A "placement of death" approach for studies of treatment effects on ICU length of stay

Winston Lin, Scott D. Halpern, Meeta Prasad Kerlin, and Dylan S. Small*

July 19, 2014

Abstract

Length of stay in the intensive care unit (ICU) is a common outcome measure in randomized trials of ICU interventions. Because many patients die in the ICU, it is difficult to disentangle treatment effects on length of stay from effects on mortality; conventional analyses depend on assumptions that are often unstated and hard to interpret or check. We adapt a proposal from Rosenbaum (2006) that addresses concerns about selection bias and makes its assumptions explicit. A composite outcome is constructed that equals ICU length of stay if the patient was discharged alive and indicates death otherwise. Given any preference ordering that compares death with possible lengths of stay, we can estimate the intervention's effects on the composite outcome distribution. Sensitivity analyses can show results for different preference orderings.

We discuss methods for constructing approximate confidence intervals for treatment effects on quantiles of the outcome distribution or on proportions of patients with outcomes preferable to various cutoffs. Strengths and weaknesses of possible primary significance tests (including the Wilcoxon–Mann–Whitney rank sum test and a heteroskedasticity-robust variant due to Brunner and Munzel [2000]) are reviewed. An illustrative example reanalyzes a randomized trial of an ICU staffing intervention.

Keywords: Length of stay, intensive care unit, censoring by death, quantile treatment effects, bootstrap confidence intervals, rank tests, Wilcoxon rank sum test, Mann–Whitney U-test, Brunner–Munzel test, nonparametric Behrens–Fisher problem.

*Lin: Department of Political Science, Columbia University (from Sept. 2014). Halpern and Kerlin: Perelman School of Medicine, University of Pennsylvania. Small: Department of Statistics, The Wharton School, University of Pennsylvania. Corresponding author: Winston Lin (Linston@gmail.com). Forthcoming in *Statistical Methods in Medical Research*.

1 Introduction

Length of stay (LOS) in the intensive care unit (ICU) is a common outcome measure used as an indicator of both quality of care and resource use.[1] Longer ICU stays are associated with increased stress and discomfort for patients and their families, as well as increased costs for patients, hospitals, and society. Recent randomized-trial reports that estimate treatment effects on LOS include Lilly et al.[2] and Mehta et al.[3] LOS was the primary outcome for the Study to Understand Nighttime Staffing Effectiveness in a Tertiary Care ICU (SUNSET-ICU),[4] a randomized trial of 24-hour staffing by intensivist physicians in a medical ICU, compared to having intensivists available in person during the day and by phone at night.

Because a significant proportion of patients die in the ICU, conventional analytic approaches may confound an intervention's effects on LOS with its effects on mortality. Analyzing only survivors' stays is problematic: if the intervention saves the lives of some patients, but those patients have atypically long LOS, then the intervention may spuriously appear to increase survivors' LOS. It is also potentially misleading to pool the LOS data of survivors and non-survivors: a reduction in average LOS could be achieved either by helping survivors to recover faster or by shortening non-survivors' lives. Finally, time-to-event analysis can attempt to account for death by treating non-survivors' stays as censored, but this typically involves dubious assumptions and concepts (such as the existence of a latent LOS that exceeds the observed values for non-survivors and is independent of time till death). Freedman[5] and Joffe[6] critique the assumptions underlying conventional time-to-event analyses.

These issues are related to the "censoring by death" problem discussed from different perspectives by Rubin[7] and Joffe[6]. Rubin's exposition uses the hypothetical example of a randomized trial where the outcome is a quality-of-life (QOL) score, some patients die before QOL is measured, and treatment may affect mortality. In a comment on Rubin's paper, Rosenbaum[8] proposes an analysis of a composite outcome that equals the QOL score if the patient was alive at the measurement time and indicates death otherwise. Death need not be valued numerically; given any preference ordering that includes death and all possible QOL scores, Rosenbaum's method gives confidence intervals for treatment effects on order statistics of the distribution of the treated patients' outcomes. He notes that although researchers cannot decide the appropriate placement of death relative to the QOL scores, we can offer analyses for several different placements, "and each patient could select the analysis that corresponds to that patient's own evaluation."

Rubin[9] notes that Rosenbaum's proposal is "deep and creative" but may be "difficult to convey to consumers." Our goal in this paper is to adapt Rosenbaum's approach to provide inferences about quantities that may be more easily understood by "consumers" such as critical care researchers interested in an intervention's effects on ICU LOS. Using a composite outcome that equals the LOS if the patient was discharged alive and indicates death otherwise, we can make inferences about

treatment effects on the median and other quantiles of the outcome distribution, or about effects on the proportions of patients whose outcomes are considered better than various cutoff values of LOS. Sensitivity analyses can show how the results vary according to whether death is treated as the worst possible outcome or as preferable to extremely long ICU stays. Because the approach (like Rosenbaum's) compares the entire treatment group with the entire control group, it avoids the selection bias problem that can arise when only survivors' outcomes are analyzed.

Our approach allows researchers to explore treatment effects on more than one quantile of the composite outcome distribution or on proportions below more than one cutoff. For protection against data dredging, it may be desirable to choose a primary significance test before outcome data are available. We discuss the properties of several possible primary tests, including the Wilcoxon–Mann–Whitney rank sum test and a heteroskedasticity-robust variant due to Brunner and Munzel.[10]

Section 2 explains Rosenbaum's proposal and our modified approach and presents simulation evidence on the validity of bootstrap percentile confidence intervals for quantile treatment effects. Section 3 discusses the choice of a primary significance test and reasons to prefer the Brunner–Munzel test to the Wilcoxon–Mann–Whitney, with both a review of the literature and new simulations. Section 4 reanalyzes the SUNSET trial data as an illustrative example. Section 5 discusses benefits and limitations of the approach and directions for further research.

2 Estimating treatment effects

2.1 Rosenbaum's original proposal

Rosenbaum[8] considers a completely randomized experiment: out of a finite population of N patients, we assign a simple random sample of fixed size to treatment and the remainder to control. Patients' QOL scores take values in a subset Q of the real numbers. For those who have died before the time of QOL measurement, the outcome is "D," indicating death, instead of a real number. The analysis requires a "placement of death" determining, for each $x \in Q$, either that x is preferred to D or vice versa. For example, two possible placements are "Death is the worst outcome" and "Death is worse than x if $x \geq 2$, but better than x if $x < 2$." (The framework could easily be modified to allow placements such as "Death is equivalent to a QOL score of 2.") Any placement of death, together with the assumption that higher QOL scores are preferred to lower scores, defines a total ordering of $Q \cup \{D\}$.

Rosenbaum derives exact, randomization-based confidence intervals for order statistics of the distribution of outcomes that the treatment group patients would have experienced if they had been assigned to control. For example, his method enables statements of the form: "Ranking the 400 treatment group patients' outcomes from best to worst, the 201st value was a QOL score of 4.2. We estimate that if the same

400 patients had not received the intervention and we ranked their outcomes from best to worst, the 201st value would lie in the range $[x, y]$ (95% confidence interval)." Here x and y could be real numbers, or one or both of them could be D.

In the example above, slightly complicated language is needed to describe the quantity being estimated. A statement is being made about the treatment group patients (and since they are a random set, the estimand is a random variable). We know their actual outcome distribution, and we are constructing a confidence interval for an order statistic of the distribution that would have been observed had they been assigned to control. With 400 treatment group patients, the median is not an order statistic, so the example uses the 201st value instead. These unusual features of the approach allow the derivation of exact confidence intervals.

2.2 Alternative estimands

We borrow Rosenbaum's use of placements of death and his suggestion to offer multiple analyses corresponding to different placements, but we explore alternative estimands that may be more familiar to applied audiences. Our confidence intervals for those estimands will be approximate instead of exact.

