An optimal Wilcoxon–Mann–Whitney test of mortality and a continuous outcome

Journal Title
XX(X):1–29

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Roland A. Matsouaka[1,2], Aneesh B. Singhal[3] and Rebecca A. Betensky[4]

Abstract

We consider a two-group randomized clinical trial, where mortality affects the assessment of a follow-up continuous outcome. Using the worst-rank composite endpoint, we develop a weighted Wilcoxon–Mann–Whitney (WMW) test statistic to analyze the data. We determine the optimal weights for the WMW test statistic that maximize its power. We derive a formula for its power and demonstrate its accuracy in simulations. Finally, we apply the method to data from an acute ischemic stroke clinical trial of normobaric oxygen therapy.

Keywords

Missing data, survivor bias, multiple endpoints, weighted Wilcoxon-Mann-Whitney test, censored-by-death, composite endpoints.

1 Introduction

In many randomized clinical trials, the difference between treatment groups is evaluated using measurements of an outcome of interest after a pre-specified follow-up time. However, for some participants, follow-up measurements may be missing if a disease-related event, such as

[1]Department of Biostatistics and Bioinformatics, Duke University, Durham, USA.
[2]Duke Clinical Research Institute, Duke University, Durham, USA.
[3]Department of Neurology, Massachusetts General Hospital, Boston, USA.
[4]Department of Biostatistics, Harvard School of Public Health, Boston, USA.

Corresponding author:
Roland A. Matsouaka, Duke Clinical Research Institute, 2400 Pratt Street, Durham, NC 27705, USA.
Email: roland.matsouaka@duke.edu

Prepared using sagej.cls *[Version: 2015/06/09 v1.01]*

death (or withdrawal due to worsening disease condition), has occurred prior to the end of follow-up time.Our motivating example is a clinical trial of acute ischemic stroke conducted at Massachusetts General Hospital in Boston, MA. In this trial, patients who had acute ischemic stroke were randomized to either normobaric oxygen (NBO) therapy or room air and assessed serially to monitor their functional ability. Among other measures, patients' neurological recovery was assessed and quantified using the NIH Stroke Scale (NIHSS) score, a function rating scale used to quantify neurological deficit due to stroke.[1;2] However, investigators were confronted with early deaths, which precluded measurements of NIHSS scores for some participants at the end of the 3-month follow-up period. Any analysis of the data that includes solely the subjects who survived would be biased and give spurious results.[3]

One approach to handle this issue is to combine the primary endpoint and mortality into a single composite endpoint: the *worst-rank* composite endpoint. It is calculated by considering death as the worst outcome on the same scale as the measured outcome and analyzed using the ranks of these combined outcomes.[4–6] Unlike traditional analyses of composite endpoints that treat all of the component endpoints equally and focus on each study participant's first occurring event, worst-rank composite endpoints incorporate a hierarchical ranking of these individual outcomes based on their clinical importance, frequency of occurrence or severity. Moreover, in contrast to the typical "time-to-event" analyses, worst-rank composite endpoints allow us to combine individual outcomes from multiple clinical domains, while accounting for their heterogeneity. Such outcomes could include both clinical events (e.g., death), continuous continuous variables, or other clinical measurements (e.g., biomarker or quality-of-life measures.[7])

Ranking individual outcomes that characterize various aspects of patients' disease experience based on a pre-specified hierarchy of various components suggest the existence of an implicit weighting scheme. In fact, several authors have suggested the use of *a priori* determined utility (or sometimes severity) weights to reflect the relative importance of the components of composite outcomes and add another layer of discrimination beyond hierarchical ordering alone.[8;9] Such weighting may be based on subjective criteria or elicitation of experts. However, deriving such *a priori* weights and finding a consensus about them have proven to be difficult.[10–14]
Building upon our previous work on this topic[6], and assuming there is a pre-specified hierarchy of various components of a composite outcome, we introduce an optimal approach that not only acknowledges such a hierarchy, but also estimates the weights so as to maximize the power to detect globally any treatment effect when present.

The use of multivariate tests to compare treatment effects from multivariate outcomes has gain interest in clinical trials of multifaceted complex diseases where the clinical course of the

disease is manifested in complex ways through a host of clinical outcomes. A global test statistic for composite endpoints that accounts for the complexity of the disease, rather than evaluating individual components, provides a comprehensive method to evaluate more effectively and more efficiently the efficacy of a treatment [15;16]. Tests such as O'Brien test [17], Wei and Johnson's test [18], Finkelstein and Schoenfeld's test [19], Moyé et al.'s test [20;21] are rank-based tests developed using U-statistics. Some of these tests of combined endpoints are weighted tests where the optimal weights are determined by maximizing the power of the test statistic under a particular alternative hypothesis: this is the framework we will focus on in this paper.

In this paper, we use the given hierarchy of outcomes to construct a *worst-rank* composite endpoint such that death (or a missing continuous outcome due to worsening of the disease condition) is considered a worse outcome than any observed primary endpoint measurement. Furthermore, two subjects who died are ranked with respect to their survival times. [4–6] In Section 2, we give the rationale for the weighted WMW test statistic for such a worst-rank composite endpoint. We then derive data-based optimal weights that maximize the power of the weighted WMW test statistic along with its analytical power formula. We demonstrate that the *optimal* weighted WMW test statistic has greater power than the ordinary WMW test statistic. We illustrate the accuracy of our results through simulation studies (Section 3). Finally, we apply the procedures to the clinical trial of the normobaric oxygen (NBO) therapy for acute ischemic stroke patients.

2 Weighted Wilcoxon–Mann–Whitney test

2.1 Notations

In this section, we present the ordinary WMW test for the worst-rank composite outcome and its analytical power formula that we previously derived. [6] Then, we motivate its extension to a weighted WMW test through a decomposition of the WMW U-statistic.

Consider a randomized clinical trial in which m and n subjects are assigned respectively to the control treatment (group 1) and the active treatment (group 2) and then followed for time period T. For subject j in group i, X_{ij} denotes the value of the continuous endpoint at the end of the follow-up time, t_{ij} denotes the time to death or disease-related withdrawal (for simplicity, we will refer to both as death), $\delta_{ij} = I(t_{ij} \leq T)$ indicates early death (i.e., before T), and $p_i = E(\delta_{ij}) = P(t_{ij} \leq T)$ the probability of early death for subjects in group i.

If the subject died before T, X is unknown. Thus, following the assumed hierarchy of outcomes, this subject is assigned a *worst-rank* score equal to $\eta + t_{ij}$, which is a function of his or her survival time, where $\eta = min(X) - 1 - T$.

Without loss of generality, we assume larger values of X correspond to better health outcome. For each subject, the worst-rank composite endpoint is thus

$$\widetilde{X}_{ij} = (1 - \delta_{ij})X_{ij} + \delta_{ij}(\eta + t_{ij}),\ i = 1, 2 \text{ and } j = 1, \ldots, N. \tag{1}$$

Let F_i and G_i be, respectively, the cumulative conditional distributions of the informative event times and observed non-fatal outcome for patients in group i, i.e. $F_i(v) = P(t_{ij} \leq v | 0 < t_{ij} \leq T)$ and $G_i(x) = P(X_{ij} \leq x | t_{ij} > T)$. The distribution of $\widetilde{X}_i$ is given by

$$\widetilde{G}_i(x) \quad = \quad p_i F_i(x - \eta) I(x < \zeta) + (1 - p_i) G_i(x) I(x \geq \zeta), \quad \zeta = min(X) - 1. \tag{2}$$

We would like to test the null hypothesis that the two treatment are equivalent with respect to both survival and the non-fatal outcome

$$H_0 : G_1(x) = G_2(x) \text{ and } F_1(t) = F_2(t) \text{ for all } x \text{ and } t \tag{3}$$

against the uni-directional alternative hypothesis that the active treatment is at least as effective as the control treatment for both mortality and the non-fatal outcome and is not harmful for either, i.e.,

$$H_1 : G_1(x) \geq G_2(x) \text{ and } F_1(t) \geq F_2(t), \text{ for some } x \text{ and/or } t \tag{4}$$

with both $G_1(x) = G_2(x)$ and $F_1(t) = F_2(t)$ not occurring simultaneously for all x and t.

2.2 Ordinary Wilcoxon–Mann–Whitney test

We will now define the ordinary WMW test using the framework of the worst-rank composite endpoint $\widetilde{X}$ of the previous section. The ordinary Wilcoxon–Mann–Whitney (WMW) U-statistic is defined by

$$U = (mn)^{-1} \sum_{k=1}^{m} \sum_{l=1}^{n} I(\widetilde{X}_{1k} < \widetilde{X}_{2l}). \tag{5}$$

Using (1), we note that $I(\widetilde{X}_{1k} < \widetilde{X}_{2l})$ is equal to

$$\delta_{1k}\delta_{2l} I(t_{1k} < t_{2l}) + \delta_{1k}(1 - \delta_{2l}) + (1 - \delta_{1k})(1 - \delta_{2l}) I(X_{1k} < X_{2l}). \tag{6}$$

Therefore,

$$\begin{aligned} \mu_1 &= E(U) = \pi_{U1}, \\ \sigma_1^2 &= Var(U) = (mn)^{-1}\left[\pi_{U1}\left(1-\pi_{U1}\right) + (m-1)(\pi_{U2}-\pi_{U1}^2) + (n-1)(\pi_{U3}-\pi_{U1}^2)\right] \end{aligned} \tag{7}$$

where

$$\begin{aligned} q_i &= 1-p_i, \quad \pi_{t1} = P(t_{1k} < t_{2l} | t_{1k} \leq T,\ t_{2l} \leq T), \\ \pi_{t2} &= P(t_{1k} < t_{2l}, t_{1k'} < t_{2l} | t_{1k} \leq T,\ t_{1k'} \leq T,\ t_{2l} \leq T), \\ \pi_{t3} &= P(t_{1k} < t_{2l}, t_{1k} < t_{2l'} | t_{1k} \leq T,\ t_{2l} \leq T,\ t_{2l'} \leq T), \\ \pi_{x1} &= P(X_{1k} < X_{2l}),\ \pi_{x2} = P(X_{1k} < X_{2l}, X_{1k'} < X_{2l}), \\ \pi_{x3} &= P(X_{1k} < X_{2l}, X_{1k} < X_{2l'}) \\ \pi_{U1} &= p_1 p_2 \pi_{t1} + p_1 q_2 + q_1 q_2 \pi_{x1} \\ \pi_{U2} &= p_1^2 q_2 + p_1^2 p_2 \pi_{t2} + 2 p_1 q_1 q_2 \pi_{x1} + q_1^2 q_2 \pi_{x2} \\ \pi_{U3} &= p_1 q_2^2 + p_1 p_2^2 \pi_{t3} + 2 p_1 p_2 q_2 \pi_{t1} + q_1 q_2^2 \pi_{x3} \end{aligned}$$

(see the proof in Appendix A).

Under the null hypothesis (H_0) of no difference between the two treatment groups, $\mu_0 = E_0(U) = 1/2$ and $\sigma_0^2 = Var_0(U) = (n+m+1)/(12mn)$. The distribution of the ordinary WMW test statistic

$$Z = \frac{U - E_0(U)}{\sqrt{Var_0(U)}} \tag{8}$$

converges to the standard normal distribution $N(0,1)$ as m and n tend to infinity, and $m/n \longrightarrow \rho$, $0 < \rho < 1$.

