2=\pi_{ib}(1-\pi_{ib})$ in the ith group ($i=E,C$). The covariance of the two in the ith group is provided by

$$\begin{aligned}\psi_{iab}&=Cov(X_{ia},X_{ib})=E(X_{ia}X_{ib})-E(X_{ia})E(X_{ib})\\&=\pi_{ib}(\mu_{ia(1)}-\mu_{ia})\end{aligned}\tag{53}$$

where $\mu_{ia(1)}=E(X_{ia}|X_{ib}=1)$ is the mean of the quantitative variable X_{ia} among those where the binary variable $X_{ib}=1$. Then $\hat{\delta}_a=(\bar{X}_{Ca}-\bar{X}_{Ea})$ with variance $\sigma_a^2=\psi_a^2\left(\frac{1}{n_{Ea}}+\frac{1}{n_{Ca}}\right)$, assuming homoscedasticity, and $\hat{\delta}_b=(p_{Cb}-p_{Eb})$ with variance $\sigma_b^2=\psi_{Eb}^2/n_{Eb}+\psi_{Cb}^2/n_{Cb}$. The covariance is

$$\begin{aligned}\sigma_{ab}&=Cov(\hat{\delta}_a\ \hat{\delta}_b)=Cov(\bar{X}_{Ea},p_{Ea})+Cov(\bar{X}_{Ca},p_{Cb})\\&=\frac{Cov(X_{Ea},X_{Eb})n_{Eab}}{n_{Ea}n_{Eb}}+\frac{Cov(X_{Ca},X_{Cb})n_{Cab}}{n_{Ca}n_{Cb}}\\&=\frac{\psi_{Eab}n_{Eab}}{n_{Ea}n_{Eb}}+\frac{\psi_{Cab}n_{Cab}}{n_{Ca}n_{Cb}}.\end{aligned}\tag{54}$$

To conduct the test these variances and covariances can be estimated consistently from the corresponding sample estimates. Sample size and power can then be evaluated as above.

For example, assume that we wish to test the difference between groups in the mean level of LDL and the prevalence of hypertension. Assume a SD $\psi_a=20$ in both groups and that the difference of interest is $\delta_a=5$ that corresponds to a 0.25 SD difference. While it is not necessary to specify the actual mean values within each group to compute δ_a, it is necessary to compute the covariance. Within each group assume that the overall mean values are $\mu_{Ea}=170$ and $\mu_{Ca}=175$ (corresponding to $\delta_a=5$), and a greater treatment effect among those who are hypertensive with mean values $\mu_{Ea(1)}=175$ and $\mu_{Ca(1)}=185$. Assume that the probabilities of being hypertensive are $\pi_{Eb}=0.60$ and $\pi_{Cb}=0.70$, yielding $\delta_b=0.1$. Then the variance components are $\psi_{Eb}^2=(0.6)(0.4)=0.24$, $\psi_{Cb}^2=(0.7)(0.3)=0.21$, $\psi_{Eab}=(0.6)(175-170)=3.0$, and $\psi_{Cab}=(0.7)(185-175)=7$.

Assuming equal sized groups with no missing data, then $\hat{\Delta}=(\hat{\delta}_a\ \hat{\delta}_b)'$ is asymptotically normally distributed with covariance matrix $\Sigma=\Omega/N$ and

$$\begin{aligned}\Omega&=\begin{bmatrix}\phi_a^2 & \phi_{ab}\\ \phi_{ab} & \phi_b^2\end{bmatrix}=\begin{bmatrix}4\psi_a^2 & 2(\psi_{Eab}+\psi_{Cab})\\ 2(\psi_{Eab}+\psi_{Cab}) & 2(\psi_{Eb}^2+\psi_{Cb}^2)\end{bmatrix}\\&=\begin{bmatrix}4(20)^2 & 2(3+7)\\ 2(3+7) & 2[0.24+0.21]\end{bmatrix}=\begin{bmatrix}1600 & 20\\ 20 & 0.9\end{bmatrix}\end{aligned}\tag{55}$$

and $\sigma_S^2=1600.9+2(20)=1640.9$. Thus, the required sample size for a one-sided test at the 0.05 level and 90% power is provided by

$$N=\left[\frac{(1.645+1.282)\sqrt{1640.9}}{5.1}\right]^2=540.5.\tag{56}$$

Z-Based Test. Alternately, since the scale-based test is not invariant under transformations, it would be more appropriate to employ a combination of the Z-tests. In this case,

$$\begin{aligned}\frac{\delta_a}{\phi_a}&=\frac{\delta_a}{\psi_a\sqrt{\frac{1}{\xi_{Ea}}+\frac{1}{\xi_{Ca}}}};\\ \frac{\delta_b}{\phi_b}&=\frac{\delta_b}{\sqrt{\frac{\psi_{Eb}^2}{\xi_{Eb}}+\frac{\psi_{Cb}^2}{\xi_{Cb}}}}=\frac{\pi_{Cb}-\pi_{Eb}}{\sqrt{\frac{\pi_{Eb}(1-\pi_{Eb})}{\xi_{Eb}}+\frac{\pi_{Cb}(1-\pi_{Cb})}{\xi_{Cb}}}}\end{aligned}$$

$$Cov(Z_a,Z_b)=Corr(\hat{\delta}_a\ \hat{\delta}_b)=\frac{\phi_{ab}}{\phi_a\phi_b}=\frac{\frac{\psi_{Eab}\xi_{Eab}}{\xi_{Ea}\xi_{Eb}}+\frac{\psi_{Cab}\xi_{Cab}}{\xi_{Ca}\xi_{Cb}}}{\psi_a\sqrt{\frac{1}{\xi_{Ea}}+\frac{1}{\xi_{Ca}}}\sqrt{\frac{\psi_{Eb}^2}{\xi_{Eb}}+\frac{\psi_{Cb}^2}{\xi_{Cb}}}}\tag{57}$$

