Biometrics 59, 1089–1098
December 2003

Incorporation of Clustering Effects for the Wilcoxon Rank Sum Test: A Large-Sample Approach

Bernard Rosner,* Robert J. Glynn, and Mei-Ling Ting Lee

Channing Laboratory, Harvard Medical School, 181 Longwood Avenue, Boston, Massachusetts, U.S.A.
**email:* bernard.rosner@channing.harvard.edu

Summary. The Wilcoxon rank sum test is frequently used in statistical practice for the comparison of measures of location when the underlying distributions are far from normal or not known in advance. An assumption of the ordinary rank sum test is that individual sampling units are independent. In many ophthalmologic clinical trials, the Early Treatment for Diabetic Retinopathy Scale (ETDRS) is a principal endpoint used for measuring the level of diabetic retinopathy. This is an ordinal scale, and it is natural to consider the Wilcoxon rank sum test for the comparison of the level of diabetic retinopathy between treatment groups. However, under this design, unlike the usual Wilcoxon rank sum test, the subject is the unit of randomization, but the eye is the unit of analysis. Furthermore, a person will tend to have different, but correlated, ETDRS scores for fellow eyes. Thus, we propose a correction to the variance of the Wilcoxon rank sum statistic that accounts for clustering effects and that can be used for both balanced (same number of subunits per cluster) or unbalanced (different number of subunits per cluster) data, both in the presence or absence of ties, with p-value adjusted accordingly. In this article, we present large-sample theory and simulation results for this test procedure and apply it to diabetic retinopathy data from type I diabetics in the Sorbinil Retinopathy Trial.

Key words: Clustered data; Nonparametric tests; Ophthalmologic data.

1. Introduction

Much work has been considered over the past 20 years regarding the effect of clustering on levels of significance or confidence interval width. In general, in the presence of clustering, true p-values will be underestimated and confidence interval width will be too narrow when using standard statistical procedures that ignore clustering, because correlation between responses for two observations in the same cluster is usually, but not always, positive.

Most work in the area of clustered data concerns outcome variables with a normal or binomial distribution (Rosner, 1984; Liang and Zeger, 1986). A large amount of literature exists concerning clustered binary data in the context of developmental toxicity studies (Haseman and Kupper, 1978; Ryan, 1992; Regan and Catalano, 1999) and other medical specialties (Jung and Ahn, 2000; Jung, Ahn and Donner, 2001; Jung, Kang and Ahn, 2001). Some work has also focused on ordinal outcome variables (Rosner and Glynn, 1997). In a previous article, we considered the incorporation of clustering effects for the Mann-Whitney U statistic (Rosner and Grove, 1999); simulation results were presented, but large-sample theory was not considered. Also, this approach requires special software which can be computationally intensive in the presence of many tied rankings, and makes stronger assumptions in the setting of unequal cluster sizes that may not be warranted.

In Section 2 of this article, we consider a large-sample approach, where clustering effects can be easily incorporated with standard software (e.g., SAS PROC RANK) and can be used for either balanced or unbalanced clustered data, both in the presence or absence of tied rankings. Use of rank tests after stratification, by confounding variables, is also considered in Section 2. We present large-sample theory for this test procedure and in Section 3, consider simulation studies to assess finite sample properties. In Section 4, we present an example using this approach based on a comparison of change in diabetic retinopathy grade between treatment groups in the Sorbinil Retinopathy Trial for type I diabetic patients.

2. Methods

2.1 *No Clustering*

For notational purposes, we first consider the case of no clustering. Suppose there are two samples X and Y of size m and n, respectively. Define the Wilcoxon rank sum statistic by

$$W_{\text{obs}} = \sum_{i=1}^{m} \text{Rank}(X_i)$$

where $\text{Rank}(X_i)$ = rank of the ith observation in the X sample among the combined sample of $m + n$ observations. Under H_0, we assume that the X and Y samples come from the same underlying distribution. Hence, we refer to the combined X and Y samples by Z_i, $i = 1, \ldots, m + n \equiv N$. We use a randomization representation of W, obtained by randomly assigning m of the N observations to the X sample (denoted by

I) and the remaining observations to the Y sample as follows:

$$W = \sum_{i=1}^{N} \delta_i R_i \tag{1}$$

where R_i = rank of Z_i within the N observations in Z, and $\delta_i = 1$ if $i \in I$, $\delta_i = 0$, if $i \notin I$. If N is small, then we can assess significance by generating the $\binom{N}{m}$ possible values for the set I. However, for large N, we consider a large-sample approximation to the distribution of W. Under H_0,

$$E(W) = (m/N)\sum_{i=1}^{N} R_i = m(N+1)/2.$$

$$\mathrm{Var}(W) = \sum_{i=1}^{N} R_i^2 \mathrm{Var}(\delta_i) + \sum_{i \neq k}^{N} R_i R_k \mathrm{Cov}(\delta_i, \delta_k) \tag{2}$$

Since $\mathrm{Var}(\delta_i) = mn/N^2$ and $\mathrm{Cov}(\delta_i,\ \delta_k) = -mn/\{N^2(N-1)\}$, it follows from (2) that

$$\begin{aligned}\mathrm{Var}(W) &= [mn/\{N(N-1)\}] \\ &\quad \times \sum_{i=1}^{N} R_i^2 - mn(N+1)^2/\{4(N-1)\} \\ &= [mn/\{N(N-1)\}]\sum_{i=1}^{N}\left(R_i - \frac{1+N}{2}\right)^2\end{aligned}$$

If the distribution of Z is continuous with no ties, then

$$\sum_{i=1}^{N} R_i^2 = N(N+1)(2N+1)/6 \text{ and } \mathrm{Var}(W) = mn(N+1)/12 \tag{3}$$

If the distribution of Z is discrete with Q groups of tied values, then (2) can be shown to be

$$\mathrm{Var}(W) = (mn/12)\left[N+1-\sum_{q=1}^{Q}\left(t_q^3 - t_q\right)/\{N(N-1)\}\right]. \tag{4}$$

where t_q = number of observations in the qth tied group, $q = 1, \ldots, Q$.

The test statistic under either (3) or (4) is $Z_W = \{W - E(W)\}/\{\mathrm{Var}(W)\}^{1/2}$, which is asymptotically normal as $N \to \infty$ (Lehmann, 1975).

2.2 *Incorporating Clustering Effects*

To incorporate clustering effects, we assume that the data come in clusters where X_{ij} denotes the score for the jth subunit from the ith cluster in the first group, $i = 1, \ldots, m$; $j = 1, \ldots, g_i$ and Y_{kl} denotes the score for the lth subunit from the kth cluster in the second group, $k = 1, \ldots, n$; $l = 1, \ldots, h_k$. We define the clustered Wilcoxon rank sum statistic $W_{c,\mathrm{obs}}$ by

$$W_{c,\mathrm{obs}} = \sum_{i=1}^{m}\sum_{j=1}^{g_i} \mathrm{Rank}(X_{ij}) \tag{5}$$

where ranks are determined based on the combined sample of all subunits over the X and Y clusters combined. We assume the subunits for a given cluster are exchangeable. We wish to test

$$\begin{aligned}&H_0\colon Pr\{U(X_{ij}-Y_{kl})=1\} = Pr\{U(X_{ij}-Y_{kl})=0\}, \\ &\qquad\qquad \text{for any } i,j,k,l, \\ &\text{vs. } H_1\colon Pr\{U(X_{ij}-Y_{kl})=1\} \neq Pr\{U(X_{ij}-Y_{kl})=0\}, \\ &\qquad\qquad \text{for some } i,j,k,l,\end{aligned}$$

where $U(a) = 1$ if $a > 0$, $U(a) = 1/2$ if $a = 0$ and $U(a) = 0$ if $a < 0$. In words, the test is based on the probability that the score from a random subunit from the X sample is greater than the score from a random subunit from the Y sample. If there are no ties, then under H_0, this probability is 1/2, while under H_1 it is different from 1/2.

