Statistics and Its Interface Volume 2 (2009) 197–201

A weighted rank-sum procedure for comparing samples with multiple endpoints*

Qizhai Li, Aiyi Liu, Kai Yu and Kai F. Yu†

For comparing the distribution of two samples with multiple endpoints, O'Brien (1984) proposed rank-sum-type test statistics. Huang et al. (2005) extended these statistics to the general nonparametric Behrens-Fisher hypothesis problem and obtained improved test statistics by replacing the ad hoc variance with the asymptotic variance of the rank-sum statistics. In this paper we generalize the work of O'Brien (1984) and Huang et al. (2005) and propose a weighted rank-sum statistic. We show that the weighted rank-sum statistic is asymptotically normally distributed, permitting the computation of power, p-values and confidence intervals. We further demonstrate via simulation that the weighted rank-sum statistic is efficient in controlling the type I error rate and under certain alternatives, is more powerful than the statistics of O'Brien (1984) and Huang et al. (2005).

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35.
Keywords and phrases: Asymptotic normality, Behrens-Fisher problem, Case-Control, Clinical trials, Multiple endpoints, Rank-sum statistics, Weights.

1. INTRODUCTION

Comparison of two or more samples with multiple endpoints is a common statistical problem in biomedical research. As an example, O'Brien (1984) described a randomized clinical trial of two therapies for the treatment of diabetes to investigate whether the experimental therapy yields better nerve function as measured by 34 electromyographic variables. Huang et al. (2005) gave another example of a clinical trial of Coenzyme Q_{10} in treating early Parkinson's disease to slow the functional decline of the disease, as indexed by a number of outcome measures, including mentation, motor and average daily living scales. Other examples can be found in Pocock, Geller and Tsiatis (1987), Shames et al. (1998), Tilley et al. (2000), and Li, Zhao and Paty (2001), to name a few.

Hotelling's T^2 and the Bonferroni procedure are two popular approaches for comparing two multivariate samples. Hotelling's T^2 is a global test statistic and makes no distinction between variables in their direction of change. The Bonferroni procedure assigns the Type I error for each variable and then tests the null hypothesis concerning each individual variable. Noting these drawbacks of the two methods, O'Brien (1984) proposed a nonparametric procedure, a rank-sum-type test, which is based on the rank of each individual variable among the combined observations from the two samples. Under the null hypothesis that the two multivariate samples have the same distribution, O'Brien's (1984) rank-sum test statistic asymptotically is distribution-free and follows a standard normal distribution. Huang et al. (2005) noticed that under a more general null hypothesis in the Behrens-Fisher problem, e.g. Troendle (2002), O'Brien's (1984) test statistics are no longer distribution-free and can substantially inflate the Type I error rate when used for testing the general Behrens-Fisher hypothesis. Subsequently Huang et al. (2005) provided a modification of O'Brien's (1984) test by adjusting for the variances of the rank sums.

Generalizing O'Brien's (1984) rank-sum test and the modified test of Huang et al. (2005), we propose a weighted-rank-sum statistic for testing the general nonparametric Behrens-Fisher hypothesis. The weights can be chosen to be constants emphasizing the importance of the individual variables, or they can be chosen to minimize the variance of the weighted-rank-sum statistic. Under mild conditions, the weighted-rank-sum statistic is asymptotically normally distributed, thus permitting the computation of power, p-values and confidence intervals. Simulation studies demonstrate that the weighted rank-sum statistic is efficient in controlling type I error and is more powerful than the statistics of O'Brien (1984) and Huang et al. (2005) for certain alternatives.