When there are equal sample sizes between groups with no missing data for either measure then

$$\begin{aligned}\frac{\delta_a}{\phi_a}&=\frac{\delta_a}{2\psi_a};\\ \frac{\delta_b}{\phi_b}&=\frac{\pi_{Cb}-\pi_{Eb}}{\sqrt{2\psi_{Eb}^2+2\psi_{Cb}^2}}=\frac{\pi_{Cb}-\pi_{Eb}}{\sqrt{2\pi_{Eb}(1-\pi_{Eb})+2\pi_{Cb}(1-\pi_{Cb})}};\\ Cov(Z_a,Z_b)&=\frac{2(\psi_{Eab}+\psi_{Cab})}{2\psi_a\sqrt{2(\psi_{Eb}^2+\psi_{Cb}^2)}}.\end{aligned}\tag{58}$$

Then for this example

$$\frac{\delta_a}{\phi_a}=\frac{5}{2(20)}=0.125;\qquad \frac{\delta_b}{\phi_b}=\frac{0.1}{\sqrt{2[0.24+0.21]}}=0.1054\tag{59}$$

$$Cov(Z_a,Z_b)=\frac{2(3+7)}{2(20)\sqrt{2(0.24+0.21)}}=0.5271$$

and

$$N=\left[\frac{(1.645+1.282)\sqrt{2+2(0.5271)}}{0.125+0.1054}\right]^2=492.9.\tag{60}$$

Thus, the Z-based test would provide greater power in this case.

Power of Tests for Multiple Event-Times

Tests for Multiple Event-times

For right censored event time data, a member of the family of Aalen-Gill tests [16,17], also known as the G^ρ family of tests of Harrington and Fleming [18], can be used to test the hypothesis of equal hazard functions, or survival functions, between two groups. This family includes the logrank test that is asymptotically fully efficient under a proportional hazards model and is equivalent to the score test of the unadjusted group effect in a Cox Proportional Hazards model. It also includes the Peto-Peto-Prentice modified Wilcoxon test that is optimal under a survival proportional odds model. Andersen, Borgan, Gill and Keiding [19] describe a generalization of the tests for $K>2$ groups. These tests are equivalent to the family of weighted Mantel-Haenszel statistics described by Kalbfleisch and Prentice [20].

Wei and Lachin [2] describe a multivariate rank test for event times that is a generalization of the above families of tests to the case of multiple time-to-event outcomes. They also introduced the one-directional multivariate test described herein, what they termed the test of stochastic ordering, to assess whether the treatment group event times differed in a favorable direction for all of the outcomes. A SAS macro for these computations is available (see discussion). The computational details will not be provided herein.

Lakatos [21] presents a general approach to the evaluation of sample size and power for the Mantel-logrank test that allows for time varying hazard rates, proportional or non-proportional hazards, and other design features. When the hazard rates are assumed constant over time with a constant of proportionality, a simple exponential model applies in which case the methods of Rubenstein et al. [22] or Lachin and Foulkes [23] can be applied. Herein we describe the computation of sample size or power for the Wei-Lachin test for multiple event-time outcomes under the exponential model of Lachin and Foulkes that includes a generalization of the method described by Lachin [15] based on the difference in the exponential hazard rates. Freedman [24] showed that the latter expression can also be derived from the expected value of the logrank chi-square test value under a proportional hazards model. Lachin and Foulkes [23] also show that the power of the test based on the difference in the estimated hazards is virtually identical to that for a test based on the log hazard ratio.

We assume that there are two or more outcome events where no one outcome is a competing risk for the other outcomes, such as the time to development of diabetic retinopathy and time to developing diabetic nephropathy, neither of which is fatal. Let $X_{ijk}=1$ denote that the kth subject had the jth event in the ith group at time t_{ijk}, and $X_{ijk}=0$ denote right censoring at time U_{ijk} that in turn is the minimum of the loss to follow-up time and the administrative censoring time for those who remain free of the jth outcome, $i=E,\ C;\ j=a,\ b$. Then the total number of subjects with an event (called events) (D_{ij}) and total time at risk (T_{ij}) for the ith group and the jth outcome are

$$D_{ij}=\sum\nolimits_k X_{ijk} \tag{61}$$

$$T_{ij}=\sum\nolimits_k \left[X_{ijk}t_{ijk}+(1-X_{ijk})U_{ijk}\right].$$

Note that the X_{ijk} are non-iid Bernoulli variables with event probabilities that are a function of the underlying hazard rates for the event and losses to follow-up and the period of exposure U_{ijk}.

Within each group, for each outcome assume a constant hazard rate λ_{ij} that is consistently estimated as $\hat{\lambda}_{ij}=D_{ij}/T_{ij}$. Let $E(D_{ij})$ designate the expected number of events based on the assumed hazard rate λ_{ij}, sample size, periods of recruitment and follow-up, and losses-to follow-up in that group. Asymptotically,

$$\hat{\lambda}_{ij}\sim\mathcal{N}(\lambda_{ij},v_{ij}^2) \tag{62}$$

where $v_{ij}^2=\lambda_{ij}^2/E(D_{ij})$ that is consistently estimated as $\hat{v}_{ij}^2=\hat{\lambda}_{ij}^2/D_{ij}$.

