STATISTICS IN MEDICINE
*Statist. Med.* 2009; **28**:1144–1158
Published online 20 January 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3536

# A comparison of tests for restricted orderings in the three-class case

Todd A. Alonzo[1,*,†], Christos T. Nakas[2], Constantin T. Yiannoutsos[3] and Sherri Bucher[4]

[1]*Division of Biostatistics, University of Southern California, Keck School of Medicine 440 E. Huntington Dr, Suite 400, Arcadia, CA 91006, U.S.A.*
[2]*Laboratory of Biometry, School of Agricultural Sciences, University of Thessaly, Magnesia, Greece*
[3]*Division of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, U.S.A.*
[4]*Department of Pediatrics, Neonatal and Perinatal Medicine Section and IU-Kenya Program, Indiana University School of Medicine, Indianapolis, IN, U.S.A.*

SUMMARY

A variety of methods for comparing three distributions have been proposed in the literature. These methods assess the same null hypothesis of equal distributions but differ in the alternative hypothesis they consider. The alternative hypothesis can be that measurements from the three classes are distributed according to unequal distributions or that measurements between the three classes follow a specific monotone ordering, an inverse-U-shaped (umbrella) ordering, or a U-shaped (tree) ordering. This paper compares these tests with respect to power and test size under different simulation scenarios. In addition, the methods are illustrated in two applications generated by different research questions with data from three classes suggesting monotone and umbrella orders. Additionally, proposals for the appropriate application of these tests are provided. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: ANOVA; non-parametric tests; ordered alternatives; umbrella alternatives; umbrella ROC volume; volume under the ROC surface; area under ROC curve; Wilcoxon–Mann–Whitney test; Kruskal–Walis test

## 1. INTRODUCTION

A variety of methods have been proposed for comparing treatment effects or, more generally, comparing relative locations (medians) of different populations. This paper focuses on the setting with three populations because this research was motivated by projects with three populations. However, the methods discussed in this paper are also applicable to settings with more than three

*Correspondence to: Todd A. Alonzo, Division of Biostatistics, University of Southern California, Keck School of Medicine 440 E. Huntington Dr, Suite 400, Arcadia, CA 91006, U.S.A.
†E-mail: talonzo@childrensoncologygroup.org

*Received 9 June 2008*
*Accepted 17 December 2008*

populations of interest. The methods considered assess the same null hypothesis of no difference among the treatment effects but differ in the alternative hypotheses considered. These can be a general alternative hypotheses that at least two treatment effects are not equal. Otherwise, the alternative hypothesis might involve specific orderings of the classes such that measurements among the three classes follow a monotone ordering (e.g. class 1 $<$ class 2 $<$ class 3). An inverse-U-shaped (umbrella) ordering can also be considered where the hypothesis is that two classes have smaller measurements, compared with a third class without imposing an ordering between these two classes (e.g. class 1 $<$ class 2 $>$ class 3) or even a U-shaped (tree) ordering where two classes have larger measurements compared with a third class, without imposing an ordering between these two classes, (e.g. class 1 $>$ class 2 $<$ class 3). The research question under study dictates the alternative hypothesis to be tested. Monotone orderings are often observed in toxicity studies where the risk of the occurrence of adverse events is expected to rise with increasing dose levels. On the other hand, umbrella orderings are commonly encountered in efficacy studies, in which treatment efficacy is expected to increase with dose only until a maximum efficacy level is reached [1].

Recently, there has been increased focus directed toward evaluation of the accuracy of new biomarkers being developed for a variety of medical conditions. Receiver operating characteristic (ROC) curves and the summary measure area under the ROC curve (AUC) are the standard approaches for assessing the ability of biomarkers measured on a continuous scale to accurately distinguish between two disease states or classes (e.g. presence vs absence of cancer) [2]. ROC analysis has been extended to accommodate three disease states. Specifically, ROC surfaces and the corresponding summary measure volume under the ROC surface (VUS) have been proposed when a monotone ordering is of interest (see, e.g. [3]). The umbrella ROC graph and summary measure umbrella volume (UV) have been recently proposed when umbrella orderings are of interest [4]. Comparisons of UV and VUS have been made [4], but comparisons have not been made with parametric and non-parametric approaches developed to compare treatment effects or population location rather than diagnostic accuracy where particular order restrictions are of interest. Therefore, a novel contribution of this paper is to provide a comparison of UV and VUS, methods which assess the ability of continuous biomarkers to accurately distinguish between three disease states, versus parametric and non-parametric methods that test particular orderings of the locations of three populations.

Brief descriptions of the different tests are provided in Section 2. Simulation studies to evaluate the power and size of the tests are summarized in Section 3. In Section 4 the methods are applied to determine the ability of a biomarker to accurately distinguish between HIV-negative persons and HIV-positive patients with and without HIV-related neurological sequelae. A second example concerns the assessment of the effects of different doses of haloperidol on the motor activity of juvenile rats. We end with a discussion of the findings and include recommendations on the appropriate use of each method.

## 2. DESCRIPTION OF METHODS

Consider the setting when there are three classes of interest, denoted 1, 2, and 3. Let $Y_{ij}$, $i = \{1,2,3\}$, $j = 1,\ldots,n_i$ be the observed measurements for the three classes 1, 2, and 3, respectively. There are a total of $N = n_1+n_2+n_3$ measurements. For ease of presentation, it is assumed that there are $N$ unique measurements. Methods to correct for ties in the data are available but are not discussed here.

### 2.1. General alternative hypothesis

The Kruskal–Wallis (KW) test is a distribution-free test that addresses the null hypothesis that there is no treatment effect or equivalently no difference in population locations against the general alternative hypothesis that at least two treatment effects or disease classes are not equal [5]. The KW test statistic is calculated by first ordering all $N$ observations from smallest to largest. Let $r_{ij}$ denote the rank of $Y_{ij}$ in the combined data. Let $R_i$ be the sum of the ranks for group $i$ and let $R_{i.}$ be the average rank for group $i$. Then the KW test statistic is given by

$$\mathrm{KW}=\frac{12}{N(N+1)}\sum_{i=1}^{3}\left(R_{i.}-\frac{N+1}{2}\right)^2 \tag{1}$$

The null hypothesis is rejected if the test statistic (1) is larger than a value chosen to make the type I error probability equal to $\alpha$. *Post-hoc* procedures would be required to determine which class measurements differ.

A parametric approach to test the general alternative hypothesis is the one-way analysis of variance (ANOVA). ANOVA compares the means of the groups by calculating the ratio of the within sum of squares and the between sum of squares. More specifically, the $F$-statistic is calculated as

$$F=\frac{\sum_{i=1}^{3} n_i \bar{Y}_i^2 - Y_{..}^2/N}{\sum_{i=1}^{3}(n_i-1)s_i^2}$$

where $Y_{..}$ is the sum of the measurements across all groups, $\bar{Y}_i$ is the average of the measurements for group $i$, and $s_i^2$ is the sample variance for group $i$. The $p$-value is obtained by comparing the F-statistic with an $F$ distribution with 2 and $n-2$ degrees of freedom.

