Power and Sample Size Calculations for the Wilcoxon–Mann–Whitney Test in the Presence of Missing Observations due to Death

Roland A. Matsouaka^{a*} and Rebecca A. Betensky^a

We consider a clinical trial of a potentially lethal disease in which patients are randomly assigned to two treatment groups and are followed for a fixed period of time; a continuous endpoint is measured at the end of follow-up. For some patients, however, death (or severe disease progression) may preclude measurement of the endpoint. A statistical analysis that includes only patients with endpoint measurements may be biased. An alternative analysis includes all randomized patients, with rank scores assigned to the patients who are available for the endpoint measurement based on the magnitude of their responses and with “worst-rank” scores assigned to those patients whose death precluded the measurement of the continuous endpoint. The worst-rank scores are worse than all observed rank scores. The treatment effect is then evaluated using the Wilcoxon–Mann–Whitney (WMW) test. In this paper, we derive closed-form formulae for the power and sample size of the WMW test when missing measurements of the continuous endpoints due to death are replaced by worst-rank scores. We distinguish two approaches for assigning the worst-rank scores. In the tied worst-rank approach, all deaths are weighted equally and the worst-rank scores are set to a single value that is worse than all measured responses. In the untied worst-rank approach, the worst-rank scores further rank patients according to their time of death, so that an earlier death is considered worse than a later death, which in turn is worse than all measured responses. In addition, we propose four methods for implementation of the sample size formulae for a trial with expected early death. We conduct Monte Carlo simulation studies to evaluate the accuracy of our power and sample size formulae and to compare the four sample size estimation methods. Copyright © 0000 John Wiley & Sons, Ltd.

Keywords: Composite outcome; Informative missingness; Intention-to-treat; Worst-rank scores, Wilcoxon Rank Sum test

^a Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

* Correspondence to: Roland Matsouaka, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA.
E-mail: rmatsoua@hsph.harvard.edu

1. Introduction

In many randomized clinical trials, treatment benefit is evaluated through a post-randomization outcome measured at a pre-specified follow-up time (or times). For some patients, however, the post-randomization measurements may be missing due to a disease-related terminal event that occurs before the end of follow-up period. Such missing observations are informatively missing since they are related to the status of patient's underlying disease. This informative missing outcome conundrum is sometimes referred to in the literature as a truncation-by-death or censoring-by-death problem, and is handled differently from traditional missing data mechanisms [1, 2, 3]. Any statistical analysis based solely on the subset of completely observed measurements may provide a spurious estimate of the treatment effect because the subjects who survived may not be comparable between the treatment groups. For example, in a study of oxygen therapy in acute ischemic stroke (AIS) [4, 5] conducted at Massachusetts General Hospital, patients with imaging-confirmed AIS were randomized to treatment with normobaric oxygen therapy (NBO) or room air for 8 hours and assessed serially—during the follow-up period of 3 months—using the NIH stroke scale (NIHSS), a function rating scale used to quantify neurological deficit due to stroke, and MRI imaging. Unfortunately, in the case of stroke, irreversible cell death begins within minutes after stroke; the final outcome depend largely on where it occurs and its severity. The extent of penumbral tissue diminishes rapidly with time, causing serious disability or the death of some patients before they could be assessed at the follow-up times [6]. As both death from stroke and poor final NIHSS score are indicative of disease worsening, patients with missing follow-up NIHSS scores could not be excluded from the analysis. This informative censoring conundrum is referred to in the literature as

One solution to this problem of informative missingness is to include all randomized patients in the analysis, with rank scores assigned to the patients who are available for the post-randomization continuous measurement based on the magnitude of their responses and with “worst-rank” scores assigned to those patients whose death precluded the final measurement of the continuous response [7]. The worst-rank scores are worse than all observed rank scores. This is a composite outcome that combines assessment of the continuous response with death. We distinguish two approaches for assigning the worst-rank scores. In the untied worst-rank approach, the worst-rank scores further rank patients according to their time of death, so that an earlier death is considered worse than a later death, which is worse than all measured responses. In the tied worst-rank approach, all deaths are weighted equally and the worst-rank scores are set to a single value that is worse than all measured responses. The treatment effect is then evaluated using the Wilcoxon–Mann–Whitney (WMW) test.

The use of worst-rank composite endpoints is prevalent across disease areas and has become well-accepted and favored in many settings. For example, a 2013 paper advocated for the “Combined Assessment of Function and Survival (CAFS)” endpoint for ALS clinical trials [8]. The authors of this paper concluded that “CAFS is a robust statistical tool for ALS clinical trials and appropriately accounts for and weights mortality in the analysis of function.” A 2010 paper advocated for a global rank score that combines clinical events (e.g., death and hospitalization) with continuous biomarker measurements in acute heart failure studies [9]. These authors proposed a hierarchy of the individual endpoints that is incorporated into a composite rank endpoint. There are several other examples from the literature in the settings of HIV-AIDS trials [13, 14, 15], cardiovascular and heart failure trials [16, 17, 11, 9, 18, 19, 20], critical limb ischemia trials [21], and orthopedic trials [22].

The idea of assigning scores to informatively missing observations was first introduced by Gould [23] and was used by Richie et al. [24, 25] to handle cases of patients' informative withdrawal. Gould suggested using a rank-based test for the composite outcome scores. To avoid dealing with the multiple ties introduced by Gould's approach, Senn [26] proposed a modified version by assigning patients with missing observations ranks that depend on their times of withdrawal. Lachin is among those who extended the idea to settings in which disease-related withdrawal is due to death [7]. Lachin also explored properties of the WMW test applied to worst-rank score composite endpoints and demonstrated the unbiasedness of this approach for some restricted alternative hypotheses. Finally, McMahon and Harrell [27] derived power and sample

size calculations for the untied worst-rank score composite outcome of mortality and a non-fatal outcome measure.

To evaluate the power and sample size of the WMW test, it is necessary to specify the alternative to the null hypothesis and to calculate certain probabilities based on the alternative that are used in the mean and variance formulae. Generally this is a difficult task, especially when the underlying distributions are unknown [28]. Lehmann[29] and Noether [30] proposed some simplifications under the location shift alternative, when the location shift is small. Rosner and Glynn [31] considered a probit-shift alternative, which has some advantages over the shift alternative, and provided formulae for the necessary probabilities. Wang et al. [32] suggested using data from a pilot study to estimate the necessary probabilities.

To our knowledge, the paper by McMahon and Harrell [27] is the only paper that has attempted to establish the power and sample size formulae for the WMW test for a worst-rank score composite outcome. While the paper has the merit of addressing this important problem, the power formula as well as the sample size estimation approach proposed by McMahon and Harrell have some limitations. First, the variance estimator for the WMW U-statistic under the alternative hypothesis contains an error, which leads to an incorrect power estimation. Second, the sample size estimation and power calculation methods proposed in the paper rely heavily on the conservative assumption of no treatment effect on mortality (i.e., the “mortality neutral” assumption). This assumption is not tenable in practice. When there is an actual treatment effect on mortality, estimating the sample size under the mortality neutral assumption will unnecessarily and drastically inflate the sample size. In addition, even when there is not a significant difference in mortality, there may still be a moderate effect on mortality, which should be accounted for in the overall estimation of the treatment effect [33]. Finally, the paper does not provide power and sample size formulae for the tied worst-rank composite outcome, where missing observations are all set to a fixed (arbitrary) value. Our paper extends the results of McMahon and Harrell [27] as it considers both the tied and untied worst rank composite outcomes, provides an accurate variance estimator for use in power calculations, and estimates sample sizes without making the mortality neutral assumption.

In Section 2 we present the notation and hypotheses that are used throughout the paper. Next, in Section 3, we derive the analytical power formulae for the WMW test for both tied and untied worst-rank score outcomes. We also present a Monte Carlo simulation study that evaluates the accuracy of the proposed analytical formulae. In Section 4, we derive the sample size formulae for the WMW test for both tied and untied worst-rank composite outcomes. In addition, we present four different methods that have been used in the literature to estimate sample size for the traditional WMW test and extend them for estimation of sample size in the context of worst-rank composite outcomes. We report the results of Monte Carlo simulation studies to assess the validity of our sample size formulae and to evaluate and compare the accuracy of the proposed methods.

2. Notation and hypotheses

Suppose that N subjects are randomly assigned to either the control ($i = 1$) or the active treatment group ($i = 2$) and followed for a pre-specified period of time T . We denote by X_{ij} the primary outcome of interest in subject j in group i , which is the continuous response measured at time T . Without loss of generality, we assume that larger values of X correspond to improved medical condition. We denote the survival time for subject j in group i by t_{ij} , with the associated event indicator $\delta_{ij} = I(t_{ij} \leq T)$ of an event occurrence (prior to T). Note that X_{ij} is missing if $\delta_{ij} = 1$. We denote by $p_i = E(\delta_{ij}) = P(t_{ij} \leq T)$ the probability of the event (prior to T) in group i .

Under the untied worst-rank score composite outcome, any subject j in group i with missing measurement X_{ij} is assigned a value $\eta + t_{ij}$, with $\eta = \min(X) - 1 - T$. Thus, for each subject, the untied worst-rank adjusted value is given by (see Lachin [7])

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

By assigning these values, we ensure that (1) patients who have had an event prior to T are ranked appropriately by their

Statistics in Medicine

R. A. Matsouaka, R. A. Betensky

survival times and (2) patients with observed follow-up measurements are ranked above all those who had an event, based on their observed measurements.

Let F_i and G_i denote the conditional cumulative distributions of the event time and observed continuous response measure, respectively, 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 $\tilde{X}_{ij}$ is given by

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

We would like to test the null hypothesis that the two groups do not differ with respect to both survival times and the non-fatal outcome

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

versus the uni-directional alternative hypothesis

$$H_1 : G_1(x) \geq G_2(x) \text{ and } F_1(t) \geq F_2(t), \quad (2)$$

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

In the case of the tied worst-rank assignment, all the subjects with missing measurements are assigned a fixed score $\zeta = \min(X) - 1$. For each subject in the study, the tied worst-rank adjusted value is

$$\tilde{X}_{ij} = \delta_{ij} \zeta + (1 - \delta_{ij}) X_{ij}. \quad (3)$$

The cumulative distribution of $\tilde{X}_i$ is given by

$$\tilde{G}_i(x) = p_i I(x = \zeta) + (1 - p_i) G_i(x) I(x \neq \zeta).$$

The null and alternative hypotheses are defined by

$$H_0 : \tilde{G}_1(x) = \tilde{G}_2(x) \text{ versus } H_1 : \tilde{G}_1(x) \geq \tilde{G}_2(x), \quad (4)$$

$$\text{i.e.,} \quad H_0 : G_1(x) = G_2(x) \text{ and } p_1 = p_2 \text{ versus } H_1 : G_1(x) \geq G_2(x) \text{ and } p_1 \geq p_2,$$

where both $G_1(x) = G_2(x)$ and $p_1(t) = p_2(t)$ do not happen at the same time.

Throughout this paper, α denote the type I error, β the type II error, and z_p and Φ , respectively, the p th percentile and the cumulative distribution function of the standard normal distribution.

Note that we focus on a restricted alternative hypotheses (2) and (4), which are uni-directional in the sense that the active treatment effect is favorable on both the primary outcome of interest and mortality. In the case where higher values of the primary outcome of interest are indicative of better health outcome, this means subjects in the active treatment group tend to have higher values of the non-fatal outcome or longer survival times, while the distribution of the other outcome is either better in the active treatment group or (at least) equivalent in both the active and control treatment groups. These restricted alternatives are also the most considered in the literature [7, 27]. The cases in which the treatment has a favorable effect on survival, but a less favorable effect on the continuous response measure, or vice versa, do not provide a clear signal for a treatment benefit. Furthermore, regulatory agencies and investigators are reluctant to support a trial that does not emphasize mortality. The uni-directional alternative hypotheses we considered here allow us to avoid these ambiguous cases.

3. Power Calculation

In Section 3.1 we focus on the untied worst-rank test. In Section 3.1.1 we derive an analytical formula for the power of the untied worst-rank test. We evaluate the test, as well as the analytical approximation of its power in simulation studies in Section 3.1.2. In Section 3.2.1 we derive an analytical formula for the power of the tied worst-rank test. We evaluate the test, as well as the analytical approximation of its power in simulation studies in Section 3.2.2. Sections 3.2.1 and 3.2.2 contain the analogous results for the tied worst-rank test.