Then $\hat{\delta}_a=(\hat{\lambda}_{Ca}-\hat{\lambda}_{Ea})$, $\hat{\delta}_b=(\hat{\lambda}_{Cb}-\hat{\lambda}_{Eb})$ and, $\hat{\Delta}=(\hat{\delta}_a\ \hat{\delta}_b)'$ is asymptotically distributed as in (3) with expectations $\delta_a=(\lambda_{Ca}-\lambda_{Ea})$ and $\delta_b=(\lambda_{Cb}-\lambda_{Eb})$ and covariance matrix Σ with elements $V(\hat{\delta}_a),V(\hat{\delta}_b)$ and $Cov(\hat{\delta}_a\ \hat{\delta}_b)$. A test based on $\hat{\Delta}$ will have power approximately equal to that of the Wei-Lachin multivariate one-directional test using the Wei-Lachin bivariate Aalen-Gill logrank test under proportional hazards. Thus, we describe the power of the bivariate logrank test based on the test of the difference in exponential hazards. Then the scale-based test employs

$$\sigma_S^2=V(\hat{\delta}_a)+V(\hat{\delta}_b)-2Cov(\hat{\delta}_a\ \hat{\delta}_b) \tag{63}$$

$$V(\hat{\delta}_j)=v_{Ej}^2+v_{Cj}^2=\left[\frac{\lambda_{Ej}^2}{E(D_{Ej})}+\frac{\lambda_{Cj}^2}{E(D_{Cj})}\right]$$

that is consistently estimated using $\hat{\lambda}_{ij}$ and the observed D_{ij}, $j=a,b$.

File S1 shows that the covariance is expressed as

$$Cov(\hat{\delta}_a\ \hat{\delta}_b)=\frac{E[D_{Eab}]-E[D_{EabI}]}{E[T_{Ea}]E[T_{Eb}]}+\frac{E[D_{Cab}]-E[D_{CabI}]}{E[T_{Ca}]E[T_{Cb}]} \tag{64}$$

where D_{iab} is the number of subjects who experience both the A and B events and $E[D_{iabI}]$ is the expected number with both events under the assumption that the Bernoulli variables X_{iak} and X_{ibk} are independent. Each is consistently estimated from the observed numbers of events and total time at risk. The computational expression for D_{iabI} is also presented in File S1.

The resulting test as in (5) then is based on the variance estimate

$$\begin{aligned}\hat{\sigma}_S^2=&\frac{\hat{\lambda}_{Ea}^2}{D_{Ea}}+\frac{\hat{\lambda}_{Ca}^2}{D_{Ca}}+\frac{\hat{\lambda}_{Eb}^2}{D_{Eb}}+\frac{\hat{\lambda}_{Cb}^2}{D_{Cb}}-\\ &2\left[\frac{D_{Eab}-D_{EabI}}{T_{Ea}T_{Eb}}+\frac{D_{Cab}-D_{CabI}}{T_{Ca}T_{Cb}}\right]\end{aligned} \tag{65}$$

that is solely a function of the numbers of individual and joint events, the corresponding event times and the corresponding times of at risk. Accordingly, the power of the test is a function of the expected numbers of events and expected time at risk that in turn are a function of the design parameters and sample size.

Lachin and Foulkes [23] provide the expression for the probabilities of events $\{\pi_{ij}\}$ for given hazard rates for events $\{\lambda_{ij}\}$ and losses to follow-up $\{\eta_{ij}\}$, recruitment period R with recruitment shape parameter γ and total follow-up Q, and sample size $n_{ij}=N\xi_{ij}$. Then the expected number of events is obtained as

$E(D_{ij}) = N\xi_{ij}\pi_{ij}$ and likewise the expected period at risk as $E(T_{ij}) = N\xi_{ij}\tau_{ij}$. File S1 also provides expressions for $E(D_{iab})$, $E(D_{iabI})$ and $E(T_{ij})$. Then $\sigma_S^2 = \phi_S^2/N$ where

$$\begin{aligned}\phi_S^2 = \phi_a^2 + \phi_b^2 - 2\phi_{ab} = \frac{\lambda_{Ea}^2}{\xi_{Ea}\pi_{Ea}} + \frac{\lambda_{Ca}^2}{\xi_{Ca}\pi_{Ca}} + \frac{\lambda_{Eb}^2}{\xi_{Eb}\pi_{Eb}} + \frac{\lambda_{Cb}^2}{\xi_{Cb}\pi_{Cb}} \\ -2\left[\frac{\xi_{Eab}[\pi_{Eab} - \pi_{EabI}]}{\xi_{Ea}\tau_{Ea}\xi_{Eb}\tau_{Eb}} + \frac{\xi_{Cab}[\pi_{Cab} - \pi_{CabI}]}{\xi_{Ca}\tau_{Ca}\xi_{Cb}\tau_{Cb}}\right].\end{aligned} \tag{66}$$

Power and sample size are then obtained from (20) and (22).

However, to obtain an analytic solution to these equations, a specific model must be specified for the dependence of the event-times with a given correlation, such as the Marshall and Olkin [25] bivariate exponential model. Hougaard [26] provides a review of such models. Alternately, a simulation model could be implemented using a given bivariate exponential distribution. Herein, a simpler approach is described using a shared frailty.

Assume that the two event types share a common frailty with parameter λ_{iF}. Then in the simulation model, in the ith group, three random exponential times are generated as

$$t_1 \sim exponential(\lambda_{ia} - \lambda_{iF})$$

$$t_2 \sim exponential(\lambda_{ib} - \lambda_{iF})$$

$$t_3 \sim exponential(\lambda_{iF})$$

and the correlated exponential event times are then obtained as

$$t_{ia} = \min(t_1, t_3) \sim exponential(\lambda_{ia})$$

$$t_{ib} = \min(t_2, t_3) \sim exponential(\lambda_{ib}).$$

from which the probability π_{iab} of both events can be obtained.

