Win statistics (win ratio, win odds, and net benefit) can complement one another to show the strength of the treatment effect on time-to-event outcomes

Gaohong Dong[1] | Bo Huang[2] | Johan Verbeeck[3] | Ying Cui[4] |

James Song[1] | Margaret Gamalo-Siebers[5] | Duolao Wang[6] |

David C. Hoaglin[7] | Yodit Seifu[8] | Tobias Mütze[9] | John Kolassa[10]

Abstract: Conventional analysis of a composite of multiple time-to-event outcomes uses the time to the first event. However, the first event may not be the most important outcome. To address this limitation, generalized pairwise comparisons and win statistics (win ratio, win odds and net benefit) have become popular and have been applied to clinical trial practice, including supporting drug approval by health authorities. However, win ratio, win odds and net benefit have typically been used individually. In this article, we examine the use of these three win statistics jointly for time-to-event outcomes. First, we explain the relation of point estimates and variances among the three win statistics, and the relation between the net benefit and the Mann-Whitney U statistic. Then, we explain that the three win statistics are based on the same win proportions, they test the same null hypothesis of equal win probabilities in two groups; we theoretically show that the Z-values of the statistical tests for the win statistics are approximately equal, therefore, the three win statistics provide very similar p-values and statistical powers. Finally, using simulation studies and data from a clinical trial, we demonstrate that, when there is (or little) censoring (i.e., early dropout), the three win statistics complement one another to show the strength of the treatment effect. However, when the amount of censoring is not small, and without adjustment for censoring, the win odds and the net benefit may have an advantage to interpret treatment effect compared to the win ratio; with adjustment (e.g., IPCW adjustment) for censoring, the three win statistics may complement one another to show the strength of the treatment effect. We perform the calculations using the R package, WINS, available on the Comprehensive R Archive Network.

Keywords: generalized pairwise comparisons, win statistics, IPCW, Mann-Whitney U statistic, nonparametrics

----

[1] BeiGene, Ridgefield Park, New Jersey, USA
[2] DSI, I-Biostat, University Hasselt, Hasselt, Belgium
[3] Pfizer Inc., Groton, Connecticut, USA
[4] Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, USA
[5] Pfizer Inc., Collegeville, Pennsylvania, USA
[6] Liverpool School of Tropical Medicine, Liverpool, UK
[7] Department of Population and Quantitative Health Sciences, UMass Chan Medical School, Worcester, Massachusetts, USA
[8] Bristol Myers Squibb, Berkeley Heights, New Jersey, USA
[9] Statistical Methodology, Novartis Pharma AG, Basel, Switzerland
[10] Department of Statistics, Rutgers University, Piscataway, New Jersey, USA

*Correspondence: Gaohong Dong, BeiGene, 55 Challenger Road, Ridgefield Park, NJ, USA.
Email: gaohong.dong@beigene.com

1. Introduction

For the analysis of prioritized multiple time-to-event outcomes in clinical trials, the common time-to-first-event analysis does not consider the outcomes' priorities. The first event may be due to an outcome of lower clinical importance (e.g., progressive disease vs. death in oncology studies, or heart failure hospitalization vs. cardiovascular death in chronic heart failure studies). To overcome this limitation, methods that incorporate the order of clinical importance among the endpoint components into the analysis, such as the generalized pairwise comparisons (GPC)[1] and the win statistics (win ratio, win odds, and net benefit), have been proposed[1-26]. The idea of all of these methods is to compare each subject in the experimental arm with every subject in the control arm in a pairwise manner and in each pairwise comparison either a winner is declared, or the comparison is considered to be a tie. Each pairwise comparison starts with the highest priority component and only takes the second most important component into account when the pairwise comparison based on the highest priority results in a tie. A pairwise comparison based on a hierarchical composite endpoint is considered to be tied if comparisons based on each component result in ties. Thus, lower-priority outcomes do not "mask" more important outcomes just because they occur earlier. Generalized pairwise comparison method and the win statistics also have been applied in the design and analysis of Phase III clinical trials (e.g., NCT04157751, NCT04847557 and NCT04510493 as registered in ClinicalTrials.gov) and in supporting drug approval by health authorities (e.g., tafamidis for treatment of cardiomyopathy per the ATTR-ACT trial). The stratified win ratio by Dong et al. [9] has also been applied to a Phase III clinical trial in adults who are in hospital for acute heart failure[27].

