A novel high-power test for continuous outcomes truncated by death

Andreas Kryger Jensen and Theis Lange
Section of Biostatistics, Department of Public Health
University of Copenhagen

05 February, 2023

Abstract

Patient reported outcomes including quality of life (QoL) assessments are increasingly being included as either primary or secondary outcomes in randomized controlled trials. While making the outcomes more relevant for patients it entails a challenge in cases where death or a similar event makes the outcome of interest undefined. A pragmatic - and much used - solution is to assign diseased patient with the lowest possible QoL score. This makes medical sense, but creates a statistical problem since traditional tests such as t-tests or Wilcoxon tests potentially loses high amounts of statistical power. In this paper we propose a novel test that can keep the medically relevant composite outcome but preserve full statistical power. The test is also applicable in other situations where a specific value (say zero days alive outside hospitals) encodes a special meaning. The test is implemented in an R package which is available for download. The paper includes a simulation study and an application to a COVID-19 trial.

Keywords: Quality of Life, RCTs, Truncated Outcomes, Zero-inflation,

1 Introduction

In medical intervention research including in particular randomized controlled trials (RCTs) there is a trend towards increased use of patient reported outcomes; Quality of Life (QoL) scores being one of the most prominent of these. Established statistical practice is to compare these between treatment groups using non-parametric tests such as Mann–Whitney–Wilcoxon rank-sum test (Wilcoxon (1945), Mann and Whitney (1947)), hence henceforth Wilcoxon, since distributions are rarely normal or symmetric. When all patients are alive and able and willing to provide QoL scores at the scheduled measurement time this procedure works very well. Not all clinical settings, however, will have all participants alive at time of scheduled assessment of QoL. This is particularly true in trials within intensive care where mortality of approximately 30% is not uncommon and which is the setting that motivated this work. A pragmatic solution which is gaining popularity is to simply give the deceased patients the lowest possible score and then use a Wilcoxon test or similar to compare groups. This paper will demonstrate that this approach can lead to dramatic loss of statistical power. The paper will also present an alternative and novel statistical test that avoids the power loss. The novel test procedure is implemented in R and made publicly available.

The key-insight in the proposed test is to incorporate that the outcome is actually a two-dimensional outcome of a very special type, and that the constructed combined outcome follows a continuous-singular mixture distribution. This unusual distribution is why one cannot resort to non-parametric Wilcoxon tests since the singular component in the distribution of the combined outcome will get reduced to simple ties. It is noted that the handling of ties in standard statistical software varies and is opaque. However, the handling of ties is not the main reason why the Wilcoxon test suffers power loss. The main reason is that the null-hypothesis in these Wilcoxon type tests (stochastic domination) does not handle the empirical fact that treatments might influence mortality and QoL differently.

As an alternative we propose to model the binary component (i.e., survival) and the continuous part (i.e., actual QoL) separately but to construct a single test for no treatment effect on either based on a likelihood ratio. We can thus provide a single p-value for the hypothesis of no treatment effect on the extended QoL where death is given the lowest possible score. To accommodate potential non-normality of the recorded QoL scores we consider both a parametric and a semi-parametric version of the test where the semi-parametric version is as widely applicable as the Wilcoxon test. Both test procedures will provide effect estimates of mean differences between treatment groups based on the combined outcome along with confidence intervals. This is also an added benefit compared to the Wilcoxon-type testing approach.

It should be noted that while we in this paper exemplify the procedure using QoL and mortality the method is applicable in any setting where a single value of a combined outcome has a special interpretation compared to an otherwise continuous scale. Another example from intensive care research consider the outcome “days alive and out of hospital within 90 days from randomization.” Here in-hospital fatalities will all have the value 0, while everybody else will have outcomes ranging from 0 to 90. Such outcomes are also routinely analyzed using Wilcoxon-type tests. Our proposed test would increase power while also providing a mean effect estimate along with confidence interval. Note also that the developed mathematical method allows straightforwardly for the inclusion of confounding variables or other adjustment factors. The method itself is therefore just as applicable in non-randomized studies or epidemiological studies in general.

The rest of the paper is structured as follows. The next section introduces the mathematical setup as well as our novel test procedure. Section 3 describes the R implementation, and Section 4 presents a simulation study illustrating the substantial power gains. Section 5 re-analyzes a COVID-19 trial from 2021, and section 6 concludes with a discusses. Additional simulation results and mathematical proofs are contained in the supplementary material.