2.2.1 *Balanced designs.* We first consider the case of balanced data, i.e., the same number of subunits (g) for all clusters. Since scores for clusters assigned to the X and Y treatments are identically distributed under H_0, hereafter we will drop the distinction between X and Y clusters and refer to a combined set of Z clusters, where Z_{ij} = score for the jth subunit of the ith cluster, $j = 1, \ldots, g$, $i = 1, \ldots, m + n = N$. Suppose that m of the N clusters are assigned at random to the X treatment and the remaining n clusters to the Y treatment. Let $\delta_i = 1$ if $i \in I$, and $\delta_i = 0$ if $i \notin I$ denote the indicator function of the m unique values out of $\{1, \ldots, N\}$ randomly assigned to the X group, with $\Pr(i \in I) = m/N$ and $\sum_{i=1}^{N} \delta_i = m$. We can write the distribution of the clustered rank sum statistic $W_{c,\mathrm{obs}}$ in the form

$$W_c = \sum_{i=1}^{N} \delta_i R_{i+} \quad \text{where} \quad R_{i+} = \sum_{j=1}^{g} R_{ij} \tag{6}$$

and R_{ij} = rank of the jth subunit in the ith cluster among all gN subunits over all Z clusters. Thus, we can consider $\{R_{i+}, i \in I\}$ as a random sample of size m from the population $\{R_{i+},\ i = 1, \ldots, N\}$. There are $\binom{N}{m}$ elements of I. Hence, if N is small, we could generate the entire distribution of W_c from (6). However, since $\binom{N}{m}$ is usually large, we consider a large-sample approximation. We have

$$E(W_c) = (m/N)\sum_{i=1}^{N} R_{i+} = gm(gN+1)/2 \tag{7}$$

Furthermore, since $\mathrm{Var}(\delta_i) = mn/N^2$ and $\mathrm{Cov}(\delta_i,\ \delta_k) = -mn/\{N^2\ (N-1)\}$, we can write

$$\begin{aligned}\mathrm{Var}(W_c) &= (mn/N^2)\sum_{i=1}^{N} R_{i+}^2 - [mn/\{N^2(N-1)\}] \\ &\quad \times \left\{\left(\sum_{i=1}^{N} R_{i+}\right)^2 - \sum_{i=1}^{N} R_{i+}^2\right\} \\ &= [mn/\{N(N-1)\}]\sum_{i=1}^{N}\{R_{i+} - g(1+gN)/2\}^2\end{aligned} \tag{8}$$

Note that (8) can be obtained directly from sampling theory for finite populations (Hansen, Hurwitz, and Madow, 1960) by considering $\{R_{i+}, i \in I\}$ as a random sample of m clusters that could have hypothetically been assigned to the X group out of a finite population of N clusters. A natural large sample test statistic to consider based on (6), (7), and (8) is

$$Z_c = \{W_c - gm(gN+1)/2\}/\{\mathrm{Var}(W_c)\}^{1/2} \tag{9}$$

In the Appendix, we prove that Z_c is asymptotically normal if both $m \to \infty$ and $n \to \infty$.

2.2.2 *Unbalanced designs.* An assumption underlying (6)–(9) is that the sample points from which the permutation distribution is derived (i.e., the R_{i+}) are identically distributed for each cluster i. This assumption will be violated in an unbalanced design with variable cluster sizes. Therefore, for unbalanced designs, we let $(m_g,\, n_g)$ = number of clusters of size g assigned to the X and Y treatment. Denote $N_g = m_g + n_g$ for $g = 1, \ldots, g_{\max}$ and $N = \sum_{g=1}^{g_{\max}} N_g$. Let $R_{ij,g}$ = rank for the jth subunit in the ith cluster of size g, $g = 1, \ldots, g_{\max}$; $i = 1, \ldots, N_g$, $j = 1, \ldots, g$, where ranks are computed based on the total study population of $\sum_{g=1}^{g_{\max}} gN_g$ subunits. Let $I_{g,\mathrm{obs}} = \{i_1, \ldots, i_{m_g}\}, 1 \le i_1 < i_2 < \cdots < i_{m_g} \le N_g$ denote a subset of m_g unique indices selected from $\{1, \ldots, N_g\}$ corresponding to clusters of size g that are actually assigned to the X treatment. The statistic $W_{c,\mathrm{obs}}$ can then be written in the form:

$$W_{c,\mathrm{obs}} = \sum_{g=1}^{g_{\max}} \sum_{i \in I_{g,\mathrm{obs}}} R_{i+,g} \tag{10}$$

where $R_{i+,g}$ = sum of ranks of all subunits in the ith cluster of size g, $i = 1, \ldots, N_g$. Now consider the set $\underset{\sim}{I} = (I_1, \ldots, I_g, \ldots, I_{g_{\max}})$, where I_g is a random subset of clusters from the N_g clusters of size g that hypothetically could have been assigned to the X treatment. The distribution corresponding to $W_{c,\mathrm{obs}}$ is

$$W_c = \sum_{g=1}^{g_{\max}} \left(\sum_{i \in I_g} R_{i+,g} \right) \equiv \sum_{g=1}^{g_{\max}} S_g = \sum_{g=1}^{g_{\max}} \sum_{i=1}^{N_g} \delta_{i,g} R_{i+,g} \tag{11}$$

where $\delta_{i,g} = 1$ if $i \in I_g$, $\delta_{i,g} = 0$ if $i \notin I_g$ denote the indicator function of the m_g unique values out of $\{1, \ldots, N_g\}$ hypothetically assigned to the X group with $\Pr\{i \in I_g\} = m_g/N_g$ and $\sum_{i=1}^{N_g} \delta_{i,g} = m_g$. This is a direct generalization of (6) in the case of unbalanced data.

If N is small, we can generate the distribution of W_c from the $\prod_{g=1}^{g_{\max}} \binom{N_g}{m_g}$ elements of $\underset{\sim}{I}$ and evaluate the significance of $W_{c,\mathrm{obs}}$ from $p = 2 \times \min\{\Pr(W_c \le W_{c,\mathrm{obs}}), \Pr(W_c \ge W_{c,\mathrm{obs}}), 0.5\}$. Note that the distribution will be unique for each possible vector $\underset{\sim}{m} = (m_1, m_2, \ldots, m_{g_{\max}})$.

If N is large, we will consider a large-sample test. We wish to obtain the moments of W_c under H_0. For this purpose, we will assume that in general, the expected rank of a subunit is a function of cluster size. Thus,

$$E(W_c) = \sum_{g=1}^{g_{\max}} m_g(R_{++,g}/N_g) = \sum_{g=1}^{g_{\max}} E(S_g) \tag{12}$$

Also, we can consider $\{R_{i+,g},\ i \in I_g\}$ as a random sample of size m_g from the population $\{R_{i+,g},\ i = 1, \ldots, N_g\}$, $g = 1, \ldots, g_{\max}$. Furthermore, S_{g_1} and S_{g_2} as defined in (11) are independent, because independent random sampling is used to select I_{g_1}, I_{g_2} for $g_1 \ne g_2$. It follows that

$$\begin{aligned}\mathrm{Var}(W_c) &= \mathrm{Var}\left(\sum_{g=1}^{g_{\max}} S_g \right) = \sum_{g=1}^{g_{\max}} \mathrm{Var}(S_g) \\ &= \sum_{g=1}^{g_{\max}} [m_g n_g / \{N_g(N_g - 1)\}] \sum_{i=1}^{N_g} (R_{i+,g} - R_{++,g}/N_g)^2 \end{aligned} \tag{13}$$

A large-sample test statistic based on (11), (12), and (13) is

$$Z_c = \left(W_c - \sum_{g=1}^{g_{\max}} m_g R_{++,g}/N_g \right) \Bigg/ \{\mathrm{Var}(W_c)\}^{1/2} \tag{14}$$

In Theorem 1 in the Appendix, we show that Z_c converges in law to a $N(0,\, 1)$ distribution as $N = \sum_{g=1}^{g_{\max}} N_g \to \infty$, provided that (a) $g_{\max}$ = maximum cluster size $< \infty$, and (b) $\lim_{N \to \infty} m_g/N_g = \xi_g$, where $0 < \xi_g < 1$, $g = 1, \ldots, g_{\max}$.