3.1. Untied Worst-Rank Scores

3.1.1. Power Formula Let $\tilde{X}_{1k}$ and $\tilde{X}_{2l}$ denote the worst-rank scores for the k th subject in group 1 and l th subject in group 2, respectively, that is $\tilde{X}_{1k} = (1 - \delta_{1k})X_{1k} + \delta_{1k}(\eta + t_{1k})$, $\tilde{X}_{2l} = (1 - \delta_{2l})X_{2l} + \delta_{2l}(\eta + t_{2l})$, for $k = 1, \dots, m$ and $l = 1, \dots, n$, where $N = m + n$. We define the WMW U-statistic by

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

Since $\tilde{X}_{1k} < \tilde{X}_{2l}$ when $\{t_{1k} < t_{2l} \text{ and } (\delta_{1k} = \delta_{2l} = 1)\}$, $\{\delta_{1k} = 1 \text{ and } \delta_{2l} = 0\}$, or $\{X_{1k} < X_{2l} \text{ and } (\delta_{1k} = \delta_{2l} = 0)\}$, we have $I(\tilde{X}_{1k} < \tilde{X}_{2l}) = I(t_{1k} < t_{2l}, \delta_{1k} = \delta_{2l} = 1) + I(\delta_{1k} = 1, \delta_{2l} = 0) + I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0)$. Therefore, U becomes

$$U = (mn)^{-1} \sum_{k=1}^m \sum_{l=1}^n [I(t_{1k} < t_{2l}, \delta_{1k} = \delta_{2l} = 1) + I(\delta_{1k} = 1, \delta_{2l} = 0) + I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0)]. \quad (5)$$

The equality $I(\tilde{X}_{1k} < \tilde{X}_{2l}) = I(t_{1k} < t_{2l}, \delta_{1k} = \delta_{2l} = 1) + I(\delta_{1k} = 1, \delta_{2l} = 0) + I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0)$ means that a patient assigned to the active treatment group (group 2) has a better outcome (and hence a better score) than a patient assigned to the control treatment group (group 1) if:

1. both patients died before time T , but the patient in the active treatment group lived longer;
2. the patient in the control group died before time T , while the patient in the active treatment group survived and had the primary outcome interest measured;
3. both patients survived until time T and had their primary outcomes of interest measured; however, the patient in the active treatment group had a higher measure of the primary outcome of interest.

We show, in Appendix A, equations (A.1) and (A.2), that the mean and variance of U under the alternative hypothesis are:

$$\begin{aligned} \mu_1 &= E(U) = p_1 p_2 \pi_{t1} + p_1 q_2 + q_1 q_2 \pi_{x1} = \pi_{U1}, \\ \sigma_1^2 &= Var(U) = (nm)^{-1} [\pi_{U1} (1 - \pi_{U1}) + (m - 1)(\pi_{U2} - \pi_{U1}^2) + (n - 1)(\pi_{U3} - \pi_{U1}^2)], \end{aligned} \quad (6)$$

$$\begin{aligned} \pi_{U1} &= p_1 p_2 \pi_{t1} + p_1 q_2 + q_1 q_2 \pi_{x1} \\ \pi_{U2} &= p_1^2 q_2 + p_1^2 p_2 \pi_{t2} + 2p_1 q_1 q_2 \pi_{x1} + q_1^2 q_2 \pi_{x2} \\ \pi_{U3} &= p_1 q_2^2 + p_1 p_2^2 \pi_{t3} + 2p_1 p_2 q_2 \pi_{t1} + q_1 q_2^2 \pi_{x3}, \end{aligned} \quad (7)$$

where $p_i = P(t_{ij} \leq T)$, $q_i = 1 - p_i$, for $i = 1, 2$, and $\pi_{t1} = P(t_{1k} < t_{2l} | \delta_{1k} = \delta_{2l} = 1)$;

$$\pi_{t2} = P(t_{1k} < t_{2l}, t_{1k'} < t_{2l} | \delta_{1k} = \delta_{1k'} = \delta_{2l} = 1); \quad \pi_{t3} = P(t_{1k} < t_{2l}, t_{1k} < t_{2l'} | \delta_{1k} = \delta_{2l} = \delta_{2l'} = 1)$$

$$\pi_{x1} = P(X_{1k} < X_{2l}); \quad \pi_{x2} = P(X_{1k} < X_{2l}, X_{1k'} < X_{2l}); \quad \pi_{x3} = P(X_{1k} < X_{2l}, X_{1k} < X_{2l'}).$$

Under the null hypothesis of no difference between the two treatment groups, we have $p_1 = p_2$, $q_1 = q_2$, $\pi_{t1} = \pi_{x1} = 1/2$, and $\pi_{t2} = \pi_{x2} = \pi_{t3} = \pi_{x3} = 1/3$, which imply that $\pi_{U1} = 1/2$ and $\pi_{U2} = \pi_{U3} = 1/3$. Hence,

$$\mu_0 = E_0(U) = \frac{1}{2} \quad \text{and} \quad \sigma_0^2 = Var_0(U) = \frac{n + m + 1}{12mn}. \quad (8)$$

To evaluate the power of the WMW for the untied worst-rank score composite outcome, we use the asymptotic distribution of the WMW test statistic

$$Z = \frac{U - E_0(U)}{\sqrt{Var_0(U)}}, \quad (9)$$

which converges to the standard normal distribution $N(0, 1)$ as m and n tend to infinity, assuming that n/N is bounded away from 0.

Practical guidelines for how large m and n should be in order for the normal approximation such as in (9) to be valid for a traditional Wilcoxon test are provided in the literature. Siegel and Castellan [34] recommended using the approximation for either $n = 3$ or 4 and $m > 12$ or for $n > 4$ and $m > 10$. Ballera et al. [35] recommended use of a graphical display to assess the normal approximation.

The power of the WMW test statistic (9) is given by

$$\Phi\left(\frac{\sigma_0}{\sigma_1} z_{\frac{\alpha}{2}} + \frac{\mu_1 - \mu_0}{\sigma_1}\right) + \Phi\left(\frac{\sigma_0}{\sigma_1} z_{\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\sigma_1}\right) \approx \Phi\left(\frac{\sigma_0}{\sigma_1} z_{\frac{\alpha}{2}} + \frac{|\mu_1 - \mu_0|}{\sigma_1}\right). \quad (10)$$

To calculate the conditional probabilities $\pi_{t\gamma}$ and $\pi_{x\gamma}$, $\gamma = 1, 2, 3$, we need to specify the distributions of the survival times, t , and the non-fatal continuous response, X , or make some additional assumptions. For instance, we might assume that $X_{ik} \sim N(\mu_{x_i}, \sigma_{x_i})$ and $t_{ik} \sim Exp(\lambda_i)$

3.1.2. Remarks

- (1) At this point, it is worth mentioning that the U -statistic considered here is exactly the same as the one used by McMahon and Harrell [27]. In their paper, U is defined as

$$U = (mn)^{-1} \sum_{k=1}^m \sum_{l=1}^n [I(t_{1k} < t_{2l})I(\delta_{1k} = 1 \text{ or } \delta_{2l} = 1) + I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0)], \quad (11)$$

which is equivalent to (5).

Indeed, notice that

$$I(\delta_{1k} = 1 \text{ or } \delta_{2l} = 1) = I(\delta_{1k} = 1 \text{ and } \delta_{2l} = 1) + I(\delta_{1k} = 0 \text{ and } \delta_{2l} = 1) + I(\delta_{1k} = 1 \text{ and } \delta_{2l} = 0).$$

Since patients who survived have longer survival time than those who died, we have

$$I(t_{1k} < t_{2l})I(\delta_{1k} = 0 \text{ and } \delta_{2l} = 1) = 0 \times I(\delta_{1k} = 0 \text{ and } \delta_{2l} = 1) = 0; \quad (12)$$

$$I(t_{1k} < t_{2l})I(\delta_{1k} = 1 \text{ and } \delta_{2l} = 0) = 1 \times I(\delta_{1k} = 1 \text{ and } \delta_{2l} = 0) = I(\delta_{1k} = 1 \text{ and } \delta_{2l} = 0).$$

where $\delta_{1k} = 1$ indicates death, while $\delta_{1j} = 0$ indicates survival and measurement of the primary outcome of interest. Thus,

$$I(t_{1k} < t_{2l})I(\delta_{1k} = 1 \text{ or } \delta_{2l} = 1) = I(t_{1k} < t_{2l})I(\delta_{1k} = \delta_{2l} = 1) + I(\delta_{1k} = 1, \delta_{2l} = 0) \quad (13)$$

and not equal to $I(t_{1k} < t_{2l}) [1 - I(\delta_{1k} = 0, \delta_{2l} = 0)]$ as they suggested.

Plugging (13) into (11), the U -statistic becomes

$$U = (mn)^{-1} \sum_{k=1}^m \sum_{l=1}^n [I(t_{1k} < t_{2l}, \delta_{1k} = \delta_{2l} = 1) + I(\delta_{1k} = 1, \delta_{2l} = 0) + I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0)],$$

which is exactly the U -statistic we have defined in (5).

Now, using the key result (13) we can show that the probability $P(t_{1k} < t_{2l}, (\delta_{1k} = 1 \text{ or } \delta_{2l} = 1))$ is not equal to $P(t_{1k} < t_{2l} | \delta_{1k} = 1 \text{ or } \delta_{2l} = 1) [1 - P(\delta_{1k} = \delta_{2l} = 0)]$ but equal to $P(t_{1k} < t_{2l} | \delta_{1k} = \delta_{2l} = 1) P(\delta_{1k} = \delta_{2l} = 1) + P(\delta_{1k} = 1, \delta_{2l} = 0)$ instead. Therefore, the expected value $E(U)$ obtained in McMahon and Harrell's paper (and denoted here by) $\pi_{U1}^* = [1 - P(\delta_{1k} = \delta_{2l} = 0)] \pi_{t1}^* + P(\delta_{1k} = \delta_{2l} = 0) \pi_{x1} = (1 - q_1 q_2) \pi_{t1}^* + q_1 q_2 \pi_{x1}$ is incorrect, where $\pi_{t1}^* = P(t_{1k} < t_{2l} | \delta_{1k} = 1 \text{ or } \delta_{2l} = 1)$.

In addition, their results for π_{U2} , and π_{U3} (denoted here by π_{U2}^* , and π_{U3}^*)

$$\begin{aligned} \pi_{U2}^* &= [1 - P(\delta_{1k} = \delta_{1k'} = \delta_{2l} = 0)] \pi_{t2}^* + P(\delta_{1k} = \delta_{1k'} = \delta_{2l} = 0) \pi_{x2} = (1 - q_1^2 q_2) \pi_{t2}^* + q_1^2 q_2 \pi_{x2} \\ \pi_{U3}^* &= [1 - P(\delta_{1k} = \delta_{2l} = \delta_{2l'} = 0)] \pi_{t3}^* + P(\delta_{1k} = \delta_{2l} = \delta_{2l'} = 0) \pi_{x3} = (1 - q_1 q_2^2) \pi_{t3}^* + q_1 q_2^2 \pi_{x3}, \end{aligned}$$

are both also incorrect, where they defined $\pi_{t2}^* = P(t_{1k} < t_{2l}, t_{1k'} < t_{2l} | \text{at least one of } \delta_{1k}, \delta_{1k'}, \delta_{2l} \text{ is equal to } 1)$ and $\pi_{t3}^* = P(t_{1k} < t_{2l}, t_{1k} < t_{2l'} | \text{at least one of } \delta_{1k}, \delta_{2l}, \delta_{2l'} \text{ is equal to } 1)$. Indeed, unlike π_{U1}^* , π_{U2}^* , and π_{U3}^* , the probabilities π_{U1} , π_{U2} , and π_{U3} have each an additional term that depends only on p_1 and q_2 . Furthermore, π_{U2} depends also on π_{x1} , and π_{U3} depends on π_{t1} . In fact, we have used arguments similar to (12) in Appendix A to show that both π_{U2}^* and π_{U3}^* are not only different from π_{U2} and π_{U3} , but also incorrect.

- (2) For any given values of q_1 and q_2 in (7), there exist pairs (π_{t1}, π_{x1}) other than $(0.5, 0.5)$ for which the Mann-Whitney worst rank test statistic has an expected value $\pi_{U1} = (1 - q_1)(1 - q_2) \pi_{t1} + q_1 q_2 \pi_{x1} + (1 - q_1) q_2$ equal to 0.5. For instance, when the effects of treatment on the components of the worst-rank composite outcome are in opposite directions and show lack of consistency, such as a treatment that increases mortality while improving significantly the primary outcome of interest. This is the very reason we focused on a particular alternative hypothesis, the uni-directional alternative, in which the active treatment fares better with respect to the (non-fatal) primary outcome of interest and the incidence of death. In the scenario where higher values of the non-fatal outcome reflect better health condition, this means that subjects in the active treatment group tend to have higher values on the non-fatal outcome *and/or* longer survival times while not having lower values for either.

