Received: 27 March 2023 | Revised: 29 November 2023 | Accepted: 10 February 2024

DOI: 10.1002/sim.10045

RESEARCH ARTICLE

Statistics in Medicine WILEY

Use of win time for ordered composite endpoints in clinical trials

James F. Troendle[1] | **Eric S. Leifer**[1] | **Song Yang**[1] | **Neal Jeffries**[1] | **Dong-Yun Kim**[1] | **Jungnam Joo**[1] | **Christopher M. O'Connor**[2]

[1]Office of Biostatistics Research, Division of Intramural Research of the National Heart, Lung, and Blood Institute, NIH/DHHS, Bethesda, Maryland, USA

[2]Inova Heart and Vascular Institute, Fairfax, Virginia, USA

Correspondence
James F. Troendle, Office of Biostatistics Research, Division of Intramural Research of the National Heart, Lung, and Blood Institute, NIH/DHHS, Bld RLK1 Room 410, Bethesda, MD, 20892, USA.
Email: jt3t@nih.gov

Funding information
Intramural Research Program of the NIH; National Heart, Lung, and Blood Institute

Consider the choice of outcome for overall treatment benefit in a clinical trial which measures the first time to each of several clinical events. We describe several new variants of the win ratio that incorporate the time spent in each clinical state over the common follow-up, where clinical state means the worst clinical event that has occurred by that time. One version allows restriction so that death during follow-up is most important, while time spent in other clinical states is still accounted for. Three other variants are described; one is based on the average pairwise win time, one creates a continuous outcome for each participant based on expected win times against a reference distribution and another that uses the estimated distributions of clinical state to compare the treatment arms. Finally, a combination testing approach is described to give robust power for detecting treatment benefit across a broad range of alternatives. These new methods are designed to be closer to the overall treatment benefit/harm from a patient's perspective, compared to the ordinary win ratio. The new methods are compared to the composite event approach and the ordinary win ratio. Simulations show that when overall treatment benefit on death is substantial, the variants based on either the participants' expected win times (EWTs) against a reference distribution or estimated clinical state distributions have substantially higher power than either the pairwise comparison or composite event methods. The methods are illustrated by re-analysis of the trial heart failure: a controlled trial investigating outcomes of exercise training.

KEYWORDS

bootstrap, hazard, pairwise comparison, win ratio, win time

1 | INTRODUCTION

Clinical trials usually target risk reduction in one or more clinical events for their primary efficacy measure. For example, in a cardiovascular disease trial one might be interested in death, hospitalization for heart failure (HF), and myocardial infarction (MI). Some methods combine the endpoints into a composite, for example the composite of death, HF, MI. However, composite methods fail to account for the fact that often the endpoints can be ranked on an ordinal scale based on clinical importance. Some pairwise methods try to analyze the worst clinical event that occurs during follow-up. The win ratio[1,2] does this by first comparing a pair of participants on the worst event, death in the example given above. If the comparison cannot definitively be concluded in either participant's favor, the comparison moves on to the next worst event in the ordinal clinical scale until a definitive winner is declared or the pair are tied.

A major drawback of the win ratio is that for most pairwise comparisons, only the first occurrence time of the worst observed event matters. A lot of information is ignored. As a concrete example, suppose we are comparing two participants in a cardiovascular trial with up to 5 years of follow-up according to a composite clinical scale of increasing severity: MI, HF, death. Participant 1 was followed for 4 years and then died, no other events being observed. Participant 2 had an MI after 1 year, then was hospitalized for HF after 2 years, and died after 4.5 years. The standard win ratio approach considers Participant 2 to be the winner in the sense of the better outcome, since his/her death came after that of Participant 1. In this case, when both participants are observed to have the worst outcome, nothing else but the time of this outcome matters. However, the trial provided a head to head comparison based on the ordinal clinical scale for the entire 4.5 years of follow-up. We propose comparing the time each participant spent in a better ordinal clinical state than the other participant. The ordinal clinical state can simply be the worst clinical event that has occurred so far (0—no events, 1—MI only, 2—HF with or without MI, 3—death). Thus, in our example, participant 1 was in state 0 for the first 4 years, then transitioned to state 3. Participant 2 started in state 0, transitioning to state 1 after 1 year, then transitioned to state 2 after 2 years, and to state 3 after 4.5 years. Over the 4.5 years for which we can compare the participants, participant 2 was in the worse clinical state for 3 years and participant 1 was in the worse clinical state for 0.5 years. Thus, participant 1 is the winner with 2.5 excess years of being in the better clinical state (time in better state minus time in worse state). We call the excess time in a better clinical state the *win time* and the corresponding pairwise method of comparison is referred to as being based on *win time*.

Comparison based on win time is appealing because the entire patient clinical experience is accounted for, without the need for any arbitrary score being assigned to the clinical events. The only clinical judgment needed is the ordinal clinical scale that is required for any pairwise composite method. Here we have chosen to define the clinical state based on the worst clinical event to have occurred by the current time. This is a natural choice that makes the clinical state monotone non-decreasing in severity. Defining clinical state in this manner avoids any potentially arbitrary decisions about length of burden from clinical events, but other definitions could be used in general. If one places a fixed-length burden window following each clinical event (a period of time where the participant is assumed to be suffering from the event that has occurred at the start of the window), one could define clinical state based on the most severe burden window the participant currently is in. Because of our desire to avoid arbitrary decisions, for the remainder of this article we define clinical state to be the worst clinical event to have occurred by the current time.

One potential drawback to using win time is evident in the example given above. In the example, participant 1 was determined to be the winner despite dying prior to participant 2. This might be appropriate in the sense of quality of life since participant 1 did not have an MI or HF event while participant 2 did. However, with follow-up of clinical trials usually only lasting at most a few years, there is a real doubt that a participant who died before another participant could be considered to have a superior outcome. Thus, for relatively modest follow-up, as in most clinical trials, we consider a restricted version of win time. In the restricted version, a win means either a win on death (as determined in a win ratio comparison), or if no determination is made based on death, a win on win time as described above. In the above example, the restricted win time determines participant 2 is the winner and therefore agrees with the win ratio. However, it is easy to see that in general the restricted win time can differ from the win ratio when no determination can be made on death.

Three other related methods are developed directly from the win times themselves. One method calculates the average win time over all pairwise comparisons. A second method computes an outcome for each participant, which is then compared between the treatment arms. A third method uses the estimated distributions of clinical state through time to directly compare the treatment arms.

Finally, we describe a test that uses the maximum of two of the above estimates as the basis for determining treatment benefit. This combination test will be shown to have robust power for detecting treatment benefit across a broad range of alternatives.

The remainder of the article is composed as follows. The various methods are described in Section 2. Section 3.1 describes how we simulated clinical trial data to evaluate the statistical properties of the methods. The methods are then compared via simulation in Section 3.2. The methods are then illustrated in Section 4, by applying them to re-analyze the trial heart failure: a controlled trial investigating outcomes of exercise training (HF-ACTION).[3] Finally, in Section 5, we discuss the findings and give recommendations for practice.