2. WEIGHTED-RANK-SUM STATISTICS FOR THE BEHRENS-FISHER PROBLEM

Suppose our interest is to compare the distribution of two p-dimensional variables, $X = (X_1, \ldots, X_p)'$, and $Y = (Y_1, \ldots, Y_p)'$, representing the outcomes of p endpoints from subjects in, say, the standard therapy arm and the experimental therapy arm in a clinical trial, or the controls and cases in a case-control study, respectively. We assume that X and Y follow distributions F and G, with marginal distributions F_a and G_a of X_a and Y_a respectively, where

*This research has been supported by the Intramural Research Program of the National Institutes of Health and the Knowledge Innovation Program of the Chinese Academy of Sciences.
†Corresponding author.

$a = 1, \ldots, p$. Following Huang et al. (2005), we define

$$\theta_a = \Pr(X_a < Y_a) - \Pr(X_a > Y_a), \quad a = 1, \ldots, p, \tag{1}$$

and consider testing the null hypothesis

$$H_0 : \quad \theta_1 = \cdots = \theta_p = 0. \tag{2}$$

This is a nonparametric version of the Behrens-Fisher problem. The null space under H_0, $\{(F, G) : \theta_1 = \cdots = \theta_p = 0\}$, is larger than the usual null space under $H_0' : F = G$. In a clinical trial setting θ_a can be viewed as a measure of the marginal treatment efficacy (corresponding to the ath endpoint) of the experimental therapy relative to the standard therapy, assuming that larger outcomes indicate better treatment results. Thus a larger positive value of θ_a indicates better treatment results with respect to the ath outcome variable for the experimental therapy than the standard.

2.1 Rank-sum type test statistics

Let $x_i = (x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots, m$, be the outcomes for the ith subject from the X-sample and $y_j = (y_{j1}, \ldots, y_{jp})'$, $j = 1, \ldots, n$, be the outcomes of the jth subject from the Y-sample, and write $N = m+n$. For the ath outcome variable, $a = 1, \ldots, p$, we combine the two samples and rank the N observations $x_{1a}, \ldots, x_{ma}, y_{1a}, \ldots, y_{na}$, and denote by R_{xia} and R_{yja}, the midrank of x_{ia} and y_{ja}, respectively. Then we observe $S_{xi} = \sum_{a=1}^{p} R_{xia}$ for each subject, $i = 1, \ldots, m$, from the X-sample, and $S_{yj} = \sum_{a=1}^{p} R_{yja}$ for each subject, $j = 1, \ldots, n$ from the Y-sample, by summing up the ranks of the p variables. O'Brien (1984) suggested reducing the problem of comparing two multivariate distributions to one of comparing the rank sums between $\{S_{xi}, i = 1, \ldots, m\}$ and $\{S_{yj}, j = 1, \ldots, n\}$ using the usual two-sample t-tests. This yields a t-test statistic

$$T_1 = \frac{\bar{S}_y - \bar{S}_x}{\hat{\sigma}\sqrt{1/m + 1/n}}, \quad T_2 = \frac{\bar{S}_y - \bar{S}_x}{\sqrt{\hat{\sigma}_x^2/m + \hat{\sigma}_y^2/n}}, \tag{3}$$

analogous to the usual two-sample t-test with equal variances and unequal variances, respectively, where $\bar{S}_x = \sum_{i=1}^{m} S_{xi}/m$, $\bar{S}_y = \sum_{j=1}^{n} S_{yj}/n$, $\hat{\sigma}_x^2 = \sum_{i=1}^{m}(S_{xi} - \bar{S}_x)^2/(m-1)$, $\hat{\sigma}_y^2 = \sum_{j=1}^{n}(S_{yj} - \bar{S}_y)^2/(n-1)$, and $\hat{\sigma}^2 = ((m-1) \times \hat{\sigma}_x^2 + (n-1)\hat{\sigma}_y^2)/(N-2)$.