Win ratio, win odds, and net benefit have been extensively studied during the past decade. When there are no ties, the win odds reduces to the win ratio. There is a rich literature on their variance estimators and weighted and stratified analyses[5-12,19-22]. Regression analysis[22,23], sample size and power calculation[21,24], and method adjusting the win statistics for censoring[13,17,18] have become available. Brunner, Vandemeulebroecke and Mütze[20] argued that, for count, ordinal and continuous outcomes, and with some discussions on time-to-event outcomes, the win odds is preferable to the win ratio for quantifying the treatment effect in the presence of ties, because the ~~win odds~~ ties reflect that the two groups become more similar, particularly as the proportion of ties increases.

For time-to-event outcomes, the censoring-induced ties do not necessarily mean that the two patients in a pair have the same value of such outcome (see details in Section 3). Moreover, at present, win ratio, win odds, and net benefit have typically been used individually; and they have not been compared systematically. Therefore, in this article, we examine the use of these three win statistics jointly for time-to-event outcomes. We explain that the three win statistics test the same null hypothesis of equal win probabilities in the experimental and control groups; then we compare the win statistics systematically for time-to-event outcomes and discuss whether the three win statistics complement one another to show the strength of the treatment effect, through simulation studies and a large randomized cardiovascular outcome trial (the CHARM study[28]). We perform the calculations using the R package WINS by Cui and Huang[29], which is available on the Comprehensive R Archive Network.

2. Win statistics

2.1 Win statistics ~~and their statistical test and power~~

Following the generalized pairwise comparisons[1], there are three possible results from the comparison of a pair of patients (one patient from the experimental group and one patient from the control group): the patient in the experimental group wins, the control patient wins, or the two patients are tied. Let π_t, π_c and π_{tie} be the probabilities corresponding to these three results, for which $\pi_t + \pi_c + \pi_{tie} = 1$. Here we use the subscripts $_t$ and $_c$ to denote the experimental and control groups, respectively. Win ratio (*WR*), win odds (*WO*) and net benefit (*NB*) are defined as follows.

$$WR = \frac{\pi_t}{\pi_c} \tag{1a}$$

$$WO = \frac{\pi_t + 0.5\pi_{tie}}{\pi_c + 0.5\pi_{tie}} = \frac{\pi_t + 0.5(1-\pi_t-\pi_c)}{\pi_c + 0.5(1-\pi_t-\pi_c)} = \frac{\pi_t + 0.5(1-\pi_t-\pi_c)}{1-[\pi_t + 0.5(1-\pi_t-\pi_c)]} \tag{1b}$$

$$NB = \pi_t - \pi_c \tag{1c}$$

We consider a randomized clinical trial with N_t patients in the experimental group and N_c patients in the control group. Let T denote event time, C denote censoring time, $Y = \min(T, C)$ be the observed time, and $\delta = I(T<C)$ be the event indicator, where $I(\cdot)$ is the indicator function. We use $i = 1, 2, \ldots, N_t$ for patients in the experimental group and $j = 1, 2, \ldots, N_c$ for patients in the control group. We define kernel function K such as $K_{ij} = 1$ if a win for the experimental group occurs when an observed time Y_i=min(T_i, C$_i$) in this group is longer than an event time T_j in the control group, namely,

$$K_{ij} = 1 \text{ (win for the patient } i \text{ in the experimental group), if } \mathrm{Y}_i > Y_j \text{ and } \delta_j = 1$$
$$= 0\text{, otherwise.} \tag{2a}$$

Similarly, a kernel function L can be defined as below, which holds $L_{ij} = 1$ if the patient j in the control group wins over the patient i in the experimental group.

$$L_{ij} = 1 \text{ (win for the patient } j \text{ in the control group), if } Y_j > Y_i \text{ and } \delta_i = 1$$
$$= 0\text{, otherwise.} \tag{2b}$$