2 | METHODS

Consider a cardiovascular trial with interest in the clinical events in order of increasing clinical severity: (MI < HF < death). Suppose the trial randomizes participants to active treatment $Z = 1$ or control treatment $Z = 0$. The latent times to first MI, first HF, and death are denoted (T_1^*, T_2^*, T_3^*), respectively, while the censoring time is denoted T_C. Since these events may not be observed, we also have indicators that an event occurs at the given time for each type of clinical event, denoted $(\delta_1, \delta_2, \delta_3)$ and the observed (possibly censored) times (T_1, T_2, T_3). The ordinal clinical state of a participant at any time t is given according to the worst clinical event to have occurred up to time t as (0—no events, 1—MI only, 2—HF with or without MI, 3—death). As noted in Section 1, as defined the clinical state is monotone non-decreasing in time. The clinical state for any participant is known up to time T_C, and for all times if death is observed.

One natural choice for analysis of treatment is to analyze the time to the first of any of the clinical events. This is called the composite event approach. We start by describing the composite event approach, since it is often used and serves as a primary competitor to pairwise approaches like the win ratio and win time ratio developed here. After that, we describe the standard win ratio approach[1,2] before introducing the new win time ratio, restricted win time ratio, the win time difference methods and combination test.

2.1 | Composite event approach

The composite event approach uses the minimum of the clinical event times, $T = \min(T_3, T_2, T_1)$, along with the indicator δ_T that an event occurs at time T. The primary treatment effect in a cardiovascular trial is often based on a composite event logrank test,[4] or the estimated coefficient $(\hat{\beta})$ of Z from a Cox model[5] of T that includes Z as a covariate:

$$\lambda(t) = \lambda_0(t) \exp(\beta Z), \tag{1}$$

where $\lambda(t)$ is the hazard function and $\lambda_0(t)$ is a baseline hazard. Our simulations (Section 3.1) will utilize (1) as given, but in practice other baseline covariates could be included in (1). The analysis of HF-ACTION (Section 4) will include a baseline covariate in (1). Tests of treatment effect can be based on $\hat{\beta}$, the estimated log hazard ratio from a fit of (1), and its estimated variance. We will use model (1) for comparison in this article, as testing based on the estimate of β is in very close agreement with a logrank test for large sample sizes and $\exp(\hat{\beta})$ provides a useful estimate of the treatment effect.

2.2 | Win ratio

We describe the win ratio test statistic that is most similar to a hazard ratio. Suppose there are $n = n_0 + n_1$ participants randomized to either the control (n_0) or novel treatment (n_1), and we have ordered the times and indicators so that the control participants are participants $i = 1, \dots, n_0$ while the active treatment participants are $i = n_0 + 1, \dots, n$. We will use the convention that i indexes the participants so that $T_{1,i}$ is the time of MI or censoring for participant $i = 1, \dots, n$. Similar notation for the other times and indicators will be used. For any pair of participants, define the win ratio score U_{ij} to be +1 if participant i has the more favorable outcome (win) than participant j and −1 if participant i has a less favorable outcome (loss) than participant j and U_{ij} is 0 if neither is more favorable. A more favorable outcome (win) for the win ratio means either that participant j died before participant i or that the comparison on death is inconclusive and participant j has HF before participant i or that neither death nor HF is conclusive and participant j has an MI before participant i. An analogous definition for less favorable (loss) is used. A comparison is inconclusive anytime one participant in the pair is censored before the other participant is observed to have the event, or if both participants are censored without observing the event, or if the participants both have the event at the same time. The score is defined as zero if all comparisons are inconclusive. The win ratio test statistic can then be given as the ratio of loses to wins on treatment:

$$\widehat{\mathrm{WR}} = \frac{\sum_{i=n_0+1}^{n} \sum_{j=1}^{n_0} 1(U_{ij} = -1)}{\sum_{i=n_0+1}^{n} \sum_{j=1}^{n_0} 1(U_{ij} = 1)}, \tag{2}$$

where 1() is the indicator function.

As pointed out by Follmann et al,[6] a related but different version that uses the tied pairs could be used to estimate the Mann-Whitney parameter.[7] We focus here on the ratio of loses to wins because it results in a metric that closely resembles a hazard ratio, with null value of 1.0, and values less than 1 indicating treatment benefit. The estimand, WR, for $\widehat{\text{WR}}$ is the probability a randomly selected treatment participant would lose when compared to a randomly selected control participant divided by the probability a randomly selected treatment participant would win when compared to a randomly selected control participant, given the pair did not tie. Unfortunately, this estimand depends on censoring. As a practical matter this usually is not a major issue. Including more (and more prevalent) events in the composite (or in determining wins and loses), results in fewer censored times, and therefore the WR becomes less affected by censoring.

The variance of $\log\left(\widehat{\text{WR}}\right)$ can be estimated according to Bebu and Lachin.[8] Alternatively, one can use the bootstrap to generate B bootstrap datasets, and calculate $\widehat{\text{var}}\left[\log\left(\widehat{\text{WR}}\right)\right]$ as the variance of $\log\left(\widehat{\text{WR}}\right)$ from the B associated datasets. Testing of H_0 : WR = 1 is then according to the asymptotic normality of $\log\left(\widehat{\text{WR}}\right)/\sqrt{\widehat{\text{var}}\left[\log\left(\widehat{\text{WR}}\right)\right]}$, established by Bebu and Lachin.[8] We use this bootstrap variance version of testing for the win ratio since we will require the bootstrap to estimate variance in testing for some of the win time methods in the following sections.

2.3 | Win time ratio

We now define the win time ratio estimator in a similar fashion to the win ratio. The only difference between the win time ratio and that of the win ratio will be the determination of more favorable outcome (win) and less favorable (loss). For any pair of participants, we will define the win time ratio score V_{ij} to be +1 if participant i has the more favorable outcome (win) than participant j and −1 if participant i has a less favorable outcome (loss) than participant j. We now proceed to define what constitutes more or less favorable outcome for the win time ratio. When comparing two participants, i and j, we define the effective common follow-up time to be the time of earliest censoring if either participant is censored. If neither participant is censored, then the larger death time is the effective common follow-up time. The effective common follow-up time, τ_{ij}, is the last time when the clinical state of both participants are known and potentially different. Thus, at any time $t \in (0, \tau_{ij})$, one can decide which participant is in the more favorable or less favorable clinical state, as was done in the calculation of the win ratio. The difference is that now we do this at each time $t \in (0, \tau_{ij})$, obtaining a function of time, $v_{ij}(t)$, the *win function*. The function, $v_{ij}(t)$, is +1 if participant i has the more favorable outcome than participant j at time t and −1 if participant i has a less favorable outcome than participant j at time t. If the participants are tied at time t, then $v_{ij}(t) = 0$. Now we calculate the *win time difference* for the two participants,

$$\widehat{\text{WTD}}_{ij} = \int_{t=0}^{\tau_{ij}} v_{ij}(t)dt. \tag{3}$$

To make this comparable to the win ratio, we note that in case one member of a pair has an event at the same time that the other member of the pair is censored, τ_{ij} is defined to be one day larger than the time of these simultaneous events, enabling breaking of ties for such cases (as is done for the win ratio). The win time difference, $\widehat{\text{WTD}}_{ij}$, tells us the excess time that participant i is in a more favorable state than participant j over the effective common follow-up time (more favorable minus less favorable). Thus, positive values mean participant i was in a more favorable state for longer than in a less favorable state over the effective common follow-up time. We are now ready to define more and less favorable according to win time.

