

# **A SAS Macro for the Wei-Lachin Multivariate Rank Analysis**

by

John M. Lachin

The Biostatistics Center

The George Washington University

August 18, 2006

Copyright (c) 2006 by John M. Lachin

## **1 Univariate Distribution-Free Mann-Whitney Wilcoxon Analysis.**

One of the most widely used analytic methods for the comparison of two groups of subjects is the Mann-Whitney Wilcoxon test. Wilcoxon (1945) proposed that a distribution-free (non-parametric) test of group differences be based on the difference in the sums of the ranks in the two groups, or either rank sum minus its expected value under the null hypothesis of no difference. This test can be used with data on any scale – nominal, ordinal or numerical – although its variance requires an adjustment for tied values.

In general, data analysis involves some estimation of the magnitude of a group difference (treatment effect) as well as a test of significance of that effect. For univariate non-parametric analyses, distribution free measures of the extent of group differences can be provided in addition to the non-parametric test. The univariate Wilcoxon test was derived from the perspective of a distribution-free estimate of group differences by Mann and

Whitney (1947) using what is often termed the Mann-Whitney Difference. The Wilcoxon rank sum test can either be expressed as a rank sum statistic in terms of observed minus expected rank sums, or it can be expressed as the Mann-Whitney Difference, the former giving only a test of group difference, the latter also providing an understandable measure of the magnitude of the group difference. The two forms of the test are algebraically equivalent.

The following description is based on Lachin (1992), with the exception that here we focus on the probability of an observation from group 1 being greater than group 2 whereas Lachin (1992) focused on the converse. This is done so that the notation herein is consistent with the description of the SAS macro.

Consider that patients are assigned to one of the two groups at random. Then the quantity  $p_> = P(X_1 > X_2)$  is the probability that a patient will have a higher value of the measurement when assigned to group 1 than when assigned to group 2. Likewise,  $p_< = P(X_1 < X_2)$  is the converse. Each is called a *proversion* probability. Each probability can be estimated by comparing the values of each of the  $n_1$  patients assigned to group 1 with those of the  $n_2$  patients assigned to group 2, and counting the number of times  $X_1 > X_2$ . Then

$$\hat{P}(X_1 > X_2) = \frac{\#(X_1 > X_2)}{n_1 n_2}. \quad (1)$$

In an excellent discussion, Wolfe and Hogg (1971) argue that the proversion probability is a much more useful measure of group differences than is the mean difference. When there are no ties, under the null hypothesis of no difference in the distributions in the two groups, then  $P(X_1 > X_2) = P(X_1 < X_2) = 0.5$ . In principle, one might consider only one of these two probabilities.

However, it is possible that there are tied values, i.e. a patient might have exactly

the same value if assigned to either group, i.e.  $X_2 = X_1$ , in which case the expected value of  $P(X_1 > X_2)$  is not  $1/2$  under the null hypothesis of no difference. To allow for this, The Mann-Whitney Difference (*MWD*) is obtained as the difference between these probabilities,  $\hat{\theta} = p_{>} - p_{<}$ . If there is no difference between the treatments, then these two probabilities are equal and the Mann-Whitney difference is zero under the null hypothesis. If the MWD is greater than 0, this indicates that there is a greater likelihood that a patient would have a higher value if assigned to treatment group 1 than group 2. The converse is true if the  $MWD < 0$ . If the  $MWD = 0.20$ , it means that  $P(X_1 > X_2)$  is 0.20 higher than is  $P(X_2 > X_1)$ , such as 0.60 vs. 0.40.

Since the  $\#(X_1 > X_2)$  is a function of the Wilcoxon rank sum statistic, then the variance of the Mann-Whitney difference can be obtained directly from the variance of the Wilcoxon rank sum statistic, and vice versa. Thus the  $Z$  or chi-square Wilcoxon test equals the Mann-Whitney test.

For example, a value  $p_{>} = 0.60$  means that there is a probability of 0.60 that a subject in population 1 has a value greater than that of a subject in population 2. When there are no tied values, it follows that the converse probability is  $p_{<} = 0.40$ , and the Mann-Whitney difference  $\theta = 0.20$ . Also, under the null hypothesis,  $H_0 : F_1 = F_2$ , there is an equal chance that a subject in one population has a value greater than that of a subject in the other population, i.e.  $E(\hat{p}_{>}|H_0) = 0.50$ . Thus,  $p_{>} - 0.5$  ( $= 0.10$  in this case) is the excess probability above chance that a subject in the first population has a value greater than that of a subject in the second population and,  $\theta$  is equal to twice this excess probability. If  $\theta$  is negative, the interpretation is the same but the populations are reversed; i.e. the second population has the larger values.