2 Method

We consider random variables (Y, A, R, X) where Y is the continuous outcome, A is a binary variable being equal to 1 if Y is observed and equal to 0 if Y is unobserved/undefined, R is a binary treatment indicator, and X is a p-dimensional vector of additional covariates. The objective is to focus on the bivariate distribution of $(Y \mid A = 1), A \mid R, X$ under the primary null-hypothesis of no treatment effect given by $(Y \mid A = 1), A \perp R \mid X$. Even though we model the outcome as a bivariate random variable, a combined outcome is often used in practice. The combined outcome is derived such that it is equal to Y if $A = 1$ (and Y is observed), and otherwise (when $A = 0$) it is equal to some predetermined value. The combined outcome can be written as

$$\widetilde{Y} = Y\mathbb{1}(A = 1) + \mathcal{E}\mathbb{1}(A = 0) \tag{1}$$

where $\mathcal{E}$ is a fixed atom assigned as the outcome value when Y is unobserved, and $\mathbb{1}$ denotes the indicator function. The semi-continuous distribution of $\widetilde{Y}$ is therefore a probabilistic mixture of a singular distribution at $\mathcal{E}$ and a continuous distribution over the domain of the random variable $Y \mid A = 1$. Thus the statistical challenge can be rephrased as assert whether the treatment affects the distribution of $\widetilde{Y}$.

The conditional expected value of the combined outcome in Equation (1) is given by

$$E[\widetilde{Y} \mid R, X] = E[Y \mid R, X, A = 1]P(A = 1 \mid R, X) + \mathcal{E}P(A = 0 \mid R, X)$$

From this expression one can show that a treatment comparison expressed in terms of a contrast of the expectation of $\widetilde{Y}$ can be zero even though the distribution of $((Y \mid A = 1), A)$ depends on R. To see this, let $E[Y \mid A = 1, R = 0, X] = \mu(X)$ and $E[Y \mid A = 1, R = 1, X] = \mu(X) + \mu_\delta(X)$ and assume without loss of generalization that $\mathcal{E} = 0$. Then the average treatment effect for the combined outcome conditional on baseline covariates is

$$\begin{aligned}\Delta(X) &= E[\widetilde{Y} \mid R = 1, X] - E[\widetilde{Y} \mid R = 0, X] \\ &= \mu(X)\left[P(A = 1 \mid R = 1, X) - P(A = 1 \mid R = 0, X)\right] + \mu_\delta(X)P(A = 1 \mid R = 1, X)\end{aligned} \tag{2}$$

If we auspiciously let $\mu_\delta(X) = \mu(X)\left[P(A = 1 \mid R = 0, X)P(A = 1 \mid R = 1, X)^{-1} - 1\right]$ then $\Delta(X) = 0$ for all values of X even though $\mu_\delta(X) \neq 0$ and $P(A = 1 \mid R = 0, X) \neq P(A = 1 \mid R = 1, X)$. On the other hand, if $\mu_\delta(X) = 0$ and $P(A = 1 \mid R = 1, X) = P(A = 1 \mid R = 0, X)$ then $\Delta(X)$ is necessarily equal to zero. This illustrates that a significance test of no treatment effect must have two degrees of freedom. Such a test which we develop in the subsequent sections is conceptually difference than a test for the null-hypothesis $\Delta(X) = 0$. Noe that in RCTs one would typically not include any variables X since these are balanced by design.

We propose to test the null-hypothesis of no treatment effect on the combined outcome by a likelihood ratio test of the joint distribution of $Y_i \mid A_i$ and A_i. This has the advantage that it increases the efficiency compared the Wilcoxon test and it yields a single p-value appropriate for testing a primary outcome in a clinical trial. Let (Y_i, A_i, R_i, X_i) for $i = 1, \dots, n$ be independent and identically distributed random variables. We can write our model in a general form as

$$\begin{aligned}Y_i \mid A_i = 1, R_i, X_i &\sim F(Y_i; \mu_i(R_i, X_i), \Psi) \\ A_i \mid R_i, X_i &\sim \text{Bernoulli}(A_i; \pi_i(R_i, X_i))\end{aligned}$$

for some mean functions π_i and μ_i and a distribution function F characterizing the distribution of the observed continuous outcomes with possible nuisance parameter Ψ. The joint likelihood function for the combined outcome conditional on treatment and baseline covariates is

$$\begin{aligned}L_n(\mu_i(R_i, X_i), \pi_i(R_i, X_i), \Psi) &= \prod_{i=1}^{n} f(Y_i; \mu_i(R_i, X_i), \Psi)^{A_i} P(A_i = a_i; \pi_i(R_i, X_i)) \\ &= \prod_{i=1}^{n} f(Y_i; \mu_i(R_i, X_i), \Psi)^{A_i} \pi_i(R_i, X_i)^{A_i}(1 - \pi_i(R_i, X_i))^{1-A_i}\end{aligned}$$