The number of wins for the experimental group can be counted as $n_t = \sum_{i=1}^{N_t}\sum_{j=1}^{N_c} K_{ij}$ and $n_c = \sum_{i=1}^{N_t}\sum_{j=1}^{N_c} L_{ij}$ is the number of wins for the control group. Then the win probabilities π_t and π_c can be estimated by $\hat{\pi}_t = P_t = n_t/N_tN_c$ and $\hat{\pi}_c = P_c = n_c/N_tN_c$, respectively, where P_t and P_c are win proportions in the experimental and control groups, respectively. Therefore, win ratio, win odds and net benefit can be estimated by

$$\widehat{WR} = \frac{P_t}{P_c}\,, \tag{3a}$$

$$\widehat{WO} = \frac{P_t+0.5P_{tie}}{P_c+0.5P_{tie}} = \frac{P_t+0.5(1-P_t-P_c)}{P_c+0.5(1-P_t-P_c)} = \frac{P_t+0.5(1-P_t-P_c)}{1-[P_t+0.5(1-P_t-P_c)]}\,, \tag{3b}$$

$$\widehat{NB} = P_t - P_c\ . \tag{3c}$$

Hence, the win ratio is a ratio of win proportions, the win odds is an odds of win proportions, and the net benefit is a difference in win proportions. Because win ratio, win odds and net benefit are derived using the same win proportions, they test the null hypotheses: *WR* = 1, *WO* = 1 and *NB* = 0, respectively, which are equivalent to the testing of the null hypothesis of equal win probabilities in the two treatment groups, H$_0$: $\pi_t = \pi_c$.

The statistics P_t and P_c are U-statistics and are asymptotically normally (*AN*) distributed. Therefore, n_t and n_c are also asymptotically normal,

$$\binom{n_t}{n_c} \sim AN\left(\begin{bmatrix}\theta_t \\ \theta_c\end{bmatrix}, \begin{bmatrix}\sigma_t^2 & \sigma_{tc} \\ \sigma_{tc} & \sigma_c^2\end{bmatrix}\right). \tag{4}$$

By the delta method, $log(WR)$, $log(WO)$ and *NB* are asymptotically normally distributed with the following variances:

$$\sigma^2_{log(WR)} = \frac{\sigma_t^2}{(\theta_t)^2} + \frac{\sigma_c^2}{(\theta_c)^2} - \frac{2\sigma_{tc}}{\theta_t\theta_c} \tag{5a}$$

$$\sigma^2_{log(WO)} = (\sigma_t^2 - 2\sigma_{tc} + \sigma_c^2)\left(\frac{1}{\gamma} + \frac{1}{N_tN_c-\gamma}\right)^2/4, \tag{5b}$$

$$\sigma^2_{NB} = (\sigma_t^2 - 2\sigma_{tc} + \sigma_c^2)/(N_tN_c)^2 \tag{5c}$$

where $\gamma = \theta_t + 0.5(N_tN_c - \theta_t - \theta_c)$.

Under the null hypothesis H_0: $\pi_t = \pi_c$, θ_t and θ_c can be estimated as

$$\hat{\theta}_t = \hat{\theta}_c = (n_t + n_c)/2. \quad (6)$$

Then the variances of $log(WR)$, $log(WO)$ and NB can be estimated under the null hypothesis by,

$$\hat{\sigma}^2_{log(WR)} = \frac{(\hat{\sigma}_t^2 - 2\hat{\sigma}_{tc} + \hat{\sigma}_c^2)}{[(n_t+n_c)/2]^2}, \quad (7a)$$

$$\hat{\sigma}^2_{log(WO)} = \frac{\hat{\sigma}_t^2 - 2\hat{\sigma}_{tc} + \hat{\sigma}_c^2}{(N_t N_c/2)^2}, \quad (7b)$$

$$\hat{\sigma}^2_{NB} = \frac{\hat{\sigma}_t^2 - 2\hat{\sigma}_{tc} + \hat{\sigma}_c^2}{(N_t N_c)^2}. \quad (7c)$$

The calculations for $\hat{\sigma}_t^2$, $\hat{\sigma}_c^2$ and $\hat{\sigma}_{tc}$ can be found in Dong et al.[8,9] and Bebu and Lachin[7].

It should be noted that (2a) and (2b) are for a single time-to-event outcome. The setting for prioritized multiple outcomes and for inverse-probability-of-censoring weighting (IPCW) adjustment for censoring can be formulated similarly (see details in Dong et al.[17,18]).

2.2 Point estimate

From the point estimate perspective, since the win odds considers a tie as a half win for the experimental group and a half win for the control group as defined in (1b), the win odds is always closer to the null value of 1.0 compared to the win ratio as explained in Dong et al[3] and Brunner, Vandemeulebroecke and Mütze[20]. When there are no ties, the win odds reduces to the win ratio.