In the LOS context, for each patient i, let Y_i denote a composite outcome that equals her LOS if she was discharged alive from the ICU and takes the value D otherwise. We allow D to be either a real number (meaning that death and some length of stay D are considered equally undesirable) or a special nonnumeric value that is considered greater (i.e., worse) than any possible LOS. Using the potential outcomes framework,[11–14] let Y_{1i} denote the outcome that would occur if patient i were assigned to treatment. If she is actually assigned to treatment, then $Y_i = Y_{1i}$; otherwise, Y_{1i} is a counterfactual. Similarly, let Y_{0i} denote the outcome that would occur if she were assigned to control.

Assume that each pair (Y_{1i}, Y_{0i}) is an independent observation from a probability distribution with marginal distribution functions $F_1(x) = P(Y_{11} \leq x)$ and $F_0(x) = P(Y_{01} \leq x)$. An intuitive interpretation of this assumption is that the patients in the trial are a random sample from an infinite population of interest.[15] We make this assumption for mathematical convenience and compatibility with the literature cited below, but it is probably not crucial, as standard errors, significance tests, and confidence intervals that are valid from the infinite-population perspective are typically conservative from the finite-population perspective (in which the N patients in the trial are the population of interest).[11,16–18]

Define the *treatment effect on the p-quantile* (where $0 < p < 1$) as

$$\mathrm{QTE}_p = \min\{x : F_1(x) \geq p\} - \min\{x : F_0(x) \geq p\}$$

if both terms on the right-hand side are real numbers; if either term is a nonnumeric placement of death, QTE_p is undefined. (A nonnumeric placement means that death is

the worst possible outcome, but need not imply that the difference between death and a 30-day ICU stay is considered greater than the difference between 30 and 3 days. Thus, the former difference is undefined, not infinite.) For example, $\mathrm{QTE}_{0.5}$ (the treatment effect on the median) is the difference between the population medians of Y_{1i} and Y_{0i}, if both are real numbers.

Define the *cutoff treatment effect at cutoff c* as

$$\mathrm{CTE}_c = P(Y_{11} \geq c) - P(Y_{01} \geq c).$$

For example, if LOS is measured in days, then CTE_{20} is the treatment effect on the proportion of patients with outcome at least as bad as a 20-day LOS. If death is the worst possible outcome, then CTE_D is the treatment effect on the mortality rate.

Quantile treatment effects and cutoff treatment effects are different ways of summarizing effects on the outcome distribution. QTEs may be undefined in the highest quantiles (if death is considered the worst possible outcome but is not assigned a numeric value), but CTEs are defined at all cutoffs. On the other hand, there is perhaps more danger of data dredging with CTEs, since researchers may have more leeway to choose cutoffs that yield results they like than to focus on, say, the treatment effect on the 0.53-quantile instead of the median. In the very different context of educational test scores, Holland[19] argues that for measuring changes over time in the gap between two distributions, analyses of differences in proportions below a cutoff score can easily mislead. He prefers analyses of differences in quantiles and recommends supplementary graphical displays. Whether analogous issues arise in the LOS context (e.g., in comparing treatment effects for different subgroups or different interventions) is a worthwhile topic for future research.

CTEs can be estimated by differences in sample proportions, with normal-approximation confidence intervals or the finite-sample improvements recommended by Agresti and Caffo[20] or Brown and Li.[21] QTEs can be estimated by differences in sample quantiles; we have used the version of sample quantiles recommended by Hyndman and Fan[22] ["Definition 8," which is median-unbiased of order $o(1/\sqrt{N})$]. Below we explore the use of bootstrap percentile confidence intervals for QTEs.

2.3 Confidence intervals for QTEs

The treatment–control difference in sample quantiles is a special case of a quantile regression estimator. Hahn[23] shows that bootstrap percentile confidence intervals for quantile regression coefficients have correct asymptotic coverage probabilities, under regularity conditions that in our case imply that the distributions of Y_{1i} and Y_{0i} are continuous and their densities are bounded away from zero near the quantiles of interest. In practice, we expect some discreteness in the distributions, in part because LOS data may be rounded, but most importantly because many values will be tied at the placement of death D.

Another wrinkle is that if D is nonnumeric, then the difference in sample quantiles is undefined when one or both of the sample quantiles equal D, and thus the bootstrap percentile CI is undefined when any bootstrap replication yields a treatment or control group sample quantile equal to D. (One can still report the sample quantiles from the original data and, in some cases, a CI for one of the population quantiles.)

To examine these issues empirically, we simulated a hypothetical trial with 1,500 patients, assigning 750 to treatment and 750 to control (slightly smaller sample sizes than the SUNSET trial's). On each of 10,000 replications of the trial:

1. We generated patients' outcomes assuming that the probability of death in the ICU was 20% for control group patients and 10% for treatment group patients. Survivors' LOS values were sampled with replacement from the data for SUNSET control group patients who survived their ICU stays. (The SUNSET data are rounded to the nearest tenth of an hour.)

2. Nominal 95% confidence intervals were constructed using the bootstrap percentile method with 1,000 bootstrap replications. We resampled the treatment group and control group independently with fixed sample sizes.

Step 1 implies the population quantiles of Y_{1i} and Y_{0i} shown in Table 1, if D is either nonnumeric or a number of days no less than 204.3 (the highest LOS value for survivors in the SUNSET control group). On each replication of the hypothetical trial, we observe the treatment and control sample quantiles, which are estimates of the population quantiles.

One might consider Hahn's[23] asymptotic results least reassuring near and above the 0.8-quantile, both because the population distributions of Y_{0i} and Y_{1i} put 20% and 10% probabilities on point masses at D, and because just below the 0.8- and 0.9-quantiles, there are nonnegligible gaps between the highest numeric LOS values in the distributions (e.g., the two highest values are 204.3 and 37.8 days). Table 2 assumes $D = 204.3$ days and shows a below-nominal CI coverage rate (88%) at the 0.8-quantile, but the effect is localized and not severe.

In Table 3, D is nonnumeric (death is the worst possible outcome). The CIs appear to be valid at the 0.7-quantile and below. At the 0.75-quantile, the bootstrap CI is undefined in 48% of the trial replications, because the control group's 0.75-quantile equaled D on at least one bootstrap replication. At the 0.8- and higher quantiles, this situation occurs frequently, and the CI is always or almost always undefined.

(The empirical coverage rates shown in Tables 2 and 3 are estimates of the true coverage probabilities and are subject to sampling error, but since they are based on 10,000 trial replications, the likely amounts of sampling error are small. It is straightforward to compute a margin of error at the 95% confidence level for the estimated coverage probability. When the empirical coverage rate is 50%, the margin of error is 1 percentage point. In all other cases, the margin of error is smaller. When the empirical coverage rate is 95% or 96%, the margin of error is 0.4 percentage

point.)

These results suggest that bootstrap percentile CIs for QTEs are likely to have approximate validity near the median (as long as mortality rates are well below 50%), but caution is warranted in the upper tail of the distribution, near the placement of death. Further research with more advanced methods such as BC_a bootstrap CIs[24,25] or subsampling[26] may be worthwhile.

3 Choosing a primary significance test

Recommended practice for analysis of clinical trials includes pre-specification of a primary outcome measure. As stated in the CONSORT explanation and elaboration document, "Having several primary outcomes ... incurs the problems of interpretation associated with multiplicity of analyses ... and is not recommended."[27] The same principle may suggest that before analyzing QTEs or CTEs, one quantile or cutoff should be designated as primary. The median may seem a natural choice, but some interventions may be intended to shorten long ICU stays without necessarily reducing the median. It may be difficult to predict which points in the outcome distribution are likely to be affected.

