NONPARAMETRIC MANOVA APPROACHES FOR NON-NORMAL MULTIVARIATE OUTCOMES WITH MISSING VALUES

Fanyin He[1], **Sati Mazumdar**[2], **Gong Tang**[2], **Triptish Bhatia**[3], **Stewart J. Anderson**[2], **Mary Amanda Dew**[1], **Robert Krafty**[2], **Vishwajit Nimgaonkar**[1], **Smita Deshpande**[3], **Martica Hall**[1], and **Charles F. Reynolds III**[1]

[1]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 15213, U.S.A

[2]Department of Biostatics, University of Pittsburgh, Pittsburgh, PA 15261, U.S.A

[3]Dr Ram Manohar Lohia Hospital, Delhi, India

Abstract

Between-group comparisons often entail many correlated response variables. The multivariate linear model, with its assumption of multivariate normality, is the accepted standard tool for these tests. When this assumption is violated, the nonparametric multivariate Kruskal-Wallis (MKW) test is frequently used. However, this test requires complete cases with no missing values in response variables. Deletion of cases with missing values likely leads to inefficient statistical inference. Here we extend the MKW test to retain information from partially-observed cases. Results of simulated studies and analysis of real data show that the proposed method provides adequate coverage and superior power to complete-case analyses.

1. INTRODUCTION

Comparisons between several treatment groups play a central role in clinical research. As these comparisons often entail many potentially correlated dependent variables, the classical multivariate general linear model has been accepted as a key tool for this endeavor. The widely applied statistical procedures, univariate and multivariate analysis of variance (ANOVA and MANOVA) are subsumed under this model. For practitioners, the use of these statistical procedures does not pose any difficulties under normality assumptions due to the wide availability of software (SAS Institute Inc., 2014; IBM Corp., 2013; StataCorp., 2011). However, application of these procedures is problematic if the assumption of normality is violated or treatment groups may not only differ in means but also in higher order moments. The ranked-based multivariate Kruskal-Wallis (MKW) test (Puri and Sen, 1969; May and Johnson, 1997) and permutation tests on either the data or rank transformed data provide robust alternatives when the normality assumption may not hold or the higher order moments vary across treatment groups (Pesarin, 2001; Basso, Pesarin, Salmaso, and Solari, 2009).

Correspondence to: Gong Tang, Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, U.S.A. got1@pitt.edu.

Missing data often occur in clinical trials. Usually the missingness for a given subject only occurs in a subset of the variables being measured. However, standard tests in the MANOVA-like framework cannot utilize information in partially observed cases. In readily available software algorithms such as the SAS/STAT® software MANOVA procedure, cases with missing values in response variables are automatically deleted. This is a major shortcoming of the standard MKW test.

In Section 2, we propose an extension of the MKW test for correlated multivariate non-normal data with missing values. This extension pertains to outcomes measured at a fixed time point, either continuous or ordinal, and retains information in partially observed cases. We call this test E-MKW. Applications illustrating the proposed method with both simulated and actual data from a psychiatric clinical trial are presented in Section 3. We conclude with discussion in Section 4.

2. METHODOLOGY

2.1. Multivariate Kruskal-Wallis (MKW) Test

The MKW test is a rank-order procedure in which the n observations on each of the p variables are ranked separately. Tied observations are assigned the mean of the total ranks for the tied observations. It should be noted that this procedure of assigning ranks poses no difficulty if the number of scores for the different variables are not equal and it works well when there are few tied observations. It becomes problematic when there are many tied observations. The null hypothesis is that the distribution of each variable is the same across different groups. Under this null hypothesis, it is implied that for each variable, the expected values of the mean ranks are equal for different groups. Large sample theory suggests that the MKW statistic is approximately χ^2 distributed (Puri and Sen, 1969). However, in small samples, permutation methods are needed to get the appropriate critical value for rejecting the null hypothesis. The alternative hypothesis implies that the mean ranks differ between at least two different groups.

Katz and McSweeney (1980) provided an explicit description of this MKW test. They also provided computational formulas and post-hoc techniques which could be used to isolate sources of differences if the null hypothesis is rejected. However, the testing procedure discussed in their paper was based on large sample properties of the statistic. May and Johnson (1997) constructed a SAS macro that computes the probability values and tabulates the exact distributions for both the univariate and multivariate Kruskal-Wallis tests.