2.2.3 *Stratification.* It is often the case in observational studies that the primary comparison groups (i.e., the X and Y groups) are not balanced on other important confounding variables. Also, in multicenter clinical trials, stratification by center is common and methods that control for the center effect are important. It is desirable, in this case, to modify W_c in equations (6) and (11) to control for confounding variables. Suppose the set of relevant confounding variables can be summarized in terms of V strata. Let $(m_{g,v},\, n_{g,v})$ = number of clusters of size g in stratum v assigned to the X and Y treatment, and let $N_{g,v} = m_{g,v} + n_{g,v}$ denote the number of clusters of size g in stratum v, $g = 1, \ldots, g_{\max}$, $v = 1, \ldots, V$. Let $R_{i+,g,v}$ be the rank sum for the subunits in the ith cluster of size g in the vth stratum. We define:

$$W_{c,\mathrm{obs}} = \sum_{g=1}^{g_{\max}} \sum_{v=1}^{V} \left(\sum_{i \in I_{g,v,\mathrm{obs}}} R_{i+,g,v} \right) \tag{15}$$

where $I_{g,v,\mathrm{obs}}$ is the observed subset of $m_{g,v}$ unique indices selected from $\{1, \ldots, N_{g,v}\}$, corresponding to clusters of size g in stratum v that are actually assigned to the X treatment. We now consider the set $\underset{\sim}{I} = (I_{1,1}, \ldots, I_{g,v}, \ldots, I_{g_{\max},V})$, where $I_{g,v}$ is a random subset of $m_{g,v}$ clusters from the $N_{g,v}$ clusters of size g in stratum v that hypothetically might have been assigned to the X treatment. The distribution corresponding to $W_{c,\mathrm{obs}}$ is

$$\begin{aligned} W_c &= \sum_{g=1}^{g_{\max}} \sum_{v=1}^{V} \sum_{i \in I_{g,v}} R_{i+,g,v} \equiv \sum_{g=1}^{g_{\max}} \sum_{v=1}^{V} S_{g,v} \\ &= \sum_{g=1}^{g_{\max}} \sum_{v=1}^{V} \sum_{i=1}^{N_{g,v}} \delta_{i,g,v} R_{i+,g,v} \end{aligned} \tag{16}$$

where $\delta_{i,g,v} = 1$ if $i \in I_{g,v}$, $\delta_{i,g,v} = 0$ otherwise and $\sum_{i=1}^{N_{g,v}} \delta_{i,g,v} = m_{g,v}$. Let $N = \sum_{v=1}^{V} \sum_{g=1}^{g_{\max}} N_{g,v}$ denote the total number of clusters.

If N is small, we can generate the distribution of W_c from the $\prod_{g=1}^{g_{\max}} \prod_{v=1}^{V} \binom{N_{g,v}}{m_{g,v}}$ elements of $\underset{\sim}{I}$ and evaluate the significance of $W_{c,\mathrm{obs}}$ from $p = 2 \times \min\{\Pr(W_c \le W_{c,\mathrm{obs}}), \Pr(W_c \ge W_{c,\mathrm{obs}}), 0.5\}$. For large N, we employ a large-sample test similar to (11)–(14) as follows:

$$E(W_c) = \sum_{g=1}^{g_{\max}} \sum_{v=1}^{V} m_{g,v} R_{++,g,v}/N_{g,v} = \sum_{g=1}^{g_{\max}} \sum_{v=1}^{V} E(S_{g,v})$$

$$\begin{aligned}\mathrm{Var}(W_c) = &\sum_{g=1}^{g_{\max}} \sum_{v=1}^{V} [m_{g,v} n_{g,v}/\{N_{g,v}(N_{g,v}-1)\}] \\ &\times \sum_{i=1}^{N_{g,v}} (R_{i+,g,v} - R_{++,g,v}/N_{g,v})^2 \end{aligned} \tag{17}$$

with test statistic

$$Z_c = \{W_c - E(W_c)\}/\{\mathrm{Var}(W_c)\}^{1/2} \tag{18}$$

In Theorem 2 in the Appendix, we show that Z_c converges in law to a $N(0,\, 1)$ distribution as $N \to \infty$, provided that (a) $g_{\max}$ = maximum cluster size $< \infty$, (b) $V < \infty$ and (c) $\lim_{N\to\infty} m_{g,v}/N_{g,v} = \xi_{g,v}$, where $0 < \xi_{g,v} < 1$, $g = 1, \ldots, g_{\max}$, $v = 1, \ldots, V$.

2.2.4 *Relationship between the clustered Wilcoxon test statistic and the Rosner-Grove clustered Mann-Whitney U statistic.* In a previous report (Rosner and Grove, 1999), denoted by RG, we introduced the clustered Mann-Whitney U Statistic $U_{c,\mathrm{RG}}$ defined by $U_{c,\mathrm{RG}} = \sum_{(i,j,k,l)} U(X_{ij} - Y_{kl})$ which, for a balanced design with g subunits per cluster, differs from $W_{c,\mathrm{obs}}$ in (5) only by a constant given by $W_{c,\mathrm{RG}} = W_{c,\mathrm{obs}} = U_{c,\mathrm{RG}} + (gm)(gm+1)/2$. Thus, $E(W_{c,\mathrm{RG}}) = E(U_{c,\mathrm{RG}}) + (gm)(gm+1)/2 = gm(gN+1)/2$. Furthermore,

$$\begin{aligned}\mathrm{Var}(U_{c,\mathrm{RG}}) = \mathrm{Var}(W_{c,\mathrm{RG}}) = \big\{&mng^2 + mng^2(g-1)^2\rho_1 \\ &+ 2mng^2(g-1)\rho_2 + mn(m+n-2)g^3(g-1)\rho_3 \\ &+ mn(m+n-2)g^3\rho_4\big\}\{1 - \Pr(Z_{ij} = Z_{kl})\}/4 \end{aligned} \tag{19}$$

where $\rho_1 = \mathrm{Corr}\{(U(Z_{ij_1} - Z_{kl_1}), U(Z_{ij_2} - Z_{kl_2})\}$, $\rho_2 = \mathrm{Corr}\{U(Z_{ij} - Z_{kl_1}), U(Z_{ij} - Z_{kl_2})\}$, $\rho_3 = \mathrm{Corr}\{U(Z_{ij_1} - Z_{k_1 l_1}), U(Z_{ij_2} - Z_{k_2 l_2})\}$, $\rho_4 = \mathrm{Corr}\{U(Z_{ij} - Z_{k_1 l_1}), U(Z_{ij} - Z_{k_2 l_2})\}$; ρ_1, ρ_2, ρ_3, and ρ_4 are estimated from a method-of-moments approach and $\Pr(Z_{ij} = Z_{kl})$ is estimated by $\sum_{i,j,k,l} \mathbf{1}\{U(Z_{ij} - Z_{kl}) = \frac{1}{2}\}/\{N_g(N_g - 1)\}$. It can be shown, after extensive algebra, that the variance expression in (19) is identical to $\mathrm{Var}(W_c)$ in (8). It follows that $Z_{c,\mathrm{RG}} = \{W_{c,\mathrm{RG}} - E(W_{c,\mathrm{RG}})\}/\{\mathrm{Var}(W_{c,\mathrm{RG}})\}^{1/2} = Z_c$ in (9). Hence, the test procedures using the two approaches are identical in the case of a balanced design. The advantage of the approach in Section 2.2.1 is that $\mathrm{Var}(W_c)$ in (8) is computationally trivial and the test can be easily implemented using standard SAS software (i.e., located by clicking on "Data Sets/Computer Codes" on the Biometrics website, `http://stat.tamu.edu/Biometrics`), while $\mathrm{Var}(W_{c,\mathrm{RG}})$ in (19) requires special software to estimate ρ_1, ρ_2, ρ_3, and ρ_4.