Instead of designating a primary quantile or cutoff, one could pre-specify that the primary significance test is a rank test with some sensitivity to effects throughout the outcome distribution. Rank tests do not require a numeric placement of death (unlike, e.g., the two-sample t-test). Rubin,[9] modifying Korn's[28] proposal, comments that the Wilcoxon–Mann–Whitney (WMW) rank sum test test could be combined with Rosenbaum's[8] approach. More generally, Rosenbaum[29,30] has extensively explored the use of rank tests in causal inference, and Imbens and Wooldridge[31] suggest the WMW test "as a generally applicable way of establishing whether the treatment has any effect" in randomized experiments.

The WMW test is often recommended because it is believed to have more robustness of efficiency (power) than tests based on the difference in mean outcomes; Lehmann[32] gives a helpful overview of results that support this view. However, when the classical assumption of a constant additive treatment effect is relaxed, power comparisons vary with the nature of the anticipated treatment effect,[33] and an even more fundamental issue is the need to carefully consider what hypothesis would be useful to test.[34,35] The WMW test is still valid for the strong null hypothesis that treatment has no effect on any patient (or for the hypothesis that treatment does not change the outcome distribution), but whether researchers should be satisfied with a test of the strong null is debatable.[36] The Mann–Whitney form of the test statistic naturally suggests a weaker null hypothesis, and there is an interesting, somewhat neglected literature on testing the weak null. We next discuss this literature from a causal inference perspective, using the potential outcomes framework. (The literature is not explicitly causal; it assumes two independent random samples from two infinite

populations and has both causal and descriptive applications.)

3.1 Rank tests and null hypotheses

Suppose m patients are assigned to treatment and $n = N - m$ to control. Let T and C denote the sets of indices of the treated and control patients. The Wilcoxon rank sum statistic is $\sum_{i \in T} R_i$, where R_i is the rank of Y_i among the N observations (in ascending order). Ties are often handled by the midrank method: each member of a group of tied observations is given the average of the ranks they would have if they were not tied. The rank sum statistic can be rewritten as $U + m(m+1)/2$, where U is the Mann–Whitney statistic

$$U = \sum_{i \in T} \sum_{j \in C} \left[I(Y_{1i} > Y_{0j}) + \frac{1}{2} I(Y_{1i} = Y_{0j}) \right]$$

and $I(A)$ equals 1 if A occurs and 0 otherwise.[37]

The WMW test compares the observed value of the rank sum statistic $\sum_{i \in T} R_i$ (or, equivalently, the Mann–Whitney statistic U) with its distribution under the null hypothesis that the potential outcomes Y_{1i} and Y_{0i} have identical distribution functions:[15]

$$H_0^d \ : \ F_1(x) = F_0(x) \qquad \text{for all } x,$$

where F_1 and F_0 are the marginal distribution functions, as in Section 2.2. In the causal inference literature, permutation tests (including the WMW and other rank tests) are often viewed as tests of a more restrictive hypothesis:

$$H_0^s \ : \ Y_{1i} = Y_{0i} \qquad \text{for all } i.$$

H_0^s says that assignment to treatment has no effect on any patient's outcome. Following Gail et al.,[36] we call H_0^s the strong null hypothesis.

The WMW test is a valid test of H_0^d, in the sense that its rejection probability under the null hypothesis is no greater than the nominal significance level. It is therefore also a valid test of the strong null hypothesis H_0^s. However, as our simulations in Section 3.2 (supporting Pratt's[38] asymptotic predictions) illustrate, the test is sensitive to certain kinds of departures from H_0^d but not others, and under some scenarios where H_0^d is false, the test is *less* likely to reject than it would if H_0^d were true. (In the terminology of theoretical statistics, the WMW test is a valid but biased test of H_0^d or H_0^s against a general alternative. Fay and Proschan[39] give a helpful discussion of desirable properties of tests.)

What kinds of departures from H_0^d can the WMW test detect? The Mann–Whitney statistic U provides a clue. The WMW test is equivalent to a test using the statistic

$$\frac{U}{mn} - \frac{1}{2} \ = \ \frac{1}{mn} \sum_{i \in T} \sum_{j \in C} I(Y_{1i} > Y_{0j}) + \frac{1}{2} \frac{1}{mn} \sum_{i \in T} \sum_{j \in C} I(Y_{1i} = Y_{0j}) - \frac{1}{2},$$

which is a consistent estimate of

$$P(Y_{1i} > Y_{0j}) + \frac{1}{2}P(Y_{1i} = Y_{0j}) - \frac{1}{2} \;=\; \frac{P(Y_{1i} > Y_{0j}) - P(Y_{1i} < Y_{0j})}{2}, \qquad i \neq j.$$

An extreme positive test statistic is evidence that $P(Y_{1i} > Y_{0j}) > P(Y_{1i} < Y_{0j})$—that is, if we sample the treated and untreated potential outcome distributions independently, it is more likely that a random treated value will exceed a random untreated value than the other way around. Similarly, an extreme negative test statistic is evidence that $P(Y_{1i} > Y_{0j}) < P(Y_{1i} < Y_{0j})$.

Thus, the WMW test can be reexamined as a test of the weak null hypothesis

$$H_0^w \;:\; P(Y_{1i} > Y_{0j}) = P(Y_{1i} < Y_{0j}), \qquad i \neq j.$$

H_0^w says that a random patient's outcome under treatment is equally likely to be better or worse than another random patient's outcome in the absence of treatment. To understand this null hypothesis, it may help to first consider an example from descriptive (non-causal) inference. McGraw and Wong[40] estimated a probability of 0.92 that a random young adult man was taller than a random young adult woman in the United States. In that context, we could consider the null hypothesis that a random man is equally likely to be taller or shorter than a random woman. In the descriptive inference problem, we are comparing the male and female populations' height distributions. In the causal inference problem, there is only one population, but each member of the population has two potential outcomes: Y_1, which would occur under treatment, and Y_0, which would occur in the absence of treatment. H_0^w is a statement comparing the population distributions of Y_1 and Y_0.

Pratt[38] shows that the WMW test is not an asymptotically valid test of H_0^w, in part because heteroskedasticity can distort the significance level, and more generally because the test is based on the distribution of the test statistic under H_0^d, not the weaker null hypothesis H_0^w. Pratt's Table 2 implies that if $m = n$, the size of a two-tailed WMW test (assuming no ties) at the nominal 5% level tends to a limit between 5% and 11%. If $m \neq n$, this range widens in both directions.

Brunner and Munzel[10] (BM) derive an asymptotically valid test of H_0^w by studentizing $U/mn - 1/2$ (i.e., dividing by a consistent estimate of its standard error). The BM test allows ties (the distributions of Y_{1i} and Y_{0i} can be of any nondegenerate form). The test statistic (which can be computed from the overall ranks R_i and the ranks within the treatment and control groups, and is implemented in the R lawstat package) is asymptotically $N(0,1)$ under H_0^w; to improve small-sample performance, BM suggest using the t-distribution with degrees of freedom from a Welch–Satterthwaite approximation. Neubert and Brunner[41] propose a permutation test based on the BM statistic and prove its asymptotic validity. Chung and Romano[35] derive a general theory for constructing asymptotically valid permutation tests based on two-sample U-statistics, discuss misapplications of the WMW test, and provide a studentized permutation version (for the case without ties) whose critical values can be tabled. Fay and Proschan[39] assess the validity and consistency of the WMW and related tests

under many different sets of assumptions. Ho[42] gives a helpful discussion of a related literature on nonparametric methods for comparing test score distributions, trends, and gaps.