### 2.2. Monotone ordering

The Jonckheere–Terpstra (JT) test is a distribution-free test for ordered alternative hypothesis that the treatment effects are in a specified monotone ordering, e.g. $Y_1<Y_2<Y_3$ [6, 7]. To calculate the JT test statistic, one calculates the three Mann–Whitney counts $U_{12}$, $U_{13}$, and $U_{23}$, where $U_{ij}$ is the number of measurements with disease class $i$ that are smaller than measurements for disease class $j$. The null hypothesis is rejected if the test statistic is larger than a value chosen to make the type I error probability equal to $\alpha$. A corresponding summary measure, $S_1$, has recently been proposed for the JT test to measure the accuracy of classifying into the correct classes [8]. $S_1$ ranges from 1 for perfect classification in the monotone ordering, say $Y_1<Y_2<Y_3$, to $-1$ where the classification is in the opposite ordering $Y_1>Y_2>Y_3$.

Terpstra and Magel proposed a non-parametric test for the same monotone ordered alternative hypothesis as the JT test, but the Terpstra and Magel (TM) test is based on comparing measurements from all three classes at the same time, rather than performing pairwise comparisons as the JT test does. More specifically, the TM test is based on the test statistic

$$T=\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} I(Y_{1i},Y_{2j},Y_{3k}) \tag{2}$$

where $I(Y_1, Y_2, Y_3)$ equals one if $Y_1, Y_2, Y_3$ are in the correct order (i.e. $Y_1 < Y_2 < Y_3$) and zero otherwise. The TM test is performed by comparing

$$\frac{T - n_1 n_2 n_3/6}{\sqrt{\mathrm{Var}(T)}}$$

with a standard normal distribution, where $\mathrm{Var}(T)$ is given in [9].

The VUS has been proposed as a summary measure for the ROC surface (see [3] for a summary). VUS is equal to the probability that three measurements, one from each class, will be classified in the correct monotone order, e.g. $Y_1 < Y_2 < Y_3$ [10]. The VUS test statistic is $(\widehat{\mathrm{VUS}} - \frac{1}{6})/\sqrt{\mathrm{Var}(\widehat{\mathrm{VUS}})}$, where $\widehat{\mathrm{VUS}}$ is the fraction of times the measurements are in the correct ordering $Y_1 < Y_2 < Y_3$ [11]. Specifically,

$$\widehat{\mathrm{VUS}} = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} I(Y_{1i}, Y_{2j}, Y_{3k}) \tag{3}$$

where $I(Y_1, Y_2, Y_3)$ equals one if $Y_1, Y_2, Y_3$ are in the correct order (i.e. $Y_1 < Y_2 < Y_3$) and zero otherwise. VUS takes the value $\frac{1}{6}$ when the three distributions completely overlap. Variance of the VUS can be estimated using a U-statistics approach [11] or by using the bootstrap [3]. By comparing (3) and (2), it is clear that $\widehat{\mathrm{VUS}}$ is equivalent to the TM test statistic. Therefore, only VUS is considered in the remainder of this paper.

Cuzick proposed an extension of the Wilcoxon test that is based on the test statistic $C = \sum_{i=1}^{N} z_i r_i$, where $N$ is the total number of observations in the combined sample, $r_i$ is the rank of the $i$th observation in the combined sample, and $z_i$ is the group number that the $i$th observation belongs [12]. The Cuzick test is performed by comparing

$$\frac{C - (N+1)(\sum_{j=1}^{3} z_j n_j)/2}{\sqrt{\mathrm{Var}(C)}} \tag{4}$$

with a standard normal distribution, where $\mathrm{Var}(C)$ is $(N^2(N+1)/12)(\sum_{i=1}^{3} z_i^2 n_i/N - \sum_{j=1}^{3} z_j n_j/N)$.

Le proposed a test for monotone ordered alternatives that has form similar to the KW test [13]. Specifically, the Le test is based on the test statistic

$$W = \sum_{i=1}^{3} n_i (L_i - M_i) \bar{R}_i$$

where $L_i$ is the total number of observations in all the groups to the left of the $i$th group in the monotone group ordering, $M_i$ is the total number of observations in all the groups to the right of the $i$th group in the monotone group ordering, and $\bar{R}_i$ is the average rank value for group $i$. The Le test is performed by comparing

$$\frac{W}{\sqrt{(N(N+1)/12)\sum_{i=1}^{3} n_i (L_i - M_i)}} \tag{5}$$

with a standard normal distribution. Interestingly, formulas (4) and (5) are equal when the three groups have the same number of measurements (i.e. $n_1 = n_2 = n_3$).

All the monotone ordering tests described previously in this subsection have been non-parametric approaches. Next, we describe a parametric test, the modified $F$ test for monotone ordering (denoted by $\bar{F}_{\mathrm{m}}$) (see, e.g. [14]). The $\bar{F}_{\mathrm{m}}$-statistic is a modification of the usual $F$-statistic in order to account for the monotone ordering of interest. Specifically

$$\bar{F}_{\mathrm{m}} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2 - \sum_{i=1}^{3}\sum_{j=1}^{n_i}(Y_{ij}-\tilde{Y}_i)^2}{\sum_{i=1}^{3}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2/(N-3)} \tag{6}$$

where $\bar{Y}$ is the overall mean and $(\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3)$ is the point at which $\sum_{i=1}^{3}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$ is minimized subject to the constraint $Y_1 \leqslant Y_2 \leqslant Y_3$. The null distribution of $\bar{F}_{\mathrm{m}}$ can be computed using a simple Monte Carlo algorithm [14].

### 2.3. Umbrella ordering

The Mack–Wolfe (MW) test has been proposed when the umbrella alternative hypothesis is of interest [15]. This approach computes $p(p-1)/2$ pairwise comparisons for the $p$ classes or treatments. When there are $p=3$ classes, so that the alternative hypothesis of interest is $Y_1 < Y_2 > Y_3$, the MW test is equivalent to a standard two-sample Wilcoxon–Mann–Whitney (WMW) test where the values for $Y_1$ and $Y_3$ are pooled [16]. The WMW test is equivalent to a test that the AUC is significantly greater than 0.5, i.e. the distributions completely overlap.

The UV test has recently been proposed as a non-parametric approach to test the alternative hypothesis that there is an umbrella ordering such as $Y_1 < Y_2 > Y_3$ [4]. The UV test statistic is $(\widehat{\mathrm{UV}} - \frac{1}{3})/\sqrt{\mathrm{Var}(\widehat{\mathrm{UV}})}$, where $\widehat{\mathrm{UV}}$ is the fraction of times the measurements are in the umbrella ordering of interest and $\mathrm{Var}(\widehat{\mathrm{UV}})$ is provided in [4]. In other words, for the umbrella ordering $Y_1 < Y_2 > Y_3$

$$\widehat{\mathrm{UV}} = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} I_U[Y_{1i}, Y_{2j}, Y_{3k}]$$

where $I_U(Y_1, Y_2, Y_3)$ equals one if $Y_1 < Y_2 > Y_3$ and zero otherwise. UV is equal to $\frac{1}{3}$ when the three distributions completely overlap. Similar to the TM and VUS statistics, the UV statistic is based on comparing measurements from all three classes at the same time rather than pairwise as the MW statistic does. The UV test can also be applied to tree orderings [4].