The MKW test transforms the original data to ranked data, and therefore it is distribution-free. The ranking is performed separately for each outcome variable, and is across groups. Let Y_{ijk} be the original observation of the kth variate for the jth subject from the ith group, where $k = 1, …, p; j = 1, …, n_i; i = 1, …, g$. Denote R_{ijk} as the rank corresponding to Y_{ijk} and $\mathbf{R}_{ij} = (R_{ij1}, …, R_{ijp})'$. In case of ties, the mean rank is used. Let

$$\overline{R}_{i.k}=\sum_{j=1}^{n_i}\frac{R_{ijk}}{n_i},$$

then, $E(\bar{R}_{i.k}) = m = (n + 1)/2$. The vector $\mathbf{U_i} = (\bar{R}_{i.1} - m, ..., \bar{R}_{i.p} - m)'$ denotes the vector of average ranks for the ith group corrected for the overall average rank for each variate. $\mathbf{U_i}$ is a measure of directed distance from the mean vector of ranks for the ith group. An estimate of the pooled within-group covariance matrix is

$$\mathbf{V}=\frac{1}{n-1}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{R_{ij}}-m\mathbf{1_p})(\mathbf{R_{ij}}-m\mathbf{1_p})'.$$

Under the null hypothesis that there is no difference in group means for the p variables,

$$E(\mathbf{U_i})=\mathbf{0}_p.$$

The MKW test is expressed as

$$W^2=\sum_{i=1}^{g}n_i\mathbf{U_i'V^{-1}U_i}.$$

In large samples, W^2 is approximately χ^2 distributed with $p(g-1)$ degrees of freedom when all the n_is are fairly large. The alternative for the MKW W^2 is that the mean ranks are not the same for at least two groups i and h, and some $1 \le k \le p$: $E(\bar{R}_{i.k}) \ne E(\bar{R}_{h.k})$.

When there are too many possible permutations of the data to allow complete enumeration in a reasonably time-efficient manner, an asymptotically equivalent permutation-based test can be created by generating the exact distribution under the null of no difference across groups through Monte Carlo sampling as the following (Pesarin, 2001; Edgington and Onghena, 2007):

a. Calculate the statistic for the data, and denote it as W^{2*}.

b. Randomly permute the group labels for all subjects, and calculate the new test statistic W^2 for the permuted data.

c. Independently repeat (b) M times to get the permutation distribution of W^2 under null hypothesis.

d. Calculate the $p\text{- value}=\frac{\text{number of } W^2 \geq W^{2*}}{M}$.

2.2. Extended MKW Test (E-MKW)

The MKW test assumes that the data are fully observed. All incomplete cases are deleted before the MKW test is performed, which means that the information in partially observed cases is lost. To retain this information, we propose a method that we call the Extended MKW test (E-MKW). We first develop a test to accommodate data that are missing completely at random (MCAR), where the missingness is independent from the observed coviariates and the outcome variables that are subject to missing values. Under MCAR, each subset represented by a missing data pattern is a random sample of the original data and the joint distributions of the observed outcome variables are preserved. We propose to construct a MKW test for each missing data pattern with sufficient number of observations and aggregate those tests at the end. Then we extend to circumstances where the missingness of outcome values may depend on the fully observed covariates. When data are not missing at random, the dependence structures among outcome variables in general vary across different missing data patterns and our extensions cannot be applied in such circumstances.

The observation vector $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijp})$ for the jth subject in the ith group has a corresponding missing indicator vector $\mathbf{r}_{ij} = (r_{ij1}, \ldots, r_{ijp})$, where $r_{ijk} = 1$ if the kth variate is missing, and 0 if it is observed. $\mathbf{r}_{ij}$s are often used to form missing data patterns in each treatment group that subjects who have equivalent missing indicator vectors belong to the same pattern. $\mathbf{r}_{ij}$ is a vector of length p, with each element valued at 0 or 1. For example, if $p = 4$, and if the first and second outcomes are observed and the third and fourth outcomes are missing from a subject, then this subject belongs to the missing data pattern represented as $(0, 0, 1, 1)$. In a dataset with p variables, there are in total 2^p possible distinct missing data patterns, including the pattern where all variables are observed and the pattern where all variables are missing.

Suppose there are L distinct missing patterns in the data ($L \leq 2^p$). Let S_l denote the set of cases with the missing pattern l, $l = 1, \ldots, L$, and let m_l denote the number of observations in S_l, and then $n = \sum_{l=1}^{L} m_l$. Let p_l denote the number of observed variables in the missing pattern l, $l = 1, \ldots, L$. Let m_{il} denote the number of observations in group i, missing pattern l, $i = 1, \ldots, g; l = 1, \ldots, L$.

We assume that $m_l > p_l$, $l = 1, \ldots, L$. If the number of observations with the missing pattern l is small ($m_l \approx p_l$), the corresponding covariance matrix estimate V for the group-average ranks becomes singular and an MKW test cannot be constructed for that pattern. Therefore we delete the cases in those S_ls with $m_l \approx p_l$ from the total sample before performing the method. We assume that the estimated variance-covariance matrix within the missing pattern is nonsingular and hence, the the test statistic can be calculated. An example is provided below. In a two-group, two-outcome case, suppose two observations are $(0.5, 2)$ and $(1, 1.5)$. The associated rank variables are $(1, 2)$ and $(2, 1)$.

The statistic W_l^2 in each S_l with regard to observed variables can be calculated from the standard MKW test. The proposed test statistic is

$$W^2 = \sum_{l=1}^{L} t_l W_l^2,$$

where the $t_l \geq 0$, $l = 1, \ldots, L$ are weights and $\Sigma \; t_l = 1$. It is noted that only missing data patterns with sufficient number of observations to construct the MKW tests for those patterns, specifically non-singular covariance matrices of the average ranks in those patterns, contribute towards W^2.