A few authors[43–47] discuss a nontransitivity paradox associated with rank-based tests. If a test gives evidence that $P(Y_{1i} > Y_{0j}) > P(Y_{1i} < Y_{0j})$, one may be tempted to say that the treated potential outcome distribution tends to have higher values than the untreated distribution. But, using $\succ$ to denote this sense of "tends to have higher values," there exist sets of three distributions F, G, and H such that $F \succ G$, $G \succ H$, and $H \succ F$. Lumley[43] argues that nontransitivity is an example of a more generally troubling property of rank tests: because they avoid explicit valuation of tradeoffs when a treatment makes some people better and others worse, they "can just be misleading." We agree that a rank test should not be the sole criterion for evaluating an intervention. However, as Aldous[48] writes, the most useful role of a significance test is "to prevent you from jumping to conclusions based on too little data." Thus, the BM and related tests can serve as restraining devices: it is appropriate for the analysis of a clinical trial to focus on estimated effect sizes and comparisons at multiple points in the outcome distribution, but a failure to reject H_0^w can prevent a premature conclusion that treatment generally improves or generally worsens the outcome distribution.

3.2 Simulation evidence on test validity

Table 4 shows the rejection rates of the WMW and BM tests (two-tailed, at the nominal 5% level) in simulations of nine null-hypothesis scenarios with 1,500 patients and 250,000 replications. For the WMW test, we used the large-sample normal approximation.[49] In each panel, we show results for a balanced design (i.e., with a 1:1 treatment:control allocation ratio), and two imbalanced designs (with 9:1 and 1:9 allocation ratios).

The first panel shows rejection rates under the strong null hypothesis that treatment has no effect on any patient's outcome. For both the treatment group and the control group, the data-generating process for outcomes is identical to that used for the control group in Table 3: the probability of death is 20%, and survivors' LOS values are sampled with replacement from the SUNSET trial's control group data. Death is placed as the worst possible outcome. As expected, the WMW and BM tests have rejection rates close to the nominal 5% significance level.

For the second and third panels, we simulated scenarios in which H_0^w holds but the strong null does not. In each case, treatment shrinks the spread of a symmetric outcome distribution without shifting its center. The second panel assumes continuous distributions (the case analyzed by Pratt[38]), while the third panel allows a substantial number of ties.

In the second panel, the treated and control patients' outcomes are drawn from the continuous uniform distributions on [12.5, 27.5] and [5, 35], respectively. As a test of

H_0^w, the WMW test rejects somewhat too often (6.4%) with a 1:1 treatment:control allocation ratio, far too often (14.7%) with a 9:1 ratio, and rarely (0.3%) with a 1:9 ratio; these rates are very close to the asymptotic limits implied by Pratt's[38] Table 1. In contrast, the BM test's rejection rates are always close to the nominal 5% level.

For the third panel, outcomes are drawn from mixed discrete/continuous distributions. For treated patients, the distribution puts 20% probability on a point mass at 2.5, 60% on the uniform distribution on [12.5, 27.5], and 20% on a point mass at 37.5. The control patients' distribution is similar but the point masses are at 0 and 40 and the uniform distribution's range is [5, 35]. Again, the BM test maintains the nominal significance level but the WMW test does not (its rejection rates are 5.7%, 11.4%, and 1.2%).

In sum, the WMW test is not a valid test of H_0^w. It is valid for the strong null, but it is sensitive to certain kinds of departures from the strong null and not others. For example, it is more likely to reject the null when treatment narrows the spread of the outcome distribution and there are more treated than control patients, or when treatment widens the spread and there are more control than treated patients. It is less likely to reject when the opposite is true. These properties complicate the test's interpretation and are probably not well-known to most of its users. In contrast, the BM test is an approximately valid test of H_0^w in sufficiently large samples, and a failure to reject H_0^w can be understood to mean there is not enough data to infer a general tendency for treated patients' outcomes to be better or worse than those of untreated patients.

On the other hand, it is not clear whether these issues are likely to be empirically important in most clinical trials. With a balanced design, the WMW test's overrejection of H_0^w in Table 4 is only slight, and the simulated scenarios are perhaps extreme (e.g., in the second panel, treatment halves the standard deviation of the outcome).

3.3 Simulation evidence on power

Table 5 compares the abilities of three tests to detect beneficial treatment effects (i.e., reducing LOS or mortality) in various scenarios. The tests are the WMW, the BM, and the significance test for $\text{QTE}_{0.5}$ (the treatment effect on the median) constructed by inverting the bootstrap percentile confidence interval (based on 1,000 bootstrap replications). In each case we used a two-tailed test (at the 5% level) but assumed that if the null hypothesis was rejected, researchers would infer the direction of the effect from the sign of (i) the difference between the Wilcoxon rank sum statistic and its expected value, (ii) the BM statistic, or (iii) the CI limits for $\text{QTE}_{0.5}$. The top half of the table shows the rates of correctly inferring a beneficial treatment effect (in 10,000 replications of a clinical trial); the bottom half shows the rates of incorrectly inferring a harmful effect, which are very low.

In each scenario, the treatment:control allocation ratio is 1:1 and death is placed as the worst possible outcome. We first simulated Settings A–F (described below) with 1,500 patients. However, it turned out that a much larger sample size was needed for any of the tests to have reasonable power to detect the weaker treatment effects of Settings D–F, so we also simulated those settings with 30,000 patients.

In Setting A, the probability of death in the ICU is 5% for both the treatment group and the control group, but treatment and control group survivors' LOS values are sampled from two different distributions. The control group LOS distribution is just the empirical distribution for the SUNSET trial's control group survivors. The treatment group LOS distribution substitutes $g(x)$ for each value x in the control group distribution, where $g(x) = x$ if $x \leq 2$ (in days) and $g(x) = x/2 + 1$ if $x > 2$. The underlying idea is that the intervention is not expected to affect the shortest ICU stays, because bed space availability limits the speed at which patients can be moved from the ICU to other hospital units.

The WMW and BM tests detected a beneficial treatment effect in 54% of the replications of Setting A, while the corresponding rate for the $\text{QTE}_{0.5}$ test was only 10%. In this setting, the true QTE is very small at the median (the population median is 2.07 days with the intervention, compared to 2.13 without it) but larger in the upper half of the composite outcome distribution (e.g., the 90th percentile is 6.9 days with the intervention, compared to 11.9 without it), except in the upper 5% tail, which represents death.

Setting B raises the probability of death to 20% but is otherwise identical to Setting A. Because death now occupies a larger area at the upper tail of the composite outcome distribution, the median values with and without the intervention are now higher, and the intervention reduces the population median from 2.7 to 2.4 days. Thus, the true QTE at the median is higher than in Setting A, and the $\text{QTE}_{0.5}$ test has more power, detecting a beneficial effect in 61% of the replications. The corresponding rates for the WMW and BM tests have fallen to 31%; these tests lose power when there are many ties. A possible remedy is to perform a WMW test after removing an equal and maximal number of observations at the extremum (here, death) from each group.[50,51] However, we do not know of a way to use this approach to construct a test that would share the BM test's property of asymptotic validity for a weak null hypothesis.

In Setting C, the intervention reduces both LOS and mortality. We assume each patient in the population belongs to one of three principal strata:[7,52] 17.5% are "never-survivors," who would die in the ICU with or without the intervention; 80% are "always-survivors," who would survive with or without the intervention; and 2.5% are "responders," who would die in the ICU without the intervention but would survive with the intervention. (For simplicity, we assume the intervention does not cause any patients to die in the ICU, although this may be unrealistic.) The control group's outcomes are generated exactly as in Setting B. The treatment group's outcome-generating process puts 17.5% probability on death, 80% on the same LOS distribution as in Settings A and B, and 2.5% on the uniform distribution with range 14 to 28 days (thus assuming that responders have atypically long ICU stays). The

WMW and BM tests have somewhat higher power than in Setting B, while the $QTE_{0.5}$ test's power is unchanged. (The mortality effects are irrelevant to $QTE_{0.5}$ here, since responders' outcomes are worse than the median with or without the intervention.)