For example, consider a $Q = 5$ year study with linear (constant) recruitment over a $R = 3$ year interval allowing for a loss-to-follow-up hazard rate of 0.05 per year and with equal size groups. Within the control group assume that the hazard rates are $\lambda_{Ca} = 0.2/year$ and $\lambda_{Cb} = 0.3/year$ and that the experimental therapy yields risk reductions of $RR_a = 0.8$ and $RR_b = 2/3$, or hazard rates of $\lambda_{Ea} = 0.16/y$ and $\lambda_{Eb} = 0.2/y$ so that $\delta_a = 0.04$ and $\delta_b = 0.10$. To allow for a correlation of the event times we assume shared frailties of $\lambda_{EF} = 0.08$ and $\lambda_{CF} = 0.1$. For a given sample size, the simulation model (herein with 10,000 replications) provides direct computation (within a small degree of error) of the expected quantities ($E(D_{ij})$, etc.) from which power is computed. By a simple search it was found that a n of 197 per group provides a one-sided one-directional test with 90% power.

For this sample size, the expected number of events marginally are $E(D_{Ca}) = 90.3$, $E(D_{Cb}) = 116.9$, $E(D_{Ea}) = 76.9$, and $E(D_{Eb}) = 90.4$; and the expected patient-years at risk are $E(T_{Ca}) = 451.6$, $E(T_{Cb}) = 389.8$, $E(T_{Ea}) = 480.6$, and $E(T_{Eb}) = 451.8$. The numbers of subjects with both events with the shared frailty are $E(D_{Cab}) = 67.4$ and $E(D_{Eab}) = 51.6$, and those expected under independence (by chance) are $E(D_{CabI}) = 54.8$ and $E(D_{EabI}) = 36.3$. These yield $\sigma_a^2 = 7.76\text{E–}4$, $\sigma_b^2 = 12.1\text{E–}4$, and $\sigma_{ab} = 1.43\text{E–}4$, that provides $corr(\hat{\delta}_a, \hat{\delta}_b) = 0.147$ and $\sigma_S^2 = 22.7\text{E–}4$. Substituting into (20), yields $Z_{1-\beta} = 1.292$ and power = 0.902.

A similar computation using (25) shows that an n of 197 per group would provide power = 0.885 using the Z-based test, indicating that in this setting the Z-based test would have less power than the original scale based test.

Generalizations

It is also possible to obtain a test based on the combination of group differences in hazard rates and differences in proportions or means. As in the preceding sections this requires the derivation of the covariance of the measures within each treatment group.

Alternately, a multivariate one-directional test can be obtained using multiple regression models as now described.

Model-Based Analysis of Multiple Outcomes

The preceding sections describe the application of the Wei-Lachin test to a combination of the group differences in means or proportions or hazard rates. In each case the covariance of the group differences, or of the corresponding Z-values, is described. The test statistic can then be computed using a consistent sample estimate of the variances and covariance(s), and the expression for power can be obtained using specified values for these parameters. In principle it is possible to construct a test for combinations of other types of outcomes, such as the difference in rates (counts) of events under a Poisson model, and to derive the equations to assess the power of the tests. However, it is more convenient to provide model-based generalizations of this approach.

From basic principles, Pipper, Ritz and Bisgaard [27] describe the joint distribution of parameter estimates from multiple models, not necessarily all of the same type. Consider two models for each of two outcomes, each with K_j parameters and coefficient estimates $\hat{\theta}_j = (\hat{\theta}_{j1} \ldots \hat{\theta}_{jK_j})'$. Arbitrarily, assume that the first parameter estimate $\hat{\theta}_{j1}$ represents the difference between groups on some scale, no difference represented by a value of zero, and the remaining K_j estimates represent the intercept (if any) and other covariate effects. Then $\mathbf{U}_{\ell j}(\hat{\theta}_j) = \left[U_{\ell j1}(\hat{\theta}_j) \ldots U_{\ell jK_j}(\hat{\theta}_j)\right]'$ is the score vector for the ℓth subject and $\mathbf{I}_j(\hat{\theta}_j)$ is the model based estimate of the expected information for the jth outcome. Also, let $\mathrm{U}_j(\hat{\theta}_j)$ denote the $K_j \times N$ matrix where the ℓth column is the score vector $\mathbf{U}_{\ell j}(\hat{\theta}_j)$. Then the generalization of the information sandwich robust estimate of the covariance matrix of the joint set of estimates $\underline{\hat{\theta}} = (\hat{\theta}'_a \hat{\theta}'_b)'$ is provided by

$$\boldsymbol{\Sigma}_R(\underline{\hat{\theta}}) = \begin{bmatrix} \boldsymbol{\Sigma}_R(\hat{\theta}_a) & \boldsymbol{\Sigma}_R(\hat{\theta}_a, \hat{\theta}_b) \\ \boldsymbol{\Sigma}_R(\hat{\theta}_a, \hat{\theta}_b) & \boldsymbol{\Sigma}_R(\hat{\theta}_b) \end{bmatrix} \tag{67}$$

where

$$\boldsymbol{\Sigma}_R(\hat{\theta}_j) = \mathbf{I}_j(\hat{\theta}_j)^{-1} \mathrm{U}_j(\hat{\theta}_j) \mathrm{U}_j(\hat{\theta}_j)' \mathbf{I}_j(\hat{\theta}_j)^{-1}, \qquad j = a, b \tag{68}$$

$$\boldsymbol{\Sigma}_R(\hat{\theta}_a, \hat{\theta}_b) = \mathbf{I}_a(\hat{\theta}_a)^{-1} \mathrm{U}_a(\hat{\theta}_a) \mathrm{U}_b(\hat{\theta}_b)' \mathbf{I}_b(\hat{\theta}_b)^{-1}.$$