The power of this WMW test is given by

$$\Phi\left(\frac{\sigma_0}{\sigma_1} z_{\frac{\alpha}{2}} + \frac{\mu_1 - \mu_0}{\sigma_1}\right) + \Phi\left(\frac{\sigma_0}{\sigma_1} z_{\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\sigma_1}\right) \approx \Phi\left(\frac{\sigma_0}{\sigma_1} z_{\frac{\alpha}{2}} + \frac{|\mu_1 - \mu_0|}{\sigma_1}\right) \tag{9}$$

where $\mu_1 = E(U)$ and $\sigma_1^2 = Var(U)$ under the alternative hypothesis (H_1) (see the proof in[6]).

2.3 Weighted Wilcoxon–Mann–Whitney test

To motivate our weighted test, we now write the WMW U-statistic applied to the worst-rank scores (5) as a sum of three dependent WMW U-statistics. Then, we demonstrate that to optimally compare two treatment groups using worst-rank scores, we need to use a weighted statistic that takes into account the dependence that exists among the three statistics.

Assume there exists weights $\mathbf{w} = (w_1, w_2)$, $w_1 + w_2 = 1$, such that (1) becomes

$$\widetilde{X}_{ij} = w_1\delta_{ij}(\eta + t_{ij}) + w_2(1-\delta_{ij})X_{ij},\ i = 1,2 \text{ and } j = 1,\ldots,N. \tag{10}$$

The U-statistic (5) then becomes $U_w = w_1^2U_t + w_1w_2U_{tx} + w_2^2U_x$, where U_t, U_{tx} and U_x are defined by

$$\begin{aligned} U_t &= (mn)^{-1}\sum_{k=1}^{m}\sum_{l=1}^{n}\delta_{1k}\delta_{2l}I(t_{1k} < t_{2l}) \\ U_{tx} &= (mn)^{-1}\sum_{k=1}^{m}\sum_{l=1}^{n}\delta_{1k}(1-\delta_{2l}) \\ U_x &= (mn)^{-1}\sum_{k=1}^{m}\sum_{l=1}^{n}(1-\delta_{1k})(1-\delta_{2l})I(X_{1k} < X_{2l}). \end{aligned} \tag{11}$$

Using vector notation, we can write U_w as $U_w = \mathbf{c}'\mathbf{U}$ where we define $\mathbf{U}' = (U_t, U_{tx}, U_x)$ and $\mathbf{c}' = (c_1, c_2, c_3) = (w_1^2, w_1w_2, w_2^2)$. Notice that $c_1 + 2c_2 + c_3 = (w_1 + w_2)^2 = 1$.

Using the results in Appendix B, we have

$$\begin{aligned} \mu_{1w} &= E(U_w) = \mathbf{c}'\left(p_1p_2\pi_{t1}, p_1q_2, q_1q_2\pi_{x1}\right)' \\ \sigma_{1w} &= Var(U_w) = \mathbf{c}'\Sigma\mathbf{c}, \end{aligned}$$

where $\Sigma = Var(\mathbf{U})$ is a 3×3 matrix given in the Appendix B.

Under the null hypothesis,

$$\begin{aligned} \mu_{0w} &= E_0(U_w) = \frac{1}{2}\mathbf{c}'\left(p^2, 2pq, q^2\right)' \\ &= \frac{1}{2}\left[w_1^2p^2 + 2w_1w_2pq + w_2^2q^2\right] = \frac{1}{2}\left[w_1p + w_2q\right]^2 \\ \sigma_{0w} &= Var_0(U_w) = \mathbf{c}'\Sigma_0\mathbf{c}, \end{aligned}$$

with $\Sigma_0 = Var_0(\mathbf{U})$ a 3×3 matrix given in the Appendix B.

2.3.1 Pre-specified weights When there are pre-specified weights, usually determined as to reflect the relative importance or the severity of component outcomes, they can be used to calculate the weighted WMW test statistic

$$Z_w = \frac{U_w - E_0(U_w)}{\sqrt{Var_0(U_w)}}. \tag{12}$$

Z_w converges to the standard normal distribution $N(0,1)$ as m and n tend to infinity, and $m/n \longrightarrow \rho$, $0 < \rho < 1$.

The correspondong power is given by

$$\Phi\left(\frac{\sigma_{0w}}{\sigma_{1w}} z_{\frac{\alpha}{2}} + \frac{\mu_{1w} - \mu_{0w}}{\sigma_{1w}}\right) + \Phi\left(\frac{\sigma_{0w}}{\sigma_{1w}} z_{\frac{\alpha}{2}} - \frac{\mu_{1w} - \mu_{0w}}{\sigma_{1w}}\right) \approx \Phi\left(\frac{\sigma_{0w}}{\sigma_{1w}} z_{\frac{\alpha}{2}} + \frac{|\mu_{1w} - \mu_{0w}|}{\sigma_{1w}}\right). \tag{13}$$

For instance, after survying a panel of clinical investigators, Bakal et al.[9] used pre-specified weights in a study that used a composite endpoints of death, cardiogenic shock (Shock), congestive heart failure (CHF), and recurrent myocardial infarction (RE-MI). The weights were 1 for death, 0.5 for Shock, 0.3 for hospitalization for CHF, and 0.2 for RE-MI, i.e., in this context $\mathbf{w} = \frac{1}{2}(1, 0.5, 0.3, 0.2)$. In another example[22] the composite outcome consisted of events weighted according to their severity: recurrent myocardial infarction (weight $w_1 = 0.415$), congestive heart failure that required the use of open-label angiotensin-converting enzyme (ACE) inhibitors (weight $w_2 = 0.17$), and hospitalization to treat congestive heart failure (weight $w_3 = 0.415$).

Although the use of pre-specified weights provides a more nuanced approach to the importance of individual endpoints of a composite outcome, recognizes the potential underlying differences that exists among them, and facilitates the results interpretation compare to traditional composite endpoints, the selection of appropriate weights is not straightforward since inherently subjective[22–24]. However, when they exist, failing to use such utility (or severity) weights to highlight clinical importance of the component outcomes of a composite endpoint implies that we assume equal weights, which is sometimes even worse.[23–25]

We note that when the weights w_1 and w_2 are equal, i.e., $c_1 = c_2 = c_3 = w_1^2$, the test statistic Z_w coincides with the (ordinary) Wilcoxon–Mann–Whitney test statistic Z given in (8). Indeed, in that case, $\mathbf{c}'\mathbf{U} = w_1^2[U_t + U_{tx} + U_x] = w_1^2 U$ with U given by the equation (5). Thus, $\mathbf{c}' E_0(\mathbf{U}) = w_1^2 E_0(U)$ and $Var_0(\mathbf{c}'\mathbf{U}) = w_1^4 Var_0(U)$, which implies that $Z = Z_w$.

2.3.2 Optimal weights Now we want to estimate the optimal weights w for the weighted WMW test statistic

$$Z_{\mathbf{c}} = \frac{\mathbf{c}'\left(\mathbf{U} - E_0(\mathbf{U})\right)}{\sqrt{Var_0(\mathbf{c}'\mathbf{U})}} = \frac{\mathbf{c}'\left(\mathbf{U} - E_0(\mathbf{U})\right)}{\sqrt{\mathbf{c}'Var_0(\mathbf{U})\mathbf{c}}}, \tag{14}$$

with $\mathbf{U}' = (U_t, U_{tx}, U_x)$ and $\mathbf{c}' = (c_1, c_2, c_3) = (w_1^2, w_1 w_2, w_2^2)$. Optimal weights c_1, c_2, and c_3 for the test statistic Z_w are those that maximize its power.

We will use the power formula of $Z_{\mathbf{c}}$, to derive its optimal weights. Then, we introduce the *optimal weighted WMW test statistic* Z_{opt} and highlight some of its properties and characteristics.

From the definition of $\mathbf{U}$, we show in Appendix B that

$$\begin{aligned} E(\mathbf{U}) &= \left(E(U_t), E(U_{tx}), E(U_x)\right)' \\ &= \left(\pi_{t1}p_1p_2, p_1q_2, \pi_{x1}q_1q_2\right)'. \end{aligned} \tag{15}$$

and $Var(\mathbf{U}) = \mathbf{\Sigma}$, where $\mathbf{\Sigma} = (mn)^{-1}\left(\Sigma_{ij}\right)_{1\le i,j\le 3}$ is a 3×3 matrix.

Under the null hypothesis of no difference between the two groups, with respect to both survival and non-fatal outcome, we have $p_1 = p_2 = p$, $q_1 = q_2 = q = 1-p$, $\pi_{t1} = \pi_{x1} = 1/2$, and $\pi_{t2} = \pi_{x2} = \pi_{t3} = \pi_{x3} = 1/3$. Thus,

$$E_0(\mathbf{U}) = \frac{1}{2}\left(p^2, 2pq, q^2\right)' \quad \text{and} \quad Var_0(\mathbf{U}) = \mathbf{\Sigma_0}, \tag{16}$$

where $\mathbf{\Sigma_0} = (mn)^{-1}\left(\Sigma_{0ij}\right)_{1\le i,j\le 3}$ is a symmetric matrix with

$$\begin{aligned} \Sigma_{011} &= \frac{p^2}{12}A(p),\ \Sigma_{012} = \Sigma_{021} = -\frac{p^2q^2}{4}(n+m-1),\ \Sigma_{013} = \Sigma_{031} = \frac{p^2q}{2}\left((n-1)q - mp\right) \\ \Sigma_{022} &= \frac{q^2}{12}A(q),\ \Sigma_{023} = \Sigma_{032} = \frac{pq^2}{2}\left((m-1)p - nq\right),\ \Sigma_{033} = pq\left(nq^2 + mp^2 + pq\right), \\ A(x) &= 6 + 4(n+m-2)x - 3(n+m-1)x^2. \end{aligned}$$

Moreover, since $Var_0(\mathbf{U}_w) = Var_0(\mathbf{c}'\mathbf{U}) = \mathbf{c}'\mathbf{\Sigma_0}\mathbf{c} \ge 0$ by definition, the matrix $\mathbf{\Sigma_0}$ is semi-positive definite.