Settings D, E, and F are identical to A, B, and C, respectively, except that the treatment group LOS distribution only substitutes $g(x)$ for a random 25% of the values x in the control group distribution. Thus, the intervention has dramatic effects on some patients, but little or no effect on most. With a sample size of 1,500 patients, all three tests have very low power in these settings. Even with 30,000 patients, power is never above 71%, and the comparison between the rank-based and $QTE_{0.5}$ tests varies with the nature of the heterogeneity in the treatment effect.

These results suggest three general conclusions. First, in large samples, it seems appropriate to prefer the BM test to the WMW test, since they have approximately equal power in Table 5 and the BM test has much more robustness of validity in Table 4. Second, power comparisons between the BM test and the $QTE_{0.5}$ test vary with the nature of the treatment effect. (Arguably, in the absence of any prior information about the anticipated effect, the BM test is a more robust choice, since it has some sensitivity to effects throughout the outcome distribution, and an intervention can have practically significant effects without affecting the median.) Third, the sample sizes needed for detection of treatment effects can depend crucially on how widespread the effects are. (Further research may be worthwhile to see if power can be improved using other tests besides the ones considered here. See Section 5 for discussion.)

3.4 Comparison with two-part tests

A different approach to a primary significance test than using all observations in a rank test is a two-part test as discussed by Lachenbruch[53,54]. In the settings considered by Lachenbruch, the outcome has a positive probability of being zero and then a continuous distribution conditional on being greater than zero, e.g., hospitalization costs in a health insurance plan. The two-part test of no treatment effect combines two tests: (i) the treatment has no effect on the probability of being zero; (ii) the treatment has no effect on the conditional distribution of the nonzero values. The two-part test statistic is $X^2 = B^2 + T^2$ where B is an asymptotically standard normal statistic from a test of (i) and T is an asymptotically standard normal statistic from a test of (ii), such as a Wilcoxon rank sum test, t-test, or Kolmogorov–Smirnov test comparing only the nonzero outcomes of treated and control subjects. Under the null hypothesis of no treatment effect, B and T are asymptotically independent and X^2 has asymptotically a χ^2 distribution with 2 degrees of freedom. The two-part test rejects for large values of X^2.

For treatment effects on ICU length of stay, the two-part test can be adapted by replacing "zero" with "death." In particular, we consider the following two-part test: let B be the t-statistic from the unequal-variances (Welch) two-sample t-test applied to the outcome of whether or not a subject died, and let T be the test statistic from the

Brunner–Munzel test applied to the LOS outcome for survivors. Table 6 simulates the rejection rates for the two-part test in null-hypothesis scenarios that are comparable to Table 4 (again with 1,500 patients and 250,000 replications). The "strong null" scenarios are identical to those in Table 4. The "weak null" scenarios are modified because a two-part test would not make sense for the corresponding scenarios in Table 4 (e.g., in the second panel of Table 4, outcomes are drawn from continuous uniform distributions, so there is no mass point at "zero" or "death"). In the "weak null" scenarios for Table 6, treated and control patients have a 20% probability of death and 80% probability of having their outcomes drawn from the same distributions as in the weak null scenarios of Table 4. For balanced designs, the empirical rejection rate is approximately equal to the nominal 5% level. For imbalanced designs, the empirical rejection rates in Table 4 are 5.5% or 5.6% (which are significantly higher than would be expected in 250,000 replications if the true rejection probability were 5%), but when we increased the sample size from 1,500 to 15,000, the empirical rejection rates (not shown in the table) were all between 5.0% and 5.1%.

Table 7 examines the power for the two-part test for the same scenarios as in Table 5. Comparing the two tables, we see that for a sample size of 1,500, the two-part test was more powerful than the one-part BM test for Settings B, E and F, less powerful for Settings A and C, and comparable in power for Setting D. For a sample size of 30,000, the two-part test was more powerful than the one-part BM test for settings E and F, but less powerful for setting D. In Table 7, we did not split out "infer beneficial effect" and "infer harm" as in Table 5 because the test statistic X^2 for the two-part test does not suggest a direction of effect. One attractive feature of the one-part tests is that they do suggest a direction of effect.

The two-part tests we have considered are nonparametric. A related approach is to use a parametric mixture model in which survivors' lengths of stay are modeled with a parametric distribution. The August 2002 issue of *Statistical Methods in Medical Research* contains several papers on mixture models[54–56].

4 Illustrative example

SUNSET-ICU[4] was conducted in the medical ICU of the Hospital of the University of Pennsylvania (a 24-bed ICU). The trial enrolled patients who were admitted between September 12, 2011, and September 11, 2012. Within each two-week block during this period (except a winter holiday block), one week was randomly assigned to the intervention staffing model and the other to the control model. In both models, daytime staff included two intensivists (attending physicians who were board-certified or board-eligible in critical care medicine), and nighttime staff included three medical residents, who were expected to review all new admissions and critical events with an intensivist or critical care fellow by phone or in person. On control nights, two intensivists were available by phone (usual care). On intervention nights, one intensivist was present in the ICU.

The staffing model on the night of admission (or the night after a daytime admission) determined whether each patient was considered a member of the treatment group or the control group. In other words, the analysis estimates the effects of being admitted during an intervention week vs. a control week. Most patients experienced only one staffing model (the median LOS was about 2 days), but patients could experience both models if they stayed in the ICU long enough. After sample exclusions detailed in Kerlin et al.,[4] there were 820 patients in the treatment group and 778 in the control group. The ICU LOS in the SUNSET data are rounded to the nearest tenth of an hour.

Using a proportional hazards model with death treated as a censoring event, Kerlin et al.[4] found no effect of intervention week admission on ICU LOS. There was also no discernible effect on ICU deaths: 18.8% of the treatment group and 17.9% of the control group died in the ICU, and the difference is not statistically significant. A supplementary rank-based analysis, with death coded as the longest possible LOS, also found no evidence of an effect. In this section, we present more detailed results from the placement-of-death approach. (The trial has a matched-pair, cluster-randomized design:[57] the patients admitted during a week are a cluster, and each two-week block is a matched pair. For simplicity, we follow Kerlin et al.[4] in analyzing the data as if individual patients were randomized without blocking.)

With death placed as the worst possible outcome, the Brunner–Munzel test does not reject the hypothesis that a random patient's outcome under the intervention is equally likely to be better or worse than another random patient's outcome in the absence of the intervention. (the P-value in a two-tailed test is 0.29). The associated 95% CI for $P(Y_{1i} < Y_{0j}) + 0.5\ P(Y_{1i} = Y_{0j})$—that is, the probability that a random treatment group patient's outcome is better than a random control group patient's, plus one-half the probability that they are equally desirable (or undesirable)—is $[0.456, 0.513]$. The results are similar when death and a 28-day LOS are considered equally undesirable. A two-part test of no treatment effect gives a P-value of 0.31, similar to the P-value of 0.29 from the Brunner-Munzel test.

The top panel of Table 8 shows estimated quantile treatment effects and 95% confidence intervals (using the bootstrap percentile method with 1,000 replications), with death placed as the worst outcome. There is no evidence that the intervention affected the median outcome or any of the other quantiles examined. The CIs for the treatment effects on the 25th to 75th percentiles of the outcome distribution all include zero. Our method is unable to perform inference for treatment effects at the 80th percentile and above. The 80th percentile outcome for the treatment group is a 30-day LOS, compared to a 17.8-day LOS for the control group, but one or both values are death on some of the bootstrap replications, so the method cannot produce a confidence interval without additional assumptions about how to value the difference between death and a numeric LOS. The 90th percentile is death in both groups. For a general audience, one might present the results for the 0.25- to 0.75-quantiles together with a CI for the intervention's effect on mortality, which will be given below.