Huang et al. (2005) noticed that under the more restricted null hypothesis, $H_0' : F = G$, both T_1 and T_2 asymptotically follow the standard normal distribution. However, when $F \neq G$ these two statistics remain asymptotically normally distributed, but with non-unit variances. When used to test the Behrens-Fisher hypothesis H_0, these test statistics can substantially inflate the Type I error rate, as demonstrated in Huang et al. (2005). To make O'Brien's (1984) test suitable for testing the null hypothesis H_0, Huang et al. (2005) derived the asymptotic variances of the two statistics and suggested using the following two modified test statistics for H_0:

$$T_{1a} = \frac{\bar{S}_y - \bar{S}_x}{\hat{\sigma}\sqrt{\hat{h}_1(1/m + 1/n)}}, \quad T_{2a} = \frac{\bar{S}_y - \bar{S}_x}{\sqrt{\hat{h}_2(\sigma_x^2/m + \sigma_y^2/n)}}, \tag{4}$$

where $\hat{h}_1$ and $\hat{h}_2$ are consistent estimates of

$$h_1 = \frac{\sum_{a=1}^{p}\sum_{b=1}^{p}(1+\lambda)^2(c_{ab} + d_{ab}\lambda)}{\sum_{a=1}^{p}\sum_{b=1}^{p}(e_{ab}\lambda^3 + (d_{ab} + 2f_{ab})\lambda^2 + (c_{ab} + 2\eta_{ab})\lambda + \xi_{ab})},$$

and

$$h_2 = \frac{\sum_{a=1}^{p}\sum_{b=1}^{p}(1+\lambda)^2(c_{ab} + d_{ab}\lambda)}{\sum_{a=1}^{p}\sum_{b=1}^{p}(d_{ab}\lambda^3 + (e_{ab} + 2\eta_{ab})\lambda^2 + (\xi_{ab} + 2f_{ab})\lambda + c_{ab})},$$

respectively, with $\lambda = m/n$, $c_{ab} = \mathrm{Cov}(G_a(X_a), G_b(X_b))$, $d_{ab} = \mathrm{Cov}(F_a(Y_a), F_b(Y_b))$, $e_{ab} = \mathrm{Cov}(F_a(X_a), F_b(X_b))$, $f_{ab} = \mathrm{Cov}(F_a(X_a), G_b(X_b))$, and $\xi_{ab} = \mathrm{Cov}(G_a(Y_a), G_b(Y_b))$, $\eta_{ab} = \mathrm{Cov}(G_a(Y_a), F_b(Y_b))$.

Huang et al. (2005) further showed that, under the null hypothesis H_0, the two test statistics asymptotically follow the standard normal distribution and thus the Type I error rate can be controlled at significance level α by rejecting H_0 if the magnitude of the test statistic exceeds the critical value of $\Phi^{-1}(1 - \alpha/2)$, where Φ is the standard normal distribution function.

2.2 Weighted rank-sum statistics

O'Brien's (1984) rank-sum test and the modified version of Huang et al. (2005) gave equal weights to the rank of each individual variable. In many situations unequal weights are desirable so that different emphasis can be assigned to different variables. Moreover statistical optimization requires that the weights be proportional to the reciprocals of the variances of the variables to be combined, when the variables are mutually independent. Taking these arguments into consideration, we propose using weighted rank-sum statistics. Let $w_a \geq 0$, $a = 1, \ldots, p$, be a constant or a random variable. The weighted rank for the ith subject in the X-sample is defined as $R_{xi} = \sum_{a=1}^{p} w_a R_{xia}$, $i = 1, \ldots, m$, and the weighted rank for the jth subject in the Y-sample is $R_{yj} = \sum_{a=1}^{p} w_a R_{yja}$, $j = 1, \ldots, n$. Setting $w_a = 1$ for every a, $a = 1, \ldots, p$ leads to the test statistics considered by O'Brien (1984) or Huang et al. (2005). Moreover, if only a subset of the p variables are of interest, we can set the weights to be one for variables in the subset and zero for variables not in the subset.