The power formula for the weighted WMW, similar to equation (9), is

$$\Phi\left(\frac{\sigma_{0w}}{\sigma_{1w}}z_{\frac{\alpha}{2}} + \frac{\mu_{1w}-\mu_{0w}}{\sigma_{1w}}\right) + \Phi\left(\frac{\sigma_{0w}}{\sigma_{1w}}z_{\frac{\alpha}{2}} - \frac{\mu_{1w}-\mu_{0w}}{\sigma_{1w}}\right) \approx \Phi\left[\frac{\sigma_{0w}}{\sigma_{1w}}\left(z_{\frac{\alpha}{2}} + \frac{|\mu_{1w}-\mu_{0w}|}{\sigma_{0w}}\right)\right], \tag{17}$$

where $\mu_{1w} = \mathbf{c}'E(\mathbf{U})$, $\mu_{0w} = \mathbf{c}'E_0(\mathbf{U})$, $\sigma_{1w} = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$, and $\sigma_{0w} = \mathbf{c}'\boldsymbol{\Sigma_0}\mathbf{c}$.
Under the assumptions that

(1) n/m converges to a constant ρ $(0 < \rho < 1)$,

(2) both $\sqrt{N}\{F_1(t) - F_2(t)\}$ and $\sqrt{N}\{G_1(x) - G_2(x)\}$ are bounded, i.e., $\dfrac{\sigma_{0w}}{\sigma_{1w}}$ converges to 1 as $N = m + n \to \infty$,

a weight-vector $\mathbf{c}$ maximizes the power (17) if and only if it maximizes $|\mu_{1w} - \mu_{0w}|/\sigma_{0w}$.
We prove in Appendix C that the optimal-weight vector $\mathbf{c}_{opt}$ is given by

$$\mathbf{c}_{opt} = \frac{\Sigma_0^{-1}\mu}{\mathbf{b}'\Sigma_0^{-1}\mu}, \tag{18}$$

for $\mathbf{b}' = (1, 2, 1)$ and $\mu = E(\mathbf{U}) - E_0(\mathbf{U}) = \left(\pi_{t1}p_1p_2 - \frac{1}{2}p^2, p_1q_2 - pq, \pi_{x1}q_1q_2 - \frac{1}{2}q^2\right)'$. Therefore, from equation (14), the corresponding optimal test statistic Z_w (denoted here Z_{opt}) is then given by

$$Z_{opt} = \frac{\mathbf{c}'_{opt}\left(\mathbf{U} - E_0(\mathbf{U})\right)}{\sqrt{\mathbf{c}'_{opt}\boldsymbol{\Sigma}_0\mathbf{c}_{opt}}} = \frac{\mu'\Sigma_0^{-1}\left(\mathbf{U} - E_0(\mathbf{U})\right)}{\sqrt{\mu'\Sigma_0^{-1}\mu}}. \tag{19}$$

2.3.3 Remarks

(i) The test statistic Z_{opt} given by (19) encompasses the contributions of the effects of treatment on both mortality (via U_t) and the non-fatal outcome (via U_x) as well as the corresponding proportions of deaths and survivors in both treatment groups (via U_{tx}) and their relative importance and magnitude, where each component is weighted accordingly through $\mathbf{c}_{opt}$.

(ii) As demonstrated, the ordinary WMW test statistic is a special case of a weighted WMW test statistics (corresponding to a weighted WMW test statistic with equal weights). This implies that both the ordinary and the optimal weighted WMW test statistics belong to same family of weighted WMW tests.

(iii) Note that the optimal weight vector $\mathbf{c}_{opt} = \Sigma_0^{-1}\mu$ depends on unknown population parameters π_{t1}, π_{x1}, p_1, p_2, and p which must be estimated in practice (since they are not available from the observed sample data). A good estimation method of these unknown parameters is needed to calculate the test statistic Z_{opt} given by equation (19):

(a) When the distributions of the primary endpoint, X, and the survival time, t, are known approximately, we can estimate analytically the probabilities π_{t1} and π_{x1}, p_1, p_2 (as we have done in Appendix D for our simulation studies) and calculate an estimate of the probability p under the null hypothesis (H_0) as $\hat{p} = (m\hat{p}_1 + n\hat{p}_2)/(m+n)$ (pooled sample proportion).

In general, the distributions of both the primary endpoint and the survival time are not known. Optimal weights are estimated using either data from a pilot study (or from previous studies, when available) or the data at hand.

(b) If we have data from prior studies, we can leverage them to estimate these parameters. Using Bayesian methods, we can elicit expert opinions to define prior distributions associated with Σ_0 and μ that best reflect the characteristics of the disease under study and determine posterior distributions to provide a more accurate assessment of the optimal weights.[26] Alternatively, if the data is structured such that we have multiple strata available (e.g., different enrollment periods or different clinical centers for patients), we can use an adaptive weighting scheme to estimate Σ_0 and μ.[27;28]

(c) In absence of data from prior studies, it is recommended to use a bootstrap approach to estimate the weights. To do this, we generate B bootstrap samples (e.g., B=500, 1000, or 2000) and, for each bootstrap sample, we estimate the corresponding optimal weight vector $\mathbf{c}_{opt}$. Then, we compute the average weights from the B estimates. Finally, using these average weights, we compute the test statistic Z_{opt} on the original sample with the average weights estimated in the first part and test the null hypothesis.

(d) With the data at hand, we can also use a $K-$fold cross-validation. In that regard, we divide the data into K subsets of roughly equal size and estimate the weights $\mathbf{c}_{opt,k}$ and the test statistic $Z_{opt,k}$ exactly K times. At the k-th time, $k = 1, \ldots, K$, we use the $k-$th subset as *validation data* to calculate the weights $\mathbf{c}_{opt,k}$ and combine the remaining $K-1$ subsets as *training data* to estimate the test statistic $Z_{opt,k}$ using the weights defined at the validation stage. Then, we estimate the test statistic Z_{opt} by averaging over all the K test statistics Z_{opt_k}, $k = 1, \ldots, K$ and run the hypothesis test.

3 Simulation Studies

We conducted simulation studies to assess the performance of the weighted test statistic. We generated data set to follow the pattern seen in stroke trials, where the outcome of interest (patient's improvement on the NIH stroke scale score over a 3-month period) may be missing for some patients due to death. We simulated death times under a proportional hazards model with $t_{1k} \sim Exp(\lambda_1)$, $t_{2l} \sim Exp(\lambda_2)$, such that $q_2 = \exp(-\lambda_2 T)$ and $HR = \lambda_1/\lambda_2$ with $T = 3$ months, $HR = 1.0, 1.2, 1.4, 1.6, 2.0, 2.4, 3.0$ and $q_2 = 0.6, 0.8$. For the non-fatal outcome, $X_{1k} \sim N(0,1)$, $X_{2l} \sim N(\sqrt{2}\Delta_x, 1)$, $k = 1, \ldots, m$; $l = 1, \ldots, n$ with $\Delta_x = (\mu_{x_2} - \mu_{x_1})/(\sigma_{x_1}\sqrt{2}) = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$. The conditional probabilities, $\pi_{t\gamma}$ and $\pi_{x\gamma}$, $\gamma = 1, 2, 3$, are given in Appendix D. We computed power for the weighted WMW test for $n = m = 50$ patients, using the analytical power formula (17) and a two-sided $\alpha = 0.05$. In addition, we estimated power empirically by averaging over 10,000 simulated data sets.

The results, given in Table 1, illustrate the accuracy of the analytical power formula (17). They indicate also that the weighted WMW test statistic is more powerful than the ordinary WMW test for the worst-rank score composite outcome. The largest differences are seen in two different scenarios:

1. the standardized difference in the non-fatal outcome Δ_x is small ($\Delta_x < 0.3$) and the difference in mortality is moderate or high (HR ≥ 1.2)

2. the difference in mortality is small (HR< 1.2) and the standard difference in the non-fatal outcome Δ_x is moderate or high ($\Delta_x \geq 0.3$).

Overall, these results mean that whenever the effect on the primary outcome is small, the larger difference in mortality is diluted when assessing the overall difference through the ordinary WMW where mortality and the non-fatal outcome are weighted equally. Likewise, if the difference in mortality is small, but the difference in the non-fatal outcome is moderate or high, the ordinary WMW test on the composite outcome has less power than the weighted WMW.

4 Application to a Stroke Clinical Trial

A clinical trial of normobaric oxygen therapy (NBO) was conducted at Massachusetts General Hospital for patients who had an acute ischemic stroke.[1;2] In this trial, 85 patients were randomly assigned to either NBO therapy (43 patients) or to room air (control) for 8 hours and assessed serially with clinical function scores. The primary efficacy and safety endpoints were, respectively, the mean change in NIHSS from baseline to 4 hours (during therapy) and 24 hours (after

Table 1. Power comparisons for a continuous outcome under proportional hazards for time to death

	$q_2 = 60\%$							$q_2 = 80\%$						
	HR							HR						
Δ_x	1.0	1.2	1.4	1.6	2.0	2.4	3.0	1.0	1.2	1.4	1.6	2.0	2.4	3.0
(A) Analytical Power for the Weighted WMW Test														
0.0	0.05[a]	0.11	0.24	0.41	0.73	0.90	0.98	0.05[a]	0.08	0.15	0.24	0.45	0.68	0.87
0.1	0.08	0.12	0.25	0.42	0.73	0.90	0.98	0.09	0.12	0.18	0.28	0.51	0.70	0.88
0.2	0.15	0.19	0.30	0.46	0.75	0.91	0.98	0.21	0.24	0.30	0.38	0.58	0.75	0.90
0.3	0.27	0.30	0.40	0.53	0.78	0.92	0.98	0.39	0.41	0.46	0.53	0.69	0.82	0.93
0.4	0.41	0.44	0.51	0.61	0.82	0.93	0.98	0.59	0.61	0.64	0.69	0.79	0.88	0.95
0.5	0.55	0.57	0.62	0.70	0.86	0.94	0.99	0.76	0.77	0.79	0.81	0.87	0.92	0.97
0.6	0.68	0.68	0.72	0.77	0.89	0.95	0.99	0.88	0.88	0.89	0.90	0.93	0.96	0.98
(B) Empirical Power for the Weighted WMW Test														
0.0	0.05[a]	0.10	0.23	0.40	0.72	0.91	0.99	0.05[a]	0.08	0.15	0.24	0.45	0.67	0.87
0.1	0.08	0.12	0.24	0.41	0.73	0.90	0.99	0.09	0.12	0.18	0.28	0.51	0.70	0.89
0.2	0.15	0.19	0.29	0.47	0.75	0.91	0.99	0.21	0.24	0.30	0.38	0.58	0.76	0.91
0.3	0.26	0.30	0.40	0.53	0.78	0.92	0.99	0.39	0.41	0.46	0.54	0.69	0.83	0.94
0.4	0.39	0.43	0.51	0.63	0.81	0.93	0.99	0.59	0.61	0.65	0.71	0.81	0.89	0.96
0.5	0.54	0.56	0.63	0.71	0.87	0.94	0.99	0.76	0.78	0.81	0.83	0.90	0.94	0.98
0.6	0.67	0.68	0.73	0.79	0.89	0.96	0.99	0.89	0.89	0.91	0.92	0.95	0.97	0.99
(C) Empirical Power for the Ordinary WMW Test on Worst Rank Scores														
0.0	0.05	0.09	0.17	0.31	0.62	0.84	0.98	0.05	0.06	0.09	0.13	0.29	0.48	0.74
0.1	0.06	0.12	0.22	0.38	0.67	0.87	0.98	0.08	0.11	0.17	0.24	0.42	0.61	0.82
0.2	0.07	0.16	0.30	0.44	0.74	0.90	0.99	0.14	0.21	0.29	0.37	0.56	0.72	0.89
0.3	0.12	0.22	0.37	0.53	0.78	0.93	0.99	0.26	0.33	0.43	0.53	0.70	0.83	0.94
0.4	0.16	0.29	0.44	0.59	0.82	0.94	0.99	0.40	0.50	0.59	0.66	0.81	0.89	0.96
0.5	0.22	0.36	0.52	0.66	0.86	0.96	0.99	0.57	0.66	0.73	0.79	0.89	0.95	0.98
0.6	0.30	0.44	0.59	0.70	0.88	0.97	0.99	0.71	0.78	0.84	0.88	0.94	0.97	0.99

[a] The weights are equal and fixed to 1.