Not everyone will agree with an analysis that places death as more undesirable than all possible lengths of stay; some people may consider a long stay in the ICU as a

"fate worse than death." Our approach, following Rosenbaum's idea, makes it possible to see how the results change as we change the placement of death. The bottom panel of Table 8 repeats the analysis with death and a 28-day LOS considered equally undesirable. The results for the 25th to 75th percentiles are unchanged. At the 80th percentile, a CI can now be constructed, and it cannot rule out a strong beneficial effect (shortening LOS by 10.8 days), a strong harmful effect (lengthening LOS by 16.3 days), or no effect. The 90th percentile outcome is 28 days (the placement of death) in both the treatment group and the control group and the CI excludes all nonzero values. The bottom panel of Table 8 shows that even if we place death as equivalent to a 28-day LOS, the results do not change much from the top panel of Table 8 where death is the worst possible outcome. We anticipate that most people would want to place death as being equivalent to a longer LOS than 28 days or as the worst outcome, so the fact that the results do not change much as we vary the placement of death from a 28-day LOS to the worst outcome suggests that the results are robust to the placement of death.

The top panel of Table 9 shows estimated effects on the proportions of patients with outcomes at least as bad as various cutoff values, with 95% CIs based on the normal approximation. Death in the ICU is placed as the worst outcome, and the last row of the panel shows the estimated effect on mortality (the point estimate is 0.9 percentage point, but the CI ranges from -2.9 to 4.7 percentage points). All the CIs include zero, so there is no evidence that the intervention affected the ICU death rate or any of the other proportions. The bottom panel repeats the analysis with death and a 28-day LOS considered equally undesirable; the results are similar.

In sum, our analysis finds no evidence that the intervention affected the distribution of patients' outcomes, regardless of whether death is considered the worst possible outcome or placed as comparable to an LOS as short as 28 days. Since there was little difference in ICU mortality between the treatment and control groups, it is not surprising that the time-to-event analysis in Kerlin et al.[4] and the placement-of-death analysis presented here yield similar conclusions. However, the example illustrates the types of analyses that could be presented to address concerns about selection bias due to mortality effects in other trials.

5 Summary and discussion

We adapted Rosenbaum's[8] proposal for addressing the "censoring by death" problem[6,7] and showed how it could be applied to randomized trials where ICU length of stay is an outcome of interest. Our approach estimates treatment effects on the distribution of a composite outcome measure based on ICU mortality and survivors' LOS. Modifying Rosenbaum's proposal, we explored methods for inference about effects on quantiles of the outcome distribution or proportions of patients with outcomes better than a cutoff value, which may be easier to understand than Rosenbaum's original estimands.

Since it focuses on the composite outcome distribution, our approach does not estimate treatment effects on LOS per se. Researchers may understandably want to disentangle effects on LOS from effects on mortality, but opinions may differ on whether this can be done persuasively, since stronger assumptions would be needed.[6,7] Thus, the placement-of-death approach does not answer all relevant questions, but it may be a useful starting point. It addresses concerns about selection bias by comparing the entire treatment group with the entire control group, and it can provide evidence of an overall beneficial or harmful effect.

The approach allows sensitivity analysis with alternative placements of death, but it does make some restrictive assumptions about valuations of LOS and death. For example, it awards no credit for reducing long ICU stays of patients who would die in the ICU with or without the intervention, although such an effect may be in accordance with some patients' wishes. Extending the approach to accomodate more complicated valuations may be a useful direction for further work. Alternatively, a cost-benefit analysis could be considered from the societal perspective, assigning costs to death, time spent in the ICU, and other relevant considerations. However, the placement-of-death approach may be more appealing to some audiences because it avoids the need to assign a numeric value to death.

Of the significance tests we studied, the Brunner–Munzel test (or a permutation test based on the BM statistic) may be a reasonable choice. Some other rank tests may have more power when there are many ties[50,51] or when a small fraction of treated patients experience large treatment effects.[29] It is not clear that any of those other tests can be easily converted into robust tests of weak null hypotheses, but further investigation may be worthwhile.

Extension of the approach to cover cluster-randomized trials would also be valuable. Rosenbaum's[8] original approach provided exact confidence intervals in experiments with complete random assignment of individuals, but more complex designs create difficulties for exact inference.

Adjustment for treatment–control imbalances in the distributions of baseline covariates may be desired. One option that could be investigated is to combine the placement-of-death approach with inverse propensity score weighting.

A reviewer brought up the good point that by only looking at LOS in the ICU, we are ignoring the possibility of discharge and readmission a short time later. Brown et al.[58] showed that the appropriate cut point for defining ICU readmissions is 2 calendar days if the goal is to capture those readmissions most likely to have been due to ICU practices rather than patient characteristics. In the SUNSET-ICU trial, only 48 (3%) of the 1,598 patients had an ICU readmission within 48 hours. In studies where there are more readmissions, it might be better to define the LOS to include the original LOS and the LOS on a readmission if the readmission was within 48 hours.

Following Kerlin et al.[4], we have categorized patients who were transferred to inpatient hospice as having died at the time of discharge from the ICU and patients

who were transferred to home hospice or a long-term acute care hospital (LTACH) as alive at the time of discharge. These transfers could in some circumstances be considered competing risks to other live discharges and death. We categorized home hospice differently from inpatient hospice because patients transferred to home hospice typically live longer and because dying at home is commonly viewed as a favorable outcome; in Kerlin et al.[4], we performed a sensitivity analysis in which we categorized transfers to home hospice as deaths, and this did not alter the results much. Some transfers to LTACHs are "useful" and "timely" (i.e., the transfer happens at the right time and the patient ultimately benefits from it). In such cases, it seems reasonable to group transfers to LTACHs together with other live discharges, as we have done. However, other LTACH transfers might be "premature" (e.g., sending a patient to an LTACH to make room in the ICU even if the patient is not ideally suited for transfer) or "useless" (i.e., the patient never regained meaningful quality of life following LTACH transfer). Methods for dealing with these premature or useless transfers are of future research interest but beyond the scope of this paper, because there is not yet a standard way to distinguish between the types of LTACH transfers.

We have considered ICU LOS, but our approach could also be applied to studies of hospital LOS. We focused on ICU LOS because deaths are particularly common among ICU patients. Our proposed approach focuses only on ICU or hospital LOS and deaths that terminate the LOS, and does not consider time to death if a patient is discharged alive or future hospital stays after discharge. The reason is that most currently available datasets do not provide information about readmissions to a different hospital or time to death after discharge. If such data were available, an alternative to our approach would be to fit a 4-state stochastic model with states (1) in ICU, (2) in hospital, (3) out of hospital, and (4) dead.

In summary, ICU length of stay is a common outcome measure in randomized trials of ICU interventions, but currently used methods of analyzing ICU LOS in the critical care literature do not protect against possible selection biases due to deaths in the ICU. Our approach addresses this problem with a composite outcome measure that combines information on ICU LOS and deaths, and it allows the analysis to be adjusted to reflect different preferences of patients regarding comparisons between death and possible lengths of stay.

Acknowledgements

We would like to thank (without implicating) Peter Aronow, Tamara Broderick, Beth Cooney, Nicole Gabler, Michael Harhay, Paul Rosenbaum, Terry Speed, and two anonymous reviewers for helpful discussions.

Funding

Halpern acknowledges support from the Agency for Healthcare Research and Quality [grant number K08HS01846]. Small acknowledges support from the National Science Foundation Measurement, Methodology and Statistics Program [grant number NSF SES-1260782].

References

1 Harhay MO, Wagner J, Ratcliffe SJ, et al. Outcomes and statistical power in adult critical care randomized trials. Am J Respir Crit Care Med. 2014;189:1469–1478.

2 Lilly CM, Cody S, Zhao H, et al. Hospital mortality, length of stay, and preventable complications among critically ill patients before and after tele-ICU reengineering of critical care processes. JAMA. 2011;305:2175–2183.