We are not interested here to test the null hypothesis of no difference on both mortality and the primary outcome of interest against the complete complement of the null hypothesis (the omnibus alternative hypothesis), but against the (restrictive) one-directional alternative hypothesis defined by (2). In this context, under the latter alternative hypothesis, we expect the pairs (π_{t1}, π_{x1}) to belong to the set $[0.5, 1] \times [0.5, 1] - \{(0.5, 0.5)\}$.

- (3) One may object that in practice, it is not possible to know in advance whether a specific drug would satisfy such a requirement and that we are doomed to produce misleading results when the treatment is inconsistent in the sense the overall treatment effect on the composite outcome might appear beneficial, while its harmful effect on one of the components of the composite outcome may go unnoticed. We have recently revised and re-submitted a paper to Statistics in Medicine [36] that gives practical guidelines and highlights two stepwise procedures that dealt with this issue. These procedures ensure that a treatment is declared superior only when it has a beneficial effect on either the primary outcome of interest or mortality and does not fare worse on either of them.

3.1.3. Simulation Study We conducted simulation studies to assess the accuracy of the power formula (10), using a two-sided $\alpha = 0.05$. In each of 10,000 simulated trials, we generated exponentially distributed survival times and normally distributed primary outcomes of interest. We estimated the power as a function of the hazard ratio (HR), the standardized difference in the primary outcome of interest $\Delta_x = (\mu_{x_2} - \mu_{x_1})/(2^{1/2}\sigma_{x_1})$, and the survival probability in the active treatment group, $q_2 = 1 - p_2$, for $n = m = 50$. Specifically, we generated $t_i \sim \text{Exp}(\lambda_i)$, $i = 1, 2$, such that $q_2 = \exp(-\lambda_2 T) = 0.6, 0.8$, $T = 3$, $\text{HR} = \lambda_1/\lambda_2 = 1.0, 1.2, 1.4, 1.6, 2.0, 2.4, 3.0$, $X_1 \sim N(0, 1)$ and $X_2 \sim N(2^{1/2}\Delta_x, 1)$, for $\Delta_x = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$. Calculations for the conditional probabilities $\pi_{t\gamma}$ and $\pi_{x\gamma}$ for $\gamma = 1, 2, 3$ are given in Appendix B. The analytical power values based on (10) are listed in Part (1) of Table 1. For each data set, we computed the WMW test statistic, Z , given in equation (9). We also calculated the empirical power as the proportion of simulated data sets for which $|Z| > 1.96$. For the sake of comparison, we also calculated the empirical power for the WMW test using the measured primary outcomes of interest of the survivors, i.e., reducing the sample to include survivors only. These results are presented in Part (2) of Table 1.

Table 1. Power for the Wilcoxon–Mann–Whitney test for untied worst-rank scores.

Δ_x	$q_2 = 60\%$							$q_2 = 80\%$							$q_1 = q_2 = 100\%$		
	1.0	1.2	1.4	HR	1.6	2.0	2.4	3.0	1.0	1.2	1.4	HR	1.6	2.0	2.4	3.0	
(1) Analytical power																	
0.0	0.05	0.08	0.17	0.30	0.63	0.87	0.99		0.05	0.06	0.09	0.14	0.28	0.47	0.73		0.05
0.1	0.06	0.11	0.22	0.37	0.69	0.90	0.99		0.07	0.11	0.17	0.24	0.42	0.60	0.83		0.08
0.2	0.08	0.16	0.28	0.44	0.75	0.93	1.00		0.14	0.20	0.28	0.37	0.56	0.73	0.90		0.16
0.3	0.11	0.21	0.36	0.52	0.80	0.95	1.00		0.25	0.34	0.43	0.52	0.70	0.83	0.94		0.31
0.4	0.16	0.28	0.43	0.59	0.84	0.96	1.00		0.40	0.49	0.58	0.67	0.81	0.90	0.97		0.49
0.5	0.22	0.36	0.51	0.66	0.88	0.97	1.00		0.56	0.65	0.72	0.79	0.89	0.95	0.99		0.68
0.6	0.29	0.43	0.58	0.72	0.91	0.98	1.00		0.71	0.78	0.83	0.88	0.94	0.98	1.00		0.83
(2) Survivors-only power																	
0.0	0.05	0.05	0.05	0.05	0.05	0.05	0.05		0.05	0.05	0.05	0.05	0.05	0.05	0.05		0.05
0.1	0.08	0.08	0.07	0.07	0.07	0.06	0.06		0.10	0.10	0.10	0.10	0.10	0.10	0.09		0.08
0.2	0.18	0.18	0.17	0.16	0.14	0.13	0.11		0.25	0.25	0.25	0.25	0.25	0.23	0.23		0.16
0.3	0.34	0.33	0.31	0.30	0.27	0.24	0.21		0.50	0.48	0.48	0.48	0.47	0.46	0.45		0.31
0.4	0.54	0.52	0.50	0.48	0.43	0.38	0.31		0.74	0.73	0.72	0.71	0.71	0.70	0.68		0.49
0.5	0.74	0.71	0.69	0.66	0.59	0.55	0.45		0.90	0.89	0.89	0.89	0.88	0.87	0.86		0.68
0.6	0.87	0.86	0.84	0.81	0.76	0.70	0.60		0.97	0.97	0.97	0.97	0.97	0.96	0.96		0.83

We used exponential distributions for the survival times, normal distributions for the non-fatal outcomes, with 50 subjects in each group.

Δ_x : standardized difference in the primary outcome of interest; HR: hazard ratio; q_1 (resp. q_2): survival probability in the control (resp. active) treatment group.

(1) Analytic power formula (10); (2) Empirical power obtained by reducing the sample to include survivors only.

First, the simulations indicate that for exponentially distributed survival times and normally distributed non-fatal outcomes, there is virtually no difference between the empirical power and the analytical power calculated using formula (10). The differences in power, for any value of the hazards ratio or the standardized difference in the primary outcome of interest is less than 0.01. Thus, as long as the survival probability and outcome distributions are well-approximated, our analytical power formula will be accurate in the context of untied worst-rank score outcome.

Secondly, the power of the WMW test based on the composite outcome increases considerably as the hazard ratio increases, for any given difference in the primary outcome of interest. This is due to its use of the actual death times, which contribute treatment information for large hazard ratios; this effect is larger for smaller survival probabilities (i.e., 60% versus 80%), as it translates into more early deaths. This also provides a compelling argument against the "mortality neutral" assumption in designing a study.

Thirdly, the (ordinary) WMW test in case of no deaths has more power compare to the untied worst-rank WMW test with deaths when there is a null or moderate effect of treatment on survival ($\text{HR} < 1.4$) and the difference on the primary outcome of interest is moderate or high ($\Delta_x > 0.3$). However, when the the hazard ratio of death is high ($\text{HR} \geq 2$) or when

it is moderate $1.4 \leq \text{HR} < 2$ and the difference on the primary outcome of interest is low ($\Delta_x \leq 0.2$), the WMW test on the untied worst-rank composite outcome is more powerful.

Fourthly, when subjects with missing observations due to death are excluded from the analysis, the power of detecting a treatment difference decreases as the hazard ratio increases, with lower power when the overall probability of survival is 60% versus 80%, due to a smaller remaining sample size in the former case. In addition, this power is extremely low when the standardized difference in the primary outcome of interest is low or moderate and the hazard ratio of death is moderate or high. This proves that the results are unreliable when informatively missing observations due to death are excluded from the analyses.

Although inefficient (and inappropriate in case of informatively missing observations), the “survivor-only” analysis is a natural approach that has been used in the literature [37] and may be done in practice. The simulation results highlight the loss of power associated with this approach.

3.2. Tied Worst-Rank Scores

3.2.1. Power Formula Under the tied worst-rank score assignment procedure, we set $\tilde{X}_{1k} = (1 - \delta_{1k})X_{1k} + \delta_{1k}\zeta$, $\tilde{X}_{2l} = (1 - \delta_{2l})X_{2l} + \delta_{2l}\zeta$, for $k = 1, \dots, m$, $l = 1, \dots, n$, and $\zeta = \min(X) - 1$.

We define the WMW U-statistic as

$$\tilde{U} = (mn)^{-1} \sum_{k=1}^m \sum_{l=1}^n \left[I(\tilde{X}_{1k} < \tilde{X}_{2l}) + \frac{1}{2} I(\tilde{X}_{1k} = \tilde{X}_{2l}) \right] \quad (14)$$

Note that $\tilde{X}_{1k} < \tilde{X}_{2l}$ when $\{X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0\}$ or $(\delta_{1k} = 1, \delta_{2l} = 0)$, whereas $\tilde{X}_{1k} = \tilde{X}_{2l}$ when $(\delta_{1k} = \delta_{2l} = 1)$. Thus, $I(\tilde{X}_{1k} < \tilde{X}_{2l}) + \frac{1}{2} I(\tilde{X}_{1k} = \tilde{X}_{2l}) = I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0) + I(\delta_{1k} = 1, \delta_{2l} = 0) + \frac{1}{2} I(\delta_{1k} = \delta_{2l} = 1)$. This means that a patient assigned to the active treatment group (group 2) has a better score compare to a patient assigned to the control treatment group (group 1) if:

1. both patients survive until time T , but the patient in the active treatment group has a higher measure of the primary outcome of interest;
2. the patient in the control group dies before time T , while the patient in the active treatment group survives.

In Appendix A, equations (A.3)-(A.4), we show that

$$\begin{aligned} \mu_1 &= E(\tilde{U}) = q_1 q_2 \pi_{x1} + p_1 q_2 + \frac{1}{2} p_1 p_2 = \pi_{\tilde{U}1}, \\ \sigma_1^2 &= \text{Var}(\tilde{U}) = (nm)^{-1} \left[\pi_{\tilde{U}1} (1 - \pi_{\tilde{U}1}) + (m-1) \left(\pi_{\tilde{U}2} - \pi_{\tilde{U}1}^2 - \frac{p_1^2 p_2^2}{12} \right) + (n-1) \left(\pi_{\tilde{U}3} - \pi_{\tilde{U}1}^2 - \frac{p_1 p_2^2}{12} \right) - \frac{p_1 p_2}{4} \right], \end{aligned} \quad (15)$$

where $\pi_{x\gamma}$, $\gamma = 1, 2, 3$, are defined as in (7), and

$$\begin{aligned} \pi_{\tilde{U}1} &= q_1 q_2 \pi_{x1} + p_1 q_2 + \frac{1}{2} p_1 p_2 \\ \pi_{\tilde{U}2} &= \pi_{x2} q_1^2 q_2 + 2\pi_{x1} p_1 q_1 q_2 + p_1^2 q_2 + \frac{1}{3} p_1^2 p_2 \\ \pi_{\tilde{U}3} &= \pi_{x3} q_1 q_2^2 + p_1 q_2^2 + p_1 p_2 q_2 + \frac{1}{3} p_1 p_2^2. \end{aligned}$$

The additional term $p_1 p_2 [3 + (m-1)p_1 + (n-1)p_2]/12$ in the variance formula represents the correction for ties.

Under the null hypothesis of no difference between the two treatment groups, $p_1 = p_2$, $q_1 = q_2$, $\pi_{x1} = 1/2$, and $\pi_{x2} = \pi_{x3} = 1/3$, which implies $\pi_{\tilde{U}1} = 1/2$ and $\pi_{\tilde{U}2} = \pi_{\tilde{U}3} = 1/3$. Therefore,

$$\mu_0 = E_0(\tilde{U}) = \frac{1}{2}, \quad \sigma_0^2 = \text{Var}_0(\tilde{U}) = \frac{1}{12mn} \{ (n+m+1) - p^2 [3 + (m+n-2)p] \}. \quad (16)$$

In practice, p is estimated by the pooled sample proportion $p = (mp_1 + np_2)/(m + n)$ and $Var_0(\tilde{U})$ is calculated accordingly.

As in the case of untied worst-rank scores, the distribution of the WMW test statistic,

$$\tilde{Z} = \frac{\tilde{U} - E_0(\tilde{U})}{\sqrt{Var_0(\tilde{U})}},$$

converges to a standard normal distribution as m and n tend to infinity.

The power of the WMW test is given by

$$\Phi\left(\frac{\sigma_0}{\sigma_1}z_{\frac{\alpha}{2}} + \frac{\mu_1 - \mu_0}{\sigma_1}\right) + \Phi\left(\frac{\sigma_0}{\sigma_1}z_{\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\sigma_1}\right) \approx \Phi\left(\frac{\sigma_0}{\sigma_1}z_{\frac{\alpha}{2}} + \frac{|\mu_1 - \mu_0|}{\sigma_1}\right). \quad (17)$$

3.2.2. Simulation Study We conducted simulation studies to assert the validity of the power formula (17) for the tied worst-rank test, based on exponentially distributed death times and normally distributed non-fatal outcomes, with a two-sided $\alpha = 0.05$. We considered the same parameters as in Section 3.1 and generated a total of 10,000 trials. The results are presented in Table 2.