For unbalanced designs, the procedures are not equivalent. It is still the case that $W_{c,\mathrm{RG}} = W_{c,\mathrm{obs}}$. However, an assumption under the RG approach is that under H_0, $E\{U(Z_{ij,g} - Z_{kl,h})\} = 1/2$ for all g, $h = 1, \ldots, g_{\max}$. In words, the score $(Z_{ij,g})$ for the jth subunit of the ith cluster of size g is independent of the cluster size (g). Under the approach in Section 2.2.2, this assumption is relaxed and instead $E\{U(Z_{ij,g} - Z_{kl,h})\} = \lambda_{gh}$ where, in general, $\lambda_{gh} \neq 1/2$ if $g \neq h$. This leads to different expressions for the moments of the test statistic under the two approaches. Specifically,

$$\begin{aligned} E(W_{c,\mathrm{RG}}) &= \sum_{g=1}^{g_{\max}} m_g \left[g \left\{ \left(\sum_{q=1}^{g_{\max}} qN_q \right) + 1 \right\} \Big/ 2 \right] \\ &\equiv \mu_{\mathrm{RG}} \neq E(W_c) \text{ in (12)} \\ \mathrm{Var}(W_{c,\mathrm{RG}}) &= E(W_{c,\mathrm{RG}} - \mu_{\mathrm{RG}})^2 \\ &= E\left[\sum_{g=1}^{g_{\max}} \left[S_g - m_g g \left\{ \left(\sum_{q=1}^{g_{\max}} qN_q \right) + 1 \right\} \Big/ 2 \right] \right]^2 \\ &\neq \mathrm{Var}(W_c) \text{ in (13).} \end{aligned}$$

In general, $\mathrm{Var}(W_c)$ will be smaller than $\mathrm{Var}(W_{c,\mathrm{RG}})$, due to heterogeneity of λ_{gh} from $1/2$ if $g \neq h$. Furthermore, one can compute $\mathrm{Var}(W_c)$ for unbalanced designs from (13) using standard software (e.g., PROC RANK of SAS), while $\mathrm{Var}(W_{c,\mathrm{RG}})$ requires special programming. We will compare the two approaches again in the eighth simulation design in Table 1.

3. Simulation Study

The test statistic Z_c is shown in the Appendix to be asymptotically normal as the number of clusters gets large, for both balanced and unbalanced designs, either in the presence or absence of stratification variables. In this section, we present the results of simulation studies to assess the finite sample properties of this random variable. To assess type I error of the test procedure in finite samples for a given m, n, and ρ, we generate

$$H_{ij} = T_i + e_{ij}, T_i \sim N(0, \rho), e_{ij} \sim N(0, 1-\rho), \tag{20}$$

where $i = 1, \ldots, N = m + n$; $j = 1, \ldots, g_i$. It follows that $\mathrm{Corr}(H_{ij_1}, H_{ij_2}) = \rho$.

We then computed $H^*_{ij} = \exp(H_{ij})$. This created lognormally distributed data that are typical of skewed distributions for which rank procedures are often used. We then computed the clustered Wilcoxon statistic in the H^*_{ij} scale. Note that from (20), the data will still be exchangeable in the H^*_{ij} scale. In addition, we computed the mean cluster score and performed the ordinary Wilcoxon rank sum test based on the cluster means, which we refer to as the "cluster-mean" approach. Each of these procedures was assessed in 8 different designs. Designs 1–6 are based on continuous data with no ties; designs 1–4 are balanced designs and designs 5–6 are unbalanced designs. Design 7 is similar to design 2, except that the continuous data are divided into 6 groups corresponding to the (<20th, 20th to <40th, 40th to <50th, 50th to <60th, 60th to <80th, and >=80th) percentiles of a normal distribution, with the actual values replaced by the median value of the normal distribution within the respective group and then exponentiated. Thus, design 7 should be typical of datasets with many tied values. Design 8 is a particular unbalanced design with continuous data, where the cluster-size distribution is different for the X and Y groups, and the expected score is a function of cluster size. This type of design should maximize the contrast between the clustered Wilcoxon procedure in (10)–(14) and the RG approach discussed in Section 2.2.4.

Table 1

Simulation study—type I error of clustered Wilcoxon rank sum statistic and the ordinary Wilcoxon rank sum statistic, lognormal distribution, nominal $\alpha = 0.05$, 4000 *replications per cell*

						W_c[a]				W[b]			
Design	m_2	n_2	m_4	n_4	ρ	0.05	0.2	0.5	0.8	0.05	0.2	0.5	0.8
1	20	20	—	—	$\hat{\alpha}$	0.050	0.048	0.048	0.050	0.048	0.050	0.050	0.050
					C[d]	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.01
2	50	50	—	—	$\hat{\alpha}$	0.052	0.053	0.051	0.052	0.051	0.052	0.050	0.050
					C	1.02	1.00	1.00	1.01	0.99	0.98	0.99	1.01
3	—	—	20	20	$\hat{\alpha}$	0.044	0.044	0.044	0.047	0.047	0.047	0.047	0.047
					C	0.97	0.99	1.01	1.01	0.99	1.00	1.01	1.01
4	—	—	50	50	$\hat{\alpha}$	0.050	0.045	0.048	0.049	0.048	0.048	0.051	0.051
					C	0.99	0.99	0.99	0.99	0.97	0.97	0.98	0.99
5	10	10	10	10	$\hat{\alpha}$	0.051	0.051	0.053	0.055	0.050	0.052	0.057	0.053
					C	1.00	1.01	1.02	1.02	1.01	1.03	1.04	1.03
6	25	25	25	25	$\hat{\alpha}$	0.052	0.051	0.048	0.053	0.052	0.047	0.046	0.044
					C	1.00	1.00	1.01	1.01	1.01	1.00	1.00	1.01
7	50	50	—	—	$\hat{\alpha}$	0.054	0.053	0.051	0.050	0.054	0.052	0.049	0.049
	(grouped data)				C	1.01	1.00	0.99	1.00	1.00	0.99	0.98	1.00
8[c]	30	20	20	30	$\hat{\alpha}_{W_c}$	0.053	0.054	0.048	0.054	0.190	0.178	0.157	0.133
					C_{W_c}	1.05	1.03	1.02	1.02	0.58	0.62	0.70	0.76
					$\hat{\alpha}_{RG}$	0.176	0.171	0.169	0.171	—	—	—	—
					C_{RG}	0.65	0.61	0.60	0.62	—	—	—	—

[a] Using the subunit as the unit of analysis.

[b] Using the cluster mean as the unit of analysis.

[c] $E(T_i|g_i = 2) = 0$; $E(T_i|g_i = 4) = 1$.

[d] $C = \sum_{i=1}^{4000}\{(W_c^{(i)} - \overline{W}_c)^2/3999\}/\text{mean}\{\text{Var}(W_c)\}$.

For each of the 8 designs, we assessed each procedure for $\rho =$ 0.05, 0.2, 0.5, and 0.8, with 4000 simulations for each value of ρ.

For each design, we computed the empirical type I error $\hat{\alpha}$ = empirical proportion of test statistics Z_c that exceed the nominal critical value 1.96 (at a 5% significance level) and

$$C = \sum_{i=1}^{4000}\left\{\left(W_c^{(i)} - \overline{W}_c\right)^2\Big/3999\right\}\Big/\text{mean}\{\text{Var}(W_c)\}$$

where $W_c^{(i)}$ = clustered Wilcoxon rank sum statistic for the ith simulation, $i = 1, \ldots, 4000$. C provides an estimate of the validity of the variance estimates in (8) and (13). The results are given in Table 1.

For designs 1–7 in Table 1, the range of estimated type I errors for the clustered Wilcoxon procedure is from 0.044 to 0.055, with average type I error = 0.050. The range of estimated C statistics is from 0.97 to 1.02, with average C statistic =1.00. For the cluster-mean procedure, the range of estimated type I errors is from 0.044 to 0.057, with average type I error = 0.050. The range of estimated C statistics is from 0.97 to 1.04, with average C statistic = 1.00. Thus, simulation results indicate that the large-sample nominal type I error and estimated variance for both the clustered Wilcoxon and cluster mean procedures are appropriate for datasets with ≥20 clusters per group.