In a randomized study where each patient is assigned either treatment with equal

chance, then  $p_{>}$  is the probability that a given patient would have a higher value when assigned the first rather than the second treatment, and  $p_{<}$  is the converse. The Mann-Whitney difference then is the difference between the probabilities that a patient would have a higher value when assigned either treatment, or twice the excess probability beyond chance that a patient would have a higher value after receiving the first rather than the second treatment.

## 2 Multivariate Mann-Whitney Wilcoxon Analysis.

Wei and Lachin (1984) presented a multivariate rank test which provides multivariate generalizations of a wide variety of distribution-free methods for the analysis of many different types of data. These include multi-state event-time analyses and repeated measures analyses of nominal, ordinal or quantitative measures. Because the method applies to multivariate censored event-times, and because random missing-ship can be expressed as a special case of random right censoring, the method also provides for the analysis of multivariate partially incomplete observations where some of the measurements may be missing at random. Wei and Lachin (1984) showed that the large-sample distribution of this family of multivariate rank statistics is multivariate normal with mean zero, and they presented a consistent estimator of the covariance matrix of these statistics under the assumption of randomly censored (missing) observations. This provides a variety of multivariate distribution-free tests for global (omnibus) and restricted alternatives.

The Wilcoxon rank statistic is a special case of the family of statistics covered by the Wei-Lachin multivariate rank test. In fact it can be shown that the Wei-Lachin test statistic with Wilcoxon scores can be expressed as a constant (depending on sample sizes) times  $[\#(X_1 \geq X_2) - \#(X_2 \geq X_1)]$ . In an application of the Wei-Lachin test,

Thall and Lachin (1988) showed that the Wei-Lachin Wilcoxon rank statistic readily provides an estimate of the Mann-Whitney Difference for each of the multiple measures included in the multivariate analysis. The covariance matrix of these estimates can also be obtained directly from the Wei-Lachin covariance matrix of the rank statistics. No additional adjustment for ties is needed since the Wei-Lachin variances and covariances automatically allow for ties.

Thall and Lachin (1988) also used these measure-specific Mann-Whitney Differences to obtain an overall estimate of the magnitude of the treatment group difference for all measures combined. If there are  $K$  measures, let  $\hat{\theta}_1 \dots \hat{\theta}_K$  denote the estimated Mann-Whitney Difference for each measure. Then a weighted average can be constructed such as  $\hat{\theta}_* = w_1\hat{\theta}_1 + w_2\hat{\theta}_2 + \dots + w_K\hat{\theta}_K$  using a set of weights that sum to unity. Thall and Lachin used weights inversely proportional to the row sums of the inverse covariance matrix, equivalent to using weighted least squares, to obtain the best average estimate of the overall group difference for all  $K$  measures combined. The variance of this average  $\hat{\theta}_*$  can also be obtained from the Wei-Lachin covariance matrix of the  $K$  Wilcoxon statistics, which yields an overall test of aggregate association. Lachin (1992) provides a detailed review of the Wei-Lachin and Thall-Lachin multivariate Mann-Whitney analysis.

It is also common to employ some form of an adjusted analysis which accounts for the effects of other covariates. Such analyses may be performed either to adjust for a baseline imbalance on a covariate, or a stratified analysis may be performed in order to obtain a  $p$ -value that more accurately describes the differences between groups under the design employed. In general, if a stratified randomization is employed, a like-stratified analysis provides a more appropriate  $p$ -value, especially with a small study. Lachin (1992) also describes a stratified multivariate Mann-Whitney analysis.

### 3 Technical Details

Let  $X' = (X_{ij1} \dots X_{ijk})$  be a vector of  $K$  measurements ( $1 \leq k \leq K$ ) for the  $j$ -th subject in a sample of size  $n_i$  from the  $i$ -th of two populations ( $1 \leq j \leq n_i$ ;  $i = 1, 2$ ), where some of the  $X_{ijk}$  may be missing completely at random. In the  $i$ -th population, let  $F_i(x)$  be the  $K$ -variate joint distribution function, with marginal distribution functions  $F_{ik}(x)$ , ( $1 \leq k \leq K$ ). The  $\{X_{ijk}\}$  may have any covariance structure, provided that it is of full rank.