We note that when F admits an absolutely continuous density function the likelihood function can equivalently be written solely in terms of the combined outcome $\widetilde{Y}$ in Equation (1) since $f(Y_i; \cdot)^{A_i} = f(\widetilde{Y}_i; \cdot)^{1_{\widetilde{Y}_i \neq \mathcal{E}}}$ and $\pi_i(\cdot)^{A_i} = \pi_i(\cdot)^{1_{\widetilde{Y}_i \neq \mathcal{E}}}$ almost surely. An important property of this model is that the likelihood function factorizes into two components as

$$L_n(\mu_i(R_i, X_i), \pi_i(R_i, X_i)) = L_{1,n}(\mu_i(R_i, X_i), \Psi)L_{2,n}(\pi_i(R_i, X_i)) \tag{3}$$

where $L_{1,n}(\mu_i(R_i, X_i), \Psi) = \prod_{i=1}^{n} f(Y_i; \mu_i(R_i, X_i), \Psi)^{A_i}$ and $L_{2,n}(\pi_i(R_i, X_i)) = \prod_{i=1}^{n} \pi_i(R_i, X_i)^{A_i}(1 - \pi_i(R_i, X_i))^{1-A_i}$. This implies that the parameters for the observed outcomes, $\mu_i(R_i, X_i)$, can be estimated independently of the parameters governing the probability of observing the outcome, $\pi_i(R_i, X_i)$. This factorization is a consequence of the likelihood construction and it does neither assume nor require independence between the value of the outcome and the probability of observing it.

Under the assumption of generalized linear models for $\mu_i(R_i, X_i)$ and $\pi_i(R_i, X_i)$ with additive structures we may parametrize the mean functions as

$$\begin{aligned} \mathrm{E}[Y_i \mid A_i = 1, R_i, X_i] &= \mu_0 + \mu_\delta R_i + s_\mu(X_i) && (4) \\ \operatorname{logit} P(A_i = 1 \mid R_i, X_i) &= \beta_0 + \beta_\delta R_i + s_\beta(X_i) && (5) \end{aligned}$$

where the treatment effect is quantified by bi-variate contrast $(\mu_\delta, \beta_\delta)$ and $s_\mu, s_\beta \colon \mathbb{R}^p \mapsto \mathbb{R}$ are some models for the baseline covariates. The parameter μ_δ is interpreted as the expected difference among the observed outcomes and β_δ is correspondingly the log odds-ratio of being observed.

To assess the effect of treatment on the combined outcome we propose a test statistic based on the likelihood ratio statistic

$$\begin{aligned} W_n(\mu_0, \mu_\delta, s_\mu, \beta_0, \beta_\delta, s_\beta) &= -2 \log \frac{\sup_{\Psi} L_{1,n}(\mu_0, \mu_\delta, s_\mu, \Psi) L_{2,n}(\beta_0, \beta_\delta, s_\beta)}{\sup_{\mu_0, \mu_\delta, s_\mu, \Psi} L_{1,n}(\mu_0, \mu_\delta, s_\mu, \Psi) \sup_{\beta_0, \beta_\delta, s_\beta} L_{2,n}(\beta_0, \beta_\delta, s_\beta)} && (6) \\ &= W_{1,n}(\mu_0, \mu_\delta, s_\mu) W_{2,n}(\beta_0, \beta_\delta, s_\beta) && (7) \end{aligned}$$

$$W_n^P(\mu_\delta, \beta_\delta) = \sup_{\mu_0, s_\mu} W_{n,1}(\mu_0, \mu_\delta, s_\mu) \sup_{\beta_0, \beta_s} W_{n.2}(\beta_0, \beta_\delta, s_\beta) \qquad (8)$$

use the profile likelihood ratio test (LRT) statistic which can be written as a function of the treatment effects as and the value of the LRT statistic under the null-hypothesis of no treatment effect is therefore $W_n(0, 0)$. Similar to the factorization of the likelihood function in Equation (3), the LRT statistic in Equation (6) also decomposes into the sum

$$W_n(\mu_\delta, \beta_\delta) = W_{1,n}(\mu_\delta) + W_{2,n}(\beta_\delta) \qquad (9)$$

of two LRT statistics – one for the continuous part and one for the discrete part of the combined outcome. Under very general conditions it follows that $W_n(\mu_\delta, \beta_\delta)$ is approximately χ^2 distributed with two degrees of freedom (Wilks 1938; A. B. Owen 1988).

In order to actually perform the test we still need to decide on an stochastic model for the continuous part of the combined outcome, $Y_i \mid A_i = 1$. In the following two sections we present both a parametric approach based on a normal model and a flexible semi-parametric approach that does not require distributional assumptions.

Parametric approach

If we combine the model for the expected value in Equation (4) with the assumption of normal distributed outcomes of the continuous part of the combined outcome we obtain the following sub-model

$$Y_i \mid A_i = 1, R_i \sim N\left(\mu_0 + \mu_\delta R_i, \sigma^2\right)$$

$$W_{1,n}(0) = (m_1 + m_2) \log \frac{\widehat{\sigma_0^2}}{\widehat{\sigma^2}}$$

With the generalized linear models in Equations (4) and (5) the first term is the LRT statistic with a single degree of freedom in a linear regression model, and the second term is the LRT statistic with a single degree of freedom in a logistic regression model. This makes this test very easy to perform in standard statistical software that outputs the likelihood value for a model fit without the need of special methods. The calculation of the LRT simply amounts to estimating two linear regression models and two logistic regression models – for each type a model with and a model without the binary treatment indicator as a covariate. Combining the two likelihood ratios according to Equation (9) calculates our test statistic, and the p-value for no treatment effect can be determined through the χ^2-distribution with two degrees of freedom. The critical value for rejecting the null-hypothesis of no treatment effect at the 5% level is equal to 5.99 and equal to 9.21 at the 1% level.