More favorable (win) for the win time ratio means $\widehat{\text{WTD}}_{ij}$ is positive (and then the score $V_{ij} = +1$). Less favorable (loss) for the win time ratio means $\widehat{\text{WTD}}_{ij}$ is negative (and then $V_{ij} = -1$). The score V_{ij} is defined as zero if $\widehat{\text{WTD}}_{ij}$ is 0. The win time ratio test statistic can then be given as the ratio of losses to wins on treatment:

$$\widehat{\text{WTR}} = \frac{\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0} 1(V_{ij} = -1)}{\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0} 1(V_{ij} = 1)}. \tag{4}$$

The estimand, WTR, for $\widehat{\text{WTR}}$ is the probability a randomly selected treatment participant would lose when compared to a randomly selected control participant divided by the probability a randomly selected treatment participant would win when compared to a randomly selected control participant, given the pair did not tie. Variance estimates are obtained by generating B bootstrap datasets, and calculating $\widehat{\text{var}}[\log(\widehat{\text{WTR}})]$ as the variance of $\log(\widehat{\text{WTR}})$ from the B associated datasets. Testing of H_0 : WTR = 1 is then according to the asymptotic normality of $\log(\widehat{\text{WTR}})/\sqrt{\widehat{\text{var}}[\log(\widehat{\text{WTR}})]}$ (again established in Bebu and Lachin[8]).

2.4 | Restricted win time ratio

We now define the restricted win time ratio estimator in a similar fashion to the win time ratio. The only difference between the restricted win time ratio and the win time ratio will be the determination of more favorable outcome (win) and less favorable outcome (loss).

A more favorable outcome (win) for the restricted win time ratio means either that death favors participant i or that death proved inconclusive while the $\widehat{\text{WTD}}_{ij}$ is positive (and then the score $W_{ij} = +1$). Less favorable (loss) for the restricted win time ratio means either that death favors participant j or that death proved inconclusive while $\widehat{\text{WTD}}_{ij}$ is negative (and then $W_{ij} = -1$). The restricted win time ratio test statistic can then be given as the ratio of losses to wins on treatment:

$$\widehat{\text{RWTR}} = \frac{\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0} 1(W_{ij} = -1)}{\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0} 1(W_{ij} = 1)}. \tag{5}$$

2.5 | Pairwise win time

We calculate the *pairwise win time* (PWT) as the pairwise average of the win time differences $\widehat{\text{WTD}}_{ij}$. The PWT test statistic can then be given as:

$$\widehat{\text{PWT}} = \frac{\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0} \text{WTD}_{ij}}{n_0 n_1}. \tag{6}$$

We use the bootstrap to get an estimated standard deviation of $\widehat{\text{PWT}}$ for hypothesis testing of the null hypothesis that PWT = 0. The estimand for PWT can be described as the difference in expected excess time in a better clinical state for treatment participants compared to control participants over the average *effective common follow-up time between the arms* (average of the τ_{ij}). Note that this substantially differs from the WTR since the WTR counts each win or loss as 1 or −1, whereas the PWT is averaging the actual win time differences.

2.6 | Expected win time against reference

The idea of *expected win time against reference* (EWTR) is to compare each participant's clinical state at any time to a reference group's distribution of clinical states. A natural choice for the reference group is to use the control arm. The first step in computing EWTR is to estimate the control arm's clinical state distribution (described in Section 1) at any time from baseline (time = 0) up until there are no control arm participants under follow-up (designated time = τ_C).

We consider two ways to estimate the control arm's clinical state distribution in time. One way, as proposed by Mao,[9] is to estimate Kaplan-Meier curves $\text{KM}_k(t)$ that give the probability of being free of clinical state $\geq k$ for $k = 1, 2, 3$. With estimates $\{\text{KM}_k(t), k = 1, 2, 3\}$, the state distribution is then obtained by subtraction. This method is non-parametric and yields estimates for times from 0 to the last time a control participant is still in state 0. In many applications this will not represent an important restriction, but in some it may.

Another option is to use a Markov model. One complication with fitting the Markov model arises because of censoring. Any time that the last participant in a clinical state below death (recall the clinical states with three events including death are 0, 1, 2, 3 with 3 being the absorbing state of death) is censored, the transition probabilities for that state are not estimable from the model. We extend the Markov model by simply assuming the transition probabilities from the state in question to any different state are zero, while the transition probability of staying in the state in question is one. Note that later, participants could transition into the affected state, thereby allowing later transition probabilities to be consistently estimated. We will call this method of estimating the clinical state distribution through time as the *extended Markov model* to call attention to this modification. In the presence of censoring, the estimated state distribution at a given time from an extended Markov model may not be consistent. For example, it may underestimate the probability of the death state since, due to censoring, transitions from a lower clinical state to death may have been missed and the probability estimated as zero. For a worked example of fitting the extended Markov model, see the Supplemental Material.

The next step in the calculation of EWTR is to compare each trial participant to the estimated reference clinical state distribution. The comparison for participant i can continue until the effective common follow-up time, which in this case is $\tau_i \equiv \min(\tau_C,\ \tau_i^*)$, where τ_i^* is either the censoring time for participant i or if participant i dies is ∞. Let $\hat{p}_{00}(t),\ \hat{p}_{01}(t),\ \hat{p}_{02}(t),\ \hat{p}_{03}(t)$ be the estimated clinical state distribution for the control arm from either the extended Markov model or the Kaplan-Meier method (KM), where the first subscript indicates the control arm and the second subscript refers to the state. We now consider the expected value of the win function, $v_{ij}(t)$, from Section 2.3. Since we are comparing participant i to a distribution, the win function gets replaced by the state specific win components, $w_{i0}(t),\ w_{i1}(t),\ w_{i2}(t),\ w_{i3}(t)$, where $w_{ik}(t)$ is $+1$ if participant i is in a better state at time t than state k and is -1 if participant i is in a worse state at time t than state k and is 0 otherwise for $k = 0, 1, 2, 3$. The expected value of $v_{ij}(t)$ then becomes $w_{i0}(t)\hat{p}_{00}(t) + w_{i1}(t)\hat{p}_{01}(t) + w_{i2}(t)\hat{p}_{02}(t) + w_{i3}(t)\hat{p}_{03}(t)$. This gets substituted for $v_{ij}(t)$ in Equation (3) to give the EWTR for participant i:

$$\widehat{\text{EWTR}}_i = \int_{t=0}^{\tau_i} \left\{ w_{i0}(t)\hat{p}_{00}(t) + w_{i1}(t)\hat{p}_{01}(t) + w_{i2}(t)\hat{p}_{02}(t) + w_{i3}(t)\hat{p}_{03}(t) \right\}\ dt. \tag{7}$$

Each participant gets a value for $\widehat{\text{EWTR}}_i$. We could compare the $\widehat{\text{EWTR}}_i$ values between the arms using a traditional t-test. Instead, as a final step in the method, we fit a linear model to the $\widehat{\text{EWTR}}_i$ values with an intercept and treatment indicator:

$$\widehat{\text{EWTR}}_i = \alpha + \beta_{\text{EWTR}} Z_i. \tag{8}$$

Testing is based on the fitted value of β_{EWTR} and its estimated standard deviation from (8). One advantage of using model (8) over a direct t-test comparison is that prognostic baseline covariates can be added to (8).