The estimated variances of the group coefficients in the two models is then provided by the elements $\hat{\sigma}_a^2 = \boldsymbol{\Sigma}_R(\hat{\theta}_a)_{1,1}$ and

$\hat{\sigma}_b^2 = \boldsymbol{\Sigma}_R(\hat{\theta}_b)_{1,1}$, and the covariance by $\hat{\sigma}_{ab} = \boldsymbol{\Sigma}_R(\hat{\theta}_a,\hat{\theta}_b)_{1,1}$. The scale-based test is then provided by (5) with $(\hat{\theta}_{a1},\hat{\theta}_{b1})$ substituted for $(\hat{\delta}_a,\hat{\delta}_b)$. Alternately, Z-tests of the group effect in the two models are then provided by $Z_j = \hat{\theta}_{j1}/\sigma_j,\ j = a,b$, and the correlation of these tests by $Cov(Z_a,Z_b) = Corr(\hat{\theta}_{a1},\hat{\theta}_{b1}) = \hat{\sigma}_{ab}/(\hat{\sigma}_a\hat{\sigma}_b)$. This provides the Z-based test as in (19).

Pipper et al. also describe the application of the joint models where data for a subject is missing for one of the component models (but not both). Under the assumption of missing completely at random, then the score vector elements for that subject are set to zero in the corresponding score matrix U.

It would be difficult to evaluate the sample size and power of such a model-based test. However, simple computations such as those herein could be applied, e.g. the power of a test for a difference in means and proportions when the actual analysis will employ a linear regression model and a logistic model.

Pipper et al. originally provided an *R* package *multmod* to fit multiple models and to compute the covariances of the coefficients in the models. That has since been replaced by the *R* package *multcomp*.

Example – The Diabetes Prevention Program

The Diabetes Prevention Program compared the risk of onset of type 2 diabetes and deterioration of metabolic function among participants randomly assigned to an intensive lifestyle intervention (ILS) versus treatment with the glucose lowering drug metformin and versus a placebo control with no lifestyle intervention [28]. The study showed that intensive lifestyle provided a 58% reduction in diabetes risk versus placebo and 39% versus metformin, and that metformin produced a 31% reduction versus placebo. The study also evaluated the differences among treatments in the prevalence of developing the *metabolic syndrome*, a metabolic state that is linked not only with risk of onset of diabetes but also the risk of developing cardiovascular disease. The prevalence of the metabolic syndrome is characterized by 3 or more of the following 5 criteria: abdominal obesity defined as a waist circumference >102 cm among men or >88 cm among women, serum triglycerides (a bad cholesterol) ≥150 mg/dL, HDL (a good cholesterol) <40 mg/dL among men or <50 mg/dL among women, systolic/diastolic blood pressure ≥130/85 mm Hg, and fasting glucose ≥110 mg/dL, the latter met by many of the study subjects. [29]

Of the 3234 randomized, 1388 (43%) already met the metabolic syndrome criteria. Among the remainder who were evaluated at 3 years of follow-up (i.e., free of the syndrome on entry), 22% (363 of 1673) had the syndrome present. [30] Herein we compare the prevalence of the metabolic syndrome and its components at 3 years of follow-up among those in the lifestyle versus metformin treated groups.

The classification of the metabolic syndrome is a composite outcome, i.e. a single binary trait to designate that the criteria were met. An alternative would be to construct an analysis of the 5 binary traits using the one-directional multivariate test described herein.

For two of the traits (waist circumference and HDL) there are separate criteria for men and women, and for hypertension both systolic and diastolic blood pressure are employed, whereas for the other two traits there is a single cutpoint for the corresponding quantitative measure. Thus an alternate analysis would be to used these three composite binary traits in conjunction with an analysis of the other two quantitative variables (triglycerides and glucose). Alternately, rather than use any cutpoints to construct derived binary variables, an analysis could compare the groups with respect to the six quantitative traits (including systolic and diastolic blood pressure) simultaneously.

Table 1 presents a comparison of the lifestyle versus metformin groups for each of the binary outcomes and each of the corresponding quantitative outcomes. The overall prevalence of the metabolic syndrome using the composite binary outcome does not differ significantly between groups, although the prevalence is about 2% lower in the lifestyle group.

For all variables other than HDL, higher values are worse, so that a positive difference between metformin minus lifestyle indicates a benefit for lifestyle. In order for the same to apply to HDL, the analysis employed the negative values of HDL.

All p-values are one-sided. Some of the one-sided p-values are >0.5 indicating a negative Z-value favoring metformin. However, most of these differences are close to zero. For no measure is there evidence that intensive lifestyle is worse than metformin, and all significant differences favor the lifestyle group. Thus, these data are consistent with the alternative hypothesis that lifestyle has a beneficial effect on some of the outcomes, and no adverse effect for any.

Table 2 presents the correlations among the measurements. The modest to low correlations suggest that a multivariate test will provide greater power than individual tests, especially when the latter are adjusted for multiple tests.

Table 3 then presents the Wei-Lachin scale-based and Z-based one-directional multivariate test Z and one-sided p-values for three different analyses of these data. As would be expected, the analysis of all six quantitative traits is more powerful or sensitive than the analyses involving binary traits, with p-values <0.001 using either the scale or Z-based tests. The analysis of the 5 binary indicator variables produces less significant results, and the scale-based test for these data proves to be more powerful (larger Z-value) than the Z-based test, although both are significant. An alternative would be to conduct an analysis of the three binary traits defined from multiple criteria (waist, HDL, hypertension) and the other two quantitative traits (triglycerides and glucose). This yields results intermediate to those of the analysis of all quantitative and all binary traits.