A direct consequence of Wilks' theorem (Wilks 1938) is given in the following proposition. Note that the stated conditions are trivially satisfied for an RCT with fixed randomization proportions.

Proposition 1. *Assume that $P(A_i = 1 \mid R_i = r) > 0$ for $r = 0, 1$ and let $m_j = \sum_{i=1}^{n} 1_{R_i=j}$. Then the parametric profile likelihood ratio test statistic $W_n^P(0,0)$ stated in Equation (8) for the null-hypothesis of no treatment effect on the combined outcome is asymptotically χ^2 distributed with two degrees of freedom for $m_1, m_2 \to \infty$ and $m_1/m_2 \to c$ for some finite, positive constant c.*

Semi-parametric approach

A drawback of the parametric LRT introduced in the previous section is that it requires deciding on a parametric distribution for the observed outcomes. In this section we introduce a more flexible approach based on an empirical LRT. This approach also utilizes the additive decomposition of the LRT statistic in Equation (9) but substitutes the term $W_{1,n}$ with a term that is free of any distributional assumptions. Combining this with the binomial model for A leads to a semi-empirical LRT for the treatment effect.

Let $\mathcal{I}_0 = \{i = 1, \ldots, n : R_i = 0, A_i = 1\}$ and $\mathcal{I}_1 = \{i = 1, \ldots, n : R_i = 1, A_i = 1\}$ be the sets of indices of the observed outcomes for the two treatments. The empirical LRT statistic as a function of the difference in expected value is given by

$$W_{1,n}^E(\mu_\delta^*) = 2 \sup_\mu \left(\sum_{i \in \mathcal{I}_0} \log\left(1 + \lambda_1(Y_i - \mu)\right) + \sum_{j \in \mathcal{I}_1} \log\left(1 + \lambda_2(Y_j - \mu - \mu_\delta^*)\right) \right)$$

where λ_1 and λ_2 are the solutions to the following equations

$$|\mathcal{I}_0|^{-1} \sum_{i \in \mathcal{I}_0} \frac{Y_i - \mu}{1 + \lambda_1(Y_i - \mu)} = 0, \quad |\mathcal{I}_1|^{-1} \sum_{j \in \mathcal{I}_1} \frac{Y_j - \mu - \mu_\delta^*}{1 + \lambda_2(Y_j - \mu - \beta_\delta^*)} = 0$$

We refer to the appendix for a derivation of the test statistics.

The semi-parametric LRT statistic for testing the null-hypothesis of no treatment effect is equal to $W_n^{SP}(0,0)$ where

$$W_n^{SP}(\mu_\delta^*, \alpha_\delta^*) = W_{1,n}^E(\mu_\delta^*) + W_{2,n}(\alpha_\delta^*) \tag{10}$$

By the non-parametric Wilk's theorem (A. B. Owen 1988) it follows that $W_n^{SP}(0,0) \sim \chi^2_2$ asymptotically.

Proposition 2. *Assume that $P(A_i = 1 \mid R_i = r) > 0$ for $r = 0, 1$ and let $m_j = \sum_{i=1}^n 1_{R_i=j}$. Then the semi-parametric profile likelihood ratio test statistic $W_n^{SP}(0,0)$ stated in Equation (10) for the null-hypothesis of no treatment effect on the combined outcome is asymptotically χ^2 distributed with two degrees of freedom for $m_1, m_2 \to \infty$ and $m_1/m_2 \to c$ for some finite, positive constant c.*

Confidence intervals

Confidence intervals for the treatment effects are readily obtained by inverting the likelihood ratio functions in either Equations (9) or (10) for either version of the test utilizing the duality between hypothesis testing and confidence intervals. Specifically, an $(1-\alpha)100\%$ confidence region or interval contains all parameter values that cannot be rejected according to the likelihood ratio test at level α. The bivariate confidence region for the treatment effect is therefore given by the following point set in $\mathbb{R}^2$

$$\mathrm{CI}_{1-\alpha}^{(\mu_\delta,\beta_\delta)} = \left\{(m,b) \in \mathbb{R}^2 : W_{1,n}(m) + W_{2,n}(b) \leq \chi^2_2(1-\alpha)\right\}$$

and it will asymptotically contain the true difference in means among the observed outcomes and the true log odds-ratio of being observed simultaneously with probability $(1-\alpha)100\%$. Similarly, univariate confidence intervals for μ_δ and β_δ can be computed by inversion with respect to a χ^2 distribution with one degree of freedom, e.g.,

$$\mathrm{CI}_{1-\alpha}^{\mu_\delta} = \left\{m \in \mathbb{R} : W_{1,n}(m) \leq \chi^2_1(1-\alpha)\right\}, \quad \mathrm{CI}_{1-\alpha}^{\beta_\delta} = \left\{b \in \mathbb{R} : W_{2,n}(b) \leq \chi^2_1(1-\alpha)\right\}$$

3 Software implementation

To facilitate a straightforward application of our approach we have implemented both the parametric and semi-parametric likelihood ratio tests in the R package `TruncComp` which is available at the GitHub repository (Jensen and Lange 2018). We illustrate its applicability based on an example data set also available from the package.