Using the weighted rank sum, we propose the following two test statistics:

(5)
$$T_{w1} = \frac{\bar{R}_y - \bar{R}_x}{\hat{\sigma}\sqrt{\hat{h}_{w1}(1/m+1/n)}}, \quad T_{w2} = \frac{\bar{R}_y - \bar{R}_x}{\sqrt{\hat{h}_{w2}(\hat{\sigma}_x^2/m+\hat{\sigma}_y^2/n)}},$$

where $\bar{R}_x = \sum_{i=1}^m R_{xi}/m$, $\bar{R}_y = \sum_{j=1}^m R_{yj}/n$, $\hat{\sigma}_x^2 = \sum_{i=1}^m (R_{xi}-\bar{R}_x)^2/(m-1)$, $\hat{\sigma}_y^2 = \sum_{j=1}^n (R_{yj}-\bar{R}_y)^2/(n-1)$, $\hat{\sigma}^2 = [(m-1)\hat{\sigma}_x^2 + (n-1)\hat{\sigma}_y^2]/(m+n-2)$, and $\hat{h}_{w1}$ and $\hat{h}_{w2}$ are consistent estimates (e.g. empirical estimates) of $h_{w1} =$

$$\frac{\sum_{a=1}^p \sum_{b=1}^p (1+\lambda)^2 w_a w_b (c_{ab}+d_{ab}\lambda)}{\sum_{a=1}^p \sum_{b=1}^p w_a w_b [e_{ab}\lambda^3 + (d_{ab}+2f_{ab})\lambda^2 + (c_{ab}+2\eta_{ab})\lambda + \xi_{ab}]},$$

and $h_{w2} =$

$$\frac{\sum_{a=1}^p \sum_{b=1}^p (1+\lambda)^2 w_a w_b (c_{ab}+d_{ab}\lambda)}{\sum_{a=1}^p \sum_{b=1}^p w_a w_b [d_{ab}\lambda^3 + (e_{ab}+2\eta_{ab})\lambda^2 + (\xi_{ab}+2f_{ab})\lambda + c_{ab}]},$$

respectively, with c_{ab}, d_{ab}, e_{ab}, f_{ab}, ξ_{ab} and η_{ab} as for Eq. (4).

We have the following results.

Theorem 1. *Under the null hypothesis H_0, T_{w1} and T_{w2} both converge in distribution to the standard normal distribution as $\min\{m,n\} \to \infty$ and $0 < \lambda = \lim\{m/n\} < \infty$.*

Therefore H_0 is rejected if $|T_{w1}|$ or $|T_{w2}|$ is larger than $\Phi^{-1}(1-\alpha/2)$; both tests asymptotically maintain the Type I error rate at the nominal level of α.

Proof of Theorem 1. Denote the indicator function on a set by $I_{\{\cdot\}}$. Let $\Lambda = (\lambda_{ab})$, with $\lambda_{ab} = \mathrm{Cov}(\bar{R}_{ya} - \bar{R}_{xa}, \bar{R}_{yb} - \bar{R}_{xb})$, and let $w = (w_1, \ldots, w_p)'$.

$$\begin{aligned}\bar{R}_y - \bar{R}_x &= \frac{m+n}{2mn} \sum_{a=1}^p w_a \sum_{i=1}^m \sum_{j=1}^n \left\{ I_{\{x_{ia}<y_{ja}\}} - I_{\{x_{ia}>y_{ja}\}} \right\} \\ &= \sqrt{m+n} w' U,\end{aligned}$$

where

$$U = \frac{\sqrt{m+n}}{2mn} \left(\sum_{i=1}^m \sum_{j=1}^n \left\{ I_{\{x_{i1}<y_{j1}\}} - I_{\{x_{i1}>y_{j1}\}} \right\}, \ldots, \sum_{i=1}^m \sum_{j=1}^n \left\{ I_{\{x_{ip}<y_{jp}\}} - I_{\{x_{ip}>y_{jp}\}} \right\} \right)'.$$