3 Mehta S, Burry L, Cook D, et al. Daily sedation interruption in mechanically ventilated critically ill patients cared for with a sedation protocol: A randomized controlled trial. JAMA. 2012;308:1985–1992.

4 Kerlin MP, Small DS, Cooney E, et al. A randomized trial of nighttime physician staffing in an intensive care unit. N Engl J Med. 2013;368:2201–2209.

5 Freedman DA. Survival analysis: An epidemiological hazard? In: Collier D, Sekhon JS, Stark PB, editors. Statistical models and causal inference: A dialogue with the social sciences. Cambridge: Cambridge University Press; 2010. p. 169–192.

6 Joffe M. Principal stratification and attribution prohibition: Good ideas taken too far. Int J Biostat. 2011;7(1):Article 35.

7 Rubin DB. Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death (with discussion). Stat Sci. 2006;21:299–321.

8 Rosenbaum PR. Comment: The place of death in the quality of life. Stat Sci. 2006;21:313–316.

9 Rubin DB. Rejoinder. Stat Sci. 2006;21:319–321.

10 Brunner E, Munzel U. The nonparametric Behrens–Fisher problem: Asymptotic theory and a small-sample approximation. Biometrical J. 2000;42:17–25.

11 Neyman J. On the application of probability theory to agricultural experiments (with discussion). Stat Sci. 1990;5:463–480.

12 Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66:688–701.

13 Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. J Am Stat Assoc. 2005;100:322–331.

14 Holland PW. Statistics and causal inference (with discussion). J Am Stat Assoc. 1986;81:945–970.

15 Lehmann EL. Elements of large-sample theory. New York: Springer; 1999, pp. 146-157.

16 Reichardt CS, Gollob HF. Justifying the use and increasing the power of a *t* test for a randomized experiment with a convenience sample. Psychol Methods. 1999;4:117–128.

17 Freedman DA, Pisani R, Purves R. Statistics. 4th ed. New York: Norton; 2007, pp. 508–511.

18 Lin W. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. Ann Appl Stat. 2013;7:295–318.

19 Holland PW. Two measures of change in the gaps between the CDFs of test-score distributions. J Educ Behav Stat. 2002;27:3–17.

20 Agresti A, Caffo B. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. Am Stat. 2000;54:280–288.

21 Brown L, Li X. Confidence intervals for two sample binomial distribution. J Stat Plan Infer. 2005;130:359–375.

22 Hyndman RJ, Fan Y. Sample quantiles in statistical packages. Am Stat. 1996;50:361–365.

23 Hahn J. Bootstrapping quantile regression estimators. Economet Theor. 1995;11:105–121.

24 Efron B, Tibshirani RJ. An introduction to the bootstrap. Boca Raton, FL: CRC Press; 1993.

25 Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge: Cambridge University Press; 1997.

26 Politis DN, Romano JP, Wolf M. Subsampling. New York: Springer; 1999.

27 Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869.

28 Korn EL. Comment: Causal inference in the medical area. Stat Sci. 2006;21:310–312.

29 Rosenbaum PR. Confidence intervals for uncommon but dramatic responses to treatment. Biometrics. 2007;63:1164–1171.

30 Rosenbaum PR. Design of observational studies. New York: Springer; 2010.

31 Imbens GW, Wooldridge JM. Recent developments in the econometrics of program evaluation. J Econ Lit. 2009;47:5–86.

32 Lehmann EL. Parametric versus nonparametrics: Two alternative methodologies (with discussion). J Nonparametr Stat. 2009;21:397–426.

33 White IR, Thompson SG. Choice of test for comparing two groups, with particular application to skewed outcomes. Stat Med. 2003;22:1205–1215.

34 Romano JP. Discussion of "Parametric versus nonparametrics: Two alternative methodologies" by EL Lehmann. J Nonparametr Stat. 2009;21:419–424.

35 Chung EY, Romano JP. Asymptotically valid and exact permutation tests based on two-sample U-statistics. Department of Statistics, Stanford University; 2011. Technical Report 2011-09.

36 Gail MH, Mark SD, Carroll RJ, et al. On design considerations and randomization-based inference for community intervention trials. Stat Med. 1996;15:1069–1092.

37 Gibbons JD, Chakraborti S. Nonparametric statistical inference. 5th ed. Boca Raton, FL: CRC Press; 2011, pp. 292-293.

38 Pratt JW. Robustness of some procedures for the two-sample location problem. J Am Stat Assoc. 1964;59:665–680.

39 Fay MP, Proschan M. Wilcoxon–Mann–Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Stat Surv. 2010;4:1–39.

40 McGraw KO, Wong SP. A common language effect size statistic. Psychol Bull. 1992;111:361–365.

41 Neubert K, Brunner E. A studentized permutation test for the non-parametric Behrens–Fisher problem. Comput Stat Data Anal. 2007;51:5192–5204.

42 Ho AD. A nonparametric framework for comparing trends and gaps across tests. J Educ Behav Stat. 2009;34:201–228.

43 Lumley T. Rock, paper, scissors, Wilcoxon test;. Blog post, 2013, available at `http://notstatschat.tumblr.com/post/63237480043/rock-paper-scissors-wilcoxon-test`.

44 Lumley T. Ordinal data and "ordinal" tests;. Slides, 2013, available at `http://cbe-test.cbenet.anu.edu.au/media/2753826/transitive-anu.pdf`.

45 Lumley T. Loopy comparisons: When can more than two treatments be ranked?;. Slides, available at `http://faculty.washington.edu/tlumley/vanderbilt-seminar.pdf`.

46 Brown BM, Hettmansperger TP. Kruskal–Wallis, multiple comparisons and Efron dice. Aust N Z J Stat. 2002;44:427–438.

47 Thangavelu K, Brunner E. Wilcoxon–Mann–Whitney test for stratified samples and Efron's paradox dice. J Stat Plan Infer. 2007;137:720–737.

48 Aldous D. Review of *The cult of statistical significance* by S Ziliak and D McCloskey;. Available at `http://www.stat.berkeley.edu/~aldous/157/Books/stat.html`.

49 Miller RG. Beyond ANOVA: Basics of applied statistics. New York: Wiley; 1986, p. 51.

50 Follmann D, Fay MP, Proschan M. Chop-lump tests for vaccine trials. Biometrics. 2009;65:885–893.

51 Hallstrom AP. A modified Wilcoxon test for non-negative distributions with a clump of zeros. Stat Med. 2010;29:391–400.

52 Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics. 2002;58:21–29.

53 Lachenbruch PA. Comparison of two-part models with competitors. Stat Med. 2001;20:1215–1234.

54 Lachenbruch PA. Analysis of data with excess zeros. Stat Methods Med Res. 2002;11:297–302.

55 Moulton LH, Curriero FC, Barroso PF. Mixture models for quantitative HIV RNA data. Stat Methods Med Res. 2002;11:317–325.

56 Zhou XH. Inferences about population means of health care costs. Stat Methods Med Res. 2002;11:327–339.

57 Imai K, King G, Nall C. The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation (with discussion). Stat Sci. 2009;24:29–72.

58 Brown SES, Ratcliffe SJ, Halpern SD. An empirical derivation of the optimal time interval for defining ICU readmissions. Med Care. 2013;51:706–714.

Table 1: True quantiles of potential outcome distributions for simulations in Section 2.3. The table assumes the placement of death D is no less than 204.3 days. The second and third columns show quantiles of the distributions that treatment and control group patients' outcomes are randomly sampled from. Lengths of ICU stay are given in days.

Quantile	Treatment	Control
0.25	1.2	1.4
0.5 (median)	2.3	2.7
0.6	3.0	4.0
0.7	4.5	7.1
0.75	5.6	10.2
0.775	6.7	16.6
0.8	7.6	204.3
0.825	9.1	Death
0.85	11.7	Death
0.9	204.3	Death
0.95	Death	Death

Table 2: Coverage rates (in $10,000$ replications) of nominal 95% confidence intervals (bootstrap percentile method) for quantile treatment effects, assuming placement of death $D = 204.3$ days. For details of the simulation design, see Section 2.3.