In design 8, we see that the unbalanced design approach in Section 2.2.2 adequately controls the type I error of W_c for each value of ρ. However, the RG approach has a type I error of $0.17 - 0.18$, and clearly does not preserve the size of the test under this design. The reason is that the average rank of the X and Y groups is not the same under H_0, although it is the same within cluster size-specific strata. The cluster mean procedure has type I error of $0.13 - 0.19$ and also does not preserve the size of the test for the same reason. In addition, the C statistic is close to 1 for the clustered Wilcoxon approach, but ranges from $0.60 - 0.65$ for the RG approach and from $0.58 - 0.76$ for the cluster mean approach, indicating inappropriate variance estimation for the latter two approaches under this design.

Another issue in comparing the clustered Wilcoxon and cluster-mean approaches is power. For this purpose, we repeated the analyses in designs 1–7 of Table 1, but in (20), set $T_i \sim (0.4, \rho)$ for the observations in group 2, and $T_i \sim N(0, \rho)$ for the observations in group 1. The resulting power analyses are shown in Figure 1.

For balanced designs for each method, power increased both with an increasing number of clusters and with an increasing number of subunits per cluster. Power for unbalanced designs was generally intermediate between the power for the two corresponding balanced designs. Also, for each of the 7 designs in Figure 1, power decreased with increasing intraclass correlation (ρ). In 23 out of 28 cases, the clustered Wilcoxon procedure had more power than the cluster-mean procedure, in some instances, substantially so. The largest differences in power (about 5–8%) occurred when the number of subunits

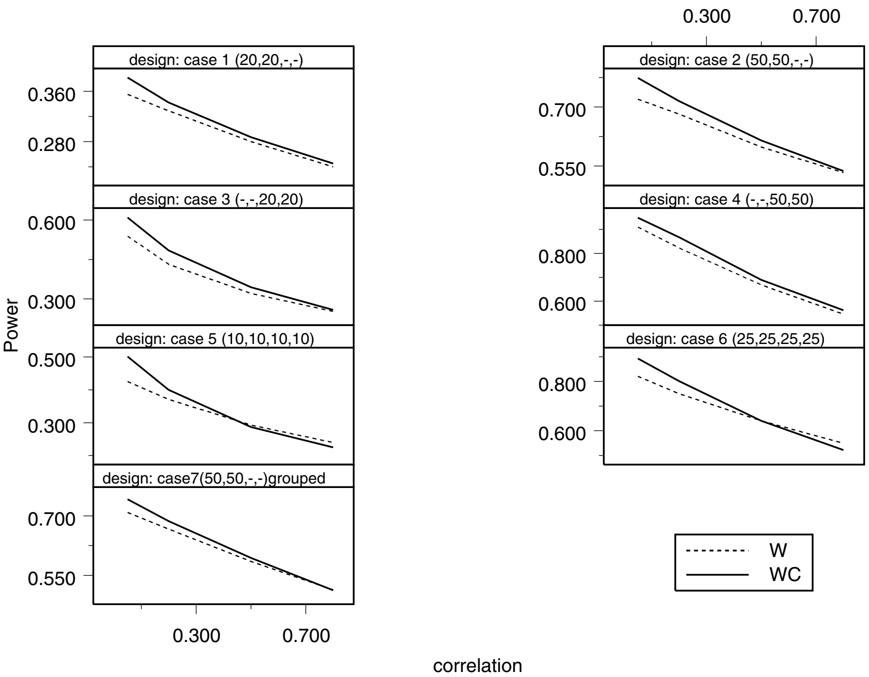

Figure 1. Power comparison of clustered Wilcoxon rank sum statistic (W_c) versus the ordinary Wilcoxon rank sum statistic (W) as a function of the intraclass correlation (ρ = 0.05, 0.2, 0.5, 0.8) for designs 1–7 described in Table 1. The values (m_2, n_2, m_4, n_4) are listed for each design. Each power estimate is based on 4000 replications using a lognormal distribution, and a nominal $\alpha = 0.05$.

Table 2

Simulation study—type I error of ordinary Wilcoxon rank sum test in the presence of clustering with the subunit as the unit of analysis, nominal α = .05, 4000 replications per cell

			ρ		
m_2	n_2		0.2	0.5	0.8
20	20	$\hat{\alpha}$	0.073 ± 0.004	0.106 ± 0.005	0.148 ± 0.006
		C	1.21	1.50	1.81
50	50	$\hat{\alpha}$	0.075 ± 0.004	0.111 ± 0.005	0.146 ± 0.006
		C	1.19	1.48	1.80

All type I errors are significantly different from $\alpha = 0.05 (p < 0.001)$.

$$C = \left\{ \sum_{i=1}^{4000} (W_{c,i} - \overline{W}_c)^2/3999 \right\} \Big/ \{mn(m+n+1)/12\}$$

was large (g = 4) and/or the intraclass correlation was low (≤0.2). For very high intraclass correlation (ρ = 0.8), there was little difference in power between the two procedures. In general, the clustered Wilcoxon procedure has appropriate type I error for at least 20 clusters per group and has a superior power profile to the cluster-mean procedure. Differences between the procedures can be expected to widen as the number of subunits per cluster increases.

We also studied the effect of using the ordinary Wilcoxon rank sum test, ignoring the clustering when clustering was actually present. Both the empirical type I error and C statistic (the design effect) were computed. Balanced designs with (m, n) = (20, 20), (50, 50), g = 2 and ρ = (0.2, 0.5, 0.8) were considered. The results are given in Table 2.

We see that the effect of ignoring the clustering is substantial, even for $\rho = 0.2$. The type I error is about 7–8% for $\rho = 0.2$, 11% for $\rho = 0.5$, and 15% for $\rho = 0.8$. Similarly, the design effect (C) is approximately 1.2 for $\rho = 0.2$, 1.5 for $\rho = 0.5$, and 1.8 for $\rho = 0.8$. Hence, the standard Wilcoxon rank sum test is inappropriate for clustered ranked data, even for levels of correlation as low as 0.2. Furthermore, it would be expected that both the type I error and the design effect would increase even further if the number of subunits per cluster was >2.

4. Example

The Sorbinil Retinopathy Trial was conducted among type I diabetic patients who had little or no evidence of retinopathy at baseline. Four hundred ninety-seven patients were randomized to either Sorbinil, an aldose reductase inhibitor, or placebo and were seen at 1-year and then at 9-month intervals, up to 48 months. Additionally, all subjects had a scheduled final visit at the end of the trial (maximum = 56 months). Sixteen of the patients provided no follow-up and an additional 3 patients were missing important baseline covariates; this resulted in 478 patients being used for the analyses in this article, of whom 237 were randomized to Sorbinil and 241 to placebo. The analyses here are based on the diabetic retinopathy grade at maximum follow-up, minus the diabetic retinopathy grade at randomization. The diabetic retinopathy grade at a visit is based on the ETDRS (Early Treatment for Diabetic Retinopathy Study) grading system with grades of 10, 20, 30, 41, 45, 55, or 61 in each eye. Higher ratings indicate more severe retinopathy, with 10 indicating no diabetic retinopathy and 61 indicating proliferative (very severe) diabetic retinopathy. In the primary analyses for this study

Table 3

Changes in ETDRS diabetic retinopathy grade in the Sorbinil Retinopathy Trial by treatment group