Note that U is a p-dimensional vector with each element being a U-statistic. It follows from standard asymptotic theory on U-statistics that, under the null hypothesis H_0, U converges to a p-variate normal distribution with mean $\mathbf{0} = (0, \ldots, 0)'$ and variance-covariance matrix $\triangle = \mathrm{Cov}(U)$ as $\min\{m,n\} \to \infty$ and $0 < m/n < \infty$. Therefore $(\bar{R}_y - \bar{R}_x)/\sqrt{w'\triangle w}$ asymptotically follows a standard normal distribution under the null hypothesis H_0. Hence, it suffices to show that $\hat{\sigma}\sqrt{\hat{h}_{w1}(1/m+1/n)}$ and $\sqrt{\hat{h}_{w2}(\hat{\sigma}_x^2/m + \hat{\sigma}_y^2/n)}$ are consistent estimates of $\sqrt{w'\triangle w}$, which can be derived following the proof of Theorem 1 and Theorem 2 in Huang et al. (2005). □

3. SELECTION OF WEIGHTS

The weights w can be chosen to meet practical needs, for example, to exclude some variables by setting the weights to be zero. In other situations the weights can be so determined that the variance of the weighted rank-sum is minimized at certain parameter values in the null or alternative hypothesis space. Here we search for the w that minimizes the variance $V(w)$ of $\bar{R}_y - \bar{R}_x$ under the null hypothesis. To this end, we first give the following definitions. For any $a \in \{1, \ldots, p\}$, define $R_y(x_{ia})$ to be the midrank of x_{ia} among $\{x_{ia}, y_{1a}, \ldots, y_{na}\}$, $R_x(x_{ia})$ the midrank of x_{ia} among $\{x_{1a}, \ldots, x_{ma}\}$, $R_x(y_{ja})$ the midrank of y_{ja} among $\{x_{1a}, \ldots, x_{ma}, y_{ja}\}$, $R_y(y_{ja})$ the midrank of y_{ja} among $\{y_{1a}, \ldots, y_{na}\}$. Let I_k be the identity matrix of order k, J_k be the column vector of order k whose elements are 1, and define $H_k = I_k - J_k J_k'/k$. Then following Huang et al. (2005), we can obtain consistent estimates of $\hat{h}_{w1}$ and $\hat{h}_{w2}$ as

$$\hat{h}_{w1} = \frac{s^2}{mn} \times \frac{w'(P'P + U'U)w}{w'[(P+Q)'(P+Q) + (U+V)'(U+V)]w},$$

and

$$\hat{h}_{w2} = s^2 \times \frac{w'(P'P + U'U)w}{w'[n^2(P+Q)'(P+Q) + m^2(U+V)'(U+V)]w},$$

respectively, where $P = (p_{ia})_{m\times p}$ with $p_{ia} = 2R_y(x_{ia}) - 2 - n + n\hat{\theta}_a$, $Q = (q_{ia})_{m\times p}$ with $q_{ia} = 2R_x(x_{ia}) - 1 - m$, $U = (u_{ja})_{n\times p}$ with $u_{ja} = 2R_x(y_{ja}) - 2 - m - m\hat{\theta}_a$, and $V = (v_{ja})_{n\times p}$ with $q_{ja} = 2R_y(y_{ja}) - 1 - n$, where $i = 1, \ldots, m$, $a = 1, \ldots, p$, $j = 1, \ldots, n$, and $\hat{\theta}_a = \sum_{i=1}^m \sum_{j=1}^n \{I_{\{x_{ia}<y_{ja}\}} - I_{\{x_{ia}>y_{ja}\}}\}/(mn)$.