The respective parametric test suitable for an umbrella alternative is based on the $\bar{F}_{\mathrm{u}}$-statistic which has the same form as equation (6)

$$\bar{F}_{\mathrm{u}} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2 - \sum_{i=1}^{3}\sum_{j=1}^{n_i}(Y_{ij}-\tilde{Y}_i)^2}{\sum_{i=1}^{3}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2/(N-3)}$$

but $(\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3)$ is the point at which $\sum_{i=1}^{3}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$ is minimized subject to the constraint $Y_1 \leqslant Y_2 \geqslant Y_3$. The null distribution of $\bar{F}_{\mathrm{u}}$ can be computed using Monte Carlo methodology analogous to that for $\bar{F}_{\mathrm{m}}$.

### 2.4. Tree ordering

The Fligner–Wolfe (FW) test is a distribution-free test to test whether treatments differ from a control [17]. The null hypothesis is still that all treatment effects are equal for the treatment and

Table I. List of methods for assessing the ability of continuous measurements to correctly classify three classes in a particular order. All approaches consider the null hypothesis that measurements from the three classes follow the same distribution but different alternative hypotheses.

| Test | Approach | Alternative | Notes |
|---|---|---|---|
| Kruskal–Wallis (KW) | Non-para | General | |
| ANOVA | Para | General | |
| Jonckheere–Terpstra (JT) | Non-para | Monotone | |
| Terpstra–Magel (TM) | Non-para | Monotone | Essentially equivalent to VUS |
| VUS | Non-para | Monotone | Essentially equivalent to TM |
| Cuzick | Non-para | Monotone | Equivalent to Le for equal class sizes |
| Le | Non-para | Monotone | Equivalent to Cuzick for equal class sizes |
| $\bar{F}_{\mathrm{m}}$ | Para | Monotone | |
| Mack–Wolfe (MW) | Non-para | Umbrella | Equivalent to WMW for 3 classes |
| Umbrella volume (UV) | Non-para | Umbrella | |
| $\bar{F}_{\mathrm{u}}$ | Para | Umbrella | |
| Fligner–Wolfe (FW) | Non-para | Tree | |

control groups while the alternative hypothesis is that the effect for at least one treatment group is different from the control group. The FW test statistic is calculated by first ordering all $N$ observations from smallest to largest. Then the FW test statistic is the sum of the joint ranks for the non-control treatments. This test statistic is equivalent to the two-sample WMW test statistic computed for the control observations and the combined treatment observations.

### 2.5. *Summary of methods*

All methods considered in this section assess the null hypothesis that measurements from the three classes follow the same distribution, i.e. $Y_1 = Y_2 = Y_3$. However, there are differences in the alternative hypotheses considered and the test statistics used (Table I). Some of the approaches are non-parametric while others make parametric assumptions. The KW and ANOVA tests consider the general alternative hypothesis that $Y_i \neq Y_j$ for some $(i, j) \in \{1, 2, 3\}$ and $i \neq j$. Conversely, MW, UV, and $\bar{F}_{\mathrm{u}}$ test the alternative hypothesis that the classes follow an umbrella ordering, e.g. $Y_1 < Y_2 > Y_3$. In this case, the MW test is equivalent to the WMW test where $Y_1$ and $Y_3$ are pooled. The VUS, TM, Cuzick, Le, $\bar{F}_{\mathrm{m}}$, and JT tests consider the alternative hypothesis that the classes follow a monotone ordering, e.g. $Y_1 < Y_2 < Y_3$. The Cuzick and Le tests are equivalent when the three classes have the same number of measurements. Finally, the FW test considers the alternative hypothesis of a tree ordering.

## 3. SIMULATION STUDIES

A simulation study was performed to evaluate the power and size of the tests described in Section 2. In the simulations, the Cuzick, Le, $\bar{F}_{\mathrm{m}}$, JT, and VUS tests tested the alternative hypothesis that the measurements follow the monotone ordering $Y_1 < Y_2 < Y_3$. The UV and $\bar{F}_{\mathrm{u}}$ tests considered the umbrella ordering $Y_1 < Y_2 > Y_3$ while MW tests $Y_2$ is greater than $Y_1$ and $Y_3$ pooled. The FW test tested the hypothesis that $Y_1 > Y_2$ or $Y_2 < Y_3$. The simulations were replicated 1000 times for each of nine scenarios with balanced sample size of 10, 20, and 40 measurements in each class (Table II)

DOI: 10.1002/sim

Table II. Results of simulations to study size (Scenario 1) and power (Scenarios 2–9) for $n_i$ number of measurements in each class. Scenarios are presented for the null hypothesis (Scenario 1) as well as for monotone orderings (Scenarios 2–3), umbrella orderings (Scenarios 4–7), and tree orderings (Scenarios 8–9).