The example data set can be loaded by writing `data("TruncCompExample")` after loading the package. The data set contains two variables, Y and R, where Y is the outcome and R is the binary treatment indicator. There are 25 observations in each treatment group, and truncated observations in Y have been assigned the atom $\mathcal{E} = 0$. Figure 1 shows histograms of the outcome for each treatment group. Visually there appears to be a difference between the two groups both in terms of the frequency of the atom at zero and a location shift in the continuous part.

The observed difference in means for the combined outcome, Δ in Equation (2), is 0.018 and both a two-sample t-test and a Wilcoxon rank sum test show highly insignificant effects of the treatment with p-values of 0.984 and 0.696 respectively. In order to analyse the data using proposed method we use the function `truncComp` as follows

```
model <- truncComp(Y ~ R, atom = 0, data = TruncCompExample, method = "SPLRT")
```

where the formula interface is similar to other regression models implemented in R. The argument `atom` identifies the value assigned to the truncated outcomes, and `method` can be `SPLRT` or `LRT` for either the semi-parametric or the parametric likelihood ratio test respectively. In this example we

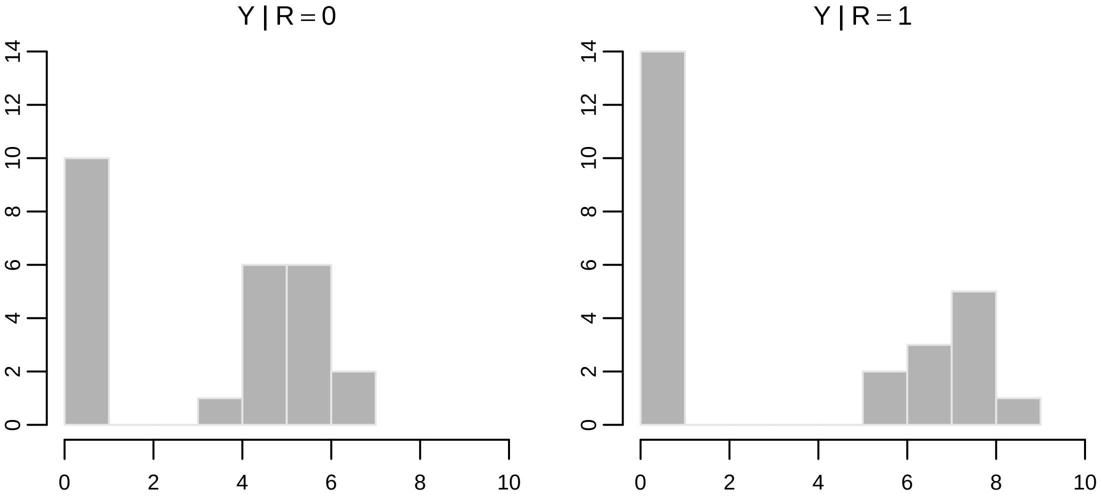

Figure 1: Histograms of the outcome for the example data stratified by treatment group.

have opted for the semi-parametric version to show how easily this additional flexibility is included. We obtain the results of the estimation by a calling the function `summary` on the model:

```
summary(model)
```

```
Estimation method: Semi-empirical Likelihood Ratio Test
Confidence level = 95%

Treatment contrasts
                                        Estimate  CI Lower CI Upper
Difference in means among the observed: 1.8564296 1.1638863 2.480132
Odds ratio of being observed:           0.5238095 0.1660407 1.596820

Joint test statistic: W = 31.09545
p-value: p = 1.768924e-07
```

The output from the call to `summary` displays the estimates for the two treatment contrasts corresponding to μ_δ and $\exp(\alpha_\delta)$ in Equations (4) and (5). These contrasts quantify the difference in means among the observed outcomes and the odds ratio of being observed respectively in accordance with the model specification. Each estimated treatment contrast is accompanied by a 95% confidence interval, and finally the output displays the joint likelihood ratio test statistic and the associated p-value for the joint null-hypothesis of no treatment effect.

From the output we see that the semi-parametric likelihood ratio analysis reports an extremely low p-value for null-hypothesis of no joint treatment effect. This strongly contradicts the conclusions from both the previous t-test and Wilcoxon test. The confidence intervals for the two treatment contrasts indicate that the average value for the observed outcome in the group defined by $R = 1$ is significantly higher than the average value for the observed outcome in the group with $R = 0$.

The confidence intervals can also be obtained by calling the function `confint` on the model object. This command reports both the marginal confidence intervals for the two treatment contrasts as well as a their simultaneous confidence region. To obtain the simultaneous confidence region we write

```
confint(model, type = "simultaneous", plot = TRUE, offset = 1.4, resolution = 50)
```

where `resolution` is the number of discrete grid points on which the likelihood surface is evaluated for plotting. Figure 2 shows a heat-map of the semi-parametric likelihood surface as well as the 95% simultaneous confidence region. The point $(0, 0)$ is far outside of the joint confidence region which corresponds to the strong rejection of the joint null hypothesis of no treatment effect.