Regardless of which of these options might have been chosen as the basis for the analysis, all would have provided a statistically significant result whereas the analysis of the composite metabolic syndrome outcome failed to demonstrate a beneficial effect of lifestyle versus metformin (Table 1, $p = 0.22$).

Discussion

A number of multivariate one-directional or one-sided tests have been described. Virtually all were developed to apply to a multivariate test of the difference in means between two groups for a multivariate outcome, such as repeated measures. These are also described for the case of two measures with group differences $\hat{\delta}_a$ and $\hat{\delta}_b$ as described above.

For a test based on multivariate normal observations, such as K repeated measures, Kudo [31] described the multivariate one-sided likelihood ratio test (LRT) of the K-variate generalization of the ordered hypotheses in (2) assuming that the covariance matrix Σ is known, and Pearlman [32] described the LRT when the estimated covariance matrix is employed. For the case of the two statistics herein, Pearlman's LRT is based on the statistic

$$S_{LR} = \min[\hat{\delta}_a \vee 0),\ (\hat{\delta}_b \vee 0)] = \max[0,\ \min(\hat{\delta}_a,\hat{\delta}_b)] \qquad (69)$$

Table 1. Differences between the DPP intensive lifestyle (ILS, $n=571$) versus metformin (MET, $n=557$) treated patients at three years of follow-up with respect to quantitative trait components of the metabolic syndrome, and binary indicators of abnormal levels, and the overall incidence of the metabolic syndrome among those free of the syndrome on entry.

	Mean (SE)			**%**		
Characteristic	**ILS**	**MET**	***p***	**ILS**	**MET**	***p***
Waist (cm)	97 (0.61)	99 (0.60)	0.0030	54.6	63.4	0.0015
Triglycerides (mg/dl)	115 (2.5)	125 (2.9)	0.0017	19.3	25.3	0.0074
HDL (mg/dl)	51.3 (0.53)	50.7 (0.53)	0.10	36.6	37.9	0.33
BP hypertension				9.5	9.3	0.53
Systolic (mm Hg)	120 (0.64)	122 (0.60)	0.0046			
Diastolic (mm Hg)	74 (0.40)	76 (0.37)	0.0001			
Glucose (mg/dL)	104 (0.49)	103 (0.53)	0.59	24.1	23.5	0.60
Metabolic Syndrome				18.2	20.1	0.22

Analysis restricted to those free of the metabolic syndrome at entry. One-sided p-values computed from a t-test for quantitative measures and chi-square test for binary variables.

doi:10.1371/journal.pone.0108784.t001

where “$\vee$” designates the maximum of the two quantities. Thus, if either $\hat{\delta}$ is negative the resulting test statistic quantity is zero. However, the distribution of S_{LR} is computationally difficult and the test is not convenient for practical use.

Tang, Gnecco and Geller [33] proposed a computationally simpler approximation to the LRT. Their approximate or ALR test is not an approximation in the sense, say, of a series expansion, but rather is an approximation in the sense that the alternative hypothesis parameter space is an approximation of that of the LRT. Their statistic is of the form

$$S_{ALR} = (\tilde{Z}_a \vee 0) + (\tilde{Z}_b \vee 0)] \qquad (70)$$

where $\tilde{Z}_a$ and $\tilde{Z}_b$ are uncorrelated standardized Z-statistics obtained as linear transformations of the $\hat{\boldsymbol{\Delta}}$ vector. Under the assumption that the covariance matrix is known, then $\tilde{\mathbf{Z}} = (\tilde{Z}_a\ \tilde{Z}_b)' = \mathbf{A}'\hat{\boldsymbol{\Delta}}$ where $\mathbf{A}$ is a square matrix such that $\mathbf{A}'\mathbf{A} = \boldsymbol{\Sigma}^{-1}$ and $\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} = \mathbf{I}$, such as is obtained from a Choleski decomposition. The distribution of this statistic is a simplified Chi-bar-squared distribution [34], though still requiring some computation to obtain a p-value. However, when an estimate of the covariance matrix is employed to provide the $\mathbf{A}$ transformation matrix, various authors have shown that the test can be serverely liberal, i.e. has an inflated type I error probability. In this case, Tamhane and Logan [35] described an accurate approximation to the distribution of the resulting test using a mixture of F-distributions, that also requires some computation to determine levels of significance.

However, this test has the unsavory feature that if either $\hat{\delta}$ value is negative, regardless how greatly so, the value is set to zero in the computation of the test statistic. Thus, for example if $\tilde{Z}_a = -1000$ and $\tilde{Z}_b = 10$, then $S_{ALR} = 10$, and depending on the estimated covariance values, could reject H_{0S} in favor of H_{1S}, even though it is clear that H_{1S} does not apply. In a recent overview, Tamhane and Logan [36] have suggested that “If several endpoints show moderate negative differences or even if a few show very large negative differences, then these tests should not be used because the *a priori* assumption of positive treatment effects in all endpoints is questionable.” However, to apply this recommendation in practice violates the principle that the test statistic for a study be specified *a priori*. In effect, the recommended practice could be viewed as a two-stage inference process - first determine if the differences are positive, and if so conduct the test. This would clearly inflate the type I error probability.

Other tests have been proposed that are based in part on Hotelling's T^2 statistic that is equivalent to the expression in (8) and is distributed as T^2 on $K\ df$ under the assumption of multivariate normality of the observations. Under this assumption, T^2 provides an optimal test of the null hypothesis against the global alternative presented in (7). Follman [37] describes a test of H_0 versus H_{1+}: $(\delta_a + \delta_b) > 0$ that is not the same as H_{1S} above. His X^2_+ test rejects H_0 in favor of H_{1+} if T^2 is significant at level 2α and $(\hat{\delta}_a + \hat{\delta}_b) > 0$. This test also could lead to rejection of H_0 when either the true δ_a or δ_b is a large negative value and the other an even larger positive value.