| Scenario—$Y_1, Y_2, Y_3$ | $n_i$ | KW | ANOVA | JT | VUS | Cuzick | Le | $\bar{F}_{\mathrm{m}}$ | MW | UV | $\bar{F}_{\mathrm{u}}$ | FW |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $1-\mathrm{N}(0,1),\mathrm{N}(0,1),\mathrm{N}(0,1)$ | 10 | 0.053 | 0.058 | 0.067 | 0.048 | 0.066 | 0.066 | 0.058 | 0.073 | 0.052 | 0.054 | 0.051 |
| $(Y_1=Y_2=Y_3)$ | 20 | 0.042 | 0.042 | 0.037 | 0.025 | 0.036 | 0.036 | 0.037 | 0.077 | 0.051 | 0.051 | 0.046 |
| | 40 | 0.070 | 0.058 | 0.065 | 0.050 | 0.063 | 0.063 | 0.055 | 0.050 | 0.048 | 0.049 | 0.062 |
| $2-\mathrm{N}(0,1),\mathrm{N}(0.5,1),\mathrm{N}(1,1)$ | 10 | 0.409 | 0.447 | 0.660 | 0.530 | 0.672 | 0.672 | 0.631 | 0.062 | 0.034 | 0.377 | 0.043 |
| $(Y_1<Y_2<Y_3)$ | 20 | 0.751 | 0.771 | 0.894 | 0.806 | 0.896 | 0.896 | 0.885 | 0.045 | 0.019 | 0.685 | 0.051 |
| | 40 | 0.983 | 0.983 | 0.999 | 0.986 | 0.999 | 0.999 | 0.997 | 0.046 | 0.012 | 0.947 | 0.039 |
| $3-t_3,t_3+0.5,t_3+1$ | 10 | 0.280 | 0.237 | 0.524 | 0.439 | 0.529 | 0.529 | 0.403 | 0.070 | 0.027 | 0.219 | 0.044 |
| $(Y_1<Y_2<Y_3)$ | 20 | 0.581 | 0.453 | 0.793 | 0.692 | 0.790 | 0.790 | 0.635 | 0.052 | 0.029 | 0.386 | 0.046 |
| | 40 | 0.885 | 0.674 | 0.969 | 0.917 | 0.969 | 0.969 | 0.802 | 0.071 | 0.024 | 0.604 | 0.055 |
| $4-\mathrm{N}(0,1),\mathrm{N}(1,1),\mathrm{N}(0,1)$ | 10 | 0.534 | 0.572 | 0.025 | 0.010 | 0.034 | 0.034 | 0.233 | 0.834 | 0.763 | 0.698 | 0.000 |
| $(Y_1<Y_2>Y_3)$ | 20 | 0.890 | 0.902 | 0.034 | 0.010 | 0.048 | 0.048 | 0.458 | 0.977 | 0.964 | 0.949 | 0.000 |
| | 40 | 0.996 | 0.996 | 0.033 | 0.000 | 0.034 | 0.034 | 0.725 | 1.000 | 1.000 | 0.998 | 0.000 |
| $5-t_3,t_3+1,t_3$ | 10 | 0.356 | 0.277 | 0.038 | 0.017 | 0.041 | 0.041 | 0.140 | 0.683 | 0.620 | 0.396 | 0.000 |
| $(Y_1<Y_2>Y_3)$ | 20 | 0.671 | 0.480 | 0.044 | 0.011 | 0.044 | 0.044 | 0.196 | 0.870 | 0.848 | 0.612 | 0.000 |
| | 40 | 0.941 | 0.773 | 0.003 | 0.003 | 0.032 | 0.032 | 0.379 | 0.989 | 0.986 | 0.854 | 0.000 |
| $6-\mathrm{N}(0,0.25),\mathrm{N}(1,9),\mathrm{N}(0,4)$ | 10 | 0.152 | 0.179 | 0.036 | 0.003 | 0.044 | 0.044 | 0.050 | 0.286 | 0.403 | 0.250 | 0.012 |
| $(Y_1<Y_2>Y_3)$ | 20 | 0.271 | 0.308 | 0.034 | 0.002 | 0.045 | 0.045 | 0.062 | 0.417 | 0.657 | 0.394 | 0.002 |
| | 40 | 0.496 | 0.578 | 0.057 | 0.000 | 0.063 | 0.063 | 0.154 | 0.621 | 0.894 | 0.672 | 0.000 |
| $7-\mathrm{U}[0.2,1.2],\mathrm{N}(1.3,1),\chi_1^2$ | 10 | 0.300 | 0.248 | 0.022 | 0.019 | 0.023 | 0.023 | 0.136 | 0.555 | 0.567 | 0.331 | 0.000 |
| $(Y_1<Y_2>Y_3)$ | 20 | 0.538 | 0.350 | 0.036 | 0.018 | 0.030 | 0.030 | 0.316 | 0.730 | 0.768 | 0.450 | 0.000 |
| | 40 | 0.882 | 0.710 | 0.040 | 0.034 | 0.023 | 0.023 | 0.679 | 0.954 | 0.966 | 0.831 | 0.000 |
| $8-\mathrm{N}(1,1),\mathrm{N}(0,1),\mathrm{N}(1,1)$ | 10 | 0.531 | 0.575 | 0.023 | 0.007 | 0.028 | 0.028 | 0.225 | 0.000 | 0.000 | 0.000 | 0.775 |
| $(Y_1>Y_2<Y_3)$ | 20 | 0.891 | 0.910 | 0.024 | 0.004 | 0.028 | 0.028 | 0.427 | 0.000 | 0.000 | 0.000 | 0.975 |
| | 40 | 0.995 | 0.998 | 0.033 | 0.000 | 0.038 | 0.038 | 0.733 | 0.000 | 0.000 | 0.000 | 0.998 |
| $9-t_3+1,t_3,t_3+1$ | 10 | 0.359 | 0.308 | 0.045 | 0.028 | 0.049 | 0.049 | 0.145 | 0.000 | 0.000 | 1.000 | 0.601 |
| $(Y_1>Y_2<Y_3)$ | 20 | 0.710 | 0.536 | 0.025 | 0.005 | 0.026 | 0.026 | 0.192 | 0.000 | 0.000 | 0.002 | 0.887 |
| | 40 | 0.943 | 0.779 | 0.047 | 0.005 | 0.048 | 0.048 | 0.366 | 0.000 | 0.000 | 0.000 | 0.993 |

and unbalanced sample sizes $(n_1, n_2, n_3)$ of (10,10,20), (10,10,40), and (10,20,40) (Table III). All tests were one-sided with 0.05 type I error.

In Scenario 1 sampling was performed under the null hypothesis where $Y_1, Y_2$, and $Y_3$ were simulated from a standard normal distribution. The simulated size for some of the tests deviated substantially from the nominal level for small sample sizes. However, the simulated size is generally closer to the nominal level for $n_1 = n_2 = n_3 = 40$ and are much closer to the nominal level for $n_1 = n_2 = n_3 = 80$ (results not provided). The Le test has unacceptably large size for unbalanced data scenarios and thus, the Le test is not recommended when there is unequal number of measurements in the classes. Since the Le test clearly is not appropriate for unbalanced data settings and it is equivalent to the Cuzick test for balanced data in the three-class case, the performance of the Le test is not discussed for the other simulated scenarios.

In Scenario 2 a monotone ordering was considered for the class measurements with $Y_1 \sim \mathrm{N}(0,1)$, $Y_2 \sim \mathrm{N}(0.5,1)$, and $Y_3 \sim \mathrm{N}(1,1)$. The Cuzick test was the most powerful for all the sample sizes considered, but power for the JT test was only slightly smaller. The next powerful tests were the $\bar{F}_\mathrm{m}$ test and the VUS test. As expected, all the tests that test a monotone ordering had larger power than the general alternative tests (KW and ANOVA). As desired, the MW and UV tests, which test umbrella orderings, had low power to detect the monotone ordering. Conversely, it is undesired that the $\bar{F}_\mathrm{u}$ test, which tests umbrella orderings, had relatively high power to detect the monotone ordering. Similar results are observed for Scenario 3 where $Y_1 \sim t_3, Y_2 \sim t_3 + 0.5$, and $Y_3 \sim t_3 + 1$. It is not surprising that the power is larger for the non-parametric KW test than the parametric ANOVA test in Scenario 3 and the reverse is observed in Scenario 2 where the data are simulated from Normal distributions.

Scenario 4 simulates an umbrella alternative where $Y_1 \sim \mathrm{N}(0,1), Y_2 \sim \mathrm{N}(1,1)$, and $Y_3 \sim \mathrm{N}(0,1)$ while in Scenario 5 $Y_1 \sim t_3, Y_2 \sim t_3 + 1$, and $Y_3 \sim t_3$. In both of these scenarios the MW test appears to be the most powerful followed by UV and $\bar{F}_\mathrm{u}$, which are all more powerful than the general alternative approaches. The large power for MW likely results from the large effective sample size due to pooling of measurements for $Y_1$ and $Y_3$ which follow identical distributions. The monotone ordering approaches are not sensitive to the umbrella alternative in this scenario, except for $\bar{F}_\mathrm{m}$.

In Scenario 6, simulated distributions have the same locations as in Scenario 4 but they have different scale parameters with $Y_1 \sim \mathrm{N}(0, 0.25)$, $Y_2 \sim \mathrm{N}(1,9)$, and $Y_3 \sim \mathrm{N}(0,4)$. Here the UV test is the most powerful for all sample sizes considered. This finding is repeated in Scenario 7 where measurements come from distributions with different shapes, i.e. $Y_1 \sim \mathrm{Unif}[0.2, 1.2]$, $Y_2 \sim \mathrm{N}(1.3, 1)$, and $Y_3 \sim \chi_1^2$. Again the UV test is the most powerful with MW having close, but smaller, power for balanced sample sizes. The difference in power is greater for unbalanced data.