The standard MKW test is a special case of the proposed test, when t_l is assigned to 1 if S_l is the set of complete cases, and 0 otherwise. Two weighting schemes are proposed:

1. Unweighted: $t_l = 1/L$, $l = 1, \ldots, L$. Then W^2 is the arithmetic mean of W_l^2s.

2. Weighted: $t_l = m_l/n$, $l = 1, \ldots, L$. Then each W_l^2 contributes to W^2 proportional to the number of cases in its missing pattern Under large samples, that is, $m_l \rightarrow \infty$, W_l^2 is approximately χ^2 distributed with degrees of freedom $v_l = p(g-1)$, $l = 1, \ldots, L$. As W^2 is a linear combination of the L independent χ^2 distributed statistics we can generate $W_{l1}^2, \ldots, W_{lM}^2$ as random samples from χ^2 distribution with v_l degrees of freedom, where $l = 1, \ldots, L$ and M is a large integer and setting

$$W^2 = \sum_{l=1}^{L} t_l W_{lm}^2, m = 1, \ldots M,$$

An empirical distribution of W^2 under the null hypothesis can be obtained by permuting the group labels among the whole data set and we can obtain a p-value by comparing the test statistic and its empirical distribution under the null. With comparing W^2 with its empirical distribution under the null hypothesis, the validity of the proposed E-MKW test will not be limited by the requirements of large m_ls.

When all m_ls are large and the numbers of observed outcome variables and m_ls are roughly equal across missing data patterns, the unweighted version is preferred for its simplicity because W^2 is approximately χ^2 distributed then the generation of its empirical distribution is not necessary, and the performance of the test is only slightly compromised. When there are large difference among the m_ls, the weighted version should be considered and the patterns with larger sample size should be given more weights.

2.3. E-MKW test when data are MAR

Here we consider circumstances when the missingness of $(X_1, \ldots, X_M)$ may depend on the fully observed covariates and data are missing at random (Mazumdar et al., 1999). For example, in stratified randomized clinical trials, the missingness within each stratum is completely at random but may be differential across randomization strata. This occurs when younger participants are more responsive in completing questionnaires and older participants are reluctant to provide answers for certain sensitive items. Suppose the missingness

depends on covariates $C_1, \dots, C_q$, that are categorical with $s_1, \dots, s_q$ levels respectively. We stratify the dataset by combinations of the covariate levels to render $S = \Pi_{a=1}^{q} s_a$ strata. Per our MAR assumption, we can claim now within each stratum defined by the covariates, the missingness does not depend on treatment and data are missing completely at random. We can apply E-MKW test within each stratum and get a statistic W_b^2, $b = 1, \dots, S$, and then sum up all the statistics to get the global test statistic W^2. We can also use weighted-sum with stratum-specific statistics as detailed in the previous section.

3. APPLICATIONS

We illustrate the proposed method with both simulated and actual data. We first investigate the performance of E-MKW test followed by an application of this method to a clinical intervention study examining yoga as an adjunctive cognition remediation strategy for schizophrenia (Bhatia et al., 2012). All analyses were carried out using codes we developed using the R software Platform (R Core Team, 2014).

3.1 Multivariate Effect Size

Effect sizes are commonly used for power analysis and to design experiments. In hypothesis testing, the effect size is an index reflecting the degree to which the null hypothesis is false, or the discrepancy between the null hypothesis and the alternative hypothesis (Cohen, 1992), without the influence of sample sizes. One of the widely used effect sizes index in one-way ANOVA setting is Cohen's f^2, the ratio of the variance of the group means to the variance of the values within groups (Cohen, 1988). Cohen's f^2 is defined as

$$f^2 = \frac{R^2}{1 - R^2}$$

where R^2 is the squared multiple correlation.

Cohen (1988) suggested a generalization of f^2 based on Wilks' λ as following::

$$f^2 = \lambda^{-1/r} - 1 = \frac{\sqrt[r]{\det(\mathbf{H}+\mathbf{E})} - \sqrt[r]{\det(\mathbf{E})}}{\sqrt[r]{\det(\mathbf{E})}},$$

where

$$r = \sqrt{\frac{p^2(g-1)^2 - 4}{p^2 + (g-1)^2 - 5}}$$

p is the number of response variables, g is the number of groups, and $\mathbf{E}$ and $\mathbf{H}$ refer to the population error and hypothesis matrices.

We note that f^2 is a ratio of signal to noise: the ratio of variance of the model to the variance of errors. f^2 is a non-increasing function of p and g, which means that for a given sample size (number of participants) when we have more groups or variables the effect size becomes smaller. For two-group cases, $r = 1$ and f^2 reduces to $\lambda^{-1} - 1$. For 3-group cases, $r = 2$ and f^2 reduces to $\lambda^{-1/2} - 1$. If these two cases have the same Wilks' λ, the latter case will have a smaller effect size. Cohen (1988) also suggested "small", "medium" and "large" f^2 values to be 0.02, 0.15 and 0.35, respectively.