Hence, for T_{w1} and T_{w2}, we have the estimated variances of $\bar{R}_y - \bar{R}_x$,

$$\widehat{V_1(w)} = \frac{(m+n)^3}{m^2n^2(m+n-2)} \times \frac{w'[M_x'H_mM_x + M_y'H_nM_y]ww'(P'P+U'U)w}{w'[(P+Q)'(P+Q)+(U+V)'(U+V)]w},$$

and

$$\widehat{V_2(w)} = (m+n)^2 \times \frac{w'\left[\frac{M_x'H_mM_x}{m(m-1)} + \frac{M_y'H_nM_y}{n(n-1)}\right]ww'(P'P+U'U)w}{w'[n^2(P+Q)'(P+Q)+m^2(U+V)'(U+V)]w},$$

where $M_x = (R_{xia})$ and $M_y = (R_{yja})$, the rank matrix for the X-sample and Y-sample, respectively.

The optimal weights w_1 (or w_2) are those that minimize the variance of $\bar{R}_y - \bar{R}_x$, and they can be estimated by

$$\hat{w}_1 = \text{argmin}_{w' J_p = 1, w \geq 0} \widehat{V_1(w)},$$
$$\hat{w}_2 = \text{argmin}_{w' J_p = 1, w \geq 0} \widehat{V_2(w)}.$$

The weights and their estimates can be computed only numerically, since there are no closed forms. Furthermore numerical results show that $\hat{w}_1 \approx \hat{w}_2$. This is understandable since, as pointed out earlier, both $\hat{\sigma}\sqrt{\hat{h}_{w1}(1/m + 1/n)}$ and $\sqrt{\hat{h}_{w2}(\hat{\sigma}_x^2/m + \hat{\sigma}_y^2/n)}$ are consistent estimates of $\sqrt{w' \triangle w}$.

4. SIMULATION STUDY AND REAL DATA EXAMPLE

4.1 Simulation studies

In this section, we conduct a simulation study to evaluate the type I error rate and power of the proposed tests, T_{w1} and T_{w2}, for comparison with those of O'Brien (1984), T_1 and T_2, and Huang et al. (2005), T_{h1} and T_{h2}. To this end, we consider $X = (X_{i1}, X_{i2})'$, $i = 1, \ldots, m$, random samples from a bivariate normal distribution with mean $(0, 0.5)'$ and variance-covariance matrix $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$, and $Y = (Y_{j1}, Y_{j2})'$, $j = 1, \ldots, n$, random samples from a bivariate normal distribution with mean $(0, 0.5)'$ and variance-covariance matrix $\begin{pmatrix} 4 & 7.2 \\ 7.2 & 16 \end{pmatrix}$. Clearly the null hypothesis holds with these two distributions, i.e., for any i and j, $Pr(X_{i1} < Y_{j1}) - Pr(X_{i1} > Y_{j1}) = Pr(X_{i2} < Y_{j2}) - Pr(X_{i2} > Y_{j2}) = 0$. We generate 10,000 replicates for each pair of m and n selected from $\{50, 100, 200\}$. For each replicate the optimal weights are estimated from the simulated data using the method described in Section 3. The simulated Type I error is the proportion of the null hypothesis H_0 being rejected at nominal significance level of 0.05 (two-sided).

The simulated power is obtained similarly under the same settings except that the mean vector of the X-samples is set to $(0, -0.5)'$, and the variance-covariance matrices are set to $\begin{pmatrix} 1 & 1.13 \\ 1.13 & 2 \end{pmatrix}$ for the X-sample and $\begin{pmatrix} 2 & 2.55 \\ 2.55 & 4 \end{pmatrix}$ for the Y-sample.