	Right eye		Left eye		Both eyes	
Difference scores[a]	S[b] n(%)	P[b] n(%)	S n(%)	P n(%)	S n(%)	P n(%)
−21	2(1)	0(0)	0(0)	0(0)	2(0)	0(0)
−20	0(0)	0(0)	0(0)	1(0)	0(0)	1(0)
−11	2(1)	0(0)	1(0)	1(0)	3(1)	1(0)
−10	20(8)	10(4)	17(7)	18(7)	37(8)	28(6)
0	125(53)	122(51)	123(52)	126(52)	248(52)	248(51)
4	1(0)	0(0)	1(0)	0(0)	2(0)	0(0)
10	56(24)	71(29)	59(25)	69(29)	115(24)	140(29)
11	1(0)	1(0)	2(1)	2(1)	3(1)	3(1)
15	0(0)	2(1)	0(0)	0(0)	0(0)	2(0)
20	10(4)	13(5)	10(4)	6(2)	20(4)	19(4)
21	7(3)	10(4)	9(4)	7(3)	16(3)	17(4)
25	5(2)	3(1)	5(2)	4(2)	10(2)	7(1)
31	5(2)	7(3)	6(3)	7(3)	11(2)	14(3)
35	1(0)	2(1)	4(2)	0(0)	5(1)	2(0)
41	2(2)	0(0)	0(0)	0(0)	2(0)	0(0)
Total	237	241	237	241	474	482
W[c]	53, 638.5		57, 615.5		222, 253	
$E(W)$	56, 761.5		56, 761.5		226, 809.0	
Var(W)	1, 921, 623		1, 912, 739		15, 321, 756	
Z_W	−2.253		0.617		−1.164	
p-value	0.024		0.537		0.245	

[a] Grade at the last available follow-up visit minus grade at the randomization visit. Positive changes indicate worsening; negative changes indicate improvement.

[b] S=Sorbinil group; P = placebo group.

[c] Observed rank sum in the Sorbinil group.

(Sorbinil Retinopathy Trial Research Group, 1990), the person was used as the unit of analysis based on a composite grade, using the joint retinopathy status of the right and left eyes; a binary outcome was used based on a worsening by 2 or more levels on this person-specific scale. Because potentially information is lost by (a) collapsing eye-specific grades into a single person-specific grade and (b) dichotomizing the outcome as a change of 2+ levels, we used a different approach. Specifically, in this article, we used the eye as the unit of analysis, with the change in retinopathy status for an eye computed as a difference score = diabetic retinopathy grade at the last available follow-up visit, minus the diabetic retinopathy grade at randomization. Since the distribution of maximum follow-up time was similar for the Sorbinil and placebo groups, this resulted in minimal bias in estimating the treatment effect. The above difference score is an ordinal variable and thus the Wilcoxon rank sum test is a natural method of analysis in this setting. However, since the difference scores for two eyes of an individual are correlated, the clustered Wilcoxon rank sum test was used. The balanced design approach in (6)–(9) was used because each individual with at least one follow-up visit provided difference scores (Z_{ij}) for each eye with no missing data, where $i = 1, \ldots, 478$ denotes the patient and $j = 1, 2$ denotes the right and left eye, respectively. In Table 3, we present the distribution of difference scores for right and left eyes, separately and combined, as well as the standard Wilcoxon rank sum test.

There were significant differences between Sorbinil and placebo for the right eye ($p = 0.024$), in the direction of benefit (i.e., smaller change scores for the Sorbinil group). However, for the left eye, no significant differences were found ($p = 0.54$). For both eyes combined, based on 956 eyes, the p-value was 0.25. However, this latter analysis is flawed, because it does not take the clustering between difference scores for fellow eyes into account. For this purpose, we present results using the clustered Wilcoxon test in Table 4.

The observed rank sum in the Sorbinil group (222, 253) and its expected value under H_0(226, 809) are identical for both the standard and clustered Wilcoxon tests. However, $\mathrm{Var}(W_c)/\mathrm{Var}(W) \approx 1.38$, reflecting the increased variance due to clustering. This results in higher p-values once clustering is accounted for (Z_c in (9) $= -0.989$, $p = 0.32$ vs. Z_W in Section 2.1 $=-1.164$, $p = 0.25$).

There were several covariates that were predictive of the difference score. The most important was baseline total glycosylated hemoglobin (TGH), an indicator of diabetic control, with higher values reflecting poorer control. Thus, to control for possible confounding by TGH, we stratified the sample at the approximate median ($\geq$12 vs. $<$ 12%). There was a slightly higher percent of persons with TGH $\geq$12% in the Sorbinil group (43%) vs. the placebo group (39%) (Table 4). Thus, we used the clustered Wilcoxon test after controlling for TGH (equations (15)–(18)). The results indicate that $E(W_c)$ increased after stratification (crude, 226,809;

Table 4

Use of the Clustered Wilcoxon Rank Sum test to compare changes in ETDRS diabetic retinopathy grade in the Sorbinil Retinopathy Trial

	Total Population	TGH^{++}		Total Population
	(unadjusted analysis)	<12%	≥12%	(adjusted for glycosylated hemoglobin)
$n(S, P)^*$	(237, 241)	(134, 146)	(103, 95)	(237, 241)
W_c	222, 253$^+$	117, 356.5	104, 896.5	222, 253$^+$
$E(W_c)$	226, 809	118, 503.1	109, 152.7	227, 655.8
$\text{Var}(W_c)$	21, 218, 545	9, 335, 256	11, 002, 385	20, 337, 641
Z_c	$-0.989^\dagger$	—	—	$-1.198^\ddagger$
p-value	0.323	—	—	0.231

$^{++}$TGH = total glycosylated hemoglobin. In this example, all subjects had 2 eyes available (i.e., $g = 2$). There were two strata (i.e., $V = 2$), with $v = 1$ indicating glycosylated hemoglobin <12% and $v = 2$ indicating glycosylated hemoglobin ≥12%. Hence, $S_{2,1}$, $E(S_{2,1})$, $\text{Var}(S_{2,1})$ are given in the 2nd, 3rd and 4th, rows of the glycosylated hemoglobin <12% column and $S_{2,2}$, $E(S_{2,2})$, $\text{Var}(S_{2,2})$ are given in the corresponding rows of the glycosylated hemoglobin ≥12% column.

*S = Sorbinil group; P = placebo group.

$^+$Rank sum in the Sorbinil group over 474 eyes from 237 subjects.

†Based on equation (9).

‡Based on equation (18).

adjusted, 227,656). Also, $\text{Var}(W_c)$ decreased by about 4% after controlling for TGH. The resulting p-value was slightly smaller (unadjusted, $p = 0.32$; adjusted, $p = 0.23$), but the results were still not statistically significant.

5. Discussion

In this article, we have generalized the standard Wilcoxon rank sum test to allow one to incorporate clustering effects for ranked data. The generalized variance formulas can be used for either balanced data (equation [8]) or unbalanced data (equation [13]), both in the presence or absence of tied values. In addition, covariate effects can be controlled for by stratification (equations [15]–[18]). These test statistics were shown to be asymptotically normal, as the number of clusters gets large if the maximum cluster size is bounded and the number of strata are finite. The simulations indicate that the large-sample procedure has appropriate type I error if the number of clusters is ≥20 in each group. In the case of a balanced design, the approach in this article can be shown to be identical to the clustered Mann-Whitney U approach reported previously (Rosner and Grove, 1999). In the case of an unbalanced design, the approach in Section 2.2.2 is more general, since it allows the expected rank to be a function of cluster size, while the clustered Mann-Whitney U approach does not. With the latter procedure, one can obtain invalid type I errors if the cluster size distribution differs by treatment group and the expected rank is a function of cluster size. Furthermore, the large-sample test approach is computationally trivial, with easy implementation in readily available software (e.g., SAS); thus, it does not require the special software needed for the previous implementation. A sample SAS program (cluswilcox.sas) is found on the Biometrics web page (`http://stat.tamu.edu/Biometrics`), which can be used for either balanced (equations [6]–[9]) or unbalanced (equations [10]–[14]) designs without additional covariates. The output from this program for the Sorbinil Retinopathy Trial data in Table 4, column 1 is provided after the program. An additional sample SAS program (stratify.cluswilcox.sas) is also provided at the same web site for stratified designs (equations [15]–[18]).