Quantile	Coverage rate (percent)
0.25	95.9
0.5 (median)	95.8
0.6	95.9
0.7	95.7
0.75	95.0
0.775	95.1
0.8	88.2
0.825	96.4
0.85	95.8
0.9	97.2
0.95	100.0

Table 3: Empirical properties (in 10,000 replications) of nominal 95% confidence intervals (bootstrap percentile method) for quantile treatment effects, assuming death is the worst possible outcome. For details of the simulation design, see Section 2.3.

Quantile	% of confidence intervals that:		
	Cover true value	Miss true value	Are undefined
0.25	95.9	4.2	0.0
0.5 (median)	95.8	4.2	0.0
0.6	95.9	4.1	0.0
0.7	95.6	4.3	0.2
0.75	50.3	1.3	48.4
0.775	5.1	2.3	92.6
0.8	0.0	0.2	99.8
0.825	NA	NA	100.0
0.85	NA	NA	100.0
0.9	NA	NA	100.0
0.95	NA	NA	100.0

Table 4: Rejection rates (in 250,000 replications) of the Wilcoxon–Mann–Whitney and Brunner–Munzel tests in nine null-hypothesis scenarios. All tests are two-tailed with nominal significance level 5%. For details of the scenarios and simulation design, see Section 3.2.

Scenario	Rejection rate (%)	
	Wilcoxon–Mann–Whitney	Brunner–Munzel
Strong null		
Balanced design	5.0	5.1
90% treated	5.0	5.0
10% treated	5.0	5.0
Weak null (no ties)		
Balanced design	6.4	4.9
90% treated	14.7	5.0
10% treated	0.3	5.0
Weak null (with ties)		
Balanced design	5.7	5.0
90% treated	11.4	5.0
10% treated	1.2	5.0

Table 5: Rejection rates (in $10,000$ replications) of three significance tests in nine alternative-hypothesis scenarios. WMW = Wilcoxon–Mann–Whitney; BM = Brunner–Munzel. The $\text{QTE}_{0.5}$ test rejects if and only if the bootstrap percentile CI for the treatment effect on the median excludes zero. All tests are two-tailed with nominal significance level 5%. For details of the scenarios and simulation design, see Section 3.3.

	WMW	BM	$\text{QTE}_{0.5}$
Reject, correctly infer beneficial effect (%)			
Sample size $= 1,500$			
Setting A	53.9	53.7	9.9
Setting B	30.8	30.7	60.7
Setting C	39.0	39.0	60.7
Setting D	7.5	7.5	3.6
Setting E	5.2	5.2	7.4
Setting F	8.2	8.2	7.1
Sample size $= 30,000$			
Setting D	62.3	62.3	7.5
Setting E	37.1	37.1	68.8
Setting F	71.1	71.1	71.0
Reject, incorrectly infer harm (%)			
Sample size $= 1,500$			
Setting A	0.0	0.0	0.4
Setting B	0.0	0.0	0.0
Setting C	0.0	0.0	0.0
Setting D	0.6	0.6	1.4
Setting E	1.0	1.0	0.5
Setting F	0.6	0.6	0.5
Sample size $= 30,000$			
Setting D	0.0	0.0	0.3
Setting E	0.0	0.0	0.0
Setting F	0.0	0.0	0.0

Table 6: Rejection rates (in $250,000$ replications) of a two-part test in nine null-hypothesis scenarios (see Section 3.4 for details). All tests are two-tailed with nominal significance level 5%. Compare to Table 4.

Scenario	Rejection rate (%)
Strong null	
Balanced design	5.1
90% treated	5.5
10% treated	5.6
Weak null (no ties)	
Balanced design	5.1
90% treated	5.5
10% treated	5.5
Weak null (with ties)	
Balanced design	5.0
90% treated	5.6
10% treated	5.5

Table 7: Rejection rates (in $10,000$ replications) of a two-part test in nine alternative-hypothesis scenarios (see Section 3.4 for details). All tests are two-tailed with nominal significance level 5%. Compare to Table 5.

Sample size $= 1,500$	
Setting A	49.0
Setting B	42.7
Setting C	29.6
Setting D	7.8
Setting E	7.0
Setting F	18.9
Sample size $= 30,000$	
Setting D	59.3
Setting E	51.8
Setting F	99.9

Table 8: Estimated quantile treatment effects in the SUNSET trial. The second and third columns show the sample quantiles for the treatment and control groups (lengths of ICU stay are given in days). The fourth column shows their difference, the estimated QTE. The fifth column shows a 95% confidence interval (bootstrap percentile method) for the QTE. The top panel assumes death is the worst possible outcome. The bottom panel assumes death and a 28-day ICU stay are considered equally undesirable.

Quantile	Treatment	Control	Difference	95% CI
Death is worst outcome				
0.25	1.4	1.3	0.1	$[-0.1, 0.3]$
0.5 (median)	2.9	2.6	0.2	$[-0.1, 0.7]$
0.6	4.2	3.8	0.3	$[-0.3, 1.3]$
0.7	7.8	6.2	1.7	$[-0.3, 3.6]$
0.75	11.0	8.9	2.2	$[-1.5, 7.8]$
0.8	30.0	17.8	12.2	Undefined
0.9	Death	Death	Undefined	Undefined
Death placed at 28 days				
0.25	1.4	1.3	0.1	$[-0.1, 0.3]$
0.5 (median)	2.9	2.6	0.2	$[-0.1, 0.7]$
0.6	4.2	3.8	0.3	$[-0.3, 1.3]$
0.7	7.8	6.2	1.7	$[-0.3, 3.6]$
0.75	11.0	8.9	2.2	$[-1.5, 7.8]$
0.8	28.0	17.8	10.2	$[-10.8, 16.3]$
0.9	28.0	28.0	0.0	[0.0, 0.0]

Table 9: Estimated cutoff treatment effects in the SUNSET trial. The second and third columns show the treatment and control group sample proportions with outcomes at least as bad as the cutoff. The fourth column shows their difference, the estimated CTE. The fifth column shows a 95% confidence interval (normal approximation) for the CTE. The top panel assumes death is the worst possible outcome. The bottom panel assumes death and a 28-day ICU stay are considered equally undesirable.

Cutoff	% with outcome at least as bad as cutoff			
	Treatment	Control	Difference	95% CI
Death is worst outcome				
1 day in ICU	84.4	83.5	0.8	$[-2.8, 4.4]$
2 days	61.0	59.6	1.3	$[-3.5, 6.1]$
3 days	48.7	45.2	3.4	$[-1.5, 8.3]$
4 days	40.9	38.6	2.3	$[-2.5, 7.1]$
1 week	32.0	28.4	3.5	$[-1.0, 8.0]$
2 weeks	23.5	21.0	2.6	$[-1.5, 6.7]$
4 weeks	20.1	18.4	1.7	$[-2.1, 5.6]$
Death	18.8	17.9	0.9	$[-2.9, 4.7]$
Death placed at 28 days				
1 day in ICU	84.4	83.5	0.8	$[-2.8, 4.4]$
2 days	61.0	59.6	1.3	$[-3.5, 6.1]$
3 days	48.7	45.2	3.4	$[-1.5, 8.3]$
4 days	40.9	38.6	2.3	$[-2.5, 7.1]$
1 week	32.0	28.4	3.5	$[-1.0, 8.0]$
2 weeks	23.5	21.0	2.6	$[-1.5, 6.7]$
4 weeks	20.1	18.4	1.7	$[-2.1, 5.6]$
5 weeks	0.7	0.3	0.5	$[-0.2, 1.2]$