Table 1 presents the simulation results for the type I error and power. The results indicate that both the methods in the present paper and in Huang et al. (1984) effectively maintain the nominal type I error, with a minor discrepancy possibly due to variation in the simulation. In comparison, O'Brien's (1984) tests produce inconsistent type I error rates, mostly inflated over the nominal significance level of 0.05. For example, with $m = 200$, $n = 100$, the empirical type I error rates of O'Brien's tests are 0.101 and 0.059, while the tests in Huang et al. (2005) are 0.051 and 0.050, and the proposed two tests give 0.054 and 0.053, respectively. The power of the proposed tests is substantially higher than those of O'Brien (1984) and Huang et al. (2005). For example, with $m = 100$, $n = 50$, the power values are, respectively, 0.394 and 0.351 for O'Brien's (1984) tests, 0.360 and 0.355 for the tests of Huang et al. (2005), and 0.567 and 0.563, for the proposed tests, more than 15% higher than other tests. It is worth noting that even when O'Brien's (1984) tests produce smaller type I error ($m = 50, n = 100$ and $m = 100, n = 200$), the proposed tests still achieve a considerably higher power than other tests.

Table 1. Type I error and power results under significance level, 0.05. (10,000 replicates)

m	n	T_1	T_2	T_{h1}	T_{h2}	T_{w1}	T_{w2}
Type I error rate							
50	50	0.064	0.063	0.052	0.051	0.056	0.055
100	100	0.066	0.066	0.052	0.051	0.052	0.052
200	200	0.061	0.061	0.048	0.048	0.049	0.049
50	100	0.031	0.069	0.051	0.051	0.053	0.053
100	200	0.029	0.067	0.049	0.050	0.050	0.051
100	50	0.105	0.061	0.056	0.054	0.060	0.059
200	100	0.101	0.059	0.051	0.050	0.054	0.053
Power							
50	50	0.311	0.311	0.316	0.315	0.512	0.511
100	100	0.534	0.533	0.535	0.535	0.734	0.734
200	200	0.825	0.825	0.825	0.825	0.933	0.933
50	100	0.375	0.431	0.436	0.433	0.637	0.635
100	200	0.664	0.716	0.717	0.716	0.866	0.865
100	50	0.394	0.351	0.360	0.355	0.567	0.563
200	100	0.641	0.591	0.596	0.593	0.790	0.788

4.2 An example

The role of certain growth hormones is one major objective of The Growth and Maturation in Children with Autism or Autistic Spectrum Disorder (ASD) Study (the Autism/ASD Study), a case-control study conducted by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development in 2002–2005; see Mills et al. (2007) for details of subject enrollment and data collection. The study enrolled 81 subjects, 75 boys and 6 girls, diagnosed as having autism/ASD, and 80 age-matched controls (59 boys and 21 girls). Blood samples were assayed for insulin-like growth factors (IGF-1, IGF-2), insulin-like growth factor binding protein (IGFBP-3), and growth hormone binding protein (GHBP), as well as for dehydroepiandrosterone (DHEA) and DHEA-sulphate (DHEAS).

To illustrate the proposed methods in comparison with other approaches, we exclude from the analysis data from the girls, due to their small sample size, and four boys in the case group who did not provide blood samples, thus yielding 71 cases and 59 controls in the analysis. We confine our attention to five hormones: insulin-like growth factor-1 (IGF-1), insulin-like growth factor 2 (IGF-2), IGF binding protein (IGFBP-3), growth hormone binding protein

(GHBP), and dehydroepiandrosterone (DHEA). DHEA-sulphate (DHEAS) was not included in the analysis since its levels were undetectable in more than half of the subjects (Mills et al., 2007). To be investigated is whether the levels of a growth-related hormone, if any, differ between cases and controls.

We applied the proposed test and the tests of O'Brien (1984) and Huang et al. (2005) to the five growth-related hormone levels in cases and controls. The P-values were 2.86×10^{-6} and 1.75×10^{-6} for O'Brien's tests and 6.28×10^{-7} and 6.30×10^{-7} for the two tests of Huang et al. In contrast the proposed two tests give P-values 1.93×10^{-7} and 3.33×10^{-7}, respectively. For this example, the proposed method is shown to be more powerful than O'Brien's method, but only slightly better than the tests of Huan et al. (2005). This is partly because the differences between cases and controls all fall into the same direction, that is, for each hormone, its levels among cases are higher than among controls.