Although the extended Markov model in the presence of censoring may not give an entirely consistently estimated clinical state distribution in time, this is only used as a reference distribution for the EWTRs, which are then compared between arms. Thus the null hypothesis of $\beta_{\text{EWTR}} = 0$ is true if treatment has no effect on clinical state (regardless of censoring and its effect on the estimated reference distribution). The estimand for EWTR can be described as the treatment arm difference in average excess time in a better clinical state compared to reference participants over the average *effective common follow-up time*, defined above.

2.7 | Expected win time

The idea of EWT is to directly compare the two arm's distributions of clinical states.

Once again there are two versions depending on the approach used to estimate the state space distributions. Either an extended Markov model or KM approach can be used.

Let $\hat{p}_{10}(t),\ \hat{p}_{11}(t),\ \hat{p}_{12}(t),\ \hat{p}_{13}(t)$ be the estimated clinical state distribution in the active treatment group either from the extended Markov model or the KM method (recall the first subscript refers to the arm so that for example $p_{1k}(t)$ refers to the treatment arm probability of being in state k at time t).

The next step in calculation of EWT is to compare the estimated clinical state distributions. The comparison can continue until the effective common follow-up time between the arms, which in this case is $\tau_* \equiv \min(\tau_C,\ \tau_T)$, where τ_T is the first time when there are no active treatment participants under follow-up and τ_C is defined in Section 2.6. We again consider the expected value of the win function, $v_{ij}(t)$, from Section 2.3. It is easiest to think about the three possible win states, $v_{ij}(t) = +1,\ v_{ij}(t) = 0,\ v_{ij}(t) = -1$ with the associated estimated probabilities of those win states, $\hat{p}_{+1}(t),\ \hat{p}_0(t),\ \hat{p}_{-1}(t)$. For example,

$$\hat{p}_{+1}(t) = \hat{p}_{10}(t)\ \left\{\hat{p}_{01}(t) + \hat{p}_{02}(t) + \hat{p}_{03}(t)\right\} + \hat{p}_{11}(t)\left\{\hat{p}_{02}(t) + \hat{p}_{03}(t)\right\} + \hat{p}_{12}(t)\ \hat{p}_{03}(t).$$

Since we are comparing two distributions, the win function gets replaced by the win state components, $+1,\ 0,\ -1$, so that the expected value of $v_{ij}(t)$ becomes $\hat{p}_{+1}(t) - \hat{p}_{-1}(t)$. This gets substituted for $v_{ij}(t)$ in Equation (3) to give the EWT:

$$\widehat{\text{EWT}} = \int_{t=0}^{\tau_*} \left\{\hat{p}_{+1}(t) - \hat{p}_{-1}(t)\right\}\ dt. \tag{9}$$

We use permutations of the treatment arm indicators to approximate the null distribution of $\widehat{\text{EWT}}$ and therefore to get a critical cutoff for hypothesis testing of the null hypothesis that EWT $= 0$. Let $\widehat{\text{EWT}}^b$ for $b = 1, \dots, B$ be the values obtained from (9) for datasets created by random permutations of the treatment indicators applied to the observed dataset. We thus use the 0.975 quantile of this distribution as a critical cutoff for $\widehat{\text{EWT}}$ to create a 2.5% test of the null hypothesis. In practice, with $B = 200$ (as will be used in our simulations), this means selecting the 196th largest of the values $\{\widehat{\text{EWT}}^b, b = 1, \dots, 200\}$ as critical cutoff for $\widehat{\text{EWT}}$.

The estimand, EWT, can be described as the difference in expected excess time in a better clinical state for treatment participants compared to control participants over the *effective common follow-up time between the arms*, defined above.

The latter three procedures (PWT, EWTR, and EWT) are each different attempts to calculate an expectation of $\widehat{\text{WTD}}_{ij}$. We shall refer to these procedures collectively as *win time difference* procedures as opposed to the *pairwise comparison* methods WTR and RWTR.

2.8 | EWTR-composite max test

The idea of *EWTR-composite max test* (MAX) is to combine the test statistics of the EWTR and composite event approaches to get a testing method with a robust power profile. To this end, define

$$\widehat{MAX} = \max\left[\frac{\hat{\beta}_{\text{EWTR}}}{\widehat{\text{std}}\left(\widehat{\beta_{\text{EWTR}}}\right)}, \frac{\widehat{-\beta}}{\widehat{\text{std}}(\hat{\beta})}\right]. \tag{10}$$

Note the negative applied to $\hat{\beta}$ in (10) makes positive values reflect treatment benefit for each component. We use the bootstrap to get a critical cutoff for hypothesis testing of the intersection null hypothesis of $\{\beta_{\text{EWTR}} = 0\} \cap \{\beta = 0\}$. This differs from the use of bootstrap for variance estimation described above. In this approach, B bootstrap datasets are generated and for each the value of the standardized EWTR and composite test statistics are calculated and respectively denoted z^b_{EWTR} and $-z^b$ for $b = 1, \dots, B$. Let the standardized test statistics calculated from the observed data (and given inside the max in 10) be respectively denoted z_{EWTR} and $-z$. We then consider the collection of values $\{\max[z^b_{\text{EWTR}} - z_{\text{EWTR}}, -z^b + z], b = 1, \dots, B\}$. These values approximate the null distribution of (10). We thus use the 0.975 quantile of this distribution as a critical cutoff for (10) to create a 2.5% test of the intersection null hypothesis. In practice, with $B = 200$ (as will be used in our simulations), this means selecting the 196th largest of the bootstrapped values $\{\max[z^b_{\text{EWTR}} - z_{\text{EWTR}}, -z^b + z], b = 1, \dots, 200\}$ as critical cutoff for (10).

Following a significant finding from the MAX procedure, one could proceed in a step-down fashion to identify if either the composite or EWTR metrics show statistically significant treatment benefit. For one-sided testing of benefit, a significant MAX test at level 2.5% followed by a significant EWTR test also at level 2.5% would allow rejection of the EWTR null. Similarly, a significant MAX test at level 2.5% followed by a significant composite test also at level 2.5% would allow rejection of the composite null. This multiple testing procedure controls familywise error at one-sided level of 2.5% (by the closed testing principle of Marcus et al[10]).