Scenario 8 considers $Y_1 \sim \mathrm{N}(1,1), Y_2 \sim \mathrm{N}(0,1)$, and $Y_3 \sim \mathrm{N}(1,1)$ so that the data follow the reverse umbrella ordering or tree ordering $Y_1 > Y_2 < Y_3$. In this scenario the FW test had the highest power but the KW and ANOVA tests had reasonable power for the larger sample sizes considered. The difference in power between the FW and the ANOVA test was larger in Scenario 9 where the measurements were not generated using a Normal distribution $(Y_1 \sim t_3 + 1, Y_2 \sim t_3, Y_3 \sim t_3 + 1)$. The JT, Cuzick, Le, and $\bar{F}$ tests appear to be very sensitive even for scenarios where the alternative is not a monotone ordering. However, this is not true for the VUS test, which is only sensitive in detecting monotone orderings.

In summary, when the measurements follow a particular ordering, tests that test that particular ordering have higher power than the general alternative approaches (KW and ANOVA tests). JT and Cuzick, which is equivalent to the Le test for balanced data, have the greatest power for monotone orderings. When $Y_1$ and $Y_3$ come from the same distribution, the MW test has the

DOI: 10.1002/sim

Table III. Results of simulations to study size (Scenario 1) and power (Scenarios 2–9) for $n^{\star}$ number of measurements in the classes, where $n^{\star}$ equal to 1, 2, and 3 corresponds to $(n_1, n_2, n_3)$ equal to (10,10,20), (10,10,40), and (10,20,40), respectively. Scenarios are presented for the null hypothesis (Scenario 1) as well as for monotone orderings (Scenarios 2–3), umbrella orderings (Scenarios 4–7), and tree orderings (Scenarios 8–9).

| Scenario—$Y_1, Y_2, Y_3$ | $n^{\star}$ | KW | ANOVA | JT | VUS | Cuzick | Le | $\bar{F}_{\mathrm{m}}$ | MW | UV | $\bar{F}_{\mathrm{u}}$ | FW |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1–$\mathrm{N}(0,1), \mathrm{N}(0,1), \mathrm{N}(0,1)$ | 1 | 0.044 | 0.042 | 0.045 | 0.038 | 0.047 | 0.355 | 0.041 | 0.067 | 0.054 | 0.045 | 0.045 |
| $(Y_1 = Y_2 = Y_3)$ | 2 | 0.052 | 0.051 | 0.056 | 0.044 | 0.063 | 0.831 | 0.052 | 0.063 | 0.063 | 0.054 | 0.057 |
| | 3 | 0.045 | 0.045 | 0.043 | 0.039 | 0.047 | 0.562 | 0.043 | 0.075 | 0.062 | 0.042 | 0.043 |
| 2–$\mathrm{N}(0,1), \mathrm{N}(0.5,1), \mathrm{N}(1,1)$ | 1 | 0.591 | 0.626 | 0.798 | 0.637 | 0.813 | 0.981 | 0.789 | 0.030 | 0.033 | 0.516 | 0.103 |
| $(Y_1 < Y_2 < Y_3)$ | 2 | 0.683 | 0.707 | 0.863 | 0.695 | 0.884 | 1.000 | 0.866 | 0.008 | 0.032 | 0.631 | 0.185 |
| | 3 | 0.716 | 0.759 | 0.884 | 0.746 | 0.896 | 0.996 | 0.879 | 0.008 | 0.020 | 0.580 | 0.271 |
| 3–$t_3, t_3+0.5, t_3+1$ | 1 | 0.427 | 0.323 | 0.655 | 0.528 | 0.657 | 0.953 | 0.506 | 0.032 | 0.038 | 0.277 | 0.092 |
| $(Y_1 < Y_2 < Y_3)$ | 2 | 0.480 | 0.357 | 0.707 | 0.540 | 0.728 | 0.999 | 0.541 | 0.020 | 0.031 | 0.309 | 0.140 |
| | 3 | 0.543 | 0.377 | 0.743 | 0.601 | 0.755 | 0.988 | 0.559 | 0.003 | 0.016 | 0.270 | 0.237 |
| 4–$\mathrm{N}(0,1), \mathrm{N}(1,1), \mathrm{N}(0,1)$ | 1 | 0.608 | 0.656 | 0.005 | 0.004 | 0.012 | 0.067 | 0.160 | 0.874 | 0.854 | 0.752 | 0.000 |
| $(Y_1 < Y_2 > Y_3)$ | 2 | 0.647 | 0.704 | 0.005 | 0.008 | 0.008 | 0.160 | 0.103 | 0.883 | 0.886 | 0.789 | 0.000 |
| | 3 | 0.909 | 0.923 | 0.000 | 0.005 | 0.001 | 0.006 | 0.152 | 0.990 | 0.963 | 0.967 | 0.000 |
| 5–$t_3, t_3+1, t_3$ | 1 | 0.426 | 0.361 | 0.013 | 0.015 | 0.014 | 0.107 | 0.104 | 0.732 | 0.725 | 0.452 | 0.000 |
| $(Y_1 < Y_2 > Y_3)$ | 2 | 0.448 | 0.342 | 0.000 | 0.009 | 0.003 | 0.266 | 0.041 | 0.700 | 0.734 | 0.430 | 0.000 |
| | 3 | 0.724 | 0.556 | 0.000 | 0.012 | 0.002 | 0.020 | 0.113 | 0.885 | 0.827 | 0.654 | 0.000 |
| 6–$\mathrm{N}(0,0.25), \mathrm{N}(1,9), \mathrm{N}(0,4)$ | 1 | 0.145 | 0.194 | 0.016 | 0.001 | 0.023 | 0.196 | 0.017 | 0.311 | 0.553 | 0.271 | 0.014 |
| $(Y_1 < Y_2 > Y_3)$ | 2 | 0.144 | 0.221 | 0.004 | 0.001 | 0.007 | 0.565 | 0.002 | 0.295 | 0.617 | 0.281 | 0.008 |
| | 3 | 0.218 | 0.269 | 0.004 | 0.000 | 0.003 | 0.169 | 0.002 | 0.405 | 0.761 | 0.363 | 0.002 |
| 7–$\mathrm{U}[0.2,1.2], \mathrm{N}(1.3,1), \chi^2_1$ | 1 | 0.282 | 0.150 | 0.002 | 0.023 | 0.002 | 0.066 | 0.030 | 0.517 | 0.691 | 0.203 | 0.001 |
| $(Y_1 < Y_2 > Y_3)$ | 2 | 0.299 | 0.099 | 0.000 | 0.022 | 0.001 | 0.250 | 0.007 | 0.549 | 0.805 | 0.146 | 0.001 |
| | 3 | 0.432 | 0.149 | 0.000 | 0.025 | 0.000 | 0.036 | 0.010 | 0.737 | 0.911 | 0.215 | 0.000 |
| 8–$\mathrm{N}(1,1), \mathrm{N}(0,1), \mathrm{N}(1,1)$ | 1 | 0.596 | 0.630 | 0.119 | 0.004 | 0.085 | 0.723 | 0.324 | 0.000 | 0.000 | 0.000 | 0.837 |
| $(Y_1 > Y_2 < Y_3)$ | 2 | 0.684 | 0.714 | 0.334 | 0.014 | 0.200 | 0.998 | 0.458 | 0.000 | 0.000 | 0.005 | 0.865 |
| | 3 | 0.910 | 0.923 | 0.567 | 0.006 | 0.335 | 0.998 | 0.771 | 0.000 | 0.000 | 0.003 | 0.973 |
| 9–$t_3+1, t_3, t_3+1$ | 1 | 0.433 | 0.329 | 0.125 | 0.016 | 0.098 | 0.628 | 0.178 | 0.000 | 0.001 | 0.002 | 0.679 |
| $(Y_1 > Y_2 < Y_3)$ | 2 | 0.479 | 0.358 | 0.232 | 0.014 | 0.126 | 0.995 | 0.222 | 0.000 | 0.000 | 0.007 | 0.734 |
| | 3 | 0.719 | 0.564 | 0.446 | 0.013 | 0.278 | 0.986 | 0.439 | 0.000 | 0.000 | 0.004 | 0.887 |