We assumed the treatment is better either on both mortality and non-fatal outcome or on one outcome and not different from the control on the other outcome. We used exponential distributions for the survival times, normal distributions for the non-fatal outcome, and the same number of subjects in each group ($n_1 = n_2 = 50$). Δ_x: standardized mean difference on the non-fatal outcome of interest; HR: hazard ratio; q_2: survival probability (proportion of patients alive) at 3 months in the treatment group.

(A) Estimated using formula (9); (B) and (C) Proportion of simulated data sets for which $|Z_{opt}| > 1.96$ and $|Z| > 1.96$, respectively.

therapy).[1] For illustration purposes, we focused on the secondary endpoint and examined the mean change in NIHSS scores from baseline to 3 months or at discharge.

Twenty-four of the 85 patients died, 17 of whom were in the NBO group. Fifty-three patients (with 31 in the control group) were discharged prior to the 3-month follow up period. Subjects with missing 3-month NIHSS scores were included in the estimation of the log rank test, but

excluded in the assessment of the change in NIHSS scores. The log rank test of survival was significant ($\chi^2 = 6$ with 1 d.f., p-value $= 0.016$), indicating that the active treatment had an unfavorable effect on mortality. The ordinary WMW test applied to the survivors was not significant ($W = 572.5$, p-value $= 0.27$). Using the untied worst-rank composite endpoint of death times and NIHSS scores, we found a significant result with the ordinary WMW test ($W = 1112.5$, p-value $= 0.01$).

Finally, we applied the proposed method, estimating the weights and the test statistic Z_w using B=2000 bootstrap samples, as explained in part (iii) of the Remarks 2.3.3. The estimated weight vector $\mathbf{c}'$, the mean difference μ, the variance-covariance matrix for U under the null, and the probability p were, respectively, $\mathbf{c}' = (0.45, 0.16, 0.24)$, $\Sigma_0 = \begin{pmatrix} 0.59 & 0.50 & -0.90 \\ 0.50 & 4.77 & -1.27 \\ -0.90 & -1.27 & 5.16 \end{pmatrix}$, $\mu = -(0.016, 0.098, 0.073)$, and $p = 0.283$. This corresponds to $w_1 = 0.61$ and $w_2 = 0.39$; which means mortality was weighted more heavily (61 % of the weight) than NIHSS score, in addition to ranking death worse than any measure of the continuous outcome (NIHSS score). The optimally weighted WMW test statistic Z_{opt} was equal to 3.42 with a corresponding p-value of 6.2×10^{-4}. This result is stronger than that from the ordinary WMW test as it captures the significant difference in mortality between the two treatment groups and demonstrates the efficiency of our test statistic.

5 Discussion

In this paper, we have generalized the notion of the Wilcoxon–Mann–Whitney (WMW) test for a worst-rank composite outcome by deriving the optimally weighted WMW test. Against the null hypothesis of no difference on both mortality and continuous endpoint, we have focused on the alternative hypothesis that "the active treatment has a preponderance of positive effects on the multiple outcomes considered, while not being harmful for any" [29]. We have motivated the worst-rank composite outcome in the context of the clinical trial of a non-mortality primary outcome where the assessment of the primary outcome of interest at a pre-specified time-point may be precluded by death, any other debilitating event, or worsening of the disease condition. The corresponding composite outcome takes into account all patients enrolled in the trial, including those who had terminal events before the end of follow-up.

When there exists a hierarchy of the constituent endpoints of a composite outcome, the method we have presented in this paper enables different components of the WMW test statistic to be weighted differentially. Using weights allows for an additional level of discrimination between

the component outcomes beyond ranks alone. While the worst-rank score mechanism pertains with how the different component outcomes of the composite endpoint are aggregated, assigning weights strengthen (or lessen) the influence these prioritized component outcomes exert in the overall composite. We considered weights obtained or elicited from expert judgments (utility weights) or determined in a way that the corresponding WMW test statistic has a maximum power. Based on a U-statistic approach, we first provided the test statistic and the power of the weighted WMW test when utilities (or severity) weights, determined *a priori*, are available. We also demonstrated that the ordinary (unweighted) WMW test on the worst-rank score outcome is a special case of the weighted WMW test i.e. when the weights are all equal. Then, we derived the optimal weights such that the power of the corresponding weighted WMW test statistic is maximal. Finally, we conducted simulation studies to evaluate the accuracy of our power formula and confirmed, in the process, that the weighted WMW is more powerful than ordinary WMW test.

We applied the proposed method to the data from a clinical trial of normobaric oxygen therapy (NBO) for patients with acute ischemic stroke. Patients' improvement was assessed using the National Institutes of Health Stroke Scale (NIHSS) Scores. The results indicated a statistically significant difference between NBO therapy and room air— using either the proposed method or the ordinary WMW test on the worst-rank composite outcome of death and change in NIHSS— which we couldn't detect using the ordinary WMW on the survivors alone.

The difference between NBO therapy and room air was driven by the difference in mortality since there was a disproportionate number of NBO-treated patients who died. It is actually for this reason the trial was stopped by the Data and Safety Monitoring Board (DSMB) after 85 patients out of the projected 240 were enrolled. The stark imbalance between the two treatment group, although not attributed to the treatment, made it untenable to continue the trial. [1;30]

The end result of the NBO trial is one of the dreaded scenarios in the (traditional) analysis of composite endpoints. That the active treatment must be better than the control for one or both of the constituent outcomes (mortality and non-fatal outcome) and not worse for either of them as suggested by our alternative hypothesis H_1 (stated in equation (4)), was clearly not the case for the NBO trial. While the active treatment was equivalent to the control treatment in change in NIHSS, the data showed also that NBO therapy increased mortality. Ideally, components of a composite endpoint should have similar clinical importance, frequency, and treatment effect. However, this is rarely the case as outcomes of different levels of severity are usually combined to facilitate the interpretation of such results, several authors have suggested running complementary analyses on components of the composite outcome. [31–38]

When the impact of the active treatment on mortality is of greater clinical importance than its effect on the primary outcome of interest, the weighted WMW test statistic we have presented can be included into a set of testing procedures that ensure that the treatment is not inferior on both mortality and the outcome of interest and that it is superior on a least one of these endpoints. In the context of ischemic stroke, the clinical investigators desired a treatment that would have a positive impact on mortality while also improving survivors' functional outcomes. Testing procedures that incorporate contributions of each individual component of the composite while penalizing for any disadvantage in the active treatment when the treatment operates in opposite directions on the components of the composite outcome have been discussed.[39–42] For the analysis of NBO clinical trial, we propose two different stepwise procedures to analyze data using this weighted test: (1) two individual non-inferiority tests on mortality and non-fatal outcome, followed (if non-inferiority established) by a global test using the optimal weighted WMW test on the worst-rank composite endpoint; or (2) a global test using the optimal weighted WMW test on the worst-rank composite endpoint, then (if significant global test) two individual non-inferiority tests followed by individual superiority tests on mortality and non-fatal outcome. In either scenario, the overall type I error is preserved.[39;40;43;44]

The method presented in this paper can be applied or extended to many other settings of composite endpoints beyond the realm of death-censored observations. The rationale, advantages (and limitations), and recommendations for using composite outcomes—based on clinical information, expert knowledge or practical matters—abound in the literature.[14;35;45] One can also accommodate ties as well as non-informative censoring in the definition of the WMW U-statistic. In particular, when non-informative censoring is present (and, without loss of generality, assuming there is no ties), survival times can be assessed using Gehan's U-statistic, which is an extension of the WMW U-statistic to right censored data.[46] In this case, $I(t_{1k} < t_{2l})$ will be equal to 1 if subject l in group 2 lived longer than subject k in group 1 and 0 if it is uncertain which subject lived longer.

Our method can be applied in many disease areas in which different outcomes are clinically related and represent the manifestation of the same underlying condition. Clinical trials of unstable angina and non-ST segment elevation myocardial infarction are examples of such an application.[47;48] The method can also be applied in clinical trials where the overall effect of treatment on a disease depends on hierarchy of meaningful—yet of different importance, magnitude, and impact—heterogenous outcomes. For instance, in clinical trials of asthma or of benign prostatic hyperplasia (BPH), several outcomes are necessary to capture the multifaceted manifestations of the disease. For patients with asthma, four outcomes (forced expiratory volume in 1 second (FEV_1), peak expiratory flow (PEF) rate, symptom score, and additional rescue

medication use) are necessary to measure the different manifestations of the disease ([49]). Due to subjective nature of benign prostatic hyperplasia (BPH) symptoms, in addition to BPH symptom score index, measures to assess disease progression include: prostate specific antigen (PSA), urinary cytology, post-void residual volume (PVR), urine flow rate, cystoscopy, urodynamic pressure-flow study, and ultrasound of the kidney or the prostate.

Our method does not immediately apply to the case where the treatment effect is assessed by stratifying for a confounding variable (baseline scores, baseline disease severity, age, ...) pre-specified in the study design.[50;51]. For the NBO trial, had the investigators anticipated the imbalance between subjects on some baseline variables (e.g., large infarcts, advanced age, co-morbidities, and most importantly, withdrawal of care based on pre-expressed wishes or family preference), they could have stratified the study population with respect to these variable.[1;30] The test statistic we have proposed does not adjust for such baseline covariates as the appropriate weighted WMW test for this case must take into account the stratum specific characteristics in addition to the specificities of the worst-ranking procedure; this is a topic for future investigations.

A strong case may be made on why one should prefer analysis of covariance to the analysis of change from baseline score as we have done in this paper[52]. But in reality, issues are more nuanced and the approach to use depends closely on the nature of the data as well as the clinical question of interest.[53–58] For the difference in NIHSS scores (from baseline to 3 months), the fundamental question of interest was "on average, how much NBO-treated patients changed over 3-month period compare to patients assigned to room air?" The change-from-baseline-score paradigm assumes that the same measure is used before and after the treatment and that these two measures are highly correlated.[59;60] In the stroke literature, it is proven that change from baseline in NIHSS satisfies this assumption since baseline NIHSS is a strong predictor of outcome after stroke.[61;62] Moreover, it has been shown that change in the NIHSS score is a useful tool to measure treatment effect in acute stroke trials (see for instance the papers by Bruno et al.[63] and by Parsons et al.[64]) Hence, this justified the choice of improvement (or change) in NIHSS score as outcome of interest in this paper.

We have assumed throughout this paper that mortality is worse than any impact ischemic stroke may have on patients. Our assumption stems from the common view that ranks death as inferior to any quality-of-life measure; such a view is advocated in several medical fields.[7;8;65–70] However, some people (patients, their family members or caregivers) may argue otherwise and affirm that there are levels of stroke that are worse than death. For instance, in a study of the effects of thrombolytic therapy in reducing damage from a myocardial infarction, the hierarchy of the quality of component outcomes was "stroke resulting in a vegetative state,

death, serious morbidity requiring major assistance, serious morbidity but capable of self-care, excess spontaneous hemorrhage (≥ 3 blood transfusions), and 1-2 transfusions".[10] There are number of papers in the causal inference literature that offer an alternative approach based on Rosenbaum's proposal of using different "placements of death".[71] However, as Rubin pointed out, this elegant idea "maybe difficult to convey to consumers"[72] and we have not pursued this avenue here.