3.2 Simulation Studies

To examine the coverage and power level of the proposed E-MKW test, simulations in different scenarios are performed. Our simulation studies assumed that missingness is MCAR.

First Simulation Study—Data with $g = 2$ groups and $p = 2$ outcome variables are simulated. To generate correlated outcomes, we use a latent variable X. Two scenarios are examined. One is based on normally distributed X, and the other is based on binomial distributed X.

For the first scenario, we set $X \sim N(0,1)$. For group 1, we generated $X_1, \ldots, X_{n_1}$ as a random sample of X. Next we set $Y_{11}|X \sim N(1 + X, 2)$, $Y_{12}|X \sim N(X, 1)$ and then Generate $[(Y_{1j1}, Y_{1j2})|X_j]$ as a random sample of $[(Y_{11}, Y_{12})|X_j]$, $j = 1, \ldots, n_1$. For group 2, we similarly generated $X_1, \ldots, X_{n_2}$ as another random sample of X and set $Y_{21}|X \sim N(1 + X, 2)$, $Y_{22}|X \sim N(\ + X, 1)$ and finally generated $[(Y_{2j1}, Y_{2j2})|X_j]$ as a random sample of $[(Y_{21}, Y_{22})|X_j]$, $j = 1, \ldots, n_2$. It is to be noted that these generated samples are conditionally independent.

For the second scenario, we first set $X \sim BIN(5, 0.5)$. For group 1, we generated $X_1, \ldots, X_{n_1}$ as a random sample of X. We then set $Y_{11}|X \sim POI(1 + X)$, $Y_{12}|X \sim POI(2 + X)$ and generated $[(Y_{1j1}, Y_{1j2})|X_j]$ as a random sample of $[(Y_{11}, Y_{12})|X_j]$, $j = 1, \ldots, n_1$. Similarly for group 2, we generated $X_1, \ldots, X_{n_2}$ as another random sample of X, set $Y_{21}|X \sim POI(1 + X)$, $Y_{22}|X \sim POI(2 + \ + X)$ and generated $[(Y_{2j1}, Y_{2j2})|X_j]$ as a random sample of $[(Y_{21}, Y_{22})|X_j]$, $j = 1, \ldots, n_2$. As noted earlier, these generated samples are conditionally independent.

Letting $n_1 = n_2 = 50$, the simulated data are given as

$$
Y = \begin{bmatrix} Y_{111} & Y_{112} \\ \vdots & \vdots \\ Y_{1,50,1} & Y_{1,50,2} \\ Y_{211} & Y_{212} \\ \vdots & \vdots \\ Y_{2,50,1} & Y_{2,50,2} \end{bmatrix} = (Y_1, Y_2).
$$

When $\ $ is zero, the underlying distributions of the two outcomes are the same in the two groups providing estimated type I error rates.

is assigned a spectrum of non-zero numbers to get different effect sizes. The underlying distributions of the first outcome variable are the same in the two groups, and the underlying distributions of the second outcome variable are different across the two groups rendering examinations of power values.

There are $L = 4$ possible missing data patterns in a bivariate data set: $\mathbf{Y_1}$ and $\mathbf{Y_2}$ both observed (M1), $\mathbf{Y_1}$ observed and $\mathbf{Y_2}$ missing (M2), $\mathbf{Y_1}$ missing and $\mathbf{Y_2}$ observed (M3), and $\mathbf{Y_1}$ and $\mathbf{Y_2}$ both missing (M4). Since cases with missing pattern M4 do not contain any information on $\mathbf{Y_1}$ and $\mathbf{Y_2}$, these cases are not involved in constructing the E-MKW test and we only consider the first three missing data patterns in the simulation study. Two missing rates, medium and high, are simulated. In the medium missing rate scenario, 40% of cases are simulated with M1, and each of 30% of cases are simulated with M2 and M3. In the high missing rate scenario, 20% of cases are simulated with M1, and each of 40% of cases are simulated with M2 and M3. Missing patterns are randomly assigned to simulated data.

For each scenario, we perform the standard and the E-MKW test in $n_{sim} = 1000$ simulated incomplete datasets, and get the power values as $\frac{\text{number of } p\text{- values} < 0.05}{n_{sim}}$ when > 0, and the type I errors as $\frac{\text{number of } p\text{- values} < 0.05}{n_{sim}}$ when $= 0$.

The simulation results for type I errors are shown in Table 1. Permutation-based p-values are close to the nominal significance level 0.05, and are slightly more accurate compared with p-values based on large sample approximation. Higher missing rates imply less information. It can be seen that type I errors are closer to 0.05 in medium missing rates scenarios compared with high missing rates scenarios, either in normal data or in non-normal data.