Table 2. Power for the Wilcoxon–Mann–Whitney test for tied worst-rank scores.

Δ_x	$q_2 = 60\%$							$q_2 = 80\%$							$q_1 = q_2 = 100\%$	
	1.0	1.2	1.4	HR	1.6	2.0	2.4	3.0	1.0	1.2	1.4	HR	1.6	2.0		2.4
	(1) Analytical power															
0.0	0.05	0.08	0.17	0.30	0.61	0.84	0.97	0.05	0.06	0.09	0.14	0.28	0.47	0.73		0.05
0.1	0.06	0.12	0.23	0.37	0.67	0.87	0.98	0.07	0.11	0.17	0.24	0.42	0.60	0.82		0.08
0.2	0.08	0.17	0.29	0.45	0.73	0.90	0.99	0.14	0.21	0.28	0.37	0.56	0.73	0.89		0.16
0.3	0.12	0.23	0.37	0.53	0.79	0.93	0.99	0.26	0.34	0.43	0.52	0.70	0.83	0.94		0.31
0.4	0.17	0.30	0.45	0.60	0.83	0.94	0.99	0.40	0.50	0.59	0.67	0.81	0.90	0.97		0.49
0.5	0.23	0.37	0.53	0.67	0.87	0.96	0.99	0.57	0.65	0.73	0.79	0.89	0.95	0.99		0.68
0.6	0.31	0.45	0.60	0.73	0.90	0.97	1.00	0.71	0.78	0.84	0.88	0.94	0.98	0.99		0.83

We used exponential distributions for the survival times, normal distributions for the non-fatal outcomes, with 50 subjects in each group.

Δ_x : standardized difference in the primary outcome of interest; HR: hazard ratio; q_1 (resp. q_2): survival probability in the control (resp. active) treatment group.

(1) Analytical power formula (17);

As in the case of the untied worst-rank score test, there is virtually no difference between the empirical power and the analytical power calculated using formula (17). We also note that the power of the tied worst-rank test is nearly identical to that of the untied test for these simulation scenarios. There are instances where the actual time to early death (before the end of follow-up) may not be very important or in case of drop-out due worsening of the disease conditions, the time of drop-out may be hard to assess for practical reasons. The simulation results we have obtained are reassuring since they demonstrate that, in these cases, one does not lose much when considering using tied worst-rank scores—by lumping all informative missing observations together to shared the same tied rank value—in lieu of untied worst-rank scores. The untied worst-rank test has higher power than the tied worst-rank test when there is substantial early death. This would not be expected in most clinical trials, as if it were, a survival analysis would likely be the primary endpoint, rather than an continuous response. This is depicted in Figure 1, for $n = m = 50$, $\Delta_x = 0.1$ and survival probabilities in the active treatment group of $q_2 = 30\%, 40\%, 60\%$, and 80% .

3.2.3. Comments The result that there is only a minor computational advantage in using the tied worst-rank composite outcome instead of the untied worst-rank composite outcome is extremely important. Up to now, it was argued in the literature—without proof—that by using the untied worst-rank composite outcome in lieu of tied worst-rank composite

Figure 1. Power comparison for the Wilcoxon–Mann–Whitney test under tied (solid line) and untied (dashed line) worst-rank scores where survival times follow exponential distribution and the non-fatal outcome follow normal distribution. The survival probabilities in the active treatment group are: (a) 30%; (b) 40%; (c) 60%; (d) 80%. There are 50 subjects in each group with $\Delta_x = 0.1$.

composite outcome one may gain some efficiency (see [7].) Our simulations contradict this assertion. Even though assigning a fixed score to all missing observations—under the tied worst-rank score imputation procedure (as oppose to the untied worst-rank score imputation)—results in lower sum of ranks in the active treatment group when there is truly a treatment effect on mortality, this goes along with lower variances of the MWM test statistic under the null and alternative hypotheses (due to the multiple ties induced by the imputation procedure.) As a result, the powers under both untied and tied imputation procedures are similar.

Usually, the decision to use either tied or untied worst-rank scores is not based merely on the desire to increase power. In practice, it depends also on the nature the disease and the willingness by the investigators to weights equally all deaths—be it, for example, two days after the treatment initiation or ten months later. Therefore, the choice of the imputation method should be made by eliciting and using experts opinions. Furthermore, the use of worst-rank composite outcome has been advocated in many settings including clinical trials where the treatment effects on a disease are assessed through a host of meaningful—yet of different importance, magnitude, and impact—heterogenous clinical events (hospitalization, death, ...) and clinical endpoints that provide a more comprehensive understanding of these effects (see [11, 9]). No matter what the situation, it is reassuring to know that a Wilcoxon–Mann–Whitney test based on tied worst-rank scores is almost as efficient as the one based on untied worst-rank scores. In our knowledge, this surprisingly encouraging finding has not been established so far in the literature.

4. Sample Size Calculation

4.1. Sample Size formula

Let $N = n + m$ be the total sample size and let $s = n/N$ be the active treatment fraction. In the case of untied worst-rank scores, if we ignore the lower order terms, the equations (6)-(8) for means and variances can be re-written so that:

$$\mu_1 - \mu_0 = u, \quad \sigma_0^2 \approx \frac{1}{12s(1-s)N}, \quad \sigma_1^2 \approx \frac{v_1}{s(1-s)N},$$

with

$$u = \pi_{U1} - \frac{1}{2}, \quad v_1 = (1-s)(\pi_{U2} - \pi_{U1}^2) + s(\pi_{U3} - \pi_{U1}^2).$$

Setting the power to $1 - \beta$, from formula (10), we derive $\frac{|\mu_1 - \mu_0|}{\sigma_0} = -\left(\frac{\sigma_1}{\sigma_0} z_\beta + z_{\frac{\alpha}{2}}\right) = \left(\frac{\sigma_1}{\sigma_0} z_{1-\beta} + z_{1-\frac{\alpha}{2}}\right)$, which implies that

$$N = \left(\frac{z_{1-\frac{\alpha}{2}} + \sqrt{12v_1} z_{1-\beta}}{u \sqrt{12s(1-s)}} \right)^2. \quad (18)$$

For the tied worst-rank test, if we also ignore the lower order terms, equations (15) and (16) yield

$$\mu_1 - \mu_0 = \tilde{u}, \quad \sigma_0^2 \approx \frac{\tilde{v}_0}{12s(1-s)N}, \quad \sigma_1^2 \approx \frac{\tilde{v}_1}{s(1-s)N},$$

where $\tilde{u} = \pi_{\tilde{U}1} - \frac{1}{2}$, $\tilde{v}_0 = 1 - p^3$, $\tilde{v}_1 = (1-s)(\pi_{\tilde{U}2} - \pi_{\tilde{U}1}^2) + s(\pi_{\tilde{U}3} - \pi_{\tilde{U}1}^2) - \frac{1}{12}p_1p_2[(1-s)p_1 + sp_2]$, implying that

$$N = \left(\frac{\sqrt{\tilde{v}_0} z_{1-\frac{\alpha}{2}} + \sqrt{12\tilde{v}_1} z_{1-\beta}}{\tilde{u} \sqrt{12s(1-s)}} \right)^2 \quad (19)$$

4.2. Sample Size Estimation

When the distributions of the non-fatal outcome, X , and the survival time, t , are known approximately, it is possible to estimate the probabilities $\pi_{t\gamma}$ and $\pi_{x\gamma}$, $\gamma = 1, 2, 3$, and thus $\pi_{U\gamma}$ ($\pi_{\tilde{U}\gamma}$) $\gamma = 1, 2, 3$, and thereby derive estimates of u and v_1 ($\tilde{u}$, $\tilde{v}_0$, and $\tilde{v}_1$), as

$$\hat{u} = \hat{\pi}_{U1} - 1/2, \quad \hat{v}_1 = (1-s)(\hat{\pi}_{U2} - \hat{\pi}_{U1}^2) + s(\hat{\pi}_{U3} - \hat{\pi}_{U1}^2), \quad \hat{\tilde{u}} = \hat{\pi}_{\tilde{U}1} - 1/2, \quad \hat{\tilde{v}}_0 = 1 - \hat{p}^3,$$

$$\hat{\tilde{v}}_1 = (1-s)(\hat{\pi}_{\tilde{U}2} - \hat{\pi}_{\tilde{U}1}^2) + s(\hat{\pi}_{\tilde{U}3} - \hat{\pi}_{\tilde{U}1}^2) - \hat{p}_1\hat{p}_2[(1-s)\hat{p}_1 + s\hat{p}_2]/12;$$

and evaluate the desired sample size.

In this section, we consider four methods for estimating $\pi_{t\gamma}$ and $\pi_{x\gamma}$, $\gamma = 1, 2, 3$, based on either the availability of data from previous studies or assumptions that can be made about the underlying distributions of the death times and continuous outcome of interest.

Method A: When data from a previous pilot study are available, Wang et al. [32] recommended estimating $\pi_{t\gamma}$ and $\pi_{x\gamma}$, $\gamma = 1, 2, 3$, with their sample counterparts:

$$\hat{\pi}_{t1} = (m_1 n_1)^{-1} \sum_{k=1}^{m_1} \sum_{l=1}^{n_1} I(t_{1k} < t_{2l} | t_{1k} \leq T, t_{2l} \leq T), \quad (20)$$

$$\begin{aligned}
 \hat{\pi}_{x1} &= (m_2 n_2)^{-1} \sum_{k=1}^{m_2} \sum_{l=1}^{n_2} I(X_{1k} < X_{2l}), \\
 \hat{\pi}_{t2} &= [m_1 n_1 (m_1 - 1)]^{-1} \sum_{k \neq k'}^{m_1} \sum_{l=1}^{n_1} I(t_{1k} < t_{2l}, t_{1k'} < t_{2l} | t_{1k} \leq T, t_{1k'} \leq T, t_{2l} \leq T), \\
 \hat{\pi}_{x2} &= [m_2 n_2 (m_2 - 1)]^{-1} \sum_{k \neq k'}^{m_2} \sum_{l=1}^{n_2} I(X_{1k} < X_{2l}, X_{1k'} < X_{2l}), \\
 \hat{\pi}_{t3} &= [m_1 n_1 (n_1 - 1)]^{-1} \sum_{k=1}^{m_1} \sum_{l \neq l'}^{n_1} I(t_{1k} < t_{2l}, t_{1k} < t_{2l'} | t_{1k} \leq T, t_{2l} \leq T, t_{2l'} \leq T), \\
 \hat{\pi}_{x3} &= [m_2 n_2 (n_2 - 1)]^{-1} \sum_{k=1}^{m_2} \sum_{l \neq l'}^{n_2} I(X_{1k} < X_{2l}, X_{1k} < X_{2l'}),
 \end{aligned} \tag{21}$$

where n_1 (resp. n_2) is the number of deaths (resp. survivors) in the active treatment group, and m_1 (reps. m_2) is the number of deaths (resp. survivors) in the control group.