In the multivariate setting, Wei and Lachin (1984) present a  $K$ -variate generalization of the Wilcoxon test. They use the family of weighted Mantel-Haenszel-like rank statistics which includes the logrank and Wilcoxon test for survival data as special cases. Wei and Lachin show that this family of tests, originally derived for application to survival data under a random censoring model, could be applied to data on a nominal, ordinal or numerical scale under a missing at random model, this being a special case of random censoring. For the  $k$ th measure, let  $T_k$  designate the resulting rank statistic value, and  $\hat{\Sigma}_T$  the consistent estimate of the covariance matrix of the vector of statistics  $T = (T_1 \dots T_K)$ . They then show that

$$T \sim \mathcal{N}_k[0, \Sigma_T] \quad (2)$$

under the null hypothesis of no difference between groups for any of the  $K$  measures.

Thall and Lachin (1988), in an application of the Wei-Lachin method, show that a scale transformation of Wei-Lachin Wilcoxon test statistics yields estimates of the Mann-Whitney difference,  $\hat{\theta}_k$  ( $1 \leq k \leq K$ ) that are asymptotically normally distributed. For the  $k$ th repeated measure, the proversion probability  $Pr(X_{1k} > X_{2k})$  is denoted as  $p_{>k}$ . To allow for tied values, the Mann-Whitney Difference, is defined herein as the difference

between the two proversion probabilities

$$\theta_k = [Pr(X_{1k} > X_{2k}) - Pr(X_{1k} < X_{2k})] = p_{>k} - p_{<k} \quad (3)$$

for  $1 \leq k \leq K$ . Under  $H_0$ , since

$$E(\hat{p}_{>k} | H_0) = \frac{1}{2} Pr(X_{1k} \neq X_{2k}) = \frac{1}{2}(p_{>k} + p_{<k}) \quad (4)$$

this quantity is estimated by

$$\hat{\theta}_k = \hat{p}_{>k} - \hat{p}_{<k} = 2[\hat{p}_{>k} - \hat{E}(\hat{p}_{>k} | H_0)]. \quad (5)$$

This estimator can be used to summarize the difference between groups for data on any scale of measurement. For binary data, it follows that  $\hat{\theta}$  is simply the difference between the proportions "positive" in the two groups, say  $\hat{p}_1 - \hat{p}_2$ .

The Mann-Whitney Difference  $\theta_k$  measures the extent to which the marginal distributions of the two populations differ for the  $k$ th measure. Under the assumption that  $F_{1k}$  and  $F_{2k}$  have the same shape but differ in location.,  $\theta_k$  satisfies

$$(\theta_k = 0) \Leftrightarrow [F_{1k}(x) = F_{2k}(x)] \quad (6)$$

and  $\theta = 0$ , for  $\theta' = (\theta_1 \dots \theta_k)$ , implies the null hypothesis that the  $K$ -variate marginal distributions are jointly equal; *i.e.*,  $H_0 : [F_{1k}(x) = F_{2k}(x)]$  for  $1 \leq k \leq K$ . The parameter  $\theta_k$  is also defined so that

$$(\theta_k > 0) \Leftrightarrow [F_{1k}(x) \succ^{right} F_{2k}(x)] \quad (7)$$

where " $\succ^{right}$ " indicates stochastic ordering; *i.e.*, the values from the first population are larger than those from the second, or  $F_{1k}(x) < F_{2k}(x)$  for all  $x \in \mathbf{R}$ .

The vector of sample estimates  $\hat{\theta}' = (\hat{\theta}_1 \dots \hat{\theta}_k)$ , based on a sample of  $N = n_1 + n_2$  subjects, and the corresponding estimate of the covariance matrix,  $\hat{\Sigma}_\theta$ , are obtained from

the vector of Wilcoxon statistics as

$$\begin{aligned}\hat{\theta} &= D'T, \quad D = \text{diag}(N^{3/2}/n_{1k}n_{2k}) \\ \hat{\Sigma}_{\theta} &= D'\hat{\Sigma}_T D\end{aligned}\tag{8}$$

and the vector estimates are asymptotically normally distributed as

$$(\hat{\theta} - \theta) \sim \mathcal{N}_k[0, \Sigma_{\theta}]\tag{9}$$

where the large-sample estimated covariance matrix  $\hat{\Sigma}_{\theta}$  is based on the estimate  $\hat{\Sigma}_T$ .

The following tests are applicable.