The simulation results of power levels are shown in Tables 2 and 3. As expected, the power levels of the E-MKW test are always higher than the power levels of the standard MKW test as the latter is applied only on complete cases. The difference is larger with higher missing rates. The permutation-based tests provide higher power levels than tests based on large sample approximation. The weighted and the unweighted test statistics provide very similar power levels, and both show increase in power when the effect size increases. In three of the four simulation sets, the power levels of the E-MKW test reach 80% when effect size is "medium" (<0.3). When percentage of missingness increases, the power level decreases. The performance of the extended test in non-normal data is as powerful as in normal data (Tables 2 and 3).

Second Simulation Study—Another set of simulations were done with $g = 3$ groups and $p = 3$ outcome variables. To generate data, we set $X \sim \text{BIN}(5, 0.5)$. For group 1, we generated $Y_{11}|X \sim POI(1 + X)$, $Y_{12}|X \sim POI(2 + X)$ and $Y_{13}|X \sim POI(3 + X)$. For group 2, we generated $Y_{21}|X \sim POI(1 + X)$, $Y_{22}|X \sim POI(2 + X + {}_1)$ and $Y_{13}|X \sim POI(3 + X)$. For group 3, we generated $Y_{31}|X \sim POI(1 + X)$, $Y_{32}|X \sim POI(2 + X)$ and $Y_{33}|X \sim POI(3 + X + {}_2)$. We used 30 cases in each group ($n_1 = n_2 = n_3 = 30$).

There are eight possible missing patterns in three-outcome data set: all outcomes observed (M1), two outcomes observed and one outcome missing (M2, M3 and M4), one outcome observed and two outcomes missing (M5, M6 and M7), and all outcomes missing (M8).

Since the cases with last missing pattern (M8) do not carry any data, they are deleted, and we only consider M1–M7. Thirty percent of cases are simulated with M1. Ten percent of cases are simulated with each M2, M3 and M4 respectively, and 10% of cases are simulated with each M5, M6 and M7. Missing patterns are randomly assigned to simulated data. We perform the standard and the E-MKW test in $n_{sim} = 500$ simulated incomplete data.

The simulation results of power levels are shown in Table 4. E-MKW tests consistently perform better than standard MKW tests, and weighted method consistently performs better than unweighted method. This better performance of the weighted method is expected as the weighted method provides weight proportional to the number of observed values in various missingness patterns.

3.3 Study on the use of yoga as adjunctive cognitive remediation for schizophrenia

Data from an open non-randomized clinical trial to evaluate the impact of adjunctive yoga therapy (YT), on cognitive domains in persons with schizophrenia (SZ) are used as an illustrative example (Bhatia et al., 2012) for the statistical method described in this paper. This study evaluated whether, among persons with SZ on conventional anti-psychotic medications, adjunctive structured yoga exercises could alter cognitive domains known to be impaired among persons with SZ. All patients clinically diagnosed in the study hospital with schizophrenia who fulfilled DSM IV diagnostic and inclusion criteria for this study were invited to participate in a specific 21-day yoga protocol in addition to their usual treatment. A total of 396 patients fulfilled inclusion criteria and 207 of them agreed to participate in one hour yoga training protocol, attending daily one hour yoga classes in the department (excluding Sundays). Following baseline evaluations, some patients dropped out of the study (N=121). Among the remainder, one group found that they could not travel to the hospital daily for yoga training as required (N=23), while the remainder (N=63) completed 21 daily yoga training sessions in the hospital and continued treatment with their therapists (YT group). The former group was therefore considered as the TAU group. They received conventional pharmacological treatment from their psychiatrists throughout the study. Cognitive functioning in all patients was assessed with a Hindi version of the Penn computerized neuropsychological battery (CNB) (Gur et al, 2001a; Gur et al, 2001b). The CNB included neurocognitive domains known to be impaired among individuals with SZ. The verbal domains were available only in English. As many participants did not speak English, the verbal domains were excluded. Accuracy (reflecting the number of correct responses) and speed (reflecting the median reaction time) for eight cognitive domains were assessed. The domains were: abstraction and mental flexibility, attention, working memory, face memory, spatial memory, spatial ability, sensorimotor dexterity and emotion processing. The neuropsychological battery was assessed at baseline, 21 days post treatment and 2 months post treatment.

The trial primarily compared YT patients who completed 21 days intervention period (N=63) and TAU patients (N=24) to evaluate the impact of adjunctive YT in cognitive domains impaired in SZ. Improvements in cognitive domains at 2-month assessment point were compared between the TAU and YT groups. SZ patients who participated in YT and those who refused YT and received only TAU were found to be similar in standard

demographic and clinical characteristics with regard to age, sex, marital status and occupation excepting education and global assessment of worst point functioning scores during recent SZ episode (Bhatia et al., 2012). A large amount of missing values existed in the data. Only 10 subjects in the YT group and 9 subjects in the TAU group completed the neuropsychological battery in all domains at all assessment points. Moreover, the distributions on the cognitive measures were skewed. The researchers used univariate Kruskal-Wallis tests to compare the various cognitive domains that involve varying sample sizes, followed by corrections for multiple comparisons. The main finding consists of YT group showing significantly greater improvement with regard to measures of attention.