Finally, the null hypothesis (3) for WMW test stipulates that the treatment does not change the outcome distribution, which means that the treatment has no effect on any patient. However, some studies may require a weaker version of the null hypothesis, i.e., the treatment does not affect the average group response.[73;74] In such a case, the WMW is not an asymptotically valid test for the weaker null hypothesis.[75;76] As an alternative, one can use the Brunner-Munzel test[77] where the marginal distribution functions of the two treatment groups are not assumed to be equal and may have different shapes, even under the null hypothesis. In this paper, we have chosen the WMW test because it is simple, widely used, efficient, and robust against parametric distributional assumptions. The use of a weighted Brunner-Munzel test for analysis of the worst-rank composite outcome of death and a quality-of-life (such as the NIHSS score) warrants further investigations and is beyond the scope of this paper.

APPENDIX

Appendix A Mean and variance of the U-statistic

Consider the untied worst-rank adjusted values for subjects in the control and active treatment groups $\widetilde{X}_{1k} = (1-\delta_{1k})X_{1k} + \delta_{1k}(\eta + t_{1k})$, for $k = 1, \ldots, m$ and $\widetilde{X}_{2l} = (1-\delta_{2l})X_{2l} + \delta_{2l}(\eta + t_{2l})$, for $l = 1, \ldots, n$.
Define the WMW U-statistic

$$U = (mn)^{-1} \sum_{k=1}^{m} \sum_{l=1}^{n} U_{kl}, \quad \text{where } U_{kl} = I(\widetilde{X}_{1k} < \widetilde{X}_{2l}).$$

Since $U_{kl} = 1$ if $\{t_{1k} < t_{2l} \text{ and } \delta_{1k}\delta_{2l} = 1\}$, $\{\delta_{1k} = 1 \text{ and } \delta_{2l} = 0\}$, or $\{X_{1k} < X_{2l} \text{ and } (\delta_{1k} = \delta_{2l} = 0)\}$, we have $U_{kl} = I(t_{1k} < t_{2l},\ \delta_{1k}\delta_{2l} = 1) + I(\delta_{1k} = 1,\ \delta_{2l} = 0) + I(X_{1k} < X_{2l},\ \delta_{1k} = \delta_{2l} = 0)$.

Therefore,

$$
\begin{aligned}
E(U) &= E(U_{kl}) \qquad\qquad (A.1)\\
&= P(t_{1k} < t_{2l} | \delta_{1k}\delta_{2l} = 1)P(\delta_{1k}\delta_{2l} = 1) + P(\delta_{1k} = 1,\ \delta_{2l} = 0) + P(X_{1k} < X_{2l})P(\delta_{1k} = \delta_{2l} = 0)\\
&= p_1 p_2 \cdot P(t_{1k} < t_{2l} | \delta_{1k} = \delta_{2l} = 1) + p_1 q_2 + q_1 q_2 \cdot P(X_{1k} < X_{2l})\\
&= p_1 p_2 \pi_{t1} + p_1 q_2 + q_1 q_2 \pi_{x1} = \pi_{U1}
\end{aligned}
$$

where $q_1 = 1 - p_1$, $q_2 = 1 - p_2$, $\pi_{t1} = P(t_{1k} < t_{2l} | \delta_{1k} = \delta_{2l} = 1)$, and $\pi_{x1} = P(X_{1k} < X_{2l})$.

$$
\begin{aligned}
Var(U) &= (mn)^{-2} \left[\sum_{k=1}^{m}\sum_{l=1}^{n} Var(U_{kl}) + \sum_{k=1}^{m}\sum_{l=1}^{n}\sum_{k'=1}^{m}\sum_{l'=1}^{n} Cov(U_{kl}, U_{k'l'}) \right], \text{ with } k \neq k' \text{ or } l \neq l' \text{ or both}\\
&= (mn)^{-1} \left[Var(U_{kl}) + (m-1)Cov(U_{kl}, U_{k'l}) + (n-1)Cov(U_{kl}, U_{kl'}) \right].
\end{aligned}
$$

Note that $Cov(U_{kl}, U_{k'l'}) = E(U_{kl}U_{k'l'}) - E(U_{kl})E(U_{k'l'}) = 0$, $Cov(U_{kl}, U_{k'l}) = E(U_{kl}U_{k'l}) - E(U_{kl})E(U_{k'l})$ and $Cov(U_{kl}, U_{kl'}) = E(U_{kl}U_{kl'}) - E(U_{kl})E(U_{kl'})$, for $k \neq k'$, $l \neq l'$. In addition, because $U_{kl} = I(\widetilde{X}_{1k} < \widetilde{X}_{2l})$ follows Bernoulli distribution with probability π_{U1}, we derive the variance $Var(U_{kl}) = E(U_{kl})\left[1 - E(U_{kl})\right] = \pi_{U1}(1 - \pi_{U1})$.

$$
\begin{aligned}
E(U_{kl}U_{k'l}) &= P(U_{kl}U_{k'l} = 1)\\
&= P(\delta_{1k}\delta_{1k'} = 1,\ \delta_{2l} = 0) + P(t_{1k} < t_{2l}, t_{1k'} < t_{2l} | \delta_{1k}\delta_{1k'}\delta_{2l} = 1)P(\delta_{1k}\delta_{1k'}\delta_{2l} = 1)\\
&+ P(X_{1k'} < X_{2l})P(\delta_{1k} = 1,\ \delta_{1k'} = \delta_{2l} = 0) + P(X_{1k} < X_{2l})P(\delta_{1k} = 0,\ \delta_{1k'} = 1,\ \delta_{2l} = 0)\\
&+ P(X_{1k} < X_{2l}, X_{1k'} < X_{2l})P(\delta_{1k} = \delta_{1k'} = \delta_{2l} = 0)\\
&= p_1^2 q_2 + p_1^2 p_2 \pi_{t2} + 2p_1 q_1 q_2 \pi_{x1} + q_1^2 q_2 \pi_{x2},\\
E(U_{kl}U_{kl'}) &= P(U_{kl}U_{kl'} = 1)\\
&= P(\delta_{1k} = 1,\ \delta_{2l} = \delta_{2l'} = 0) + P(t_{1k} < t_{2l},\ t_{1k} < t_{2l'} | \delta_{1k}\delta_{2l}\delta_{2l'} = 1)P(\delta_{1k}\delta_{2l}\delta_{2l'} = 1)\\
&+ P(t_{1k} < t_{2l} | \delta_{1k}\delta_{2l} = 1,\ \delta_{2l'} = 0)P(\delta_{1k}\delta_{2l} = 1,\ \delta_{2l'} = 0)\\
&+ P(t_{1k} < t_{2l'} | \delta_{1k} = 1,\ \delta_{2l} = 0,\ \ \delta_{2l'} = 1)P(\delta_{1k} = 1,\ \delta_{2l} = 0,\ \delta_{2l'} = 1)\\
&+ P(X_{1k} < X_{2l}, X_{1k} < X_{2l'})P(\delta_{1k} = \delta_{2l} = \delta_{2l'} = 0)\\
&= p_1 q_2^2 + p_1 p_2^2 \pi_{t3} + 2p_1 p_2 q_2 \pi_{t1} + q_1 q_2^2 \pi_{x3}
\end{aligned}
$$

with $\pi_{t2} = P(t_{1k} < t_{2l}, t_{1k'} < t_{2l} | \delta_{1k} = \delta_{1k'} = \delta_{2l} = 1)$, $\pi_{x2} = P(X_{1k} < X_{2l}, X_{1k'} < X_{2l})$,
$\pi_{t3} = P(t_{1k} < t_{2l}, t_{1k} < t_{2l'} | \delta_{1k} = \delta_{2l} = \delta_{2l'} = 1)$, and $\pi_{x3} = P(X_{1k} < X_{2l}, X_{1k} < X_{2l'})$.

In summary,

$$Var(U) = (mn)^{-1}\left[\pi_{U1}\left(1-\pi_{U1}\right) + (m-1)(\pi_{U2}-\pi_{U1}^2) + (n-1)(\pi_{U3}-\pi_{U1}^2)\right], \qquad (A.2)$$

where $\pi_{U2} = p_1^2 q_2 + p_1^2 p_2 \pi_{t2} + 2p_1 q_1 q_2 \pi_{x1} + q_1^2 q_2 \pi_{x2}$ and $\pi_{U3} = p_1 q_2^2 + p_1 p_2^2 \pi_{t3} + 2p_1 p_2 q_2 \pi_{t1} + q_1 q_2^2 \pi_{x3}$.

Under the null hypothesis of no difference between the two groups, with respect to survival and non fatal outcome, we have $F_1 = F_2 = F$, $G_1 = G_2 = G$ and $p_1 = p_2 = p$, $q_1 = q_2 = q$. This implies

$$\begin{aligned}
\pi_{t1} &= P(t_{1k} < t_{2l} | t_{1k} \leq T, t_{2l} \leq T) = \frac{1}{p^2}\int_0^{'} F(t)dF(t) = \frac{1}{2p^2}\left[F(T)^2 - F(0)^2\right] = \frac{1}{2} \\
\pi_{t2} &= P(t_{1k} < t_{2l}, t_{1k'} < t_{2l} | t_{1k} \leq T, t_{1k'} \leq T, t_{2l} \leq T) = \frac{1}{p^3}\int_0^{'} F(t)^2 dF(t) \\
&= \frac{1}{3p^3}\left[F(T)^3 - F(0)^3\right] = \frac{1}{3} \\
\pi_{t3} &= P(t_{1k} < t_{2l}, t_{1k} < t_{2l'} | t_{1k} \leq T, t_{2l} \leq T, t_{2l'} \leq T)) = \frac{1}{p^3}\int_0^{T} \left[1-F(t)\right]^2 dF(t) \\
&= \frac{1}{3p^3}\left\{\left[1-F(T)\right]^3 - \left[1-F(0)\right]^3\right\} = \frac{1}{3} \\
\pi_{x1} &= P(X_{1k} < X_{2l}) = \int_{-\infty}^{\infty} G(x)dG(x) = \frac{1}{2}\left[G(x)^2\right]_{-\infty}^{\infty} = \frac{1}{2} \\
\pi_{x2} &= P(X_{1k} < X_{2l}, X_{1k'} < X_{2l}) = \int_{-\infty}^{\infty} G(t)^2 dG(t) = \frac{1}{3}\left[G(x)^3\right]_{-\infty}^{\infty} = \frac{1}{3} \\
\pi_{x3} &= P(X_{1k} < X_{2l}, X_{1k} < X_{2l'}) \int_{-\infty}^{\infty} [1-G(t)]^2 dG(t) = -\frac{1}{3}\left\{\left[1-G(x)\right]^3\right\}_{-\infty}^{\infty} = \frac{1}{3}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\pi_{U1} &= p_1 p_2 \pi_{t1} + p_1 q_2 + q_1 q_2 \pi_{x1} = \frac{1}{2}p^2 + pq + \frac{1}{2}q^2 = \frac{1}{2}(p+q)^2 = \frac{1}{2} \\
\pi_{U2} &= p_1^2 q_2 + p_1^2 p_2 \pi_{t2} + 2p_1 q_1 q_2 \pi_{x1} + q_1^2 q_2 \pi_{x2} = p^2 q + \frac{1}{3}p^3 + pq^2 + \frac{1}{3}q^3 = \frac{1}{3}(p+q)^3 = \frac{1}{3} \\
\pi_{U3} &= p_1 q_2^2 + p_1 p_2^2 \pi_{t3} + 2p_1 p_2 q_2 \pi_{x1} + q_1 q_2^2 \pi_{x3} = pq^2 + \frac{1}{3}p^3 + p^2 q + \frac{1}{3}q^3 = \frac{1}{3}(p+q)^3 = \frac{1}{3}.
\end{aligned}$$