DOI: 10.1002/sim

greatest power for the umbrella ordering; otherwise, the UV test has the greatest power. The FW test had the greatest power for tree orderings. Recommendations on the appropriate use of each method are provided in Section 5.

## 4. DATA ANALYSIS

In this section, the approaches are applied to a study of a metabolite to accurately distinguish between HIV-negative persons and HIV-positive patients with and without HIV-related neurological sequelae and to a study to assess the effects of different doses of haloperidol on the play behavior of juvenile rats.

### *4.1. Analysis of AIDS neurological sequelae*

The Human immunodeficiency virus (HIV) invades the central nervous system causing structural and metabolic changes in the brain. Patients with AIDS often show varying degrees of cognitive, motor and behavioral impairment, including dementia (AIDS dementia complex—ADC [18]). A number of studies have shown that an imaging technique called proton magnetic resonance spectroscopy ($^1$H-MRS) provides a reliable *in vivo*, non-invasive method for the assessment of HIV-associated brain injury (for example, see [19]). The area under each spectral peak is associated with the concentration of a metabolite that reflects activity in a specific cell type in response to signals in its microenvironment. A frequently measured metabolite is the ratio of myoinositol (MI) over creatine (Cr). MI is a marker of glial cells that are involved in providing nutrition, structural support of neuronal cells, and removing pathogens and damaged neurons from the brain. As a marker of glial cell proliferation, MI is an index of ongoing inflammation and injury to the brain. Creatine, on the other hand, is a marker of cellular metabolism that is assumed to be constant in most cases, including during many pathological conditions. Thus, dividing by Cr in the ratio is used as an internal standard to reduce the variability of the MI signal by accounting for difference among imaging machine technologies, brain structure localization and other considerations particular to each specific imaging procedure.

In a study of 136 individuals [20], MI/Cr ratios were measured in the white matter of 60 HIV-positive patients with neurological disease (ADC), 39 HIV-positive patients that were asymptomatic (NAS), and 37 HIV-negative controls (NEG). One objective of the study was to determine whether the MI/Cr ratio could accurately distinguish between ADC, NAS, and NEG patients. The MI/Cr ratios were lowest for the NEG patients (median 0.6, mean 0.614, standard deviation (SD) 0.073), followed by NAS patients (median 0.67, mean 0.664, SD 0.03), and ADC patients (median 0.705, mean 0.709, SD 0.134) (Figure 1).

Exploratory data analysis suggests that there may be a monotone ordering NEG<NAS<ADC in MI/Cr ratio values and possibly an umbrella ordering NEG<ADC>NAS. The Cuzick test ($C = 22103$, $p<0.0001$), VUS test (VUS$=0.4120$, $p<0.0001$), and JT test ($S_1 = 0.3955$, $p<0.0001$) all support the monotone ordering NEG<NAS<ADC. The UV test (UV$=0.5580$, $p<0.0001$) and MW test (AUC$=0.6875$, $p<0.0001$) support the umbrella ordering NEG<ADC>NAS. Given these results, it is not surprising that the KW test (KW$=22.7058$, $p<0.0001$) and ANOVA ($F = 10.28$, $p<0.0001$) are also highly significant. We conclude that MI/Cr measurements differ between groups of patients and furthermore that the data support both the umbrella (NEG<ADC>NAS) and monotone (NEG<NAS<ADC) orderings. Both conclusions have far reaching implications

DOI: 10.1002/sim

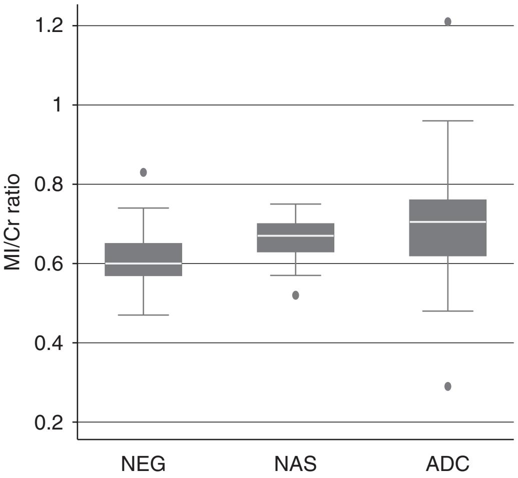

Figure 1. Box plots of MI/Cr ratios in the white matter of HIV-negative controls (NEG), HIV-positive patients that are asymptomatic (NAS), and HIV-positive patients with neurological disease (ADC).

for our understanding of HIV-related neurological disease progression. The monotone ordering implies that, despite not having overt clinical symptoms, NAS subjects do exhibit perturbations that are measurable through proton MRS. The combination of the results of both the monotone trend and umbrella ordering strengthens the impression that brain inflammation, which has been widely reported as being the hallmark of HIV infection and is thus inflammation is ubiquitous in all stages of neurological progression (e.g. [19]), persists and increases further among neurologically advanced (ADC) patients.

### *4.2. Analysis of rat play behavior*

Haloperidol, a non-selective neuroleptic medication widely used in the past to treat schizophrenia [21], binds to a wide variety of neurotransmitter receptors, but is thought to exert its primary neurochemical and behavioral effects via antagonism of dopaminergic systems in the brain. The dose-dependent effects of haloperidol have been extensively studied; at higher doses, haloperidol has sedative properties, whereas at lower doses, haloperidol has sometimes been associated with stimulatory effects [22–24]. The rodent model has been frequently used to investigate the neurochemical impact of haloperidol on the central nervous system, as well as the drug's effect on appetitive and movement behaviors thought to be significantly underpinned by central dopaminergic systems (e.g. the basal ganglia) [25, 26]. Juvenile rats engage in rough-and-tumble play (RTP) behavior during a specific period of normal development, from around days 20–40 [27]. Because RTP behavior involves vigorous motoric behaviors including chasing, wrestling, and pinning [28], it is reasonable to assume that brain areas associated with movement (e.g. cerebellum and basal ganglia), and the neurochemical systems that innervate these structures, might underlie rat play activities. Further, perturbation of these systems via either surgical or neurochemical means, and observation of the resulting effects, can provide information regarding the relative importance of these central nervous systems on behavior [29–35]. In order to investigate the manner by which dopaminergic systems might impact juvenile RTP play behavior, various doses of haloperidol were