The use of the univariate Kruskal-Wallis test followed by adjustments for multiple comparisons is a common approach in applied research in analyzing multiple outcomes. We reanalyzed this dataset with MKW and E-MKW to assess the robustness of the results, and the pros and cons of univariate and multivariate tests.

For illustrative purpose, we analyzed the improvements in the speed summary functions in four domains: abstraction and mental flexibility, attention, face memory and spatial memory as less missingness were observed in these domains. Results from univariate Kruskal-Wallis test, using complete cases for individual domains, are shown in Table 5 The speed functions in abstraction and mental flexibility and in attention are shown to improve more in the YT group than in the TAU group (p-values = 0.028 and 0.014, respectively). However, after a Hochberg adjustment for multiple comparisons, only attention remained borderline significant (p-value=0.056).

An examination of the dataset revealed that the missingness was mostly due to administrative reasons, and no covariates were involved. Hence we did not stratify the data by any covariates. Although a traditional randomization was not performed in our study, patients in the TAU and YT groups were similar on demographic and clinical characteristics. The chief point of difference was inability to travel daily for required YT participation. Therefore permuting the group labels helped us to generate the empirical distribution of the E-MKW test under the null. Table 6 presents the missing patterns in these 4 cognitive domains. We note that 37 cases have no data (missing pattern 8) and missing patterns 3, 4 and 5 could not be used in the E-MKW calculation, (m_l p_l). Results from the MANOVA, the standard MKW test and the E-MKW tests (permutation-based) are given in Table 7. We note that while the MANOVA and the standard MKW used 32 cases, the E-MKW is based on 46 cases by retaining information from partially observed data.

No significant difference between the two groups was detected by the MANOVA procedure. However, both the standard MKW test (large sample approximated p=0.038 and permutation-based p=0.030) and E-MKW tests (unweighted, p=0.031) and (weighted p= 0.034) showed significant p-values. This implies that the improvements in at least one of the four domains are different between the two groups. As indicated earlier in Table 5, the univariate Kruskal-Wallis test fails to detect the difference between the YT and the TAU groups after correction for multiple comparisons with some borderline significance for the domain attention. However, if we now consider the univariate test as a post-hoc comparison, we do not need adjustment for multiple comparisons and can conclude that indeed the two

groups are significantly different in attention domain thus confirming the previous finding where we used only univariate Kruskal-Wallis test corrected for multiple comparisons (Bhatia et al., 2012). The results of unweighted E-MKW, weighted E-MKW are seen to be similar. We attribute this to the fact that except for missing pattern 2, other missing patterns have similar proportions of missingness (Table 6).

4. DISCUSSION

In clinical trials with multivariate outcomes the classical parametric methods for group comparisons have two major drawbacks. First, they require distributional assumptions such as multivariate normality. When the sample size is small, or response variables are ordinal, the use of parametric multivariate methods seems to be problematic. Second, when the multivariate tests are performed using standard software, incomplete cases are deleted and all information is lost. Nonparametric multivariate methods are available in the statistical literature. They circumvent the distributional assumptions, but the issue with missing data remains. The usual approach is to resort to univariate nonparametric approaches followed by correction for multiple comparisons. However, with correlated multivariate data the usual corrections may not be appropriate. Hence, global tests should be considered.

In this paper, we revisited the Multivariate Kruskal-Wallis (MKW) test and proposed an extension of that test to retain information from partially observed cases. We first developed the method under the MCAR assumption on the missingness and extended to MAR data where the missingness may depend on fully observed covariates or baseline variables. Our simulation results, encompassing a broad spectrum of multivariate effect sizes, show that the proposed extended test provides higher power values than the standard MKW test. In our illustrative example, we detected a group difference with the E-MKW tests and also in post-hoc comparisons. We can note here that the overall results using univariate tests and multivariate tests are similar. This may be due to small sample size, and not very strong correlations between the response variables. Other datasets may bring out the usefulness of the MKW and E-MKW over univariate methods. The validity of the E-MKW test and the corresponding permutation procedure for deriving the p-value relies on the randomness of group assignment even though the number of observations in each missing data pattern does not have to be large. Blind application of the proposed method without justifying the random assignment may lead to wrong conclusions. Because the proposed E-MKW test is nonparametric by nature and may be less powerful in detecting restricted alternatives such as $H_1: \mu_{1k} \quad \mu_{2k}$; $k = 1,\ldots,K$ with strict inequality for at least one k. Theories on restricted alternatives have been well developed for normally distributed data and data of other parametric distributions. (Silvapulle and Sen, 2005; Basso, Pesarin, Salmaso and Solari, 2009).

Acknowledgments

References

Basso, D., Pesarin, F., Salmaso, L., Solari, A. Permutation tests for stochastic ordering and ANOVA: theory and applications in R. Springer; New York: 2009.

Bhatia T, Agarwal AS, Wood J, Richard J, Gur RE, Gur RC, Nimgaonkar VL, Mazumdar S, Deshpande SN. Adjunctive cognitive remediation for schizophrenia using yoga: an open non-randomised trial. Acta Neuropsychiatr. 2012; 24(2):91–100. [PubMed: 22661830]

Cohen, J. Statistical Power Analysis for the Behavioral Sciences. 2. Erlbaum; Hillsdale, NJ: 1988.