The mean and variance become

$$\mu_0 = E_0(U) = \pi_{U1} = \frac{1}{2};$$

$$\begin{aligned}\sigma_0^2 = Var_0(U) &= (mn)^{-1}\left[\pi_{U1}\left(1-\pi_{U1}\right)+(m-1)\left(\pi_{U2}-\pi_{U1}^2\right)+(n-1)\left(\pi_{U3}-\pi_{U1}^2\right)\right] \\ &= (mn)^{-1}\left[\frac{1}{2}\left(1-\frac{1}{2}\right)+(m-1)\left(\frac{1}{3}-\left(\frac{1}{2}\right)^2\right)+(n-1)\left(\frac{1}{3}-\left(\frac{1}{2}\right)^2\right)\right] \\ &= (mn)^{-1}\left[\frac{1}{4}+\frac{1}{12}(m-1)+\frac{1}{12}(n-1)\right] = \frac{m+n+1}{12mn}.\end{aligned}$$

Appendix B Mean and variance of the Weighted U-Statistic

Consider the weights $\mathbf{w}=(w_1,w_2)$, we define the vector $\mathbf{c}'=(c_1,c_2,c_3)=\left(w_1^2,w_1w_2,w_2^2\right)$. Let $\widetilde{X}_{1k}=w_1\delta_{1k}(\eta+t_{1k})+w_2(1-\delta_{1k})X_{1k}$, for $k=1,\ldots,m$ and $\widetilde{X}_{2l}=w_1\delta_{2l}(\eta+t_{2l})+w_2(1-\delta_{2l})X_{2l}$, for $l=1,\ldots,n$.

We define the weighted WMW U-statistic by $\mathbf{c}'\mathbf{U}=(U_t,U_{tx},U_x)$ where $\mathbf{U}'=(U_t,U_{tx},U_x)$ and

$$\begin{aligned}U_t &= (mn)^{-1}\sum_{k=1}^{m}\sum_{l=1}^{n}\delta_{1k}\delta_{2l}I(t_{1k}<t_{2l}) \\ U_{tx} &= (mn)^{-1}\sum_{k=1}^{m}\sum_{l=1}^{n}\delta_{1k}(1-\delta_{2l}) \\ U_x &= (mn)^{-1}\sum_{k=1}^{m}\sum_{l=1}^{n}(1-\delta_{1k})(1-\delta_{2l})I(X_{1k}<X_{2l})\end{aligned} \tag{B.1}$$

$$\begin{aligned}E(\mathbf{U}) &= \left(P(\delta_{1k}=1)P(\delta_{2l}=1)P(t_{1k}<t_{2l}|\delta_{1k}=\delta_{2l}=1),\ P(\delta_{1k}=1)P(\delta_{2l}=0),\right. \\ &\quad \left.P(\delta_{1k}=0)P(\delta_{2l}=0)P(X_{1k}<X_{2l})\right)' \\ &= \left(p_1p_2\cdot P(t_{1k}<t_{2l}|\delta_{1k}=\delta_{2l}=1),\ p_1q_2,q_1q_2\cdot P(X_{1k}<X_{2l})\right)' \\ &= \left(p_1p_2\pi_{t1},p_1q_2,q_1q_2\pi_{x1}\right)'\end{aligned} \tag{B.2}$$

where $q_1=1-p_1$, $q_2=1-p_2$, $\pi_{t1}=P(t_{1k}<t_{2l}|\delta_{1k}=\delta_{2l}=1)$, and $\pi_{x1}=P(X_{1k}<X_{2l})$.

$Var(\mathbf{U})=\mathbf{\Sigma}$, where $\mathbf{\Sigma}=(mn)^{-1}\left(\Sigma_{ij}\right)_{1\le i,j\le 3}$ is a 3×3 matrix such that

$$\begin{aligned}\Sigma_{11} &= E[(U_t-p_1p_2\pi_{t1})(U_t-p_1p_2\pi_{t1})] \\ &= p_1p_2\left[\pi_{t1}(1-\pi_{t1})+p_1(m-1)(\pi_{t2}-\pi_{t1}^2)+p_2(n-1)(\pi_{t3}-\pi_{t1}^2)\right. \\ &\quad \left.+\,\pi_{t1}^2(mp_1q_2+(n-1)p_2q_1+q_1)\right], \\ \Sigma_{12} &= \Sigma_{21}=E[(U_t-p_1p_2\pi_{t1})(U_{tx}-p_1q_2)]=\pi_{t1}p_1p_2q_2\left[(n-1)q_1-mp_1\right], \\ \Sigma_{13} &= \Sigma_{31}=E[(U_t-p_1p_2\pi_{t1})(U_x-q_1q_2\pi_{x1})]=-\pi_{t1}\pi_{x1}(m+n-1)p_1q_1p_2q_2,\end{aligned}$$

$$\begin{aligned}
\Sigma_{22} &= E[(U_{tx} - p_1q_2)(U_{tx} - p_1q_2p)] = p_1q_2\left[mp_1p_2 + (n-1)q_1q_2 + q_1\right] \\
\Sigma_{23} &= \Sigma_{32} = E[(U_{tx} - p_1q_2)(U_x - q_1q_2\pi_{x1})] = \pi_{x1}p_1q_1q_2\left[(m-1)p_2 - nq_2\right], \\
\Sigma_{33} &= q_1q_2\left[\pi_{x1}(1-\pi_{x1}) + q_1(m-1)(\pi_{x2} - \pi_{x1}^2) + q_2(n-1)(\pi_{x3} - \pi_{x1}^2)\right. \\
&+ \left.\pi_{x1}^2(mq_1p_2 + (n-1)q_2p_1 + p_1)\right].
\end{aligned}$$

Therefore,

$$Var(\mathbf{c}'\mathbf{U}) = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}.$$

Under the null hypothesis of no difference between the two groups, with respect to both survival and non-fatal outcome, we have $p_1 = p_2 = p,\ q_1 = q_2 = q = 1-p,\ \pi_{t1} = \pi_{x1} = 1/2,$ and $\pi_{t2} = \pi_{x2} = \pi_{t3} = \pi_{x3} = 1/3$. Thus,

$$E_0(\mathbf{U}) = \frac{1}{2}\left(p^2, 2pq, q^2\right)' \quad \text{and} \quad Var_0(\mathbf{U}) = \boldsymbol{\Sigma_0}, \tag{B.3}$$

where $\boldsymbol{\Sigma_0} = (mn)^{-1}\left(\Sigma_{0ij}\right)_{1\le i,j\le 3}$ is a symmetric matrix with

$$\begin{aligned}
\Sigma_{011} &= \frac{p^2}{12}A(p),\ \Sigma_{012} = \Sigma_{021} = \frac{p^2q}{2}\left((n-1)q - mp\right),\ \Sigma_{013} = \Sigma_{031} = -\frac{p^2q^2}{4}(n+m-1) \\
\Sigma_{022} &= pq\left(nq^2 + mp^2 + pq\right), \Sigma_{023} = \Sigma_{032} = \frac{pq^2}{2}\left((m-1)p - nq\right), \Sigma_{033} = \frac{q^2}{12}A(q), \\
A(x) &= 6 + 4(n+m-2)x - 3(n+m-1)x^2.
\end{aligned}$$

Moreover, since $Var_0(\mathbf{c}'\mathbf{U}) = \mathbf{c}'\boldsymbol{\Sigma_0}\mathbf{c} \ge 0$ by definition, the matrix $\boldsymbol{\Sigma_0}$ is positive semi-definite. In practice, p is estimated by the pooled sample proportion $\hat{p} = (mp_1 + np_2)/(m+n)$ and both $E_0(\mathbf{U})$ and $Var_0(\mathbf{U})$ are calculated accordingly.

Appendix C Optimal Weights

From equation (17), we have

$$\mu_{1w} - \mu_{0w} = c_1\left(\pi_{t1}p_1p_2 - \frac{1}{2}p^2\right) + c_2\left(p_1q_2 - pq\right) + c_3\left(\pi_{x1}q_1q_2 - \frac{1}{2}q^2\right) = \mathbf{c}'\mu$$

where $\mathbf{c}' = (c_1, c_2, c_3)$, $c_1 + 2c_2 + c_3 = 1$, and $\mu' = \left(\pi_{t1}p_1p_2 - \frac{1}{2}p^2, p_1q_2 - pq, \pi_{x1}q_1q_2 - \frac{1}{2}q^2\right)$ and p is estimated by $\hat{p} = (mp_1 + np_2)/(m+n)$.

We assume that $det(\mathbf{\Sigma_0}) > 0$ i.e. $\mathbf{\Sigma_0}$ is positive-definite. Maximizing $\frac{|\mu_{1w} - \mu_{0w}|}{\sigma_{0w}}$, subject to $c_1 + 2c_2 + c_3 = 1$, with respect to $\mathbf{c}$ corresponds to maximizing the Lagrange function

$$O(\mathbf{c}, \lambda) = |\mathbf{c}'\mu| \left(\mathbf{c}'\Sigma_0\mathbf{c}\right)^{-\frac{1}{2}} - \lambda(\mathbf{c}'\mathbf{b} - 1)$$

with respect to the vector $\mathbf{c}$ and λ, where λ is the Lagrange multiplier and $\mathbf{b}' = (1, 2, 1)$.

Let $K(\mathbf{c}) = sign(\mathbf{c}'\mu)[(\mathbf{c}'\Sigma_0\mathbf{c})^{-\frac{3}{2}}]$, we have

$$\frac{\partial}{\partial \mathbf{c}} O(\mathbf{c}, \lambda) = K(\mathbf{c}) \left[(\mathbf{c}'\Sigma_0\mathbf{c})\mu - (\Sigma_0\mathbf{c})(\mathbf{c}'\mu)\right] - \lambda\mathbf{b} = 0 \tag{C.1}$$

$$\frac{\partial}{\partial \lambda} O(\mathbf{c}, \lambda) = \mathbf{c}'\mathbf{b} - 1 = 0 \tag{C.2}$$

From (C.1) and (C.2), we have

$$0 = \mathbf{c}' \left\{K(\mathbf{c}) \left[(\mathbf{c}'\Sigma_0\mathbf{c})\mu - (\Sigma_0\mathbf{c})(\mathbf{c}'\mu)\right] - \lambda\mathbf{b}\right\} = K(\mathbf{c}) \left[(\mathbf{c}'\Sigma_0\mathbf{c})\mathbf{c}'\mu - (\mathbf{c}'\Sigma_0\mathbf{c})(\mathbf{c}'\mu)\right] - \lambda\mathbf{c}'\mathbf{b} = \lambda,$$

because both $(\mathbf{c}'\Sigma_0\mathbf{c})$ and $(\mathbf{c}'\mu)$ are scalars and $\mathbf{c}'\mathbf{b} = c_1 + 2c_2 + c_3 = 1$.