2.9 | Restricted mean survival in favor of treatment

If one picks an arbitrary timepoint, θ, at which the EWT methodology is truncated, one gets the restricted mean survival time in favor of treatment (RMT) of Mao.[9]

The estimate of RMT_θ is given by

$$\widehat{\text{RMT}}_\theta = \int_{t=0}^{\theta} \left\{\hat{p}_{+1}(t) - \hat{p}_{-1}(t)\right\} dt. \tag{11}$$

We use permutations of the treatment arm indictors to approximate the null distribution of $\widehat{\text{RMT}}_\theta$ and therefore to get a critical cutoff for hypothesis testing of the null hypothesis that $\text{RMT}_\theta = 0$. Let $\widehat{\text{RMT}}^b_\theta$ for $b = 1, \dots, B$ be the values obtained from (11) for datasets created by random permutations of the treatment indicators applied to the observed dataset. We thus use the .975 quantile of this distribution as a critical cutoff for $\widehat{\text{RMT}}_\theta$ to create a 2.5% test of the null hypothesis. In practice, with $B = 200$ (as will be used in our simulations), this means selecting the 196th largest of the values $\{\widehat{\text{RMT}}_\theta^b, b = 1, \dots, 200\}$ as critical cutoff for $\widehat{\text{RMT}}_\theta$.

The estimand, RMT_θ, can be described as the difference in expected excess time in a better clinical state for treatment participants compared to control participants over the interval of $[0, \theta]$. Note that RMT_θ is also a *win time difference* procedure.

3 | COMPARISON OF METHODS

3.1 | Simulation setup

We used simulation to compare the statistical properties of the methods described in Section 2. We consider the two-armed randomized clinical trial setting comparing an active treatment ($Z = 1$) vs standard treatment ($Z = 0$). We aim to simulate times for each of three clinical events, while allowing the times to be correlated. To do this in our simulations, we hypothesize that there is an additional relevant covariate Y which is an unobserved subject-specific frailty (that will be common across the three clinical event models). Therefore the true simulation models will depend on $\mathbf{X} = (Y, Z)$. The ith participant is followed until one of two potential competing events occurs:

(i) participant i dies at time t_i.

(ii) participant i is lost to follow-up or the trial ends at time t_i.

We simulated data using the methods of Beyersmann et al,[11] adapted to our semi-competing risk scenario. In this case there are two events (death and censoring) that compete with all other events, and two events (HF and MI) that do not compete with other events. The clinical ordering of events is assumed to be: MI < HF < death. The method specifies an active treatment vs control proportional cause specific hazard (CSH) model for each of the four events.

We simulated data from a semi-competing risk model assuming a log(CSH) for each of the four events given by a time-independent linear function of the active treatment indicator Z and the frailty variable Y. The CSHs $\lambda_1(t)$, $\lambda_2(t)$, and $\lambda_3(t)$ (this notation suppresses their dependence on $\mathbf{X}$) for the clinical events (MI, HF, death) are given by:

$$\text{MI:} \qquad \lambda_1(t) = \exp[\alpha_1 + \beta_1 * Z + \gamma_1 * Y], \tag{12}$$

$$\text{HF:} \qquad \lambda_2(t) = \exp[\alpha_2 + \beta_2 * Z + \gamma_2 * Y], \tag{13}$$

$$\text{Death:} \qquad \lambda_3(t) = \exp[\alpha_3 + \beta_3 * Z + \gamma_3 * Y], \tag{14}$$

where $Y \sim U\left(-\frac{\psi}{2}, \frac{\psi}{2}\right)$, Z is set to 0 for half of the participants and 1 for the other half, and $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3$ are real valued parameters ($U(a, b)$ refers to the uniform distribution on the interval [a,b]). Thus, the frailty Y has no effect on the event's CSHs if and only if $\psi = 0$. We assume independent censoring with hazard given by

$$\lambda_{\text{cen}}(t) = \exp[\alpha_4]. \tag{15}$$

We start our semi-competing risk simulation by generating a cause-specific event time S_1 from the all-cause hazard $\lambda_1(t) + \lambda_2(t) + \lambda_3(t) + \lambda_{\text{cen}}(t)$. To determine which type of event occurred at time S_1, we perform a multinomial experiment with probabilities determined by the hazards of each event type:

$$\Pr(\text{MI}) = \frac{\lambda_1(S_1)}{\lambda_1(S_1) + \lambda_2(S_1) + \lambda_3(S_1) + \lambda_{\text{cen}}(S_1)},$$

$$\Pr(\text{HF}) = \frac{\lambda_2(S_1)}{\lambda_1(S_1) + \lambda_2(S_1) + \lambda_3(S_1) + \lambda_{\text{cen}}(S_1)},$$

$$\Pr(\text{death}) = \frac{\lambda_3(S_1)}{\lambda_1(S_1) + \lambda_2(S_1) + \lambda_3(S_1) + \lambda_{\text{cen}}(S_1)},$$

$$\Pr(\text{censoring}) = \frac{\lambda_{\text{cen}}(S_1)}{\lambda_1(S_1) + \lambda_2(S_1) + \lambda_3(S_1) + \lambda_{\text{cen}}(S_1)}.$$

In our simulations, the administrative censoring time is 4 years so any $S_1 \geq 4$ is truncated to be an administrative censoring at 4 years. If the event at time S_1 is either death or censoring, then the simulation of events for this participant is finished. If the event at time S_1 is either HF or MI, then the simulation continues with a second event at time increment S_2 after time S_1. The all-cause hazard for S_2 is adjusted by removing either $\lambda_1(t)$ or $\lambda_2(t)$ from the all cause hazard for S_1 based on which non-competing event occurred. The multinomial probabilities to determine the type of event at time $S_1 + S_2$ are adjusted accordingly, based on which of the non-competing events has already occurred. If this results in a second non-competing event, a third time increment is generated in a similar manner, at which time one of the competing events must occur. In this way we generate (possibly censored) times for each of the clinical events, (T_1, T_2, T_3), along with the indicators of censoring introduced in Section 2.

We used the KM method of estimation of the state space distribution for EWT and $\mathrm{RMT}_{3\text{ year}}$, since our simulations indicated that using an extended Markov model adds variability which results in reduced power for those procedures which compare the arms directly based on the estimated state space distributions. In contrast, EWTR uses the estimated state space distribution differently. For EWTR, the state space distribution is merely used as a reference that each participant in either arm is compared to. Thus, it makes sense that being able to integrate further in time via an extended Markov model leads to higher power for EWTR (and thus also for MAX). Therefore, EWTR and MAX are implemented using an extended Markov model.