Cohen J. A power primer. Psychol Bull. 1992; 112(1):155–159. [PubMed: 19565683]

Edgington, E., Onghena, P. Randomization tests. CRC Press; 2007.

Gur RC, Ragland JD, Moberg PJ, Bilker WB, Kohler C, Siegel SJ, Gur RE. Computerized neurocognitive scanning: II. The profile of schizophrenia. Neuropsychopharmacol. 2001; 25:777–788.

Gur RC, Ragland JD, Moberg PJ, Turner TH, Bilker WB, Kohler C, Siegel SJ, Gur RE. Computerized neurocognitive scanning: I. Methodology and validation in healthy people. Neuropsychopharmacol. 2001; 25:766–776.

IBM Corp. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp; 2013.

Katz BM, McSweeney M. A multivariate Kruskal-Wallis test with post hoc procedures. Multivar Behav Res. 1980; 15(3):281–297.

May WL, Johnson WD. A SAS macro for the multivariate extension of the Kruskal-Wallis test including multiple comparisons: randomization and chi-squared criteria. Stat Softw Newsl. 1997; 26(2):239–250.

Mazumdar S, Liu KS, Houck PR, Reynolds CF. Intent-to-treat analysis for longitudinal clinical trials: coping with the challenge of missing values. J Psychiatr Res. 1999; 33(2):87–95. [PubMed: 10221740]

Pesarin, F. Multivariate permutation tests: with applications in biostatistics. Vol. 240. Wiley; Chichester: 2001.

Puri ML, Sen PK. A class of rank order tests for a general linear hypothesis. Ann Math Stat. 1969:1325–1343.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2014. (Available from http://www.Rproject.org/)

SAS Institute Inc. *Base SAS® 9.3 Procedures Guide*. Cary, NC: SAS Institute Inc; 2014.

StataCorp. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP; 2011.

Silvapulle, MJ., Sen, PK. Constraint statistical inference. Wiley & Sons: Hoboken; New Jersey: 2005.

Zeng C, Pan Z, MaWhinney S, Baron AE, Zerbe GO. Permutation and F distribution of tests in the multivariate general linear model. Am Stat. 2011; 65(1):31–36.

Table 1

Simulation results of Type I errors

Distribution	Missingness	Unweighted [b] Statistic	Weighted[c] Statistic
Normal[d]	Medium[a]	0.056	0.05
	High[a]	0.062	0.054
Poisson[e]	Medium[a]	0.044	0.046
	High[a]	0.062	0.062

[a]Medium: M1=40%, M2=M3=30%. High: M1=20%, M2=M3=40%. $n_1 = n_2 = 50$

[b]$t_I = 1/3$, $I = 1,2,3$ (See text)

[c]$t_1 = 0.4$, $t_2 = t_3 = 0.3$ (See text)

[d]$X \sim N(0,1)$. $Y_{i1}|X \sim N(1 + X, 2)$, $Y_{i2}|X \sim N(X, 1)$, $i = 1,2$

[e]$X \sim BIN(5,0.5)$. $Y_{i1}|X \sim POI(1 + X)$, $Y_{i2}|X \sim POI(2 + X)$, $i = 1,2$

Table 2

Power simulation results for Normal[e] outcomes, varying missingness and effect sizes[a]

Effect Size	Missing Rates[a]	Standard MKW Test (Deleting All Missing Data)	Extended MKW Test (Partially Observed Data)		Standard MKW Test In Original Data (Assuming No Missing)
			Unweighted[c]	Weighted[d]	
0.08	medium	0.21	0.24	0.24	0.49
0.12	medium	0.34	0.40	0.43	0.73
	high	0.16	0.34	0.32	0.73
0.18	medium	0.51	0.61	0.64	0.93
	high	0.26	0.54	0.55	0.93
0.24	medium	0.67	0.81	0.82	0.99
	high	0.33	0.72	0.71	0.98
0.33	high	0.40	0.84	0.82	0.998

[a]Simulations based on 1000 replications

[b]Medium: M1=40%, M2=M3=30%. High: M1=20%, M2=M3=40%. $n_1 = n_2 = 50$.

[c]$t_l = 1/3$, $l = 1, 2, 3$

[d]$t_1 = 0.4$, $t_2 = t_3 = 0.3$

[e]$X \sim N(0,1)$.