The probabilities π_{U1} , π_{U2} and π_{U3} (resp. $\pi_{\tilde{U}}$, $\pi_{\tilde{U}2}$, $\pi_{\tilde{U}3}$, and p) can be estimated as:

$$\hat{p}_1 = \frac{m_1}{m}; \quad \hat{p}_2 = \frac{n_1}{n}; \quad \hat{q}_1 = 1 - \hat{p}_1; \quad \hat{q}_2 = 1 - \hat{p}_2; \quad \hat{\pi}_{U1} = \hat{p}_1 \hat{p}_2 \pi_{t1} + \hat{p}_1 \hat{q}_2 + \hat{q}_1 \hat{q}_2 \hat{\pi}_{x1} \tag{22}$$

$$\hat{\pi}_{U2} = \hat{p}_1^2 \hat{q}_2 + \hat{p}_1^2 \hat{p}_2 \hat{\pi}_{t2} + 2\hat{p}_1 \hat{q}_1 \hat{q}_2 \hat{\pi}_{x1} + \hat{q}_1^2 \hat{q}_2 \hat{\pi}_{x2}; \quad \hat{\pi}_{U3} = \hat{p}_1 \hat{q}_2^2 + \hat{p}_1 \hat{p}_2^2 \hat{\pi}_{t3} + 2\hat{p}_1 \hat{p}_2 \hat{q}_2 \hat{\pi}_{t1} + \hat{q}_1 \hat{q}_2^2 \hat{\pi}_{x3}$$

$$\hat{\pi}_{\tilde{U}1} = \hat{q}_1 \hat{q}_2 \hat{\pi}_{x1} + \hat{p}_1 \hat{q}_2 + \frac{1}{2} \hat{p}_1 \hat{p}_2; \quad \hat{\pi}_{\tilde{U}2} = \hat{\pi}_{x2} \hat{q}_1^2 \hat{q}_2 + 2\hat{\pi}_{x1} \hat{p}_1 \hat{q}_1 \hat{q}_2 + \hat{p}_1^2 \hat{q}_2 + \frac{1}{3} \hat{p}_1^2 \hat{p}_2 \tag{23}$$

$$\hat{\pi}_{\tilde{U}3} = \hat{\pi}_{x3} \hat{q}_1 \hat{q}_2^2 + \hat{p}_1 \hat{q}_2^2 + \hat{p}_1 \hat{p}_2 \hat{q}_2 + \frac{1}{3} \hat{p}_1 \hat{p}_2^2; \quad \hat{p} = \frac{(m \cdot \hat{p}_1 + n \cdot \hat{p}_2)}{m + n} = \frac{m_1 + n_1}{m + n}.$$

Method B: Noether [30] suggested assuming $\sigma_0 = \sigma_1$. The formulae (18) and (19) then become, respectively,

$$N = \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{u \sqrt{12s(1-s)}} \right)^2 \quad \text{and} \quad N = \tilde{v}_0 \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\tilde{u} \sqrt{12s(1-s)}} \right)^2. \tag{24}$$

Hence, we need to estimate π_{U1} (resp. $\pi_{\tilde{U}1}$) in equations (22) (resp. (23)) via estimates of π_{t1} and π_{x1} from previous similar studies. When a pilot study exists, this estimate of π_{U1} (resp. $\pi_{\tilde{U}1}$) can be simply derived using (20) (resp. (21)).

Method C: One particular alternative hypothesis that has been extensively considered in the literature is the location shift alternative. In our setting, this translates into

$$H_1 : G_1(x - \Lambda_x) = G_2(x) \text{ and } F_1(t - \Lambda_t) = F_2(t), \tag{25}$$

for the untied worst-rank scores. For the tied worst-rank scores, it translates into

$$H_1 : G_1(x - \Lambda_x) = G_2(x) \text{ and } p_1 \geq p_2. \tag{26}$$

When the shift Λ_x is small, Lehmann[29] assumed $\sigma_0 = \sigma_1$ and used the approximation

$$\hat{\pi}_{x1} = 1/2 + \Lambda_x g^*(0) \tag{27}$$

to estimate π_{x1} , where $g^*(0)$ is the density of G^* evaluated at 0. G^* denotes the distribution of difference of two independent random variables, each with distribution G . For example, $g^*(0) = 1/2$ for the exponential distribution with mean one, and $g^*(0) = \pi^{-1/2}/2$ for the standard normal distribution. When g^* is unknown, as it is often the case, one can use a normal approximation $g^*(0) = \pi^{-1/2}/(2\sigma_x)$, with σ_x being the standard deviation of the distribution of X .

Method D: In practice, the underlying distributions G_i , $i = 1, 2$, are almost always unknown, which makes it difficult to estimate the sample size under the alternative shift hypothesis. In addition, the meaning of Λ_x in the location shift alternative varies for each G_1 . For these reasons, Rosner and Glynn [31] proposed the use of the probit shift alternative.

Let X_{2c} be defined as the counterfactual random variable obtained if each of the active treatment group subjects received the control treatment. Let $H_{X_1} = \Phi^{-1}(G_1)$, and $H_{X_2} = \Phi^{-1}(G_2)$. By definition $G_1 = G_{2c}$, so that $H_{X_{2c}} = H_{X_1}$. We can then consider the class of probit shift alternatives for the primary outcome of interest given by

$$H_0 : H_{X_1} = H_{X_2} = H_{X_{2c}} \text{ and } F_1(t) = F_2(t) \text{ versus } H_1 : H_{X_2} = H_{X_{2c}} - \nu_x \text{ and } F_1(t) \geq F_2(t).$$

Therefore, the null and alternative hypotheses become $H_0 : G_1(x) = G_2(x)$ and $F_1(t) = F_2(t)$, versus

$$H_1 : \Phi [\Phi^{-1}\{G_1(x)\} - \nu_x] = G_2(x) \text{ and } F_1(t) \geq F_2(t), \quad (28)$$

for the untied worst-rank scores. In the case of the tied worst-rank scores, we have

$$H_0 : G_1(x) = G_2(x) \text{ and } p_1 = p_2, \text{ versus } H_1 : \Phi [\Phi^{-1}\{G_1(x)\} - \nu_x] = G_2(x) \text{ and } p_1 \geq p_2. \quad (29)$$

For any random variable X_1 , H_{X_1} has a standard normal distribution. Thus, under the alternative (28), $H_{X_1} \sim N(0, 1)$ and $H_{X_2} \sim N(-\nu_x, 1)$. We can show (see Rosner and Glynn [31]) that

$$\pi_{x1} = \Phi\left(\frac{\nu_x}{\sqrt{2}}\right) \text{ and } \pi_{x2} = \pi_{x3} = P\left(Z < \frac{\nu_x}{\sqrt{2}}, Z' < \frac{\nu_x}{\sqrt{2}}\right), \text{ for } (Z, Z') \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}\right). \quad (30)$$

The probit shift alternatives (28) and (29) have an advantage over the location shift alternatives (25) and (26) since we need simply to either (a) estimate π_{x1} from (21), then derive ν_x , π_{x2} and π_{x3} from (30) or (b) prespecify ν_x and derive π_{x1} , π_{x2} , and π_{x3} from (30). Moreover, according to Rosner and Glynn [38], the rationale of using a probit transformation to define a shift alternative in (28) and (29) is that

1. any continuous distribution G_i has a corresponding underlying probit transformation H_{X_i} ;
2. it is natural to define location shifts between distributions in terms of a normally distributed scale, which by definition is satisfied by the probit transformation;
3. ν_x can be interpreted as an effect sizes on the probit scale, which is very useful for sample size and power calculations.

For the standard WMW test applied to a continuous outcome, methods A and B are used when data are available from previous pilot studies, while methods C and D are used when data from both groups are limited to summary statistics [39]. Method D can also be used for data not satisfying the location shift alternative or for an exploratory or a pilot study [31]. In general, the computations of π_{t2} , π_{t3} , π_{x2} , and π_{x3} are more involved than those of π_{t1} and π_{x1} . Methods B and C circumvent the evaluation of π_{t2} , π_{t3} , π_{x2} , and π_{x3} and only necessitate estimation of the probabilities π_{t1} and π_{x1} .

4.3. Simulation Study

To evaluate the performances of the four sample size methods presented in Section 4.2 and to assess the accuracy of the formulae (18) and (19), we conducted extensive simulation studies, similar to those conducted by Zhao [40]. First, we estimated – via these formulae – the sample sizes required to detect a specific difference in the primary outcome for the nominal power fixed at 80%. Then, we examined the accuracy of these sample size formulae by comparing these results to the (estimated) true simple sizes obtained by inverting the empirical power, as outlined in the algorithm below.

Moreover, for each sample size obtained in the first step, we calculated the corresponding empirical power and evaluated its discrepancy from the target, nominal power.

We set the active treatment fraction to be $s = 1/2$ (i.e., $m = n$) and the two-sided significance level to be $\alpha = 0.05$. We generated survival times from exponential (mean 1), Weibull (with shape parameter equal to 1.2), and log-logistic (with shape parameter 1) distributions. We fixed the follow-up time T at 3 months, the survival probability q_2 in the active treatment group at 60% and 80%, and the hazard ratio (HR) for mortality (control vs. active treatment group) at 1.0, 1.5, and 3.0. For the log-logistic distribution, the odds ratio (OR) of surviving longer for subjects in the active treatment group were fixed also at 1.0, 1.5, and 3.0. We generated non-fatal outcomes such that

$$X_{ij} = \mu_{xi} + \epsilon_x \quad (31)$$

with mean response $\mu_{x1} = 0$ (control group) and $\mu_{x2} = 0.5$ (active treatment group). We considered three distributions for ϵ_x : normal, lognormal (each with mean 0 and variance 1), and t-distribution with 3 degrees of freedom.

For each method, we estimated the relevant probabilities from among $\pi_{t1}, \pi_{t2}, \pi_{t3}, \pi_{x1}, \pi_{x2}$, and π_{x3} as well as π_{U1}, π_{U2} and π_{U3} (resp. $\pi_{\tilde{U}1}, \pi_{\tilde{U}2}$, and $\pi_{\tilde{U}3}$) using the methods described in Section 4.2. More precisely, we generated a total of 1000 simulated data sets, each with a sample size of 10,000, and estimated these probabilities. We then used these estimates to calculate the sample sizes under both the untied (18) and tied (19) worst-rank missing data assignment. For method B, we used the same π_{t1} and π_{x1} from method A. For method C, we used π_{t1} obtained from method A. In addition, we used the normal approximation in (27) to obtain $\hat{\pi}_{x1} = 1/2 + \Lambda_x(2\sigma_x)^{-1}\pi^{-1/2}$. We estimated the common standard deviation for non-fatal outcomes σ_x using the pooled standard deviation. We estimated both $\Lambda_x = \mu_{x2} - \mu_{x1}$ and σ_x by their sample counterparts. Finally, for method D, we considered $\pi_{t1}, \pi_{t2}, \pi_{t3}$, and π_{x1} calculated through method A. We estimated $\nu_x = 2^{1/2}\Phi^{-1}(\pi_{x1})$, and derived probabilities π_{x2} and π_{x3} from (30).

To evaluate these estimated sample sizes relative to the true sample sizes, we used the following algorithm:

- (1) Begin with a pre-specified sample size N .
- (2) Set counter $C = 0$.
- (3) Generate $\tilde{X}_{ij}$ given in equation (1) (resp. equation (3)) for the untied (resp. tied) worst rank scores, with $t_{ij} \sim \text{exponential}, \text{Weibull}, \text{or log-logistic}$ and X_{ij} as defined in equation (31).
- (4) Increment C by 1 if $|Z| > 1.96$ (resp. $|\tilde{Z}| > 1.96$).
- (5) Repeat steps (3) and (4) for $R = 10,000$ simulated datasets.
- (6) Calculate $\tau = C/R$. If $\tau = 60\%$ (reps. $\tau = 80\%$), stop and choose N as the estimate of the true sample size. This corresponds to 60% (reps. 80%) power. If $\tau < 60\%$ (reps. $\tau < 80\%$), increase N and return to step (2). If $\tau > 60\%$ (reps. $\tau > 80\%$), decrease N and return to step (2).

We also evaluated the discrepancy between the nominal power and the corresponding (empirically) estimated power based on the estimated sample size. The adequacy of power estimations under these methods is measured by the relative percentage error, which we define as

$$\text{Relative percentage error} = \frac{\text{estimated power} - \text{nominal power}}{\text{estimated power}} \times 100\%.$$

The estimated power is obtained by following the steps (1) to (6) of the algorithm above, with the exception that at step (6), instead of re-calibrating N , we simply calculate the estimated power, τ .

The results are given in Tables 3 for the untied worst-rank test and in Table 4 for the tied worst-rank test. Results for the Weibull and log-logistic distributions are given in Appendix C.

Table 3. Sample size estimation and corresponding relative percentage error for the Wilcoxon–Mann–Whitney test applied to untied worst-rank scores (exponential survival)

Distr. ^a	HR	$q_2 = 60\%$					$q_2 = 80\%$				
		A	B	C	D	Truth	A	B	C	D	Truth
normal	1.0	1071 (1.0)	1071 (1.0)	1033 (-1.4)	1071(1.0)	1070	327 (-1.4)	329 (-1.2)	317 (-2.2)	327 (-1.4)	330
	1.5	210 (-0.3)	214 (-0.2)	211(-0.4)	209 (-0.3)	210	173 (-1.6)	176 (0.3)	172 (-0.3)	173 (-1.6)	174
	3.0	44 (-0.4)	48 (3.9)	48 (3.9)	44 (-0.4)	44	60 (-0.9)	63 (1.8)	63 (1.8)	60 (-0.9)	60
t_3	1.0	1619 (0.8)	1619 (0.8)	3024 (17.5)	1619 (0.8)	1616	497 (-0.2)	499 (-0.8)	937 (17.6)	496 (-0.6)	502
	1.5	242 (0.0)	245 (0.3)	292 (7.7)	242 (0.0)	242	224 (0.1)	227 (0.2)	326 (13.1)	224 (0.1)	224
	3.0	46 (-0.2)	49 (1.0)	51 (3.5)	46 (-0.2)	46	67 (-1.3)	70 (2.7)	81 (8.3)	67 (-1.3)	68
lognormal	1.0	814 (-0.6)	812 (0.0)	4721 (20.2)	812 (0.0)	812	260 (0.9)	261 (1.3)	1594 (20.0)	258 (0.8)	250
	1.5	191 (-1.0)	194 (0.7)	321 (15.9)	191 (-1.0)	192	146 (0.4)	148 (0.2)	411(19.8)	145 (-1.2)	146
	3.0	43 (-2.2)	47 (2.4)	52 (8.0)	43 (-2.2)	44	55 (-2.1)	59 (1.2)	86 (14.7)	54 (-2.1)	56

^a Distribution of the primary outcome of interest; t_3 denotes Student's t distribution with 3 degrees of freedom. The nominal power is equal to 80%.