3.2 | Results

We simulated 1000 clinical trials (10 000 for the null case) using the methodology described in Section 3.1 with $n = 2000$, $\psi = 0.5$, $\alpha_1 = \alpha_2 = -0.5$, $\alpha_3 = -1.6$, $\gamma_1 = \gamma_2 = \gamma_3 = 1$, and $\alpha_4 = 0.0$. The value of $\alpha_3 = -1.6$ was chosen to have about half as many fatal events as each of the non-fatal events. Several different possible configurations of cause specific treatment effect log hazard ratios for each of the three clinical events were used and the results are shown in Table 1. In Table 1, k indexes the various alternative scenarios.

TABLE 1 Estimated rejection rates (%) for simulated clinical trials from cause specific models ($n = 2000$, $\psi = 0.5$, $\alpha_1 = \alpha_2 = -0.5$, $\alpha_3 = -1.6$, $\gamma_1 = \gamma_2 = \gamma_3 = 1$, $\alpha_4 = 0.0$).

	MI	**HF**	**Death**									
k	β_1	β_2	β_3	$\mathbf{RR}_c$[a]	$\mathbf{RR}_{\mathrm{wr}}$[b]	$\mathbf{RR}_{\mathrm{wt}}$[c]	$\mathbf{RR}_{\mathrm{rwt}}$[d]	$\mathbf{RR}_{\mathrm{pwt}}$[e]	$\mathbf{RR}_{\mathrm{ewtr}}$[f]	$\mathbf{RR}_{\mathrm{ewt}}$[g]	$\mathbf{RR}_{\mathrm{max}}$[h]	$\mathbf{RMT}_{3\text{ year}}$[i]
Null	**0.0**	**0.0**	**0.0**	**2.00**	**2.37**	**2.29**	**2.34**	**2.21**	**2.71**	**2.53**	**2.21**	**2.50**
1	−0.17	−0.17	−0.17	82.0	78.1	77.7	77.6	70.7	65.6	68.5	80.1	69.1
2	−0.20	−0.15	−0.10	78.1	63.3	66.4	65.5	51.6	38.0	45.6	73.3	46.3
3	−0.10	−0.20	−0.30	81.8	90.7	87.7	89.1	92.4	94.0	95.8	93.4	96.3
4	−0.20	−0.20	0.00	80.6	56.0	62.5	57.3	31.1	15.9	26.3	71.4	27.2
5	−0.24	−0.16	0.00	80.8	49.6	58.8	53.9	28.2	13.0	20.5	71.3	21.1
6	0.10	0.10	−0.35	0.3	6.5	3.4	6.3	28.6	67.2	37.6	56.7	37.4
7	−0.35	0.10	−0.35	61.6	44.3	50.2	58.6	71.8	84.5	62.6	82.1	63.1
8	0.10	0.00	−0.35	1.9	23.5	14.4	20.9	50.5	80.1	61.3	71.3	61.7
9	0.00	0.00	−0.35	10.3	34.5	25.2	32.9	61.1	84.4	69.4	75.8	69.6

Note: Estimated over 1000 replicated trials (10 000 for null case); one-sided target rate is 2.5% under the null hypothesis (with an average of 410 deaths, 764 HF, and 765 MI events).

[a] Test based on $\hat{\beta}$ from Section 2.1.

[b] Test based on $\log\left(\widehat{\mathrm{WR}}\right)$ from Section 2.2 with $B = 200$ bootstraps for std estimate.

[c] Test based on $\log(\widehat{\mathrm{WTR}})$ from Section 2.3 with $B = 200$ bootstraps for std estimate.

[d] Test based on $\log(\widehat{\mathrm{RWTR}})$ from Section 2.4 with $B = 200$ bootstraps for std estimate.

[e] Test based on $\widehat{\mathrm{PWT}}$ from Section 2.5 with $B = 200$ bootstraps for std estimate.

[f] Test based on $\widehat{\mathrm{EWTR}}$ using an extended Markov model from Section 2.6.

[g] Test based on $\widehat{\mathrm{EWT}}$ using KM from Section 2.7 with $B = 200$ permutations for critical cutoff.

[h] Test based on (10) using an extended Markov model from Section 2.8 with $B = 200$ bootstraps for critical cutoff.

[i] Test based on $\widehat{\mathrm{RMT}}_{3\text{ year}}$ using KM from Section 2.9 with $B = 200$ permutations for critical cutoff.

Note first that all nine testing methods approximately control the type I error rate at a one-sided 2.5%. Scenarios 1-5 (with approximate proportional hazards for the composite event) are cases where one might expect the composite event approach to dominate power. The cause specific treatment log hazard ratios for scenarios 1-5 were chosen to approximately give 80% power for the composite event approach. As expected, the composite event approach does perform well over scenarios 1-5. However, in scenario 3 where the largest treatment effect is on death, the composite method has notably lower power than the other methods. Scenarios 6-9 are cases where treatment is either null or slightly harmful for at least one non-fatal event and possibly beneficial for a different non-fatal event (scenario 7) as well as death. Scenarios 6-9 are more difficult to evaluate scientifically since overall treatment benefit requires weighing pros and cons; in these cases it seems safe to conclude there is a strong overall benefit. Scenarios 3, 8, and 9 show that the hierarchical pairwise comparison approaches can have somewhat higher power than the composite event approach when the treatment benefit is largest on death. Scenario 6 is particularly noteworthy as an example where treatment is slightly harmful for MI and HF, but much more beneficial for death. In general, the win time difference methods have highest power whenever treatment benefit is substantial on death. This is primarily because the pairwise comparison methods treat any win or loss as the same whereas the win time difference methods allow the effect of death to dramatically impact the metric since death is the worst clinical state and there is a longer effective common follow-up when death occurs. The EWTR and especially the EWT (or RMT) have even longer effective common follow-up times than does the PWT. This is because when a participant dies, a longer period of time dead can be used in the calculation of EWTR or EWT (or RMT) than for PWT. Effectively this means that the EWTR and EWT (or RMT) methods more highly weight death in the treatment arm comparison. In contrast, the composite event method considers all events equal while the pairwise methods do not explicitly weight the length of time a participant is dead during follow-up. The result is that the EWTR and EWT (or RMT) methods are most sensitive to treatment benefit on death.

Power should not be the only deciding factor in choosing a primary analysis estimand or testing procedure. The main factor should be using a meaningful metric that best represents the combined treatment effect on important health events. The win time difference methods naturally incorporate the importance of death, leading to a metric that appropriately highly weights death. The power of the win time difference methods is tied to the overall treatment benefit, determined by summing the patient experience over an effective common follow-up time. Overall, the simulations show that power depends critically on the CSHs for treatment. However, Table 1 confirms that in cases where overall treatment benefit is clear and the treatment effect on death is substantial (scenarios 6-9), the win time difference methods are substantially more sensitive than the composite event approach. Moreover, the MAX test achieves a relatively robust power profile across all of the alternatives.