### 3.1 Omnibus Test.

The traditional  $T^2$ -like multivariate test is a test of the null hypothesis of the joint equality of the marginal distributions,  $H_0 : F_{1k}(x) = F_{2k}(x)$  for  $1 \leq k \leq K$ . Under  $H_0$ , from basic principles it follows that

$$X_O^2 = \hat{\theta}' \hat{\Sigma}_{\theta}^{-1} \hat{\theta} = (T'D)(D\hat{\Sigma}_T D)^{-1}(DT) = T\hat{\Sigma}_T^{-1}T\tag{10}$$

is asymptotically distributed as chi-square on  $K$  *df*. Note that the test is the same whether using the Wilcoxon statistics or the Mann-Whitney Difference estimates. This test is called the omnibus or global test because it is directed towards the omnibus alternative  $H_{1O} : F_{1k}(x) \neq F_{2k}(x)$  for some  $k$ ; that is, towards a difference between the populations of any magnitude in any direction for any of the  $K$  measures.

In some cases, it may not be important to perform a test designed to detect any type of difference between groups for any of the  $K$  measures. Rather, it may be important to detect population differences which follow some specified pattern. For example, rather than use the omnibus test which is directed towards all points  $\theta$  in the parameter space



$\mathbf{R}^k$  away from the origin, it may be desired to use a test directed towards a restricted alternative hypothesis represented by a sub-region of  $\mathbf{R}^k$ . In this case, the omnibus test will not be as powerful as a test of such a restricted alternative, when that alternative is true.

Lachin (1992), see also Lachin (2000), describes some useful restricted alternatives and tests which are especially useful when the  $K$  measures are inter-related. The most common case is where each of the  $K$  variates is a repeated measure (replicate) obtained from each subject under different conditions or at different times. Another case is where each variate is a different measure of the same phenomenon, such as where  $K$  different biochemical measures of liver function are assessed simultaneously.

### 3.2 Stochastic Ordering.

One useful test is a multivariate generalization of the ordinary one-sided univariate test, termed the test of stochastic ordering. Here the test of  $H_0$  is directed towards the restricted ordered alternative

$$H_{1S} : F_{1k}(x) \succ^{right} F_{2k}(x) \quad (11)$$

for  $1 \leq k \leq K$  and all  $x \in \mathbf{R}$ . This alternative  $H_{1S}$  specifies that the values in the first population tend to be greater than (or at least as great as) those in the second for all of the  $K$  measures. For a vector of rank statistics, Wei and Lachin (1984) suggested the following simple test statistic

$$Z = \frac{J'T}{[J'\hat{\Sigma}_T J]^{1/2}} \quad (12)$$

where  $J' = (1 \dots 1)$ , and where  $Z$  is asymptotically normally distributed under  $H_0$ . Note that the test is based on the simple sum (or unweighted mean) of the  $\{T_k\}$ .

Alternately, a test can be based on the sum of the distribution-free Mann-Whitney

Difference estimates as

$$Z = \frac{J'\hat{\theta}}{[J'\hat{\Sigma}_{\theta}J]^{1/2}}. \quad (13)$$

This test would lead to rejection of  $H_0$  for  $Z \geq z_{1-\alpha}$  at level  $\alpha$ . For a multivariate alternative in the opposite direction,  $H_{1S} : F_2 \succ^{right} F_1$ , the test would reject for  $Z \leq z_{\alpha}$ . For a two-sided test where it is of interest to claim that one population dominates the other, but where the direction is not of interest, the test would reject for  $|Z| \geq z_{1-\alpha/2}$ .

The two in general differ slightly. Frick (1995) shows that this is an optimal robust test (see also Lachin, 2000).

An N-weighted test, unpublished, has also been proposed for the case where the sample sizes for the repeated measures vary substantially, such as in a longitudinal study with declining sample sizes due to administrative curtailment of follow-up due to staggered patient entry. Rather than using the unit vector  $J$  in (13), this test employs weights equal to the total sample size ( $N_k$ ) at each time,  $N' = (N_1 \dots N_K)$  to yield the test statistic

$$Z = \frac{N'\hat{\theta}}{[N'\hat{\Sigma}_{\theta}N]^{1/2}}. \quad (14)$$

This test has not been published or explored in the literature. It is a special case of the test based on (13) in Lachin (1992). We have used this test for analyses of repeated measures when the sample sizes decline markedly over time due to staggered entry.