An assumption of the methods in the article is that all subunits within a cluster are exchangeable. This is usually appropriate for eye-specific outcomes measured at the same time, but may be invalid in other settings. Furthermore, in this article, the unit of randomization is the cluster, while the unit of analysis is the subunit. While this may be appropriate in many clinical trial settings, in observational studies, it is common to have both subunit-specific outcome and exposure variables (e.g., if the presence of cataract or elevated intraocular pressure in an eye is used to predict visual field in the same eye in glaucoma patients).

Acknowledgements

This work was supported by grant EY12269 from the National Eye Institute. We acknowledge the programming support of Marion McPhee and Ya-Hua Chen for the work in this article.

Résumé

Le test de Wilcoxon est fréquemment utilisé en pratique pour la comparaison des tendances centrales lorsque les distributions sous-jacentes sont loin d'être normales ou qu'elles sont inconnues. Une hypothèse du test ordinaire basé sur la somme des rangs est que les unités de l'échantillon sont indépendantes. Dans beaucoup d'essais cliniques ophtalmologiques, l'échelle de traitement précoce pour la rétinopathie diabétique (ETDRS) est le principal critère de jugement pour mesurer le niveau de la rétinopathie diabétique. C'est une échelle ordinale et il est naturel de considérer le test de Wilcoxon pour la comparaison des niveaux de rétinopathie diabétique entre les groupes de traitement. Cependant, avec ce plan d'étude, à la différence des situations usuelles d'utilisation du test de Wilcoxon, le sujet est l'unité de randomisation, mais l'oeil est l'unité d'analyse. De plus une personne tend des scores différends mais corrélés pour les deux yeux. Donc, nous proposons une correction de la variance du test de Wilcoxon pour les effets de groupe qui peut être utilisée pour les données équilibrées (même nombre dans chaque sous-unité du groupe) ou déséquilibrées (nombre différent de sous-unités par groupe), en présence possible d'ex-aequos, avec une p-value ajustée suivant les cas. Dans cet article, nous présentons la théorie pour les grands échantillons et des résultats de simulation pour cette procédure de test et l'appliquons à des données de rétinopathie diabétique (de diabète de type I) dans l'essai "Sorbinil Retinopathy trial".

References

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1960). *Sample Survey Methods and Theory II.* New York: Wiley.

Haseman, J. K. and Kupper, L. L. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* **34,** 69–76.

Jung, S. H. and Ahn, C. (2000). Estimation of response probability in correlated binary data: A new approach. *Drug Information Journal* **34,** 599–604.

Jung, S. H., Ahn, C., and Donner, A. (2001). Evaluation of an adjusted chi-square statistic as applied to studies involving clustered binary data. *Statistics in Medicine* **20,** 2149–2161.

Jung, S. H., Kang, X. X., and Ahn, C. (2001). Sample size calculations for clustered binary data. *Statistics in Medicine* **20,** 1971–1982.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

Liang, K. Y. and Zeger, S. W. (1986). Longitudinal data analysis using generalized linear models. *Biometrics* **73,** 13–22.

Regan, M. M. and Catalano, P. J. (1999). Likelihood models for clustered binary data and continuous outcomes: Application to developmental toxicity. *Biometrics* **55,** 760–768.

Rosner, B. (1984). Multivariate methods in ophthalmology with application to other paired data situations. *Biometrics* **40,** 1025–1035.

Rosner, B. and Glynn, R. J. (1997). Multivariate methods for clustered ordinal data with application to survival analysis. *Statistics in Medicine* **16,** 357–372.

Rosner, B. and Grove, D. (1999). Use of the Mann-Whitney U-test for clustered data. *Statistics in Medicine* **18,** 1387–1400.

Ryan, L. (1992). Quantitative risk assessment for developmental toxicity. *Biometrics* **48,** 155–162.

Sorbinil Retinopathy Trial Research Group (1990). A randomized trial of Sorbinil, an aldose reductase inhibitor, in diabetic retinopathy. *Archives of Ophthalmology* **108,** 1234–1244.

Received May 2002. *Revised June* 2003.
Accepted July 2003.

Appendix

Proof of Asymptotic Normality of W_c in Equations (6), (11), and (16). We first consider the asymptotic normality of S_g in (11) where, for the Appendix, for a dataset with N clusters, S_g is denoted by $S_{g,N}$. In the case of a balanced design, W_c in (6) is the same as S_g, and thus the Lemma will prove the asymptotic normality of W_c for balanced designs. We then extend this result to unbalanced designs in Theorem 1, and further extend it to the general case of stratified, possibly unbalanced, designs in Theorem 2. A sketch of this proof is given here. The full proof is available on the *Biometrics* web page (`http://stat.tamu.edu/Biometrics`).

Lemma: *Let $S_{g,N} = \sum_{i \in I_g} (R_{i+,g} - R_{++,g}/N_g)$ where I_g is defined in* (11). *$S_{g,N}$ is asymptotically normal as $N \to \infty$, provided that (a) $\lim_{N\to\infty} m_g/N_g = \xi_g$, where $0 < \xi_g < 1$ and (b) $N_g \to \infty$ as $N \to \infty$.*

Proof. Let $Z_{ij,g}$ denote the $\theta^{\text{th}}_{ij,g}$ percentile of Z in the reference population, i.e., $\Pr(Z \le Z_{ij,g}) = \theta_{ij,g}$. Note that we can write $R_{ij,g}$ in the form $R_{ij,g} = 1 + \sum_{(k,l)\neq(i,j)} U(Z_{ij,g} - Z_{kl,g}) + \sum_{h\neq g}^{g_{\max}} \sum_{k=1}^{N_h} \sum_{l=1}^{h} U(Z_{ij,g} - Z_{kl,h})$, and define $R^*_{ij,g} \equiv E(R_{ij,g}) = 1 + \{(\sum_{q=1}^{g_{\max}} qN_q) - 1\}\theta_{ij,g}, R^*_{i+,g} = \sum_{j=1}^{g} R^*_{ij,g}$. We consider the auxiliary statistic $S^*_{g,N} = \sum_{i=1}^{N_g} \{(R^*_{i+,g} - R^*_{++,g}/N_g)\}\delta_{i,g} \equiv \sum_{i\in I_g} V_{i,g,N_g}$, where $\delta_{i,g} = 1$ if $i \in I_g$ and $V_{i,g,N_g} = R^*_{i+,g} - R^*_{++,g}/N_g$, $i = 1, \ldots, N_g, \delta_{ij} = 0$ otherwise, and $\delta_{i,g}$ is independent of $\theta_{ij,g}$. Thus, $S^*_{g,N}$ is the randomization distribution of $\sum_{i\in I_{g,\text{obs}}} (R^*_{i+,g} - R^*_{++,g}/N_g)$, conditional on $\underset{\sim}{\theta}_g = (\theta_{11,g}, \ldots, \theta_{N_g g,g})$. We first establish the asymptotic normality of $S^*_{g,N}$ and then show asymptotic equivalence between $S^*_{g,N}$ and $S_{g,N}$. Using Lehmann (1975, Corollary 3, p. 354), we can show that $S^*_{g,N}$ is asymptotically normal as $N \to\infty$, for any $\underset{\sim}{\theta}_g$. We now must establish asymptotic equivalence between $S_{g,N}$ and $S^*_{g,N}$. From Lehmann (1975, Corollary 2, p. 349), asymptotic normality of $S_{g,N}$ will be established if

$$E(S_{g,N} - S^*_{g,N})^2/\text{Var}(S^*_{g,N}) \to 0 \quad \text{as} \quad N \to \infty. \quad \text{(A.1)}$$