DOI: 10.1002/sim

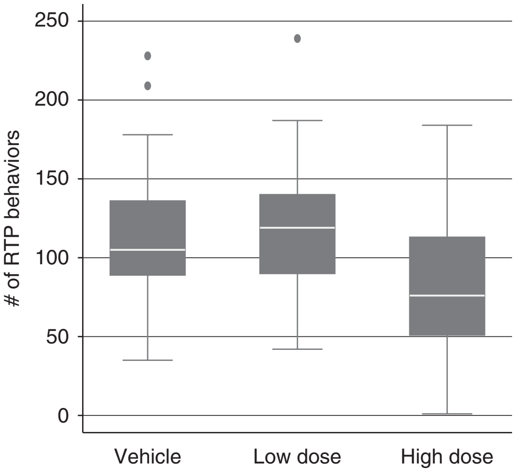

Figure 2. Box plots of number of RTP behaviors observed in pairs of rats that received vehicle, low dose haloperidol, and high dose haloperidol.

administered to juvenile rats. This analysis compares the results of tartaric acid vehicle, low dose haloperidol (0.025 mg/kg and 0.05 mg/kg), and high dose haloperidol (0.1 mg/kg and 0.2 mg/kg). Pairs of juvenile rats were videotaped, and their play interactions later analyzed by one of the authors (S.B).

Figure 2 summarizes the aggregate number of RTP behaviors (e.g. wrestle, chase, pin) for the 19 rats administered vehicle (median 105, mean 118.05, SD 48.5), 39 rats injected with low dose haloperidol (median 119, mean 117.9, SD 38.9), and 39 rats injected with high dose haloperidol (median 76, mean 83.2, SD 44.4). The UV test yields a significant umbrella ordering Vehicle < Low dose > High dose ($\mathrm{UV}=0.4481$, $p=0.0421$), as does the $\bar{F}_{\mathrm{u}}$ test ($\bar{F}_{\mathrm{u}}=15.14$, $p<0.0001$). The MW test also suggests that rats that received low dose haloperidol exhibited more RTP behaviors than rats that received vehicle or high dose haloperidol ($\mathrm{AUC}=0.6647$, $p=0.0009$). Therefore, these data support the hypothesis that at higher doses, haloperidol exerts sedative properties on juvenile rat RTP behavior, whereas at lower doses, haloperidol exerts stimulatory effects. Interestingly, the Cuzick test ($C=10387$, $p=0.0002$), JT test ($\mathrm{S1}=0.3477$, $p=0.0002$) and VUS test ($\mathrm{VUS}=0.315$, $p=0.0024$) also suggest a significant monotone ordering High dose < Vehicle < Low dose, illustrating the sedative effect of high doses of haloperidol on this particular type of vigorous motoric activity.

## 5. DISCUSSION

In this paper we compare and contrast approaches for testing hypotheses regarding treatment effects distributions, or equivalently, the accuracy of biomarkers. The specific focus is on settings and research questions where there are three classes or populations and the measurements of interest are made on a continuous scale. We agree with Terpstra and Magel [9] that a monotone or umbrella ordered test should have the following properties: (1) the size of the test should be approximately equal to the nominal size, (2) the test should have higher power than a general

DOI: 10.1002/sim

alternative test when the alternative hypothesis being tested is true, (3) the test should have low power for any alternative hypothesis that is not consistent with the true alternative. Simulations suggest that the Cuzick and JT tests, both designed to detect monotone orderings, have high power to detect a monotone ordering but the VUS is less sensitive to situations that are not consistent with a monotone ordering, especially with unbalanced data. These results are consistent with previous findings [9, 36]. Thus, the VUS test may be the preferred method of analysis in actual research practice.

Our simulation results also suggest that the UV test has higher power than the MW test for umbrella orderings except when measurements for two of the classes come from the same distribution, which is not common in real-world applications. The UV test also has the appealing property that it has lower power than MW for alternative hypotheses not consistent with an umbrella ordering. The ANOVA and KW tests yielded lower power than tests of restricted orders in all of the research scenarios considered, and hence should not serve as the initial test, unless restricted orderings are not of interest. This finding for KW is consistent with previously published results [16].

There are interesting relationships between monotone, umbrella, and tree orderings. First, a tree alternative hypothesis can be converted to an umbrella alternative hypothesis by changing the signs of the measurement values. For example, the tree ordering $Y_1>Y_2<Y_3$ is equivalent to the umbrella ordering $-Y_1<-Y_2>-Y_3$. Second, an alternative hypothesis regarding a monotone ordering can be converted to a less stringent hypothesis of a tree or umbrella ordering by changing the order of the groups. For example, an alternative hypothesis of the monotone ordering $Y_1<Y_2<Y_3$ can be converted to a less stringent hypothesis of the tree ordering $Y_2>Y_1<Y_3$ and the umbrella ordering $Y_1<Y_3>Y_2$. In real research data, however, there often is inherent ordering of the groups so that it is not possible to change the order of the groups, as is the case in the data analyzed in Section 4. Furthermore, it is in settings with an inherent ordering that the contrast between umbrella and monotone orderings is especially interesting because the orderings yield different explanations. This paper compares methods that can detect and distinguish between the two orderings.

Ideally, the research question will drive the alternative hypothesis to be tested. This will dictate the statistical test to be performed and, thus, will limit the need to worry about inflated type I error due to multiple testing. In Section 4.2, there was particular interest in testing whether juvenile rats that received low dose haloperidol exhibited greater RTP activity as compared with rats that received vehicle or high dose haloperidol. In other words, an umbrella ordering was of interest. In the neurological impact of HIV example (Section 4.1), both a monotone and an umbrella ordering were of interest. The monotone ordering is plausible, since neurological disease progression should, theoretically, be associated with higher brain inflammation. Thus, higher MI/Cr levels would be expected among more severely affected HIV-infected patients [20, 37]. On the other hand, an umbrella ordering is also plausible. NAS subjects would be expected to have higher MI/Cr levels than HIV-negative controls as brain inflammation has been reported in all stages of HIV infection regardless of whether or not patients are symptomatic for neurological deterioration [19, 37]. Results of these analyses have far reaching implications for our understanding of HIV-related neurological disease progression. The result that inflammation continues in the ADC stage is interesting (umbrella ordering) because lack of that would imply a burn-out disease situation. The result that inflammation is present among subjects without overt clinical symptoms (monotone ordering) implies two things: first, interventions in the brain are needed for those who do not exhibit clinical symptoms; second, MRS is a non-invasive early warning system that can identify those otherwise clinically asymptomatic patients most in need of this intervention. In all cases,

DOI: 10.1002/sim

exploratory data analysis can be useful if *a priori* there is not a particular alternative hypothesis of interest.

In summary, research questions may be best tested with an alternative hypothesis of a specific ordering of the groups rather than a general ordering. This paper discusses several approaches to test a monotone or umbrella ordering. Additionally, the paper illustrates that application of these approaches can yield novel findings.

## ACKNOWLEDGEMENTS

The authors would like to thank two anonymous referees for helpful comments and suggestions.