Table 2. Correlations among the component measurements obtained from the pooled within-groups covariance matrix.

	Triglycerides	**HDL**	**SBP**	**DBP**	**Glucose**
Waist (cm)	0.07	0.24	0.13	0.19	0.28
Triglycerides (mg/dl)		0.27	0.03	0.11	0.06
HDL (mg/dl)			−0.09	0.04	0.14
Systolic (mm Hg)				0.55	0.08
Diastolic (mm Hg)					0.05

doi:10.1371/journal.pone.0108784.t002

Table 3. The Wei-Lachin scale-based and Z-based one-directional multivariate test Z and one-sided p-values for three different analyses of the DPP metabolic syndrome data.

	Scale-based Test		*Z*-based Test	
Analysis	Z_s	p_s	$Z_{S,z}$	$p_{S,z}$
All quantitative (6)	3.52	0.00022	3.48	0.00025
All binary (5)	2.37	0.0089	2.22	0.0131
Mixed (5)	2.49	0.0064	2.02	0.0215

doi:10.1371/journal.pone.0108784.t003

Bloch, Lai and Tubert-Bitter [38] describe another test procedure which requires that T^2 reach significance at level α two-sided and that both individual one-sided t-tests of an indifference hypothesis be significant at level α. The indifference hypothesis is H_{0I}: $(0 \geq \delta_a > -\varepsilon)$ and $(0 \geq \delta_b > -\varepsilon)$ for some small positive value ε, and the alternative hypothesis is H_{1S} as in (2) above so that the one-sided t-test is of the form

$$t_j = \frac{(\hat{\delta}_j - \varepsilon)}{\sqrt{\hat{V}(\hat{\delta}_j)}}, \quad j = a,b. \tag{71}$$

This test was later criticized by Pearlman and Wu [39] who proposed use of the one-sided LRT of Pearlman [32] in lieu of T^2, among other improvements. The result of either test, however, depends on the specification of the value ε and thus the test may not be uniformly acceptable.

Other tests have also been applied, although not specifically designed to test H_0 against the one-sided alternative H_{1S} in (2). O'Brien [13] proposed his ordinary least squares (OLS) and weigthed least squares (WLS) tests of H_0 versus the alternative hypothesis of a common difference H_{1A}: $\delta_a = \delta_b = \delta \neq 0$. Thus the alternative hypothesis consists of the line of equality other than the origin. The one-sided version of this test will also be sensitive to alternatives where δ_a and δ_b are of similar positive magnitude, but will not be optimal against the general alternative H_{1S}. Pocock, Geller and Tsiatis [40] describe the application of these tests to the analysis of multiple outcomes in clinical trials on different scales.

For a two group comparison of a vector of repeated measures, under the usual normal errors assumptions O'Brien also suggested that his statistics were distributed as t. However, the exact small sample distribution with normal errors is not known and many authors have shown that the resulting t-statistics have an inflated type I error probability. For a vector of repeated measures in two groups, Läuter [41] shows that statistics that employ weighted averages, as in O'Brien's WLS test, are indeed distributed as t provided that the weights are functions of the empirical covariance matrix estimated from all groups combined rather than the pooled within-groups covariance matrix estimate as employed by O'Brien. He proposes a family of such weighted tests that includes the Wei-Lachin test as a trivial special case. Frick [42] also showed that O'Brien's OLS test is biased.

Thus, among the various tests that have been proposed that could be applied to the assessment of simultaneous differences between groups for multiple outcomes, the Wei-Lachin test has the advantages that it is simple to compute; can be applied to mixtures of outcomes on different scales (e.g. means and proportions); that it has a large sample normal distribution (or a t-distribution with normal errors); provides a test with type I error probabilities close to the nominal levels with generally acceptable sample sizes; is directed towards the specific multivariate one-directional alternative of interest, is maximin efficient relative to the possible true but unknowable optimal test, and readily provides for the computation of sample size and power.

Rahlfsand Vester [43] describe applications of the Wei-Lachin test to the analysis of multiple outcomes using the multivariate Mann-Whitney difference analysis described initially by Thall and Lachin [11]. The authors are affiliated with idv Data Analysis and Study Planning that also markets a program (TESTIMATE) that conducts such Wei-Lachin analyses. Pan [44] also recently presented a review of various procedures including the Wei-Lachin test (called the SUM test therein) and some of the above referenced one-directional procedures and showed by simulation that the Wei-Lachin test had good power when the outcomes tended to jointly show beneficial effects.

Programs for computations herein are available from www.bsc.gwu.edu. These include the coefficient vector L for use in (10) when Frick's condition does not apply, the simulation event time model, and the Wei-Lachin multivariate rank test.

Ethical Statement

Neither animals or human subjects were involved in this methodological research.

Supporting Information

File S1
(PDF)

Author Contributions

Conceived and designed the methodological research: JML. Analyzed the data: JML. Wrote the paper: JML.