### 3.3 Test For Homogeneity

When all of the  $K$  variates are related measures of the same phenomenon on the same scale, such as a set of repeated measurements, it may be useful to address the test towards the average difference between populations in some sense. In this approach, the null hypothesis can be stated as  $H_0 : \theta_k = \theta_* = 0$ , for  $1 \leq k \leq K$ , where  $\theta_*$  is the average value of  $\theta_k$  for all of the  $K$  measures. This hypothesis can be expressed as the intersection of two null

hypotheses: that for homogeneity on  $(K - 1)$   $df$  and that for association on 1  $df$ . The test of homogeneity addresses the null  $H_{0H} : \theta_1 = \dots = \theta_K = \theta_*$  against the alternative  $H_{1H} : \theta_k \neq \theta_\ell$  for some  $k \neq \ell$ . In terms of a simple ANOVA for repeated measures, the test of homogeneity corresponds to the groups by time interaction effect. This test, and others, can be obtained through the application of basic principles.

The test of homogeneity is a special case of a test of a contrast hypothesis  $H_{0C} : C'\theta = 0$ , where  $C'$  is an  $r \times K$  matrix,  $r \leq K$ , each row of which is a contrast on the elements of  $\theta$ . A test of  $H_{0C}$  is provided by

$$X_C^2 = \hat{\theta}' C (C' \hat{\Sigma}_\theta C)^{-1} C' \hat{\theta} \quad (15)$$

which is asymptotically distributed as chi-square on  $r$   $df$ . For a test of homogeneity,  $C$  consists of  $r = K - 1$  contrasts of successive differences among the  $\{\theta_k\}$ , or alternately consists of contrasts of each  $\theta_k$  ( $k > 1$ ) versus  $\theta_1$ . These two (and other such) formulations of  $C$  yield an equivalent test result.

### 3.4 Test For Association

The test of association addresses the null  $H_{0A} : \theta_1 = \dots = \theta_K = \theta_* = 0$  against the alternative  $H_{1A} : \theta_1 = \dots = \theta_K = \theta_* \neq 0$  and corresponds to the overall group effect. The test is obtained as a special case of a generalized least squares (GLS) test based on GLS estimates of underlying parameters. The test is obtained from the application of the theorem to the simplest possible case where  $X = J$  (the ones-vector), and  $\beta = \theta_*$  (a scalar) is the expectation of the  $\{\hat{\theta}_k\}$ . In this case, the aggregate GLS estimate,  $\hat{\theta}_*$ , of the average difference between groups combined over the  $K$  measures is simply a linear

combination of the  $\{\hat{\theta}_k\}$ , and its variance, are expressed as

$$\begin{aligned}\hat{\theta}_* &= \hat{W}'\hat{\theta} \\ \hat{\sigma}^2(\theta_*) &= \hat{W}'\hat{\Sigma}\hat{W} = (J'\hat{\Sigma}^{-1}J)^{-1}\end{aligned}\tag{16}$$

where

$$\hat{W}' = (J'\hat{\Sigma}_\theta^{-1}J)^{-1}J'\hat{\Sigma}_\theta^{-1}\tag{17}$$

and  $\hat{W}'J = 1$ ; *i.e.* the weights sum to unity. By definition, this linear combination provides a minimum variance linear estimator, and thus yields the test of association

$$X_a^2 = \hat{\theta}_*^2 / \hat{\sigma}^2(\theta_*) = \hat{W}'\hat{\theta}(\hat{W}'\hat{\Sigma}_\theta\hat{W})^{-1}\hat{\theta}'\hat{W}\tag{18}$$

which is asymptotically distributed as chi-square on 1 df.

This test is equivalent to an overall test of the "group" effect in a repeated measures analysis as in PROC GLM, MIXED or GLM. All are based on the average group difference over time as an estimate of an assumed common difference over time.

### 3.5 Comparison of Weighted Tests.

Each of the above tests will have greater power than the others in specific instances. When the  $\{\theta_k\}$  are homogeneous ( $= \theta_*$  for all  $k$ ), or nearly so, the test of association will be more powerful than either the omnibus test or the test of stochastic ordering, but not necessarily otherwise. Both the omnibus and stochastic ordering tests can also be expressed as a function of a linear estimator as in (18). The test of stochastic ordering uses equal weights ( $= 1/K$ ), which do not provide an efficient estimator of  $\theta_*$ . On the other hand, the omnibus test  $X_O^2$  on  $K$  df can also be expressed as a linear combination of the vector estimates as in (18) using weights  $\hat{W}'_O = (\hat{\theta}'\hat{\Sigma}_\theta^{-1}J)^{-1}\hat{\theta}'\hat{\Sigma}_\theta^{-1}$  which provide a

maxima for the value for the quadratic form (18). When  $\hat{\theta}_k = \hat{\theta}_*$  for all  $k$ , then  $\hat{W}_O = \hat{W}$  in (18), but not so otherwise. Thus, in specific situations where there is some heterogeneity among the  $\{\theta_k\}$ , then it is possible that a test using other weights, such as the test of stochastic ordering in (13), may yield a larger test value.