It is straight forward to derive that win ratio, win odds and net benefit have the following relationships. We demonstrate these relationships via Simulation Study 1 in Section 4.1.

$$NB = \frac{WR-1}{WR+1}\frac{n_t+n_c}{N_t N_c} = \frac{WR-1}{WR+1}(1 - P_{tie}) \quad (8a)$$

$$NB = \frac{WO-1}{WO+1} \quad (8b)$$

$$WO = \frac{1+NB}{1-NB} \quad (8c)$$

$$WO = \frac{WR+0.5P_{tie}(WR-1)}{1+0.5P_{tie}(WR-1)} \quad (8d)$$

2.3 Variance

With respect to the variance, the estimated variance for the win odds ($\hat{\sigma}^2_{log(WO)}$ per (7b)) is always smaller than or equal to that for the win ratio ($\hat{\sigma}^2_{log(WR)}$ per (7a)) because the total number of wins is always smaller than or equal to the total number of comparisons, namely, $n_t + n_c \leq N_t N_c$ (note: $n_t + n_c = N_t N_c$ only if there are no ties). Consequently, the confidence interval for the win odds is always narrower than that for the win ratio, and their confidence intervals become the same when there are no ties. For a large clinical trial with low event rates (i.e., proportion of ties is high), the point estimate of the win odds can be much closer to 1.0 and its confidence interval can be very narrow compared to the win ratio, as also shown from CHARM application in Section 5.

2.4 Net Benefit as a direct transformation of the Mann-Whitney U statistic

Without ties (i.e., $n_t + n_c = N_t N_c$), the number of wins, n_t and n_c, are Mann-Whitney U statistics[30]. With ties, $U = n_t + 0.5(N_t N_c - n_t - n_c)$ is a Mann-Whitney U statistic instead, and its variance is[31],

$$\sigma_U^2 = \frac{N_t N_c (N+1)}{12} - \frac{N_t N_c \sum_{i=1}^{k} (d_i^3 - d_i)}{12N(N-1)}, \tag{9}$$

where $N = N_t + N_c$, k is the number of distinct observations and d_i is the number of times the ith tied observation is repeated. The expression (9) clearly shows that the variance of the Mann-Whitney U statistic increases as the number of ties decreases.

Verbeeck et al.[32] explained that the net benefit is a direct transformation of the Mann-Whitney U statistic,

$$NB = \frac{2U}{N_t N_c} - 1, \tag{10}$$

where the quantity of $\frac{U}{N_t N_c}$ is an estimator of probabilistic index. Therefore, the variance of the net benefit also increases as the number of ties decreases.

2.5 Approximate equivalence of statistical tests for the three win statistics

As shown from (7a), (7b) and (7c), by applying the delta method, the asymptotic variances for log(win ratio), log(win odds) and net benefit can be derived based on the asymptotic variances and covariance of the win proportions (or numbers of wins) for the two treatment groups, the three win statistics test the same null hypothesis of equal win probabilities in the experimental and control groups, and the z-values of the three test statistics are approximately equal as we show below. Therefore, the three test provide very similar p-values and powers.

From (7a) and (7c), we can obtain $\hat{\sigma}_{NB}^2 = \frac{(n_t+n_c)^2}{4(N_tN_c)^2}\hat{\sigma}_{log(WR)}^2$. By applying (8a), we can calculate the Z-value of the statistical test for the net benefit as

$$Z_{NB} = \frac{NB}{\hat{\sigma}_{NB}} = \frac{WR-1}{WR+1}\frac{n_t+n_{tc}}{N_tN_c}\frac{1}{\hat{\sigma}_{\log(WR)}}\frac{2N_tN_c}{n_t+n_{tc}} = \frac{WR-1}{WR+1}\frac{2}{\hat{\sigma}_{\log(WR)}}.$$

Following the Taylor expansion, $\log(x) = 2\left[\frac{x-1}{x+1} + \frac{1}{3}\left(\frac{x-1}{x+1}\right)^3 + \frac{1}{5}\left(\frac{x-1}{x+1}\right)^5 +\ldots\right]$ for $x > 0$. Therefore, $\log(x) \approx \frac{2(x-1)}{x+1}$. This means $\log(WR) \approx \frac{2(WR-1)}{WR+1}$. Hence, the Z-values of the statistical tests for net benefit and log(*WR*) are approximately equal as shown below.

$$Z_{NB} \approx \frac{\log(WR)}{\hat{\sigma}_{\log(WR)}} = Z_{WR}. \tag{11a}$$

Similarly, since $NB = \frac{WO-1}{WO+1}$ and $\hat{\sigma}_{NB}^2 = \frac{1}{4}\hat{\sigma}_{log(WO)}^2$ per (7b), (7c) and (8a), the Z-values of the statistical tests for net benefit and log(*WO*) are also approximately equal.