References

1. Nathan DM, Buse JB, Kahn SE, Krause-Steinrauf H, Larkin ME, et al. (2013) Rationale and design of the Glycemia Reduction Approaches in Diabetes: A Comparative Effectiveness Study (GRADE). Diabetes Care 36(8): 2254–61.
2. Wei LJ, Lachin JM (1984) Two-sample asymptotically distribution-free tests for incomplete multivariate observations. J Amer Statist Assoc 79: 653–661.
3. Lachin JM (1992) Some large sample size distribution-free estimators and tests for multivariate partially incomplete observations from two populations. Stat Med 11: 1151–1170.
4. Frick H (1994) A maximin linear test and its application to Lachin's Data. Commun Statist - Theory and Methods 23: 1021–1029.
5. Frick H (1995) Comparing trials with multiple outcomes: The multivariate one-sided hypothesis with unknown covariances. Biom J 8: 909–917.
6. Lachin JM (2011) Biostatistical Methods: The Assessment of Relative Risks. Second Edition. New York: John Wiley & Sons, Inc.

7. Lachin JM, Wei LJ (1988) Estimators and tests in the analysis of multiple nonindependent 2×2 tables with partially missing observations. Biometrics 44: 513–528.
8. Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika: 73, 13–22.
9. Demidenko E (2005) Mixed Models: Theory and Applications. New York: John Wiley & Sons, Inc.
10. Wei LJ, Lin DY, Weissfeld L (1989) Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. J Amer Statist Assoc 84: 1065–1073.
11. Thall PF, Lachin JM (1986) Assessment of stratum-covariate interactions in Cox's proportional hazards regression model. Stat Med 5: 73–83.
12. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Statist 6: 65–70.
13. O'Brien PC (1984) Procedures for comparing samples with multiple endpoints. Biometrics 40: 1079–1087.
14. Gastwirth JL (1966) On robust procedures. J Amer Statist Assoc 61: 929–948.
15. Lachin JM (1981) Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials 2: 93–113.
16. Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical Models Based on Counting Processes. New York: Springer-Verlag.
17. Fleming TR, Harrington DP (1991) Counting Processes and Survival Analysis. New York: John Wiley & Sons, Inc.
18. Harrington DP, Fleming TR (1982) A class of rank test procedures for censored survival data. Biometrika 69: 553–566.
19. Andersen PK, Borgan O, Gill RD, Keiding N (1982) Linear nonparametric tests for comparison of counting processes, with applications to censored survival data. Int Statist Rev 50: 219–258.
20. Kalbfleisch JD, Prentice RL (1980) The Statistical Analysis of Failure Time Data. New York: John Wiley & Sons.
21. Lakatos E (1988) Sample sizes based on the log-rank statistic in complex clinical trials. Biometrics 44: 229–241.
22. Rubenstein LV, Gail MH, Santner TJ (1981) Planning the duration of a comparative clinical trial with losses to follow-up and a period of continued observation. J Chronic Dis 34: 469–479.
23. Lachin JM, Foulkes MA (1986) Evaluation of sample size and power for analyses of survival with allowance for non-uniform patient entry, losses to follow-up, non-compliance and stratification. Biometrics 42: 507–519.
24. Freedman LS (1982) Tables of the number of patients required in clinical trials using the logrank test. Stat Med 1: 121–129.
25. Marshall AW, Olkin I (1967) A multivariate exponential distribution, J Amer Statist Assoc 62: 30–44.
26. Hougaard P (2000) Analysis of Multivariate Survival Data. New York: Springer-Verlag.
27. Pipper CB, Ritz C, Bisgaard H (2012) A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. J Royal Statist Soc Series C (Applied Statistics) 61: 315–326.
28. Diabetes Prevention Program Research Group (2002) Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 346: 393–403.
29. NCEP Adult Treatment Panel III (2001) Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). JAMA 285: 2486–97.
30. Orchard TJ, Temprosa M, Goldberg R, Haffner S, Ratner R, et al. (2005) The effect of metformin and intensive lifestyle intervention on the metabolic syndrome: The Diabetes Prevention Program Randomized Trial. Ann Intern Med 142: 611–619.
31. Kudô A (1963) A multivariate analogue of the one-sided test. Biometrika 50: 403–418.
32. Perlman MD (1969) One-sided testing problems in multivariate analysis. Ann Math Stat 40: 549–567.
33. Tang DI, Gnecco C, Geller NL (1989) An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. Biometrika 76: 577–583.
34. Robertson T, Wright FT, Dykstra RL (1988) Order Restricted Statistical Inference. New York: Wiley.
35. Tamhane AC, Logan BR (2002) Accurate critical constants for the one-sided approximate likelihood ratio test of a normal mean vector when the covariance matrix is estimated. Biometrics 58: 650–656.
36. Tamhane AC, Logan BR (2003) Multiple Endpoints: An Overview and New Developments. Division of Biostatistics, Medical College of Wisconsin, Technical Report 43.
37. Follman D (1996) A simple multivariate test for one-sided alternatives. J Amer Statist Assoc 91: 854–861.
38. Bloch DA, Lai TL, Tubert-Bitter P (2001) One-sided tests in clinical trials with multiple endpoints. Biometrics 57: 1039–1047.
39. Perlman MD, Wu L (2004) A note on one-sided tests with multiple enpoints. Biometrics 60: 276–280.
40. Pocock SJ, Geller NL, Tsiatis AA (1987) The analysis of multiple endpoints in clinical trials. Biometrics 43: 487–498.
41. Läuter J (1996) Exact t and F tests for analyzing studies with multiple endpoints. Biometrics 52: 964–970.
42. Frick H (1997). A note on the bias of O'Brien's OLS test. Biometrical J. 39: 125–128.
43. Rahlfs V, Vester JC (2012) The new trend in clinical research: The multidimensional approach instead of testing individual endpoints. Pharm Med 14: 160–165 (in German, English translation available).
44. Pan Q (2013) Multiple hypotheses testing procedures in clinical trials and genomic studies. Frontiers in Public Health 1(63): 1–8. doi:10.3389/fpubh.2013.00063.