Group 1: $Y_{11}|X \sim N(1 + X, 2)$, $Y_{12}|X \sim N(X, 1)$

Group 2: $Y_{21}|X \sim N(1 + X, 2)$, $Y_{22}|X \sim N(+ X, 1)$

Table 3

Power simulation results for non-Normal outcomes[e], varying missingness and effect sizes[a]

Effect Size	Missing Rates[b]	Standard MKW Test (Deleting all missing data)	Extended MKW Test (Partially observed data)		Standard MKW Test in original data (assuming no missingness)
			Unweighted[c]	Weighted[d]	
0.08	medium	0.26	0.28	0.28	0.57
	high	0.15	0.23	0.23	0.55
0.10	medium	0.32	0.41	0.40	0.67
	high	0.16	0.33	0.34	0.72
0.12	medium	0.38	0.45	0.47	0.78
	high	0.19	0.40	0.41	0.76
0.16	medium	0.43	0.58	0.58	0.89
	high	0.24	0.52	0.54	0.89
0.19	medium	0.55	0.73	0.71	0.96
	high	0.26	0.60	0.61	0.95
0.26	medium	0.73	0.87	0.89	0.99
	high	0.38	0.81	0.79	0.99
0.36	medium	0.85	0.96	0.97	0.998
	high	0.52	0.93	0.93	1

[a] Simulations based on 1000 replications

[b] Medium: M1=40%, M2=M3=30%. High: M1=20%, M2=M3=40%. $n_1 = n_2 = 50$.

[c] $t_l = 1/3$, $l = 1,2,3$

[d] $t_1 = 0.4$, $t_2 = t_3 = 0.3$

[e] $X \sim BIN(5, 0.5)$.

Group 1: $Y_{11}|X \sim POI(1 + X)$, $Y_{12}|X \sim POI(2 + X)$

Group 2: $Y_{22}|X \sim POI(1 + X)$, $Y_{22}|X \sim POI(2 + \ + X)$

Table 4

Power Simulation results for non-Normal outcomes[c] for 3 groups, 3 outcomes, and varying effect sizes[c]

μ_1, μ_2	Effect size	standard MKW test (Deleting all missing data)	extended MKW test (Partially observed data)		standard MKW test in original data (with no missing)[c]
			unweighted[a]	weighted[b]	
2 2	0.13	0.44	0.48	0.52	0.97
2 2.5	0.18	0.57	0.61	0.63	0.99
2.5 2.5	0.18	0.65	0.69	0.73	0.99
1 3	0.24	0.48	0.50	0.56	0.99
2 3	0.24	0.65	0.69	0.73	0.997
2.5 3	0.24	0.70	0.78	0.80	1
3 3	0.23	0.80	0.85	0.87	1

[a]
$t_l = 1/7, l = 1,...,7.$

[b]
$t_l = 0.3, t_l = 0.1, l = 2,...,7.$

[c]
$X \sim BIN(5,0.5)$. Group 1, $Y_{11}|X \sim POI(1 + X)$, $Y_{12}|X \sim POI(2 + X)$, $Y_{13}|X \sim POI(3 + X)$. Group 2, $Y_{21}|X \sim POI(1 + X)$, $Y_{22}|X \sim POI(2 + X + 1)$, $Y_{23}|X \sim POI(3 + X)$. Group 3, $Y_{31}|X \sim POI(1 + X)$, $Y_{32}|X \sim POI(2 + X)$, $Y_{33}|X \sim POI(3 + X + 2)$.

30% of cases are simulated with all three outcomes observed (M1). Each of 10% of cases are simulated with two outcome observed and one outcome missing (M2, M3 and M4), and each of 10% of cases are simulated with one outcome observed and two outcomes missing (M5, M6 and M7).

Table 5

Comparisons of CNB domain improvements between YT and TAU groups by univariate Kruskal-Wallis tests

Domains	Variables	Number Of Complete Cases			P-value	Adjusted P-value
		YT (N=63)	TAU (N=24)	All		
Abstraction and Mental Flexibility	Y_1	23	21	44	0.028	0.084
Attention	Y_2	18	16	34	0.014	0.056
Face Memory	Y_3	26	22	48	0.069	0.138
Spatial Memory	Y_4	24	19	43	0.66	0.66

Table 6

Missing pattern in improvement in four selected cognitive domain scores for all patients in both YT and TAU groups (n=87), O=Observed, M=Missing

Missing Pattern	Y_1	Y_2	Y_3	Y_4	m_1	Used in MANOVA and Standard MKW Test	Used in Extended MKW Test
1	O	O	O	O	32	Yes	Yes
2	O	M	O	O	9	No	Yes
3	M	M	O	O	2	No	No
4	O	O	O	M	1	No	No
5	M	O	O	M	1	No	No
6	M	M	O	M	3	No	Yes
7	O	M	M	M	2	No	Yes
8	M	M	M	M	37	No	No

[a] Y_1 = Abstraction and Mental Flexibility, Y_2 = Attention, Y_3 = Face Memory, and Y_4 = Spatial Memory.

Table 7

Comparisons of CNB domain improvements between YT and TAU groups by MANOVA and MKW and E-MKW tests

Test		n	Large Sample Approximation	Permutation-Based
MANOVA, Wilks' λ		32	0.054	0.081[b]
Standard MKW Tests		32	0.038[a]	0.030
Extended MKW Tests	Unweighted	46	-	0.031
	Weighted	46	-	0.034

[a]Approximated by χ_4^2 distribution

[b]Based on a permutation test proposed by Zeng et al. (2011)