## REFERENCES

1. Ruberg SJ. Dose response studies II analysis and interpretation. *Journal of Biopharmaceutical Statistics* 1995; **5**:15–42.
2. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: New York, 2003.
3. Nakas CT, Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine* 2004; **23**:3437–3449.
4. Nakas CT, Alonzo TA. ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. *Biometrics* 2007; **63**:603–609.
5. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 1952; **47**:583–621.
6. Jonckheere AR. A distribution-free *k*-sample test against ordered alternatives. *Biometrika* 1954; **41**:133–145.
7. Terpstra T. The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae* 1952; **14**:327–333.
8. Flandre P, O'Quigley J. Predictive strength of Jonckheere's test for trend: an application to genotypic scores in HIV infection. *Statistics in Medicine* 2007; **26**:4441–4454. DOI: 10.1002/sim.2871.
9. Terpstra JT, Magel RC. A new nonparametric test for the ordered alternative problem. *Nonparametric Statistics* 2003; **15**:289–301.
10. Mossman D. Three-way ROCs. *Medical Decision Making* 1999; **19**:78–89.
11. Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making* 2000; **20**:323–331.
12. Cuzick J. A Wilcoxon-type test for trend. *Statistics in Medicine* 1985; **4**:87–90.
13. Le CT. A new rank test against ordered alternatives in *k*-sample problems. *Biometrical Journal* 1988; **1**:87–92.
14. Silvapulle MJ, Sen PK. *Constrained Statistical Inference*. Wiley: New Jersey, 2005.
15. Mack GA, Wolfe DA. *k*-sample rank tests for umbrella alternatives. *Journal of the American Statistical Association* 1981; **76**:175–181.
16. Hollander M, Wolfe DA. *Nonparametric Statistical Methods*. Wiley: New York, 1999.
17. Fligner MA, Wolfe DA. Distribution-free tests for comparing several treatments with a control. *Statistica Neerlandica* 1982; **36**:119–127.
18. Navia BA, Jordan BD, Price RW. The aids dementia complex: I. Clinical features. *Annals of Neurology* 1986; **19**:517–524.
19. Lopez-Villegas D, Lenkinski RE, Frank I. Biochemical changes in the frontal lobe of HIV-infected individuals detected by magnetic resonance spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America* 1997; **94**:9854–9859.
20. Chang L, Lee PL, Yiannoutsos CT, Ernst T, Marra CM, Richards T, Kolson D, Schifitto G, Jarvik JG, Miller EN, Lenkinski R, Gonzalez G, Navia BA. A multicenter in vivo proton-MRS study of HIV-associated dementia and its relationship to age. *NeuroImage* 2004; **23**:1336–1347.
21. Gaebel W, Riesbeck M, Wlwer W, Klimke A, Eickhoff M, von Wilmsdorff M, Jockers-Scherbl MC, Khn KU, Lemke M, Bechdolf A, Bender S, Degner D, Schlsser R, Schmidt LG, Schmitt A, Jger M, Buchkremer G, Falkai P, Klingberg S, Kpcke W, Maier W, Häfner H, Ohmann C, Salize HJ, Schneider F, Mller HJ. Maintenance

DOI: 10.1002/sim

treatment with risperidone or low-dose haloperidol in first-episode schizophrenia: 1-year results of a randomized controlled trial within the German Research Network on Schizophrenia. *Journal of Clinical Psychiatry* 2007; **68**:1763–1774.

22. Ito S, Mori T, Namiki M, Suzuki T, Sawaguchi T. Complicated interaction between psychostimulants and morphine in expression of phenotype of behavior in the dopaminergic system of BALB/c mice. *Journal of Pharmacological Sciences* 2007; **105**:326–333. DOI: 10.1254/jphs.FP0070653.
23. O'Neill MF, Shaw G. Comparison of dopamine receptor antagonists on hyperlocomotion induced by cocaine, amphetamine, MK-801 and the dopamine D1 agonist C-APB in mice. *Psyhcopharamacology* 1999; **145**:237–250.
24. Irifune M, Sato T, Nishikawa T, Masuyama T, Nomoto M, Fukuda T, Kawahara M. Hyperlocomotion during recovery from isoflurane anesthesia is associated with increased dopamine turnover in the nucleus accumbens and striatum in mice. *Anesthesiology* 1997; **86**:464–475.
25. Boye SM, Rompre PP. Behavioral evidence of depolarization block of dopamine neurons after chronic treatment with haloperidol and clozapine. *The Journal of Neuroscience* 2000; **20**:1229–1239.
26. Fukushiro DF, Jdo NA, Tatsu JA, de Castro JP, Chinen CC, Frussa-Filho R. Haloperidol (but not ziprasidone) withdrawal enhances cocaine-induced locomotor activation and conditioned place preference in mice. *Progress in Neuro-psychopharmacology and Biological Psychiatry* 2007; **31**:867–872. DOI: 10.1016/j.pnpbp.2007.01.025.
27. Pellis SM, McKenna M. What do rats find rewarding in play fighting?—an analysis using drug-induced non-playful partners. *Behavioural Brain Research* 1995; **68**:65–73.
28. Siviy SM, Atrens DM. The energetic costs of rough-and-tumble play in the juvenile rat. *Developmental Psychobiology* 1992; **25**:137–148.
29. Gordon NS, Kollack-Walker S, Akil H, Panksepp J. Expression of c-fos gene activation during rough and tumble play in juvenile rats. *Brain Research Bulletin* 2002; **57**:651–659.
30. Siviy SM, Baliko CN. A further characterization of alpha-2 adrenoceptor involvement in the rough-and-tumble play of juvenile rats. *Developmental Psychobiology* 2000; **37**:25–34.
31. Burgdorf J, Panksepp J, Beinfeld MC, Kroes RA, Moskal JR. Regional brain cholecystokinin changes as a function of rough-and-tumble play behavior in adolescent rats. *Peptides* 2006; **27**:172–177.
32. Siviy SM, Fleischhauer AE, Kerrigan LA, Kuhlman SJ. D2 dopamine receptor involvement in the rough-and-tumble play behavior of juvenile rats. *Behavioral Neuroscience* 1996; **110**:1168–1176.
33. Siviy SM, Line BS, Darcy EA. Effects of MK-801 on rough-and-tumble play in juvenile rats. *Physiology and Behavior* 1995; **57**:843–847.
34. Siviy SM, Fleischhauer AE, Kuhlman SJ, Atrens DM. Effects of alpha-2 adrenoceptor antagonists on rough-and-tumble play in juvenile rats: evidence for a site of action independent of non-adrenoceptor imidazoline binding sites. *Psychopharmacology* 1994; **113**:493–499.
35. Thor Jr DH, Holloway WR. Play soliciting in juvenile male rats: effects of caffeine, amphetamine and methylphenidate. *Pharmacology Biochemistry and Behavior* 1983; **19**:725–727.
36. Mahrer JM, Magel RC. A comparison of tests for the $k$-sample non-decreasing alternative. *Statistics in Medicine* 1995; **14**:863–871.
37. Yiannoutsos CT, Nakas CT, Navia BA. Assessing multiple-group diagnostic problems with multi-dimensional receiver operating characteristic surfaces: application to proton MR spectroscopy (MRS) in HIV-related neurological injury. *NeuroImage* 2008; **40**:248–255. DOI: 10.1016/j.neuroimage.2007.09.056.