Results with no correlation between events ($\psi = 0$, not shown) are similar and can be found in the Supplemental Material. Results with similar numbers of fatal events as each of the non-fatal events ($\alpha_3 = -0.5$, not shown) demonstrate that the relative power of the win time based methods are even higher and can also be found in the Supplemental Material.

We further investigated the effects of smaller sample size or higher censoring rates. The results did not substantially change the primary comparison of relative power between the methods and can also be found in the Supplemental Material.

We investigated a second set of trials with only two events to see if the relative power results changed in any substantial manner. The events are HF and death (events 2 and 3 from section 3.1). The results are presented in Table 2. The results are mostly similar in nature to those with three events. Scenario 5 is a case with treatment benefit on HF and treatment harm on death. This scenario seems to present a difficult case for judging overall treatment effect, and perhaps most would consider it unclear that treatment is beneficial overall. The interesting observation for scenario 5 is that the win time difference methods are the only methods that are insensitive to this treatment effect that includes harm on death.

4 | HF-ACTION TRIAL

The trial HF-ACTION[3] was sponsored by the National Heart, Lung and Blood Institute, and randomized participants from 82 centers in the United States, Canada, and France. A total of 2331 participants were enrolled in 2003 through early 2007 with median follow-up of 30 months. The trial randomized 1:1 into usual care (UC) vs usual care plus exercise training (ET). The protocol specified primary outcome was a composite of all-cause mortality or hospitalization, analyzed here by a Cox model with treatment group and heart failure etiology as covariates. In this re-analysis, we consider application of the pairwise methods along with the win time difference methods and the composite method to the primary outcome ordered as (hospitalization < death). The EWTR method included the etiology covariate in its linear model (8). We also

TABLE 2 Estimated rejection rates (%) for simulated clinical trials from cause specific models ($n = 2000$, $\psi = 0.5$, $\alpha_1 = -99$, $\alpha_2 = -0.5$, $\alpha_3 = -1.6$, $\gamma_1 = \gamma_2 = \gamma_3 = 1$, $\alpha_4 = 0.0$).

	HF	Death									
k	β_2	β_3	RR$_{comp}$[a]	RR$_{wr}$[b]	RR$_{wt}$[c]	RR$_{rwt}$[d]	RR$_{pwt}$[e]	RR$_{ewtr}$[f]	RR$_{ewt}$[g]	RR$_{max}$[h]	RMT$_{3\ year}$[i]
Null	**0.0**	**0.0**	**2.47**	**2.74**	**2.67**	**2.74**	**2.54**	**2.78**	**2.81**	**2.65**	**2.81**
1	−0.18	−0.18	80.2	71.8	71.4	71.8	65.6	61.8	57.8	76.4	66.3
2	−0.21	−0.10	80.4	62.6	63.8	62.6	46.4	37.0	36.8	72.5	44.3
3	−0.10	−0.21	49.5	51.6	48.4	51.6	55.3	59.1	51.4	59.7	57.7
4	0.10	−0.30	1.3	8.8	5.6	8.8	30.3	59.5	43.0	48.5	42.3
5	−0.30	0.10	81.0	40.0	47.3	40.0	11.6	2.6	4.9	72.8	8.1

Note: Estimated over 1000 replicated trials (10 000 for null case); one-sided target rate is 2.5% under the null hypothesis (with an average of 410 deaths and 764 HF events).

[a] Test based on $\hat{\beta}$ from Section 2.1.

[b] Test based on $\log\left(\widehat{\mathrm{WR}}\right)$ from Section 2.2 with $B = 200$ bootstraps for std estimate.

[c] Test based on $\log(\widehat{\mathrm{WTR}})$ from Section 2.3 with $B = 200$ bootstraps for std estimate.

[d] Test based on $\log(\widehat{\mathrm{RWTR}})$ from Section 2.4 with $B = 200$ bootstraps for std estimate.

[e] Test based on $\widehat{\mathrm{PWT}}$ from Section 2.5 with $B = 200$ bootstraps for std estimate.

[f] Test based on $\widehat{\mathrm{EWTR}}$ using an extended Markov model from Section 2.6.

[g] Test based on $\widehat{\mathrm{EWT}}$ using KM from Section 2.7 with $B = 200$ permutations for critical cutoff.

[h] Test based on (10) using an extended Markov model from Section 2.8 with $B = 200$ bootstraps for critical cutoff.

[i] Test based on $\widehat{\mathrm{RMT}}_{3\ \mathrm{year}}$ using KM from Section 2.9 with $B = 200$ permutations for critical cutoff.

TABLE 3 Re-analysis of HF-ACTION.

	$\widehat{\mathrm{HR}}$[a]	WR[b]	WTR[c]	RWTR[d]	PWT[e]	β_{EWTR}[f]	EWT[g]	MAX[h]	RMT$_{3\ year}$[i]
Primary two components (hospitalization < death)									
Effect estimate	0.925	0.934	0.932	0.934	22.4 days	36.5 days	43.8 days	–	31.8 days
P-value	0.143	0.210	0.199	0.210	0.141	0.096	0.058	0.077	0.054
Three components (AE < hospitalization < death)									
Effect estimate	0.912	0.931	0.927	0.930	24.0 days	38.4 days	46.4 days	–	34.0 days
P-value	0.081	0.182	0.165	0.178	0.118	0.082	0.052	0.075	0.042

[a] $\widehat{\mathrm{HR}}$ is $\exp(\hat{\beta})$ from Section 2.1.

[b] $\log\left(\widehat{\mathrm{WR}}\right)$ from Section 2.2 with testing based on $B = 1000$ bootstraps.

[c] $\log(\widehat{\mathrm{WTR}})$ from Section 2.3 with testing based on $B = 1000$ bootstraps.

[d] $\log(\widehat{\mathrm{RWTR}})$ from Section 2.4 with testing based on $B = 1000$ bootstraps.

[e] $\widehat{\mathrm{PWT}}$ from Section 2.5 with $B = 1000$ bootstraps.

[f] $\hat{\beta}_{\mathrm{EWTR}}$ using an extended Markov model from from Section 2.6.

[g] $\widehat{\mathrm{EWT}}$ using KM from Section 2.7 with $B = 1000$ permutations for critical cutoff.

[h] Test based on (10) using an extended Markov model from Section 2.8 with $B = 1000$ bootstraps for critical cutoff.

[i] Test based on $\widehat{\mathrm{RMT}}_{3\ \mathrm{year}}$ using KM from Section 2.9 with $B = 1000$ permutations for critical cutoff.

constructed a three component outcome that (in addition to the components of the primary) included any adverse event (AE) from a pre-specified list (worsening heart failure, myocardial infarction, unstable angina, serious arrhythmia, stroke, transient ischemic attack) ordered as AE < hospitalization < death. Results are shown in Table 3.