Table 4. Sample size estimation and corresponding relative percentage error for the Wilcoxon–Mann–Whitney test applied to tied worst-rank scores (exponential survival)

Distr. ^a	HR	$q_2 = 60\%$					$q_2 = 80\%$				
		A	B	C	D	Truth	A	B	C	D	Truth
normal	1.0	998 (0.7)	998 (0.7)	963 (-1.3)	998 (0.6)	996	324 (-0.7)	327 (-0.5)	315 (-3.7)	324 (-0.7)	322
	1.5	205 (-0.3)	209 (0.7)	205 (-0.3)	205 (-0.3)	206	172 (0.2)	175 (0.8)	171 (-0.6)	172 (0.2)	172
	3.0	45 (0.7)	48 (3.9)	48 (3.9)	45 (0.7)	44	59 (-0.9)	63 (1.9)	62 (1.9)	59 (-0.9)	60
t_3	1.0	1515 (0.9)	1514 (1.8)	2828 (17.6)	1514 (0.7)	1512	493 (-0.6)	495 (-0.4)	930 (17.5)	493 (-0.6)	498
	1.5	237 (0.5)	240 (0.4)	290 (8.8)	237 (0.5)	236	223 (0.4)	226 (-0.4)	327 (13.3)	222 (0.1)	222
	3	46 (0.5)	49 (1.9)	51 (3.9)	46 (0.5)	44	67 (-0.5)	70 (2.2)	80 (8.0)	67 (-0.5)	68
lognormal	1.0	758 (0.5)	757 (0.5)	4425 (20.1)	756 (-1.2)	758	257 (0.9)	258 (1.4)	1578 (20)	256 (1.0)	254
	1.5	187 (0.1)	189 (0.8)	320 (16.4)	186 (0.2)	186	145 (-0.23)	147 (0.5)	410 (19.9)	144 (-0.2)	146
	3.0	44 (1.7)	46 (3.2)	52 (8.3)	43 (0.2)	42	55 (-2.1)	58 (1.3)	86 (14.9)	54 (-1.1)	56

^a Distribution of the primary outcome of interest; t_3 denotes Student's t distribution with 3 degrees of freedom. The nominal power is equal to 80%.

As seen in Tables 3 and 4 for the case of exponential survival, the sample sizes estimated by methods A, B and D are fairly similar and close to the true sample sizes across all the distributions for the nominal power considered. As expected, sample sizes estimation by method C are close to the true sample size only when the non-fatal outcome is normally distributed. This method performs poorly when the distribution is the Student's t with 3 degree of freedom or lognormal, with estimates as high as three times the true sample size. Method A and D are suitable when one has data from a pilot study that provides a good estimate of sample sizes. Method D has the advantage over method A of requiring only estimates for π_{t1} and/or π_{x1} . Although method B makes one additional assumption than method A, the results are quite similar, which indicates that the equal-variance assumption may be reasonable. Lastly, we note that the sample size requirements for the untied and tied worst-rank tests are quite similar, which is not surprising given the similarity in their power seen in Tables 1 and 2. Analogous results are obtained in Appendix Tables C.1, C.2, C.3, and C.4 for the Weibull and log-logistic distributions of survival times.

5. Discussion

We have considered the use of the ordinary Wilcoxon–Mann–Whitney (WMW) test in the context of informatively missing observations, with the tied or untied worst-rank score imputation procedure for missing observations. This is important in the context of trials with primary endpoints that are measured at a pre-specified fixed time-point for highly lethal diseases.

The impact of missing observations due to disease-related events is often overlooked [41, 37, 42]. We focused on testing the null hypothesis of no difference in both mortality and the primary outcome of interest against the restricted, uni-directional alternative hypothesis that the active treatment has a favorable effect on these two outcomes. In the context where higher values of the primary outcome of interest are indicative of better health outcome, the uni-directional alternative hypothesis specifies that subjects in the active treatment tend to have higher values of the primary outcome of interest or longer survival, while the distribution of the other outcome is either better in the active treatment group or at least equivalent in both treatment groups.

We have found that these worst-rank tests can be considerably more powerful than the alternative approach of analyzing the survivors only. In particular, this is the case as long as there is a large survival advantage to the treated group, or a moderate survival advantage coupled with a small to moderate advantage in the non-fatal outcome. This reflects the tradeoff between the added power of the worst-rank tests due to their use of all subjects and the added power of the survivors-only test due to their not diluting the non-fatal outcome effect with the survival effect. This also suggests that the “mortality neutral” assumption of McMahon and Harrell [27] in their power and sample size calculations is not advisable if a treatment effect on survival is plausible in that it would lead to a much larger sample size than necessary.

We also found that there is no real power advantage to the untied worst-rank test relative to the tied worst-rank test when the survival probability is at least 60%, and thus due to the simplicity of the tied worst-rank test, that may be preferable. While there may be an advantage to the untied worst-rank test when the survival probability is lower, we did not consider this case as it is unlikely to represent a realistic scenario for a trial of a non-fatal outcome, except maybe in clinical trials of critically ill patients [43].

We have derived formulae for power calculation for the WMW test for both tied and untied worst-rank tests, using the standard normal approximation of the distribution of the WMW test. These formulae take into account rank assignment procedures, and evaluate the mean and variance of the WMW test statistic accordingly. Our simulation studies demonstrate that our formulae are correct and highly accurate and thereby improve upon formulae in the literature.

We have also extended four methods commonly used in the literature to estimate a sample size to the setting of the worst-rank tests. Three of the methods do not require an assumption of normality for the non-fatal outcome and do not assume a location shift alternative. In simulation studies, we found these three methods to be highly accurate in their estimation of the sample size. Choice of which method to use depend on the presence of pilot data and the nature of the primary outcome of interest.

The WMW test on tied and untied worst-rank composite outcome presented in this paper can be extended to more complex settings when there are multiple components of the clinical course that could be considered in assigning worst-rank scores. Examples of these are the variations of worst-rank imputations proposed by Felker et al. [11, 9] and Moyé et al. [44, 45] and the longitudinally measured non-fatal outcomes introduced by Finkelstein and Schoenfeld [46].

Appendix A: Mean and variance for U and $\tilde{U}$

A.1 Untied Worst-Rank Scores

Consider the untied worst-rank adjusted values for subjects in the control and active treatment groups

$$\tilde{X}_{1k} = (1 - \delta_{1k})X_{1k} + \delta_{1k}(\eta + t_{1k}), \text{ and } \tilde{X}_{2l} = (1 - \delta_{2l})X_{2l} + \delta_{2l}(\eta + t_{2l}), \text{ for } k = 1, \dots, m \text{ and } l = 1, \dots, n.$$

Define the WMW U -statistic by $U = (mn)^{-1} \sum_{k=1}^m \sum_{l=1}^n U_{kl} = (mn)^{-1} \sum_{k=1}^m \sum_{l=1}^n I(\tilde{X}_{1k} < \tilde{X}_{2l})$.

Since $\tilde{X}_{1k} < \tilde{X}_{2l}$ when $\{t_{1k} < t_{2l} \text{ and } (\delta_{1k} = \delta_{2l} = 1)\}$, $\{\delta_{1k} = 1 \text{ and } \delta_{2l} = 0\}$, or $\{X_{1k} < X_{2l} \text{ and } (\delta_{1k} = \delta_{2l} = 0)\}$, we

have $U_{kl} = I(\tilde{X}_{1k} < \tilde{X}_{2l}) = I(t_{1k} < t_{2l}, \delta_{1k} = \delta_{2l} = 1) + I(\delta_{1k} = 1, \delta_{2l} = 0) + I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0)$.

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

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

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

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 $U_{kl} = I(\tilde{X}_{1k} < \tilde{X}_{2l})$ follows Bernoulli distribution with probability π_{U1} , we derive the variance $Var(U_{kl}) = E(U_{kl})[1 - E(U_{kl})] = \pi_{U1}(1 - \pi_{U1})$.

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

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

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 $\pi_{U3} = p_1 q_2^2 + p_1 p_2^2 \pi_{t3} + 2p_1 p_2 q_2 \pi_{t1} + q_1 q_2^2 \pi_{x3}$

A.2 Tied Worst-Rank Scores

Let $\tilde{X}_{1k} = (1 - \delta_{1k})X_{1k} + \delta_{1k}\eta$, $\tilde{X}_{2l} = (1 - \delta_{2l})X_{2l} + \delta_{2l}\eta$, for $k = 1, \dots, m$ and $l = 1, \dots, n$ be the tied adjusted worst-rank values for subjects in control and active treatment groups. Consider the WMW U-statistic defined by

$$\tilde{U} = (mn)^{-1} \sum_{k=1}^m \sum_{l=1}^n \tilde{U}_{kl} = (mn)^{-1} \sum_{k=1}^m \sum_{l=1}^n \left[I(\tilde{X}_{1k} < \tilde{X}_{2l}) + \frac{1}{2} I(\tilde{X}_{1k} = \tilde{X}_{2l}) \right].$$

$\tilde{U}_{kl} = I(X_{1k} < X_{2l})I(\delta_{1k} = \delta_{2l} = 0) + I(\delta_{1k} = 1, \delta_{2l} = 0) + \frac{1}{2} I(\delta_{1k} = \delta_{2l} = 1)$, which implies

$$\begin{aligned} E(\tilde{U}_{kl}) &= P(X_{1k} < X_{2l})P(\delta_{1k} = \delta_{2l} = 0) + P(\delta_{1k} = 1, \delta_{2l} = 0) + \frac{1}{2} P(\delta_{1k} = \delta_{2l} = 1) = q_1 q_2 \pi_{x1} + p_1 q_2 + \frac{1}{2} p_1 p_2. \\ \text{Thus, } E(\tilde{U}) &= q_1 q_2 \pi_{x1} + p_1 q_2 + \frac{1}{2} p_1 p_2 = \pi_{\tilde{U}} \end{aligned} \quad (A.3)$$