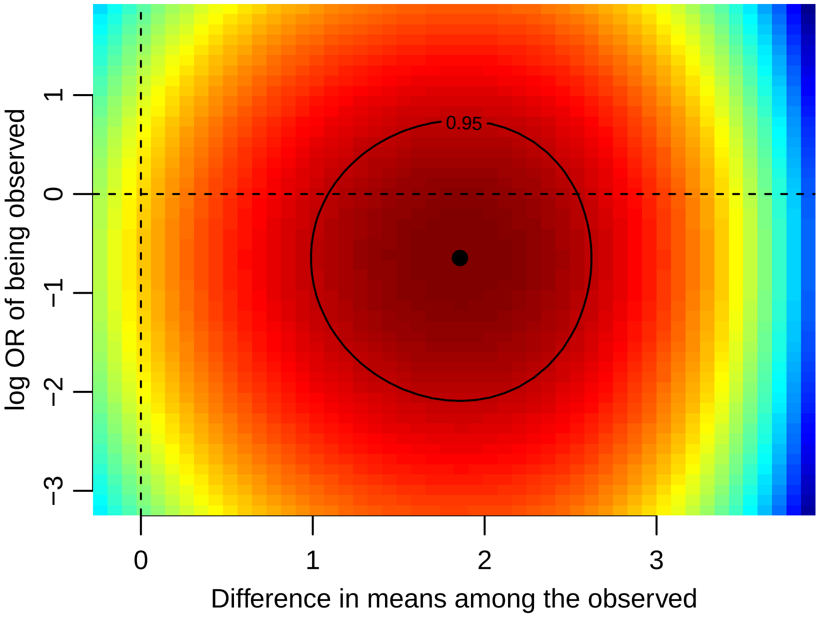

Figure 2: Simultaneous empirical likelihood ratio surface for the two treatment contrasts.

4 Simulation study

To illustrate the power benefit and small sample properties of our proposed procedure we consider four different scenario through a simulation study. In all scenarios it is assumed that treatment (denoted R_i) is randomized one to one between two groups, and we let the observed combined outcome be given by the product $\widetilde{Y}_i = A_i Y_i$, where A_i is binary and Y_i is continuously distributed. This is equivalent to observing the atom $\mathcal{E} = 0$ when $A_i = 0$ and otherwise Y_i is observed.

Table 1 shows the distributions of $A_i \mid R_i$ and $Y_i \mid A_i = 1, R_i$ for each of the four simulation scenarios. Scenario 1 represents the case where the point mass at zero is independent of the treatment but with a continuous component that is shifted between the groups. Scenario 2 is the opposite of scenario 1 where the continuous component is independent of the treatment but the probability of the point mass at zero differs between treatment groups. In scenario 3 we model a treatment effect on both components but in opposite directions of each other, and scenario 4 is similar to scenario 1 but where the continuous component is non-normal distributed.

To assess the power we vary the sample size between 50 to 350 observations in each group with increments of 25, and for each scenario we perform 25,000 simulations and estimate the power by the proportion of rejected null hypotheses at the 5% level. We compare the power of both our parametric and semi-parametric likelihood ratio tests to the two-sample t-test and the Wilcoxon test. Figure 3 shows the estimated power functions. The results are also presented in table form in

Scenario	$A_i \mid R_i$	$Y_i \mid A_i = 1, R_i$
1	Bernoulli(0.35)	$N(3 + 0.5R_i, 1)$
2	Bernoulli$(0.5 + 0.15R_i)$	$N(3.5, 1)$
3	Bernoulli$(0.4 - 0.1R_i)$	$N(3 + 0.5R_i, 1)$
4	Bernoulli(0.35)	Beta$(1, 1 - 0.7R_i)$

Table 1: Power simulation scenarios

the supplementary material.

In all scenarios except number two it is clear that we observe a large power gain compared to both the t-test and the Wilcoxon test.

In setup 1 we observe a large power gain compared to the Wilcox test. Here the Wilcox tests gets “confused” by the large number of ties in the atom. In setup 2 Wilcox test has slightly better power profile. This is to be expected as our novel test is here disadvantaged by being a two-degrees-of-freedom test where the Wilcox is only one. It is further observed that Wilcox as expected as very little power when the effects on mortality and among survivors are of opposite sign despite the two distributions being markedly different indicating clear treatment effect (setups 3 and 4). In contrast our novel methods has excellent power. In all settings with a normally distributed outcome among survivors the parametric and semi-parametric approaches are similar, but for the heavy tail setup (no. 4) the semi-parametric approach is clearly superior.