Analysis of the primary outcome, at the top of Table 3, yields similar treatment effect estimates for all of the methods based on ratios. Recall that the estimand for the protocol specified analysis is a hazard ratio, while the pairwise methods have win ratios as estimands. The estimands for the win time difference methods are based on definitions of effective common follow-up times, and expressed here in excess days in the better clinical state. In most cases the effective common follow-up time for EWT (or RMT) will be larger than the average effective common follow-up time for EWTR (similarly they will be larger for EWTR than for PWT), leading to larger effect estimates for EWT (or RMT) than EWTR (similarly they will be larger for EWTR than for PWT). Separate analysis of each component of the primary outcome (with a Cox model exactly like the primary composite model adjusting for HF etiology) reveals an estimated treatment hazard ratio of 0.921 for death and 0.933 for hospitalization. Based on the similarity of these separate estimated hazard ratios and the results of Section (3.2), it is not too surprising that the composite primary analysis yields the smallest *P*-value among the methods based on ratios. However EWTR, EWT, and RMT have smaller *P*-values than any of the ratio methods. In addition to considerations of power, the composite primary analysis leaves a question of whether or not one should regard the components of the composite as equal when there is a clear clinical hierarchy.

Analysis of the three component outcome, at the bottom of Table 3, yields different effect estimates with each method yielding a somewhat stronger estimated treatment benefit. This is explained by a separate analysis of AEs alone (with a Cox model exactly like the primary composite model) that revealed an estimated treatment hazard ratio of 0.890 for AE. In this case the composite, EWTR, and EWT all yield similar *P*-values while RMT has the smallest *P*-value of any method. With this three-level outcome, the appropriateness of including all hospitalizations and AEs as equal to deaths (as the composite method does) is even more suspect than in the composite primary analysis. Note that while for this example the CSH ratios are all in the direction of benefit, in general the win time difference methods are appropriate regardless of the direction or magnitude of the CSH ratios.

5 DISCUSSION

We have introduced a pairwise comparison method, the win time ratio, that accounts for the time spent in each clinical state during the combined common follow-up period. A second method, the restricted win time ratio, allows restriction so that death during study follow-up is more important than time spent in other states. We also introduced three methods based on win time differences. PWT, EWTR, and EWT are all based on the excess time spent in a better clinical state over an effective common follow-up time. These methods reflect the entire clinical experience of the trial participants and offer sensitive tests of overall treatment benefit. The major advantage of the win time difference methods, as compared to the win ratio or composite event method, is that they account for the patient's entire experience, captured through the time spent in a clinical state, and not just the occurrence of the clinically most important event. This is the main reason for using the win time difference methods in a primary analysis of a clinical trial. Additionally, simulations indicate that in many cases where overall treatment benefit is clear and treatment benefit is substantial on death, the win time difference methods have substantially higher power than either the win ratio or the composite event approach. Overall there is no procedure that is most powerful, and relative power depends critically on the particular CSH ratios for treatment. Finally, we created a combination hypothesis testing procedure (MAX) that has a very robust power profile across a wide range of alternatives. Regarding the power of several related procedures, Yang et al[12] have some relevant discussion.

Based on these observations, it seems that win time difference methods are well suited for trials where several clinical events are expected to occur in addition to death and when treatment is expected to provide benefit on death. We note that the EWTR has the advantage over the EWT of being able to easily include baseline covariates directly into the model. This can lead to further power gains if the baseline covariates are strongly prognostic. The PWT has the advantage of not requiring estimation of any state space distribution.

The choice of events to include in a hierarchy is always difficult and controversial. Inclusion of unimportant events makes the resulting metric less meaningful and can inflate the variance. Exclusion of important events can make the resulting metric further from an overall treatment effect.

We think the win time difference methods reflect well the desirability of outcomes, although we do not use the win time metric to rank subjects which would make the approach similar to the desirability of outcome ranking by Evans and Follmann.[13] In contrast, the win time difference methods use the win time metric itself to directly describe the treatment effect. This leads to tests that are very sensitive to treatment benefit on death. A pragmatic option is to use the MAX combination testing procedure.

The EWT method is very similar to restricted mean survival time in favor of treatment (RMT) of Mao[9] if both methods use the same approach to estimate the state space distributions. The main difference is that RMT restricts analysis to an arbitrary time cutoff. This means that RMT has an estimand that is easier to describe. In contrast, EWT uses all available data. Our simulations seem to indicate the RMT is slightly more powerful in most practical cases than EWT.

Further work could explore other definitions of clinical state. As Mao[9] did, one could define clinical state according to the number of occurrences of a recurrent event process with death given a value higher than the largest number of occurrences. This definition makes sense when there is one recurrent event process and death. Another possibility is to define a time window of burden for each clinical event, then the clinical state could be defined as the worst of the clinical burden windows currently affecting the participant. Such a definition would allow participants to transition non-monotonically to different clinical states throughout their follow-up.

ACKNOWLEDGEMENTS

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (http://biowulf.nih.gov). This research was supported in part by the Intramural Research Program of the NIH and NHLBI. The authors are employees of the National Heart, Lung, and Blood Institute. The views expressed in this article are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; National Institutes of Health; or the United States Department of Health and Human Services.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in NHLBI's biolincc at https://biolincc.nhlbi.nih.gov/home/, Reference 14.

ORCID

James F. Troendle https://orcid.org/0000-0002-7903-1015
Song Yang https://orcid.org/0000-0002-3051-5844
Dong-Yun Kim https://orcid.org/0000-0001-9405-4840

REFERENCES

1. Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. *Stat Med*. 1999;18:1341-1354.
2. Pocock SJ, Ariti AA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J*. 2012;33:176-182.
3. O'Connor CM, Whellan DJ, Lee K, et al. Efficacy and safety of exercise training in patients with chronic heart failure. *JAMA*. 2009;301:1439-1450.
4. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *J R Stat Soc Ser A*. 1972;135:185-207.
5. Cox DR. Regression models and life-tables (with discussion). *J R Stat Soc Series B Stat Methodol*. 1972;34:187-220.
6. Follmann D, Fay MP, Hamasaki T, Evans S. Analysis of ordered composite endpoints. *Stat Med*. 2020;39:602-616.
7. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat*. 1947;18:50-60.
8. Bebu I, Lachin JM. Large sample inference for a win ratio analysis of a composite outcome based on prioritized components. *Biostatistics*. 2016;17:178-187.
9. Mao L. On restricted mean time in favor of treatment. *Biometrics*. 2023;79:61-72.
10. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*. 1976;63:655-660.
11. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. *Stat Med*. 2009;28:956-971.
12. Yang S, Troendle J, Pak D, Leifer E. Event-specific win ratios for inference with terminal and non-terminal events. *Stat Med*. 2022;41:1225-1241.
13. Evans SR, Follmann D. Using outcomes to analyze patients rather than patients to analyze outcomes: a step toward pragmatism in benefit:risk evaluation. *Stat Biopharm Res*. 2016;8:386-393.
14. HF-ACTION. Heart failure: a controlled trial investigating outcomes of exercise training.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Troendle JF, Leifer ES, Yang S, et al. Use of win time for ordered composite endpoints in clinical trials. *Statistics in Medicine*. 2024;1-13. doi: 10.1002/sim.10045