In practice, therefore, there is a tradeoff between the power robustness of the omnibus test to detect group differences under the broadest possible range of alternatives, versus the sensitivity (increased power) of the other tests to detect systematic differences between the groups under specific alternatives. For  $K = 2$  measures, the omnibus test is designed to detect any group difference  $(\theta_1, \theta_2)$  of points in the  $\mathbf{R}^2$  region away from the origin. The test of stochastic ordering for positive differences is directed to points in the positive quadrant only. The test of association is addressed to points along the line  $\theta_1 = \theta_2$ . Thus, as one compares the omnibus test, the test of stochastic ordering, and the test of association, in turn, there is decreasing robustness to a range of alternative hypotheses, but increasing power to detect specific restricted alternatives. For this reason, one might wish to perform an initial test of homogeneity prior to performing a test of association. However, with very large sample sizes, it is possible that the test of homogeneity may lead to significance, suggesting that the estimates should not be combined, when in fact the differences between the estimates are negligible.

All of these tests control the type I error probability at the desired level under the null hypothesis. The difference is that each is more efficient (powerful) under specific alternatives.

Lachin (2000), Sections 4.5 - 4.9, also provides a comparison of these various tests and displays the null and alternative hypothesis spaces, and the test rejection regions, for the case of  $K = 2$  measures.

## 4 Computational Procedures

Wei and Lachin (1984) performed an extensive simulation of the Wei-Lachin multivariate procedure for logrank and Wilcoxon scores using a program they had prepared in the language *PL/I*. This simulation showed that these tests had the proper size, even with fairly small sample sizes and moderate amounts of censored (missing) observations. Subsequently, Makuch, Escobar and Merrill (1991) used the Wei-Lachin procedure for the analysis of a set of data. To do so they independently prepared a highly efficient Fortran subroutine to perform these calculations. The subroutine was later tested against the original Wei-Lachin *PL/I* program and the two performed equivalently. This Fortran program was subsequently published by Makuch, Escobar and Merrill (1991) in *Applied Statistics* as Algorithm AS 262.

The original Makuch, Escobar and Merrill program only performed the Wei-Lachin analysis. Ms Joni Evans incorporated this algorithm into a SAS Macro that also computes the Thall-Lachin multivariate Mann-Whitney analysis.

The Thall-Lachin (Wei-Lachin) variance is one of many possible approaches which yield a consistent estimate of the variance of the Mann-Whitney Difference. Recently, an independent software house, idv of Munich, Germany, has incorporated this variance estimate of the Mann-Whitney Difference into their general program of non-parametric procedures TESTIMATE 5.1.1 (idv, 1993). As part of their software development they performed comparisons of the Thall-Lachin variance against other available methods and determined that the Thall-Lachin variance yields superior confidence limits versus other methods.

## REFERENCES

Frick, H. (1995). Comparing trials with multiple outcomes: The multivariate one-sided

hypothesis with unknown covariances. *Biom. J.*, 8, 909-917.

idv — Datenanalyse und Versuchsplanung (1993). TESTIMATE Handbook: Supplement to Handbook 5.1, TESTIMATE Version 5.1.1, idv, Munich, Germany.

Lachin, J.M. (1992). Some large-sample distribution-free estimators and tests for multivariate partially incomplete data from two populations. *Statistics in Medicine*, 11:1151-1170.

Lachin, J. M. (2000). *Biostatistical Methods. The Assessment of Relative Risk*. New York: John Wiley and Sons, Inc.

Makuch, R.W., Escobar, M. and Merrill S. (1991). A two-sample test for incomplete multivariate data. *Applied Statistics*, 40:202-212.

Mann, H.B. and Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statistics*, 18:50-60.

Thall, P.F. and Lachin, J.M. (1988). Analysis of recurrent events: nonparametric methods for random interval count data. *Journal of the American Statistical Association*, 83:339-347.

Wei, L.J. and Lachin, J.M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association*, 79:653-661.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1:80-3.

Wolfe, D.A. and Hogg, R.V. (1971). On constructing statistics and reporting data. *American Statistician*, 25, 27-30.