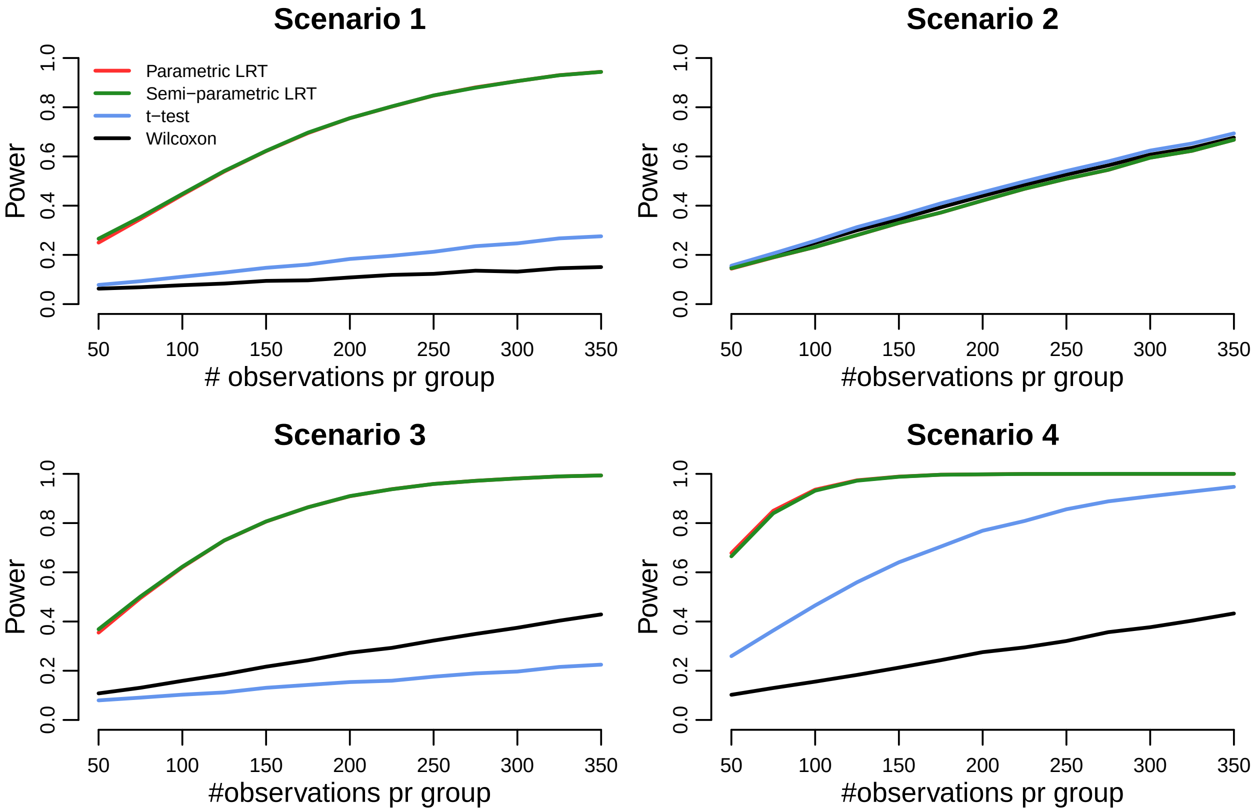

Figure 3: Simulated power as a function of sample size for each of the four scenarios.

In addition to the power simulations we also performed a simulation study for the type I error.

We considered two scenarios similar to scenarios 1 and 4 in Table 1 under a null-hypothesis of no treatment effect. Specifically, we considered $A_i \sim \text{Bernoulli}(\pi)$, $\widetilde{Y}_i \mid A_i = 1 \sim N(3,1)$ and $A_i \sim \text{Bernoulli}(\pi)$, $\widetilde{Y}_i \mid A_i = 1 \sim \text{Beta}(1,1)$. We estimated the type I error through simulation for different sample sizes and $\pi \in \{0.2, 0.4, 0.6, 0.8\}$. The results are shown in the supplementary material.

5 Application

The COVID-STEROID2 trial was a multi-center study comparing daily dose 6 mg (low) of dexamethasone for up to 10 days vs. 12 mg (high) in patients with severe and critical COVID-19. The full description of the study including all results can be found in The COVID STEROID 2 Trial Group (2021). The primary outcome of the trial was days alive and without out use of life-support in the period from randomization to day 28. In other words, that outcome is an integer between 0 and 28 with a substantial peak at day 0 coming from the patients who are never taken off life-support systems. The analyses in the main publications (The COVID STEROID 2 Trial Group (2021)) were conducted using the test proposed in this paper. In the JAMA publication the test is being referred to as the Kryger Jensen and Lange test. In this section we re-analyze the sub-group consisting of patients in need of invasive mechanical ventilation at baseline (n = 206). These can generally be taken to be the most severely ill patients.

The distribution of the outcome in each randomization group is presented by histograms in Figure 4 (left and middle panels). In our notation the outcome (days alive without life support) is re-coded such that A takes the value zero if the outcome is zero or one otherwise. Y is said to be the observed number of days without life support.