Then, equation (C.1) implies $(\mathbf{c}'\Sigma_0\mathbf{c})\mu = (\Sigma_0\mathbf{c})(\mathbf{c}'\mu)$, i.e. $\mu = (\Sigma_0\mathbf{c})\frac{(\mathbf{c}'\mu)}{(\mathbf{c}'\Sigma_0\mathbf{c})} = \Sigma_0\frac{(\mathbf{c}'\mu)}{(\mathbf{c}'\Sigma_0\mathbf{c})}\mathbf{c}$. Since we assume that the matrix Σ_0^{-1} exists, this implies

$$\Sigma_0^{-1}\mu = \frac{(\mathbf{c}'\mu)}{(\mathbf{c}'\Sigma_0\mathbf{c})}\mathbf{c} \tag{C.3}$$

and thus, $\mathbf{b}'\Sigma_0^{-1}\mu = \frac{(\mathbf{c}'\mu)}{(\mathbf{c}'\Sigma_0\mathbf{c})}\mathbf{b}'\mathbf{c} = \frac{(\mathbf{c}'\mu)}{(\mathbf{c}'\Sigma_0\mathbf{c})}$.

Replacing $\frac{(\mathbf{c}'\mu)}{(\mathbf{c}'\Sigma_0\mathbf{c})}$ by $\mathbf{b}'\Sigma_0^{-1}\mu$ in equation (C.3) yields $\Sigma_0^{-1}\mu = (\mathbf{b}'\Sigma_0^{-1}\mu)\mathbf{c}$. Therefore, the optimal weight-vector is

$$\mathbf{c}_{opt} = \frac{\Sigma_0^{-1}\mu}{\mathbf{b}'\Sigma_0^{-1}\mu}, \tag{C.4}$$

as long as $\mathbf{b}'\Sigma_0^{-1}\mu \neq 0$. In addition,

$$\begin{aligned}
\frac{\partial^2}{\partial \mathbf{c}^2}\left[O(\mathbf{c})\right]_{\mathbf{c}=\mathbf{c}_{opt}} &= sign(\mathbf{c}'\mu)(\mathbf{c}'\Sigma_0^{-1}\mathbf{c})^{-\frac{3}{2}}\left[2(\mathbf{c}'\Sigma_0)\mu - \mu'(\Sigma_0\mathbf{c}) - \Sigma_0(\mathbf{c}'\mu)\right]_{\mathbf{c}=\mathbf{c}_{opt}} \\
&\quad -3sign(\mathbf{c}'\mu)(\Sigma_0\mathbf{c})(\mu'\Sigma_0^{-1}\mu)^{-\frac{5}{2}}\left[(\mathbf{c}'\Sigma_0\mathbf{c})\mu - (\Sigma_0\mathbf{c})(\mathbf{c}'\mu)\right]_{\mathbf{c}=\mathbf{c}_{opt}} \\
&= 2sign(\mathbf{c}'\mu)(\mu'\Sigma_0^{-1}\mu)^{-\frac{3}{2}}(\mathbf{b}'\Sigma_0^{-1}\mu)^2\left[\mu\mu' - (\mu'\Sigma_0^{-1}\mu)\Sigma_0\right]
\end{aligned}$$

$$= \quad 2sign(\mathbf{b}'\Sigma_0^{-1}\mu)(\mu'\Sigma_0^{-1}\mu)^{-\frac{3}{2}}(\mathbf{b}'\Sigma_0^{-1}\mu)^2\left[\mu\mu' - (\mu'\Sigma_0^{-1}\mu)\Sigma_0\right].$$

Since Σ_0 is positive definite, we can show that the border-preserving principal minors of order $k > 2$ have sign $(-1)^k$. Therefore, $\mathbf{c} = \mathbf{c}_{opt} = \dfrac{\Sigma_0^{-1}\mu}{\mathbf{b}'\Sigma_0^{-1}\mu}$ maximizes $O(\mathbf{c})$.

Let us define two vectors $\mathbf{d_1}' = (1, 1, 0)$ and $\mathbf{d_2}' = \mathbf{b}' - \mathbf{d_1}' = (0, 1, 1)$. To calculate w_1 and w_2, we just need to consider the relationships $\mathbf{c} = (w_1^2, w_1w_2, w_2^2)$ and $w_1 + w_2 = 1$. We have $\mathbf{d_1}'\mathbf{c} = w_1^2 + w_1(1 - w_1) = w_1$. Therefore, using the result given in equation (C.4), we can deduce $w_1 = \mathbf{d_1}'\mathbf{c} = \dfrac{\mathbf{d_1}'\Sigma_0^{-1}\mu}{\mathbf{b}'\Sigma_0^{-1}\mu}$ and $w_2 = 1 - \mathbf{d_1}'\mathbf{c} = \dfrac{(\mathbf{b}' - \mathbf{d_1}')\Sigma_0^{-1}\mu}{\mathbf{b}'\Sigma_0^{-1}\mu} = \dfrac{\mathbf{d_2}'\Sigma_0^{-1}\mu}{\mathbf{b}'\Sigma_0^{-1}\mu}$.

Appendix D Conditional Probabilities

D.1: Exponential Distribution

Suppose that the death times t_1, t_2 follow exponential distributions with hazards λ_1, λ_2, respectively and denote $\theta = \dfrac{\lambda_1}{\lambda_2}$, $q_1 = q_2^\theta$, and $q_2 = e^{-T\lambda_2}$. Given that $P(\delta_{1k} = 1) = p_1$, $P(\delta_{2l} = 1) = p_2$, we have

$$\begin{aligned}
\pi_{t1} &= P(t_{1k} < t_{2l}|\delta_{1k} = \delta_{2l} = 1) = (p_1p_2)^{-1}\int_0^T (1 - e^{-\lambda_1 u})\lambda_2 e^{-\lambda_2 u}du \\
&= \frac{1}{(1 - q_2^\theta)}\left[1 - \frac{1 - q_2^{(1+\theta)}}{(1 + \theta)(1 - q_2)}\right]; \\
\pi_{t2} &= P(t_{1k} < t_{2l}, t_{1k'} < t_{2l}|\delta_{1k} = \delta_{1k'} = \delta_{2l} = 1) = p_1^{-2}p_2^{-1}\int_0^T (1 - e^{-\lambda_1 u})^2\lambda_2 e^{-\lambda_2 u}du \\
&= (1 - q_2^\theta)^{-2}\left\{1 + \frac{1}{(1 - q_2)}\left[\frac{1 - q_2^{(1+2\theta)}}{1 + 2\theta} - \frac{2(1 - q_2^{(1+\theta)})}{1 + \theta}\right]\right\} \\
\pi_{t3} &= P(t_{1k} < t_{2l}, t_{1k} < t_{2l'}|\delta_{1k} = \delta_{2l} = \delta_{2l'} = 1) = p_1^{-1}p_2^{-2}\int_0^T (e^{-\lambda_2 T} - e^{-\lambda_2 u})^2\lambda_1 e^{-\lambda_1 u}du \\
&= \left(\frac{q_2}{1 - q_2}\right)^2\left[1 + \frac{\theta(1 - q_2^{(2+\theta)})}{(2 + \theta)(1 - q_2^\theta)q_2^2} - \frac{2\theta(1 - q_2^{(1+\theta)})}{(1 + \theta)(1 - q_2^\theta)q_2}\right]
\end{aligned}$$

D.2: Normal Distribution

Suppose that the non-fatal outcomes X_1, X_2 follow normal distributions $N(\mu_{x_1}, \sigma_{x_1})$ and $N(\mu_{x_2}, \sigma_{x_2})$, respectively.

Consider $\Delta_x = \dfrac{\mu_{x_2} - \mu_{x_1}}{\sqrt{\sigma_{x_1}^2 + \sigma_{x_2}^2}}$, $\rho_{x_j} = \dfrac{\sigma_{x_j}^2}{\sigma_{x_1}^2 + \sigma_{x_2}^2}$, and $Z_{kl} = \dfrac{X_{1k} - X_{2l} - (\mu_{x_1} - \mu_{x_2})}{\sqrt{\sigma_{x_1}^2 + \sigma_{x_2}^2}}$.

We can show that

$$\pi_{x1} = P(X_{1k} < X_{2l}) = \Phi(\Delta_x),$$
$$\pi_{x2} = P(X_{1k} < X_{2l}, X_{1k'} < X_{2l}) = P(Z_{kl} < \Delta_x,\ Z_{k'l} < \Delta_x),$$
$$\pi_{x3} = P(X_{1k} < X_{2l}, X_{1k} < X_{2l'}) = P(Z_{kl} < \Delta_x,\ Z_{kl'} < \Delta_x),$$

$$(Z_{kl}, Z_{k'l}) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{x_2} \\ \rho_{x_2} & 1 \end{pmatrix}\right) \text{ and } (Z_{kl}, Z_{kl'}) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{x_1} \\ \rho_{x_1} & 1 \end{pmatrix}\right).$$

Acknowledgements

Funding

This work was supported by grants P50-NS051343, R01-CA075971, T32 NS048005, and UL1TR001117 awarded by the National Institutes of Health. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official view of the National Institutes of Health.

Conflict of Interest: None declared.

References

1. Singhal AB. Normobaric Oxygen Therapy in Acute Ischemic Stroke Trial. ClinicalTrials.gov Database, 2006. [online] http://clinicaltrials.gov/ct2/show/NCT00414726.
2. Singhal AB. A review of oxygen therapy in ischemic stroke. *Neurological research* 2007; 29(2): 173–183.
3. Little RJ and Rubin DB. *Statistical analysis with missing data*. Wiley, 2002.
4. Lachin J. Worst-rank score analysis with informatively missing observations in clinical trials. *Controlled Clinical Trials* 1999; 20(5): 408–422.
5. McMahon R and Harrell F Jr. Power calculation for clinical trials when the outcome is a composite ranking of survival and a nonfatal outcome. *Controlled Clinical Trials* 2000; 21(4): 305–312.
6. Matsouaka RA and Betensky RA. Power and sample size calculations for the wilcoxon–mann–whitney test in the presence of death-censored observations. *Statistics in Medicine* 2015; 34(3): 406–431.
7. Felker GM and Maisel AS. A global rank end point for clinical trials in acute heart failure. *Circulation: Heart Failure* 2010; 3(5): 643–646.
8. Follmann D, Wittes J and Cutler JA. The use of subjective rankings in clinical trials with an application to cardiovascular disease.

Statistics in medicine 1992; 11(4): 427–437.