5. DISCUSSION

For testing the Behrens-Fisher hypothesis, we proposed a weighted rank-sum test statistic that effectively maintains the type I error rate and possesses higher power than the tests of O'Brien (1984) and Huang et al. (2005). The optimal weights do not have closed forms and need to be estimated using available data.

All the tests discussed are nonparametric in nature and are "global" in the sense that they summarize the multi-dimensional data into one-dimensional statistics. Under the most restricted null hypothesis that the two multivariate distributions are identical these tests are asymptotically equivalent. However, under the less restrictive Behrens-Fisher hypotheses, they perform differently. It would be interesting to see how these test statistics behave under other hypotheses.

The proposed test gains its power by accumulating evidence across comparisons on each individual outcomes, but may still have low power in situations when the differences in the outcomes between the two samples, as measured by the θ_as, exist but fall into different directions (some differences are positive and some are negative or zero). In this regard, more robust tests, such as the one described in Yu et al. (2006), could serve as plausible alternatives.

ACKNOWLEDGEMENTS

We would like to thank Dr. B. J. Stone for help. The opinions expressed in the article are not necessarily those of the National Institutes of Health. The authors thank the referee, an Associate Editor and the Editor for helpful comments and suggestions.

Received 15 September 2008

REFERENCES

Huang, P., Tilley, B. C., Woolson, R. F. and Lipsitz, S. (2005). Adjusting O'Brien's test to control type I error for the generalized nonparametric Behrens-Fisher problem. *Biometrics* **61**, 532–539. MR2140925

Li, D. K., Zhao, G. J., Paty, D. W. (2001). Randomized controlled trial of interferon-beta-1a in secondary progressive MS: MRI results. *Neurology* **56**, 1505–1513.

Mills, J. L., Hediger, M. L., Molloy, C. A., Chrousos, G. P., Manning-Courtney, P., Yu, K. F., Brasington, M. and England, L. J. (2007). Elevated levels of growth-related hormones in autism and autism spectrum disorder. *Clinical Endocrinology* **67**, 230–237.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087. MR0786180

Pocock, S. J., Geller, N. L. and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487–498. MR0909756

Shames, R. S., Heilbron, D. C., Janson, S. L., Kishiyama, J. L., Au, D. S. and Adelman, D. C. (1998). Clinical differences among women with and without self-reported perimenstrual asthma. *Annals of Allergy Asthma Immunology* **81**, 65–72.

Tilley, B. C., Pillemer, S. R., Heyse, S. P., Li, S., Clegg, D. O. and Alarcó, G. S. (1999). Global statistical tests for comparing multiple outcomes in rheumatoid arthritis trials. *Arthritis and Rheumatism* **42**, 1879–1888.

Troendle, J. F. (2002). A likelihood ratio test for the nonparametric Behrens-Fisher problem. *Biometrical Journal* **44**, 813–824. MR1934965

Yu, K., Gu, C., Xiong, C. J., An, P. and Province, M. A. (2005). Global transmission/disequilibrium tests for haplotypes reconstructed from multiple genes. *Genetic Epidemiology* **29**, 323–335.

Qizhai Li
Biostatistics Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute, USA
Academy of Mathematics and Systems Science
Chinese Academy of Sciences, China
E-mail address: liqz@amss.ac.cn

Aiyi Liu
Biostatistics and Bioinformatics Branch
Eunice Kennedy Shriver
National Institute of Child Health
and Human Development, USA
E-mail address: liua@mail.nih.gov

Kai Yu
Biostatistics Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute, USA
E-mail address: yuka@mail.nih.gov

Kai F. Yu
Biostatistics and Bioinformatics Branch
Eunice Kennedy Shriver
National Institute of Child Health
and Human Development, USA
E-mail address: yukf@mail.nih.gov
url: http://www.nichd.nih.gov/about/org/despr/bms/