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

$$\begin{aligned} Var(\tilde{U}_{kl}) &= Var(I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0)) + Var(I(\delta_{1k} = 1, \delta_{2l} = 0)) + \frac{1}{4}Var(I(\delta_{1k} = \delta_{2l} = 1)) \\ &\quad + Cov(I(\delta_{1k} = 1, \delta_{2l} = 0), I(\delta_{1k} = \delta_{2l} = 1)) + Cov(I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0), I(\delta_{1k} = \delta_{2l} = 1)) \\ &\quad + 2Cov(I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0), I(\delta_{1k} = 1, \delta_{2l} = 0)) \\ &= \pi_{x1}q_1q_2(1 - \pi_{x1}q_1q_2) + p_1q_2(1 - p_1q_2) + \frac{1}{4}p_1p_2(1 - p_1p_2) - p_1^2p_2q_2 - \pi_{x1}q_1q_2p_1p_2 - 2\pi_{x1}p_1q_1q_2^2 \\ &= (\pi_{x1}q_1q_2 + p_1q_2 + \frac{1}{2}p_1p_2)(1 - \pi_{x1}q_1q_2 - p_1q_2 - \frac{1}{2}p_1p_2) - \frac{1}{4}p_1p_2 = \pi_{\tilde{U}1}(1 - \pi_{\tilde{U}1}) - \frac{1}{4}p_1p_2. \end{aligned}$$

$$Cov(\tilde{U}_{kl}, \tilde{U}_{k'l}) = E(\tilde{U}_{kl}\tilde{U}_{k'l}) - E(\tilde{U}_{kl})E(\tilde{U}_{k'l}) \text{ and } Cov(\tilde{U}_{kl}, \tilde{U}_{kl'}) = E(\tilde{U}_{kl}\tilde{U}_{kl'}) - E(\tilde{U}_{kl})E(\tilde{U}_{kl'}), \text{ for } k \neq k', l \neq l'$$

$$\begin{aligned} E(\tilde{U}_{kl}\tilde{U}_{k'l}) &= E[I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0)I(X_{1k'} < X_{2l}, \delta_{1k'} = \delta_{2l} = 0)] + E[I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0) \times \\ &\quad I(\delta_{1k'} = 1, \delta_{2l} = 0)] + \frac{1}{2}E[I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0)I(\delta_{1k'} = \delta_{2l} = 1)] \\ &\quad + E[I(\delta_{1k} = 1, \delta_{2l} = 0)I(X_{1k'} < X_{2l}, \delta_{1k'} = \delta_{2l} = 0)] + E[I(\delta_{1k} = 1, \delta_{2l} = 0)I(\delta_{1k'} = 1, \delta_{2l} = 0)] \\ &\quad + \frac{1}{2}E[I(\delta_{1k} = 1, \delta_{2l} = 0)I(\delta_{1k'} = \delta_{2l} = 1)] + \frac{1}{2}E[I(\delta_{1k} = \delta_{2l} = 1)I(X_{1k'} < X_{2l}, \delta_{1k'} = \delta_{2l} = 0)] \\ &\quad + \frac{1}{2}E[I(\delta_{1k} = \delta_{2l} = 1)I(\delta_{1k'} = 1, \delta_{2l} = 0)] + \frac{1}{4}E[I(\delta_{1k} = \delta_{2l} = 1)I(\delta_{1k'} = \delta_{2l} = 1)] \\ &= P(X_{1k} < X_{2l}, X_{1k'} < X_{2l})P(\delta_{1k} = \delta_{1k'} = \delta_{2l} = 0) + P(X_{1k} < X_{2l})P(\delta_{1k'} = 1, \delta_{1k} = \delta_{2l} = 0) \\ &\quad + P(X_{1k'} < X_{2l})P(\delta_{1k} = 1, \delta_{1k'} = \delta_{2l} = 0) + P(\delta_{1k} = \delta_{1k'} = 1, \delta_{2l} = 0) + \frac{1}{4}P(\delta_{1k} = \delta_{1k'} = \delta_{2l} = 1) \\ &= \pi_{x2}q_1^2q_2 + 2\pi_{x1}p_1q_1q_2 + p_1^2q_2 + \frac{1}{4}p_1^2p_2. \end{aligned}$$

$$\begin{aligned} E(\tilde{U}_{kl}\tilde{U}_{kl'}) &= E[I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0)I(X_{1k} < X_{2l'}, \delta_{1k} = \delta_{2l'} = 0)] + E[I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0) \times \\ &\quad I(\delta_{1k} = 1, \delta_{2l'} = 0)] + \frac{1}{2}E[I(X_{1k} < X_{2l}, \delta_{1k} = \delta_{2l} = 0)I(\delta_{1k} = \delta_{2l'} = 1)] \\ &\quad + E[I(\delta_{1k} = 1, \delta_{2l} = 0)I(X_{1k} < X_{2l'}, \delta_{1k} = \delta_{2l'} = 0)] + E[I(\delta_{1k} = 1, \delta_{2l} = 0)I(\delta_{1k} = 1, \delta_{2l'} = 0)] \\ &\quad + \frac{1}{2}E[I(\delta_{1k} = 1, \delta_{2l} = 0)I(\delta_{1k} = \delta_{2l'} = 1)] + \frac{1}{2}E[I(\delta_{1k} = \delta_{2l} = 1)I(X_{1k} < X_{2l'}, \delta_{1k} = \delta_{2l'} = 0)] \\ &\quad + \frac{1}{2}E[I(\delta_{1k} = \delta_{2l} = 1)I(\delta_{1k} = 1, \delta_{2l'} = 0)] + \frac{1}{4}E[I(\delta_{1k} = \delta_{2l} = 1)I(\delta_{1k} = \delta_{2l'} = 1)] \\ &= P(X_{1k} < X_{2l}, X_{1k} < X_{2l'})P(\delta_{1k} = \delta_{2l} = \delta_{2l'} = 0) + P(\delta_{1k} = 1, \delta_{2l} = \delta_{2l'} = 0) \\ &\quad + \frac{1}{2}P(\delta_{1k} = \delta_{2l'} = 1, \delta_{2l} = 0) + \frac{1}{2}P(\delta_{1k} = \delta_{2l} = 1, \delta_{2l'} = 0) + \frac{1}{4}P(\delta_{1k} = \delta_{2l} = \delta_{2l'} = 1) \\ &= \pi_{x3}q_1q_2^2 + p_1q_2^2 + p_1p_2q_2 + \frac{1}{4}p_1p_2^2. \end{aligned}$$

Therefore,

$$Var(U) = (nm)^{-1} \left[\pi_{\tilde{U}1}(1 - \pi_{\tilde{U}1}) + (m-1) \left(\pi_{\tilde{U}2} - \pi_{\tilde{U}1}^2 - \frac{p_1^2p_2}{12} \right) + (n-1) \left(\pi_{\tilde{U}3} - \pi_{\tilde{U}1}^2 - \frac{p_1p_2^2}{12} \right) - \frac{p_1p_2}{4} \right] \quad (\text{A.4})$$

$$\text{where } \pi_{\tilde{U}2} = \pi_{x2}q_1^2q_2 + 2\pi_{x1}p_1q_1q_2 + p_1^2q_2 + \frac{1}{3}p_1^2p_2 \text{ and } \pi_{\tilde{U}3} = \pi_{x3}q_1q_2^2 + p_1q_2^2 + p_1p_2q_2 + \frac{1}{3}p_1p_2^2.$$

Appendix B: Conditional probabilities

Suppose the death times follow exponential distributions i.e. $t_i \sim \text{Exp}(\lambda_i)$, $i = 1, 2$. Let $\theta = \frac{\lambda_1}{\lambda_2}$, $q_1 = q_2^\theta$, $q_2 = e^{-T\lambda_2}$.

Since $P(\delta_{1k} = 1) = p_1$, $P(\delta_{2l} = 1) = p_2$, we have

$$\pi_{t1} = P(t_{1k} < t_{2l} | \delta_{1k} = \delta_{2l} = 1) = (p_1 p_2)^{-1} \int_0^T (1 - e^{-\lambda_1 u}) \lambda_2 e^{-\lambda_2 u} du = \frac{1}{(1 - q_2^\theta)} \left[1 - \frac{1 - q_2^{(1+\theta)}}{(1 + \theta)(1 - q_2)} \right] \quad (\text{B.1})$$

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

$$\pi_{t3} = P(t_{1k} < t_{2l}, t_{1k} < t_{2l'} | \delta_{1k} = \delta_{2l} = \delta_{2l'} = 1) = p_1^{-1} p_2^{-2} \int_0^T (e^{-\lambda_2 T} - e^{-\lambda_2 u})^2 \lambda_1 e^{-\lambda_1 u} du \quad (\text{B.3})$$

$$= \left(\frac{q_2}{1 - q_2} \right)^2 \left[1 + \frac{\theta(1 - q_2^{(2+\theta)})}{(2 + \theta)(1 - q_2^\theta)q_2^2} - \frac{2\theta(1 - q_2^{(1+\theta)})}{(1 + \theta)(1 - q_2^\theta)q_2} \right]. \quad (\text{B.4})$$

Now, 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 $\Delta_x = \frac{\mu_{x_2} - \mu_{x_1}}{\sqrt{\sigma_{x_1}^2 + \sigma_{x_2}^2}}$, $\rho_{x_j} = \frac{\sigma_{x_j}^2}{\sigma_{x_1}^2 + \sigma_{x_2}^2}$, and $Z_{kl} = \frac{X_{1k} - X_{2l} - (\mu_{x_1} - \mu_{x_2})}{\sqrt{\sigma_{x_1}^2 + \sigma_{x_2}^2}}$. We can show that

$$\pi_{x1} = P(X_{1k} < X_{2l}) = \Phi(\Delta_x),$$

$$\pi_{x2} = P(X_{1k} < X_{2l}, X_{1k'} < X_{2l}) = P(Z_{kl} < \Delta_x, Z_{k'l} < \Delta_x), \quad (Z_{kl}, Z_{k'l}) \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{x_2} \\ \rho_{x_2} & 1 \end{pmatrix} \right)$$

$$\pi_{x3} = P(X_{1k} < X_{2l}, X_{1k} < X_{2l'}) = P(Z_{kl} < \Delta_x, Z_{kl'} < \Delta_x), \quad (Z_{kl}, Z_{kl'}) \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{x_1} \\ \rho_{x_1} & 1 \end{pmatrix} \right).$$

Appendix C: Simulation results for Weibull and log-logistic survival times

Table C.1. Sample size estimation for the Wilcoxon–Mann–Whitney test for untied worst-rank scores (Weibull survival)

Distr. ^a	HR	$q_2 = 60\%$					$q_2 = 80\%$				
		A	B	C	D	Truth	A	B	C	D	Truth
Normal	1.0	1024 (-1.3)	1025 (-1.3)	982 (-3.8)	1024 (-1.3)	1026	326 (-1.0)	329 (-0.2)	316 (-3.3)	326 (-1.0)	332
	1.5	168 (0.2)	172 (1.2)	170 (1.1)	168 (0.2)	168	152 (0.5)	156 (0.9)	152 (0.5)	152 (0.5)	152
	3.0	31 (-3.8)	35 (3.8)	35 (3.8)	31 (-3.8)	32	44 (0.0)	47 (1.9)	47 (1.9)	44 (0.0)	44
t_3	1.0	1604 (0.7)	1604 (0.7)	2995 (17.33)	1604 (0.7)	1600	497 (-0.6)	499 (0.0)	938 (17.6)	497 (-0.6)	498
	1.5	185 (-1.1)	189 (0.9)	217 (6.1)	185 (-1.1)	186	192 (-0.1)	195 (0.0)	268 (12.2)	191 (-1.4)	194
	3.0	32 (-0.6)	35 (2.9)	36 (5.1)	32 (-0.6)	32	47 (-1.6)	51 (-1.2)	56 (6.5)	47 (-2.7)	52
Lognormal	1.0	794 (-1.7)	792 (-1.4)	4586 (20)	792 (-1.3)	800	252 (-0.3)	253 (0.4)	1499 (20.0)	251 (-0.3)	252
	1.5	155 (0.14)	158 (1.0)	239 (14.5)	154 (-0.4)	154	130 (-0.4)	132 (0.7)	334 (19.6)	129 (-0.7)	130
	3.0	31 (-3.4)	35 (4.5)	37 (6.0)	31 (-3.4)	32	42 (0.7)	44 (3.2)	59 (12.2)	41 (-2.0)	42

^a Distribution of the primary outcome of interest; t_3 denotes Student's t distribution with 3 degrees of freedom. The nominal power is equal to 80%.

Table C.2. Sample size estimation for the Wilcoxon–Mann–Whitney test for tied worst-rank scores (Weibull survival)

Distr. ^a	HR	$q_2 = 60\%$					$q_2 = 80\%$				
		A	B	C	D	Truth	A	B	C	D	Truth
Normal	1.0	956 (-1.0)	956 (-1.0)	917 (-4.0)	956 (-1.0)	958	324 (-0.9)	326 (-0.2)	314 (-2.1)	324 (-0.9)	326
	1.5	166 (0.7)	169 (1.8)	167 (0.6)	166 (0.7)	164	152 (-0.2)	155 (1.0)	152 (-0.2)	152 (-0.2)	152
	3.0	32 (0.6)	35 (2.7)	35 (2.7)	32 (0.6)	32	43 (-2.3)	47 (2.0)	46 (3.0)	43 (-2.3)	44
t_3	1.0	1502 (-0.2)	1502 (-0.2)	2806 (17.4)	1502 (-0.2)	1502	494 (-0.4)	496 (-0.8)	931 (17.5)	493 (-0.9)	494
	1.5	184 (0.2)	187 (0.3)	217 (7.0)	184 (0.2)	184	192 (-0.2)	196 (0.3)	275 (12.9)	193 (0.9)	194
	3	33 (-1.5)	36 (5.1)	37 (5.3)	33 (-1.5)	34	47 (-2.0)	51 (2.3)	56 (7.2)	47 (-2.0)	48
Lognormal	1.0	742 (-1.0)	740 (-0.6)	4273 (20.0)	740 (-0.6)	740	250 (-1.5)	252 (0.4)	1435 (20.0)	249 (-0.8)	250
	1.5	151 (0.9)	153 (1.0)	237 (15.5)	150 (-0.1)	150	129 (-0.5)	132 (1.3)	334 (19.6)	128 (0.3)	128
	3.0	32 (0.9)	35 (4.1)	37 (7.3)	32 (0.9)	0.32	41 (-2.8)	44 (3.1)	59 (12.4)	42 (1.3)	43

^a Distribution of the primary outcome of interest; t_3 denotes Student's t distribution with 3 degrees of freedom. The nominal power is equal to 80%.