9. Bakal JA, Westerhout CM and Armstrong PW. Impact of weighted composite compared to traditional composite endpoints for the design of randomized controlled trials. *Statistical Methods in Medical Research* 2012; : 0962280211436004.
10. Hallstrom A, Litwin P and Douglas Weaver W. A method of assigning scores to the components of a composite outcome: an example from the MITI trial. *Controlled Clinical Trials* 1992; 13(2): 148–155.
11. Neaton J, Gray G, Zuckerman B et al. Key issues in end point selection for heart failure trials: composite end points. *Journal of Cardiac Failure* 2005; 11(8): 567–575.
12. Califf R and DeMets D. Principles from clinical trials relevant to clinical practice: Part I. *Circulation* 2002; 106(8): 1015.
13. Braunwald E, Cannon C and McCabe C. An approach to evaluating thrombolytic therapy in acute myocardial infarction. The 'unsatisfactory outcome' end point. *Circulation* 1992; 86(2): 683.
14. Moyé L. *Multiple analyses in clinical trials: fundamentals for investigators*. Springer Verlag, 2003. ISBN 038700727X.
15. Huang P, Tilley BC, Woolson RF et al. Adjusting o'brien's test to control type i error for the generalized nonparametric behrens–fisher problem. *Biometrics* 2005; 61(2): 532–539.
16. Häberle L, Pfahlberg A and Gefeller O. Assessment of multiple ordinal endpoints. *Biometrical Journal* 2009; 51(1): 217–226.
17. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984; : 1079–1087.
18. Wei L and Johnson W. Combining dependent tests with incomplete repeated measurements. *Biometrika* 1985; 72(2): 359.
19. Finkelstein D and Schoenfeld D. Combining mortality and longitudinal measures in clinical trials. *Statistics in medicine* 1999; 18(11): 1341–1354.
20. Moyé L, Davis B and Hawkins C. Analysis of a clinical trial involving a combined mortality and adherence dependent interval censored endpoint. *Statistics in Medicine* 1992; 11(13): 1705–1717.
21. Moyé LA, Lai D, Jing K et al. Combining censored and uncensored data in a u-statistic: design and sample size implications for cell therapy research. *The International Journal of Biostatistics* 2011; 7(1): 1–29.
22. Sampson UK, Metcalfe C, Pfeffer MA et al. Composite outcomes: weighting component events according to severity assisted interpretation but reduced statistical power. *Journal of clinical epidemiology* 2010; 63(10): 1156–1158.
23. Ahmad Y, Nijjer S, Cook CM et al. A new method of applying randomised control study data to the individual patient: A novel quantitative patient-centred approach to interpreting composite end points. *International Journal of Cardiology* 2015; 195: 216–224.
24. Wilson RF and Berger AK. Are all end points created equal?: The case for weighting. *Journal of the American College of Cardiology*

2011; 57(5): 546–548.

25. Armstrong PW, Westerhout CM, Van de Werf F et al. Refining clinical trial composite outcomes: An application to the assessment of the safety and efficacy of a new thrombolytic–3 (assent-3) trial. *American Heart Journal* 2011; 161(5): 848–854.
26. Minas G, Rigat F, Nichols TE et al. A hybrid procedure for detecting global treatment effects in multivariate clinical trials: theory and applications to fMRI studies. *Statistics in medicine* 2012; 31(3): 253–268.
27. Fisher LD. Self-designing clinical trials. *Statistics in medicine* 1998; 17(14): 1551–1562.
28. Ramchandani R, Schoenfeld DA and Finkelstein DM. Global rank tests for multiple, possibly censored, outcomes. *Biometrics* 2016; : s1–s10.
29. Lachin JM and Bebu I. Application of the wei–lachin multivariate one-directional test to multiple event-time outcomes. *Clinical Trials* 2015; : 1740774515601027.
30. Samson K. News from the AAN annual meeting: Why a trial of normobaric oxygen in acute ischemic stroke was halted early. *Neurology Today* 2013; 13(10): 34–35.
31. Freemantle N, Calvert M, Wood J et al. Composite outcomes in randomized trials: greater precision but with greater uncertainty? *Jama* 2003; 289(19): 2554.
32. Cordoba G, Schwartz L, Woloshin S et al. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *British Medical Journal* 2010; 341(aug18 3): c3920.
33. Tomlinson G and Detsky AS. Composite end points in randomized trials: there is no free lunch. *Jama* 2010; 303(3): 267–268.
34. Ferreira-Gonzalez I, Permanyer-Miralda G, Busse J et al. Composite outcomes can distort the nature and magnitude of treatment benefits in clinical trials. *Annals of internal medicine* 2009; 150(8): 566.
35. Ferreira-González I, Permanyer-Miralda G, Busse JW et al. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *Journal of clinical epidemiology* 2007; 60(7): 651–657.
36. Ferreira-González I, Permanyer-Miralda G, Domingo-Salvany A et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ* 2007; 334(7597): 786.
37. Lubsen J, Just H, Hjalmarsson A et al. Effect of pimobendan on exercise capacity in patients with heart failure: main results from the Pimobendan in Congestive Heart Failure (PICO) trial. *Heart* 1996; 76(3): 223.
38. Lubsen J and Kirwan BA. Combined endpoints: can we use them? *Statistics in Medicine* 2002; 21(19): 2959–2970.
39. Huque MF, Alosh M and Bhore R. Addressing multiplicity issues of a composite endpoint and its components in clinical trials. *Journal of biopharmaceutical statistics* 2011; 21(4): 610–634.
40. Mascha EJ and Turan A. Joint hypothesis testing and gatekeeping procedures for studies

with multiple endpoints. *Anesthesia & Analgesia* 2012; 114(6): 1304–1317.

41. Dmitrienko A, D'Agostino RB and Huque MF. Key multiplicity issues in clinical drug development. *Statistics in medicine* 2013; 32(7): 1079–1111.
42. Sankoh AJ, Li H and D'Agostino RB. Use of composite endpoints in clinical trials. *Statistics in medicine* 2014; 33(27): 4709–4714.
43. Logan B and Tamhane A. Superiority inferences on individual endpoints following noninferiority testing in clinical trials. *Biometrical Journal* 2008; 50(5): 693–703.
44. Röhmel J, Gerlinger C, Benda N et al. On testing simultaneously non-inferiority in two multiple primary endpoints and superiority in at least one of them. *Biometrical Journal* 2006; 48(6): 916–933.
45. Gómez G and Lagakos SW. Statistical considerations when using a composite endpoint for comparing treatment groups. *Statistics in medicine* 2013; 32(5): 719–738.
46. Gehan EA. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 1965; 52(1-2): 203–223.
47. Braunwald E, Antman EM, Beasley JW et al. ACC/AHA 2002 guideline update for the management of patients with unstable angina and non–st-segment elevation myocardial infarctionsummary article: a report of the american college of cardiology/american heart association task force on practice guidelines (committee on the management of patients with unstable angina). *Journal of the American College of Cardiology* 2002; 40(7): 1366–1374.
48. Grech E and Ramsdale D. Acute coronary syndrome: unstable angina and non-st segment elevation myocardial infarction. *BMJ (Clinical research ed)* 2003; 326(7401): 1259.
49. National Asthma Education and Prevention Program (National Heart, Lung, and Blood Institute) Third Expert Panel on the Management of Asthma. *Expert panel report 3: guidelines for the diagnosis and management of asthma*. US Department of Health and Human Services, National Institutes of Health, National Heart, Lung, and Blood Institute, 2007.
50. Van Elteren P. On the combination of independent two-sample tests of Wilcoxon. *Bull Int Stat Inst* 1960; 37: 351–361.
51. Zhao Y. Sample size estimation for the van Elteren test—a stratified Wilcoxon–Mann–Whitney test. *Statistics in medicine* 2006; 25(15): 2675–2687.
52. Senn S. Change from baseline and analysis of covariance revisited. *Statistics in medicine* 2006; 25(24): 4334–4344.
53. Fitzmaurice G. A conundrum in the analysis of change. *Nutrition* 2001; 17(4): 360–361.
54. van Breukelen GJ. Ancova versus change from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research* 2013; 48(6): 895–922.
55. Shahar E and Shahar DJ. Causal diagrams and change variables. *Journal of evaluation in clinical practice* 2012; 18(1): 143–148.
56. Pearl J. Lord's paradox revisited–(oh lord! kumbaya!). Technical report, Citeseer, 2014.

57. Oakes JM and Feldman HA. Statistical power for nonequivalent pretest-posttest designs the impact of change-score versus ancova models. *Evaluation Review* 2001; 25(1): 3–28.
58. Willett JB. Questions and answers in the measurement of change. *Review of research in education* 1988; 15: 345–422.
59. Bonate PL. *Analysis of pretest-posttest designs*. CRC Press, 2000.
60. Campbell DT and Kenny DA. *A primer on regression artifacts*. Guilford Publications, 1999.
61. Young FB, Weir CJ, Lees KR et al. Comparison of the national institutes of health stroke scale with disability outcome measures in acute stroke trials. *Stroke* 2005; 36(10): 2187–2192.
62. Adams Jr H, Davis P, Leira E et al. Baseline NIH Stroke Scale score strongly predicts outcome after stroke: a report of the Trial of Org 10172 in Acute Stroke Treatment (TOAST). *Neurology* 1999; 53(1): 126.
63. Bruno A, Saha C and Williams LS. Using change in the national institutes of health stroke scale to measure treatment effect in acute stroke trials. *Stroke* 2006; 37(3): 920–921.
64. Parsons M, Spratt N, Bivard A et al. A randomized trial of tenecteplase versus alteplase for acute ischemic stroke. *New England Journal of Medicine* 2012; 366(12): 1099–1107.
65. Brittain E, Palensky J, Blood J et al. Blinded subjective rankings as a method of assessing treatment effect: a large sample example from the systolic hypertension in the elderly program (SHEP). *Statistics in medicine* 1997; 16(6): 681–693.
66. Felker G, Anstrom K and Rogers J. A global ranking approach to end points in trials of mechanical circulatory support devices. *Journal of cardiac failure* 2008; 14(5): 368–372.
67. Allen LA, Hernandez AF, O'Connor CM et al. End points for clinical trials in acute heart failure syndromes. *Journal of the American College of Cardiology* 2009; 53(24): 2248–2258.
68. Sun H, Davison BA, Cotter G et al. Evaluating treatment efficacy by multiple endpoints in phase ii acute heart failure clinical trials: analyzing data using a global method. *Circulation: Heart Failure* 2012; : CIRCHEARTFAILURE–112.
69. Subherwal S, Anstrom KJ, Jones WS et al. Use of alternative methodologies for evaluation of composite end points in trials of therapies for critical limb ischemia. *American heart journal* 2012; 164(3): 277.
70. Berry JD, Miller R, Moore DH et al. The combined assessment of function and survival (cafs): a new endpoint for als clinical trials. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 2013; 14(3): 162–168.
71. Rosenbaum PR. Comment: the place of death in the quality of life. *Statistical Science* 2006; 21(3): 313–316.
72. Rubin DB. Rejoinder:[causal inference through potential outcomes and principal stratification: Application to studies with" censoring" due to death]. *Statistical Science* 2006; 21(3): 319–321.

73. Fay MP and Proschan MA. Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys* 2010; 4: 1.
74. Gail MH, Mark SD, Carroll RJ et al. On design considerations and randomization-based inference for community intervention trials. *Statistics in medicine* 1996; 15(11): 1069–1092.
75. Pratt JW. Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association* 1964; : 665–680.
76. Chung E and Romano JP. Asymptotically valid and exact permutation tests based on two-sample U-statistics. *Journal of Statistical Planning and Inference* 2016; 168: 97–105.
77. Brunner E and Munzel U. The nonparametric behrens-fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal* 2000; 42(1): 17–25.