After extensive algebra, and use of sampling theory for finite populations (Hansen et al., 1960), it can be shown that

$$\begin{aligned} &E(S_{g,N} - S^*_{g,N})^2/\text{Var}(S^*_{g,N}) \\ &\le \frac{\left(g^4_{\max}/4\right)}{(n_g/N_g)g\text{Var}(\theta_{ij})\{1 + (g-1)\rho_{\theta,g}\}\{\widehat{\text{Var}}(\theta_{i+,g})/\text{Var}(\theta_{i+,g})\}} \\ &\quad \times \left\{\frac{N}{(N-1)^2}\right\}\left(\frac{N_g + 1}{N_g}\right) \end{aligned}$$

where $\rho_{\theta,g} = \text{Corr}(\theta_{ij,g},\ \theta_{il,g})$, $j \neq l$ and $\widehat{\text{Var}}(\theta_{i+,g}) = \sum_{i=1}^{N_g} (\theta_{i+,g} - \overline{\theta}_{\cdot+,g})^2/(N_g - 1), \overline{\theta}_{\cdot+,g} = \sum_{i=1}^{N_g} \theta_{i+,g}/N_g$. Since $0 < \text{Var}(\theta_{ij}) < 1$, $0 < 1 + (g-1)\rho_{\theta,g} < g$, and as $N \to\infty$, n_g/N_g converges in probability to $1 - \xi_g$, where $0 < \xi_g < 1$, $\widehat{\text{Var}}(\theta_{i+,g})/\text{Var}(\theta_{i+,g})$ and $(N_g + 1)/N_g$ each converges in probability to 1, it follows that (A.1) is satisfied and the Lemma is proven; this implies that Z_c in (9) is asymptotically normal for balanced designs. For an unbalanced design, we consider the following theorem.

Theorem 1: *Suppose we have two samples X and Y consisting of m and n clusters, respectively. Let $X_{ij,g}$ = score for the jth subunit from the ith X cluster of size g, $i = 1, \ldots, m$, $j = 1, \ldots, g$; $g = 1, \ldots, g_{\max}$ and let $Y_{kl,g}$ be defined similarly.*

We wish to test the hypothesis

$$H_0: \Pr\{U(X_{ij,g} - Y_{kl,g}) = 1\} = \Pr\{U(X_{ij,g} - Y_{kl,g}) = 0\} \quad \text{for all } g = 1, \ldots, g_{\max} \text{ vs.}$$

$$H_1: \Pr\{U(X_{ij,g} - Y_{kl,g}) = 1\} \neq \Pr\{U(X_{ij,g} - Y_{kl,g}) = 0\} \quad \text{for at least one } g = 1, \ldots, g_{\max}.$$

Let W_c, $E(W_c)$ and $\mathrm{Var}(W_c)$ be defined as in (11), (12), and (13). If (a) $g_{\max}$ is finite and (b) for all cluster sizes g, $m_g/N_g \to \xi_g$, $0 < \xi_g < 1$, $g = 1, \ldots, g_{\max}$ as $N \to \infty$, then under H_0, $Z_c = \{W_c - E(W_c)\}/\{\mathrm{Var}(W_c)\}^{1/2}$ converges in law to a $N(0,\,1)$ distribution as $N \to \infty$.

Proof. In this proof, we denote the rank sum for the ith cluster of size g in a dataset with N clusters by $R_{i+,g,N}$. The overall Wilcoxon rank sum statistic is given by $W_{c,N} = \sum_{g=1}^{g_{\max}} S_{g,N}$. Let $D = \{g_1, \ldots, g_Q\}$ be the set of cluster sizes such that $N_{g_q} \to \infty$ as $N \to \infty$. Clearly, $\{S_{g_1,N}, \ldots, S_{g_Q,N}\}$ are independent, because the realizations of the randomization distributions for different cluster sizes are independently determined.

Furthermore, from the Lemma, $S_{g_q,N}$ is asymptotically normally distributed as $N \to \infty$, $q = 1, \ldots, Q$. Let $W_{c,N}^* = \sum_{q=1}^{Q} S_{g_q,N}, E(S_{g_q,N}) = \mu_{g_q,N}, \mathrm{Var}(S_{g_q,N}) = \sigma_{g_q,N}^2$, as given in (11)–(13). It follows that

$$E_N \equiv E(W_{c,N}^*) = \sum_{q=1}^{Q} \mu_{g_q,N}, \quad V_N \equiv \mathrm{Var}(W_{c,N}^*) = \sum_{q=1}^{Q} \sigma_{g_q,N}^2.$$

Therefore, based on moment-generating function methods, it can be shown that $Z_{c,N}^* = (W_{c,N}^* - E_N)/V_N^{1/2}$ converges in law to a $N(0,\,1)$ distribution as $N \to \infty$. We now consider the asymptotic distribution of $W_{c,N} = \sum_{g=1}^{g_{\max}} S_{g,N}$. We can write $W_{c,N} - W_{c,N}^* = \sum_{r=1}^{R} S_{h_r,N}$, where $\bar{D} = \{h_1, \ldots, h_R\}$ denote the set of indices in $\{1, \ldots, g_{\max}\}$ that are not included in D. Based on Lehmann (1975, Corollary 2, p. 349), to prove the asymptotic normality of $W_{c,N}$, it will be sufficient to show that $E(W_{c,N} - W_{c,N}^*)^2/\mathrm{Var}(W_{c,N}^*) \to 0$ as $N \to \infty$.

After extensive algebra, it can be shown that

$$E(W_{c,N} - W_{c,N}^*)^2/\mathrm{Var}(W_{c,N}^*) \leq \{N/(N-1)\}^2 A(h_1, \ldots, h_R) \Bigg/ \sum_{q=1}^{Q} N_{g_q} B(g_q) \quad \text{(A.2)}$$

where $A(h_1, \ldots, h_R)$ and $B(g_q)$, $q = 1, \ldots, Q$ are positive and bounded, and $N_{g_q} \to \infty$ as $N \to \infty$, $q = 1, \ldots, Q$. It follows from (A.2) that $E(W_{c,N} - W_{c,N}^*)^2/\mathrm{Var}(W_{c,N}^*) \to 0$ as $N \to \infty$. Thus, $W_{c,N}$ is asymptotically normal as $N \to \infty$, and Z_c in (14) converges in law to a $N(0,\,1)$ distribution, as $N \to \infty$ in the case of an unbalanced design. Finally, we state the following Theorem for the case of a stratified design.

THEOREM 2: *Suppose we have two samples X and Y consisting of m and n clusters, respectively. In addition, assume that the clusters are stratified into V strata according to* 1 *or more confounding variables. Let $X_{ij,g,v}$ = score for the jth subunit from the ith X cluster of size g, from the vth stratum, $i = 1, \ldots, m$; $j = 1, \ldots, g$; $g = 1, \ldots, g_{\max}$; $v = 1, \ldots, V$, and let $Y_{kl,g,v}$ be defined similarly. We wish to test the hypothesis:*

$$H_0: \Pr\{U(X_{ij,g,v} - Y_{kl,g,v}) = 1\} = \Pr\{U(X_{ij,g,v} - Y_{kl,g,v}) = 0\} \quad \textit{for all } g = 1, \ldots, g_{\max}; v = 1, \ldots, V \text{ vs.}$$

$$H_1: \Pr\{U(X_{ij,g,v} - Y_{kl,g,v}) = 1\} \neq \Pr\{U(X_{ij,g,v} - Y_{kl,g,v}) = 0\} \quad \textit{for at least one } (g, v), g = 1, \ldots, g_{\max}, v = 1, \ldots, V.$$

Let W_c, $E(W_c)$ and Var*(W_c) be defined as in* (16) *and* (17).

If (a) $g_{\max}$ is finite, (b) V is finite, and (c) for all $(g,\,v)$, $m_{g,v}/N_{g,v} \to \xi_{g,v}$, $0 < \xi_{g,v} < 1$ as $N \to \infty$, then under H_0, $Z_c = \{W_c - E(W_c)\}/\{\mathrm{Var}(W_c)\}^{1/2}$ converges by law to a $N(0,$ 1) distribution as $N \to \infty$.

Proof. The proof is identical to Theorem 1, where here $D = \{(g_1,\, v_1), \ldots, (g_Q,\, v_Q)\}$ is the set of (cluster size, stratum) pairs such that $N_{g_q,v_q} \to \infty$ as $N \to \infty$; hence, the details are omitted.