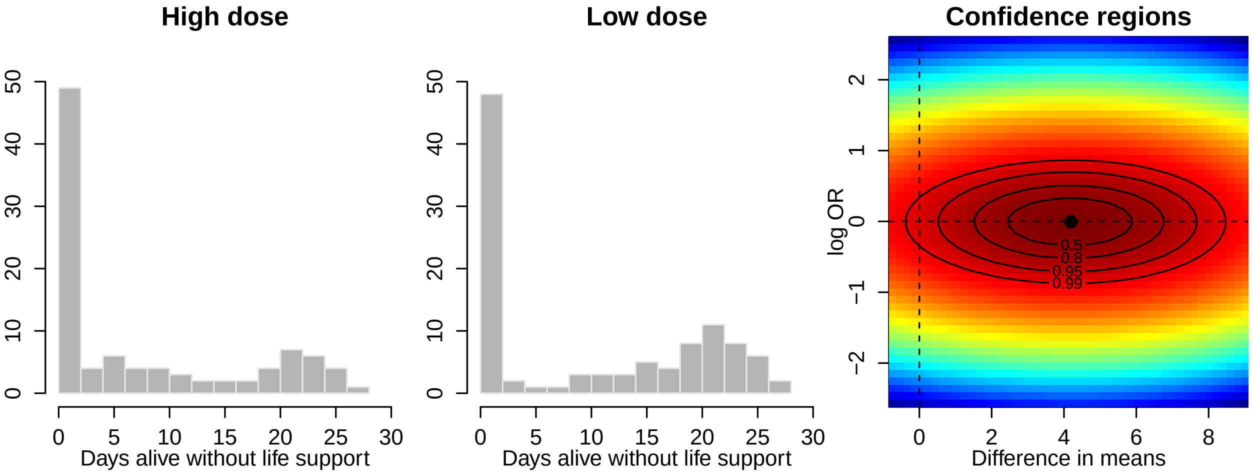

Figure 4: Histograms of days alive without life support from the The COVID-STEROID2 trial stratified by treatment dosis (left, middle) and simultaneous confidence regions based on the semi-parametric likelihood ratio test (right).

Comparing the groups by a Wilcoxon test yields a p-value of 0.242 hence non-significant at the pre-specified 5% level. Using our proposed parametric likelihood ratio test yields a p-value of 0.019 and thus leading to a rejection of the null-hypothesis. From the histograms it could be questioned whether the continuous part of the distributions are normal distributed. Accordingly, we also apply our proposed semi-parametric version of the likelihood ratio test. This yields a p-value of 0.021. In conclusion, our proposed method is able to detect the difference between the groups. Further and in contrast to the Wilcoxon test, our method also provides a way to interpret the difference between the groups. This is illustrated in the right panel of Figure 4 showing simultaneous confidence regions for the two effect parameters. In this case it is clear that the primary effect of the intervention is not on the discrete component (i.e., the proportion of patients who never get off life support). Instead, the effect appears to be on the continuous part and is between half a day and seven days.

6 Discussion

In this paper we have introduced a novel statistical test to assess treatment effect on continuous outcomes where one value has special meaning (e.g., all diseased are assigned lowest possible value). The procedure in potentially much more power-full than the current best-practice which is to use Wilcox-type tests. The proposed method includes both a fully parametric approach and a semi-parametric where the latter makes no assumptions on the distribution of the continuous part of the outcome. In all settings this new method not only provides an effect measure but also effect parameters with associated confidence intervals. The test is implemented in an R package available on GitHub.

The simulation study and in particular the real data example demonstrate the efficiency and power gain. It is noted that unlike the Wilcox test, our proposed method can easily be extended to include covariates (A. Owen 1991). It is therefore not only useful in RCT settings but also to epidemiological studies.

[**TODO: Do we need something about this is very useful when needing to prespecify a test in an RCT?**]

References

Jensen, Andreas Kryger, and Theis Lange. 2018. "The TruncComp R package for Two-Sample Comparison of Truncated Continuous Outcomes Using Parametric and Semi-Empirical Likelihood." https://github.com/aejensen/TruncComp.

Mann, Henry B, and Donald R Whitney. 1947. "On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other." *The Annals of Mathematical Statistics*, 50–60.

Owen, Art. 1991. "Empirical Likelihood for Linear Models." *The Annals of Statistics*, 1725–47.

Owen, Art B. 1988. "Empirical Likelihood Ratio Confidence Intervals for a Single Functional." *Biometrika* 75 (2): 237–49.

The COVID STEROID 2 Trial Group. 2021. "Effect of 12 mg vs 6 mg of Dexamethasone on the Number of Days Alive Without Life Support in Adults With COVID-19 and Severe Hypoxemia: The COVID STEROID 2 Randomized Trial." *JAMA* 326 (18): 1807–17. https://doi.org/10.1001/jama.2021.18295.

Wilcoxon, Frank. 1945. "Individual Comparisons by Ranking Methods." *Biometrics Bulletin*, 80–83.

Wilks, Samuel S. 1938. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." *The Annals of Mathematical Statistics* 9 (1): 60–62.