Table C.3. Sample size estimation and corresponding relative percentage error for the Wilcoxon–Mann–Whitney test on untied worst-rank scores (log-logistic survival)

Distr. ^a	HR	$q_2 = 60\%$					$q_2 = 80\%$				
		A	B	C	D	Truth	A	B	C	D	Truth
Normal	1.0	1079 (-0.1)	1080 (0.7)	1038 (-0.9)	1079 (-0.1)	178	327 (-1.6)	330 (0.3)	318 (-2.9)	327 (-1.6)	330
	1.5	272 (-0.3)	276 (0.7)	271 (-0.14)	272 (-0.3)	272	186 (-1.0)	190 (1.5)	185 (-0.9)	186 (-1.0)	186
	3.0	71 (-1.1)	74 (0.7)	74 (0.7)	71 (-1.1)	72	75 (-1.3)	79 (1.4)	78 (1.6)	75 (-1.3)	74
t_3	1.0	1617 (1.4)	1616 (1.0)	3021 (17.7)	1616 (1.0)	116	520 (2.2)	522 (1.5)	1011 (18.2)	519 (1.5)	518
	1.5	317 (-1.0)	320 (-1.0)	398 (9.0)	317 (-1.0)	318	241 (-1.7)	244 (-0.6)	354 (13.2)	240 (-0.8)	246
	3.0	76 (-0.1)	80 (2.2)	85 (4.6)	76 (-0.1)	78	86 (-1.1)	90 (1.5)	106 (8.2)	86 (-1.1)	88
Lognormal	1.0	791 (-0.7)	790 (-1.23)	4483 (20.0)	789 (-1.3)	792	253 (0.4)	254 (0.3)	1539 (20.0)	251 (0.0)	252
	1.5	240 (-0.1)	243 (-0.4)	453 (17.5)	240 (-0.5)	244	154 (-0.7)	156 (0.2)	456 (19.9)	153 (-1.6)	154
	3.0	69 (-1.2)	73 (1.5)	89 (9.5)	69 (-1.2)	72	69 (-1.1)	72 (1.6)	118 (16.2)	68 (-0.8)	70

^a Distribution of the primary outcome of interest; t_3 denotes Student's t distribution with 3 degrees of freedom. The nominal power is equal to 80%.

Table C.4. Sample size estimation and corresponding relative percentage error for the Wilcoxon–Mann–Whitney test for tied worst-rank scores (log-logistic survival)

Distr. ^a	HR	$q_2 = 60\%$					$q_2 = 80\%$				
		A	B	C	D	Truth	A	B	C	D	Truth
Normal	1.0	1010 (1.1)	1010 (1.1)	971 (-1.2)	1010 (1.1)	1008	325 (-1.1)	328 (-0.9)	315 (-2.0)	325 (-1.1)	324
	1.5	282 (0.9)	285 (0.8)	280 (0.0)	282 (0.9)	280	186 (-0.3)	190 (1.0)	184 (-1.3)	186 (-0.3)	188
	3.0	82 (-1.0)	85 (1.3)	85 (1.3)	82 (-1.0)	84	78 (-0.1)	81 (2.2)	80 (1.1)	78 (-0.1)	78
t_3	1.0	1519 (0.7)	1518 (1.5)	2841 (17.6)	1518 (1.5)	1516	515 (0.9)	517 (2.1)	1002 (18.1)	514 (1.6)	512
	1.5	331 (-1.3)	334 (-0.6)	422 (9.1)	330 (-0.6)	332	242 (-0.6)	245 (-0.2)	357 (13.1)	242 (-0.6)	244
	3.0	89 (-1.1)	92 (1.6)	100 (5.9)	89 (-1.1)	90	88 (-1.0)	92 (2.2)	110 (9.3)	88 (-1.0)	90
Lognormal	1.0	739 (-0.9)	738 (-1.2)	4178 (20.0)	737 (-1.2)	740	251 (0.0)	252 (0.9)	1531 (20)	249 (-0.4)	252
	1.5	247 (0.4)	249 (0.3)	485 (17.8)	246 (0.1)	246	154 (-0.3)	156 (1.1)	460 (19.8)	153 (-0.8)	154
	3.0	79 (-0.8)	82 (1.2)	104 (11.0)	79 (-0.8)	80	70 (-0.4)	73 (1.0)	122 (16.3)	70 (-0.4)	70

^a Distribution of the primary outcome of interest; t_3 denotes Student's t distribution with 3 degrees of freedom. The nominal power is equal to 80%.

References

1. Zhang JL, Rubin DB. Estimation of causal effects via principal stratification when some outcomes are truncated by death. *Journal of Educational and Behavioral Statistics* 2003; **28**(4):353–368.

2. Rubin DB. Causal inference through potential outcomes and principal stratification: application to studies with "censoring" due to death. *Statistical Science* 2006; **21**:299–309.
3. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; **58**(1):21–29.
4. Singhal A. Normobaric Oxygen Therapy in Acute Ischemic Stroke Trial. ClinicalTrials.gov Database December 2006. URL <http://clinicaltrials.gov/ct2/show/NCT00414726>, [online] <http://clinicaltrials.gov/ct2/show/NCT00414726>.
5. Singhal A, Benner T, Roccatagliata L, Koroshetz W, Schaefer P, Lo E, Buonanno F, Gonzalez R, Sorensen A. A pilot study of normobaric oxygen therapy in acute ischemic stroke. *Stroke* 2005; **36**(4):797.
6. Kumar S, Selim M, Caplan L. Medical complications after stroke. *The Lancet Neurology* 2010; **9**(1):105–118.
7. Lachin J. Worst-rank score analysis with informatively missing observations in clinical trials. *Controlled clinical trials* 1999; **20**(5):408–422.
8. Berry JD, Miller R, Moore DH, Cudkowicz ME, Van Den Berg LH, Kerr DA, Dong Y, Ingersoll EW, Archibald D. 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.
9. Felker G, Maisel A. A global rank end point for clinical trials in acute heart failure. *Circulation: Heart Failure* 2010; **3**(5):643–646.
10. Zannad F, Garcia AA, Anker SD, Armstrong PW, Calvo G, Cleland JG, Cohn JN, Dickstein K, Domanski MJ, Ekman I, et al.. Clinical outcome endpoints in heart failure trials: a European Society of Cardiology heart failure association consensus document. *European journal of heart failure* 2013; **15**(10):1082–1094.
11. Felker G, Anstrom K, 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.
12. Subherwal S, Ohman EM, Mahaffey KW, Rao SV, Alexander JH, Wang TY, Alexander KP, Hasselblad V, Roe MT. Incorporation of bleeding as an element of the composite end point in clinical trials of antithrombotic therapies in patients with non-ST-segment elevation acute coronary syndrome: Validity, pitfalls, and future approaches. *American heart journal* 2013; **165**(5):644–654.
13. Neaton JD, Wentworth DN, Rhame F, Hogan C, Abrams DI, Deyton L. Considerations in choice of a clinical endpoint for AIDS clinical trials. *Statistics in medicine* 1994; **13**(19–20):2107–2125.
14. Lisa AB, James SH. Rule-based ranking schemes for antiretroviral trials. *Statistics in Medicine* 1997; **16**:1175–1191.
15. Joshua Chen Y, Gould AL, Nessly ML. Treatment comparisons for a partially categorical outcome applied to a biomarker with assay limit. *Statistics in medicine* 2005; **24**(2):211–228.
16. Follmann D, Wittes J, Cutler JA. The use of subjective rankings in clinical trials with an application to cardiovascular disease. *Statistics in medicine* 1992; **11**(4):427–437.
17. Brittain E, Palensky J, Blood J, Wittes J. 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.
18. Allen LA, Hernandez AF, O'Connor CM, Felker GM. End points for clinical trials in acute heart failure syndromes. *Journal of the American College of Cardiology* 2009; **53**(24):2248–2258.
19. Allen LA, Spertus JA, et al.. End points for comparative effectiveness research in heart failure. *Heart failure clinics* 2013; **9**(1):15–28.
20. Sun H, Davison BA, Cotter G, Pencina MJ, Koch GG. Evaluating treatment efficacy by multiple end points in phase II acute heart failure clinical trials analyzing data using a global method. *Circulation: Heart Failure* 2012; **5**(6):742–749.
21. Subherwal S, Anstrom KJ, Jones WS, Felker MG, Misra S, Conte MS, Hiatt WR, Patel MR. Use of alternative methodologies for evaluation of composite end points in trials of therapies for critical limb ischemia. *American Heart Journal* 2012; **164**(3):277–284.
22. DeCoster T, Willis M, Marsh J, Williams T, Nepola J, Dirschl D, Hurwitz S. Rank order analysis of tibial plafond fractures: does injury or reduction predict outcome? *Foot & ankle international* 1999; **20**(1):44–49.
23. Gould A. A new approach to the analysis of clinical drug trials with withdrawals. *Biometrics* 1980; **36**(4):721–727.
24. Ritchie J, Cerqueira M, Maynard C, Davis K, Kennedy J. Ventricular function and infarct size: the Western Washington intravenous streptokinase in myocardial infarction trial. *Journal of the American College of Cardiology* 1988; **11**(4):689.
25. Ritchie J, Davis K, Williams D, Caldwell J, Kennedy J. Global and regional left ventricular function and tomographic radionuclide perfusion: the Western Washington Intracoronary Streptokinase in Myocardial Infarction Trial. *Circulation* 1984; **70**(5):867.
26. Senn S. *Statistical issues in drug development*. Wiley, 1997.
27. McMahon R, 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.
28. Shieh G, Jan S, Randles R. On power and sample size determinations for the Wilcoxon–Mann–Whitney test. *Nonparametric Statistics* 2006; **18**(1):33–43.
29. Lehmann E, D'Abnera H. *Nonparametrics: statistical methods based on ranks*, vol. 204. Holden-Day San Francisco, 1975.
30. Noether G. Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association* 1987; **82**(398):645–647.
31. Rosner B, Glynn R. Power and sample size estimation for the Wilcoxon rank sum test with application to comparisons of C statistics from alternative prediction models. *Biometrics* 2009; **65**(1):188–197.
32. Wang H, Chen B, Chow S. Sample size determination based on rank tests in clinical trials. *Journal of Biopharmaceutical Statistics* 2003; **13**(4):735–751.
33. Greene T, Joffe M, Hu B, Li L, Boucher K. The balanced survivor average causal effect. *The international journal of biostatistics* 2013; **9**(2):291–306.
34. Siegel S, Castellan Jr N. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill Book Company, 1988.

35. Bellera C, Julien M, Hanley J. Normal Approximations to the Distributions of the Wilcoxon Statistics: Accurate to What N? Graphical Insights. *Journal of Statistics Education* 2010; **18**(2).
 36. Matsouaka R, Singhal A, Betensky R. The Optimal Wilcoxon–Mann–Whitney Test in the Presence of Death-Censored Observations. *submitted to Statistics in medicine* 2014; .
 37. Shih W. Problems in dealing with missing data and informative censoring in clinical trials. *Current Controlled Trials in Cardiovascular Medicine* 2002; **3**(1):4.
 38. Rosner B, Glynn R. Power and Sample Size Estimation for the Clustered Wilcoxon Test. *Biometrics* 2010; .
 39. Rahardja D, Zhao Y, Qu Y. Sample size determinations for the wilcoxon–mann–whitney test: A comprehensive review. *Statistics in Biopharmaceutical Research* 2009; **1**(3):317–322.
 40. Zhao Y. Sample size estimation for the van Elteren test—a stratified Wilcoxon–Mann–Whitney test. *Statistics in medicine* 2006; **25**(15):2675–2687.
 41. Wood A, White I, Thompson S. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials* 2004; **1**(4):368.
 42. Armijo-Olivo S, Warren S, Magee D. Intention to treat analysis, compliance, drop-outs and how to deal with missing data in clinical research: a review. *Physical Therapy Reviews* 2009; **14**(1):36–49.
 43. Freeman BD, Danner RL, Banks SM, Natanson C. Safeguarding patients in clinical trials with high mortality rates. *American journal of respiratory and critical care medicine* 2001; **164**(2):190–192.
 44. Moyé L, Davis B, 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.
 45. Moyé LA, Lai D, Jing K, Baraniuk MS, Kwak M, Penn MS, Wu CO. 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.
 46. Finkelstein D, Schoenfeld D. Combining mortality and longitudinal measures in clinical trials. *Statistics in medicine* 1999; **18**(11):1341–1354.
-