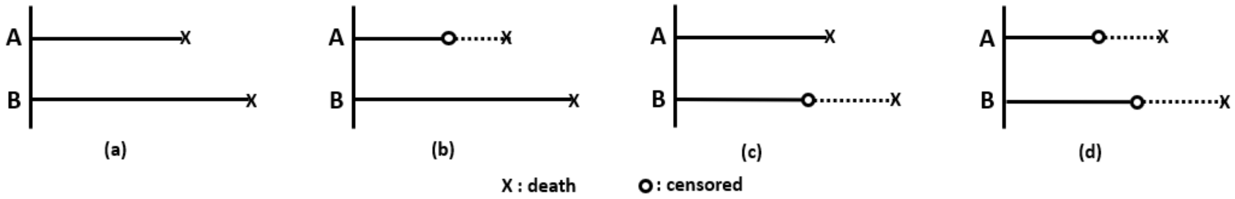
$$Z_{NB} \approx \frac{\log(WO)}{\hat{\sigma}_{\log(WO)}} = Z_{WO}. \tag{11b}$$

3. Censoring for time-to-event outcomes

When ideally there are no censoring, the win probabilities at time *x* can be calculated by $\pi_t(x) = -\int_0^x S_t dS_c$ and $\pi_c = -\int_0^x S_c dS_t$, where S_t and S_c are the survival functions of time to event in the experimental and control groups, respectively, and the true value of the three win statistics can be calculated accordingly based on $\pi_t(x)$ and $\pi_c(x)$. In practice, S_t and S_c are unknown, Kaplan-Meier estimators of S_t and S_c, can be used to estimate the win statistics[4].

As seen from the kernel functions defined in (2a) and (2b), censoring for time-to-event outcomes could impact the win statistics. Censoring can be caused by (a) some patients dropped out without experiencing an event of interest (i.e., early dropout) and (b) at the time of the data cutoff for the analysis, some patients have not experienced an event (i.e., administrative censoring or end-of-study censoring). For illustration of censoring bias, assume that both patients A and B had a death event (Figure 1a). Patient B is the winner because Patient A died earlier. However, if Patient A was censored (Figure 1b), or Patient B was censored (Figure 1c) or both patients were censored before the death of Patient A (Figure 1d), a “win” cannot be determined for this pair of patients. In the calculation of win statistics, this situation is typically considered a “tie”, which obviously introduces a bias. In fact, for time-to-event outcomes, the censoring-induced ties do not necessarily mean that the two patients in a pair have the same value of such outcome.

Figure 1 Illustration of censoring bias



Therefore, censoring has an impact on the win statistics. As demonstrated by Dong et al[16], the win probabilities can be calculated by $\pi_t(x) = -\int_0^x F^{(t)} G^{(t)}\, G^{(c)} dF^{(c)}$ and $\pi_c(x) = -\int_0^x F^{(c)} G^{(c)}\, G^{(t)} dF^{(t)}$, where $G^{(t)}$ and $G^{(c)}$ are survival function of time to censoring in the experimental and control groups, respectively. Statistical methods adjusting the win statistics for censoring are available. For example, Péron et al.[13] suggested an adaptation of Efron's scoring to adjust the win statistics, and Dong et al.[17,18] applied the inverse-probability-of-censoring weighting (IPCW) approach to adjust the win statistics for independent censoring (i.e., IPCW-adjusted win statistics) and dependent censoring (i.e., CovIPCW-adjusted win statistics).

~~For time-to-event analyses using the win statistics, censoring can cause ties. The ties due to administrative censoring, in general, mean that the two patients of a pairwise comparison are similar at the data cutoff for the analysis, hence they are less of a concern. Therefore, in this article, we focus on the censoring caused by early dropout. To ease the writing, we refer censoring to early dropout in this article.~~

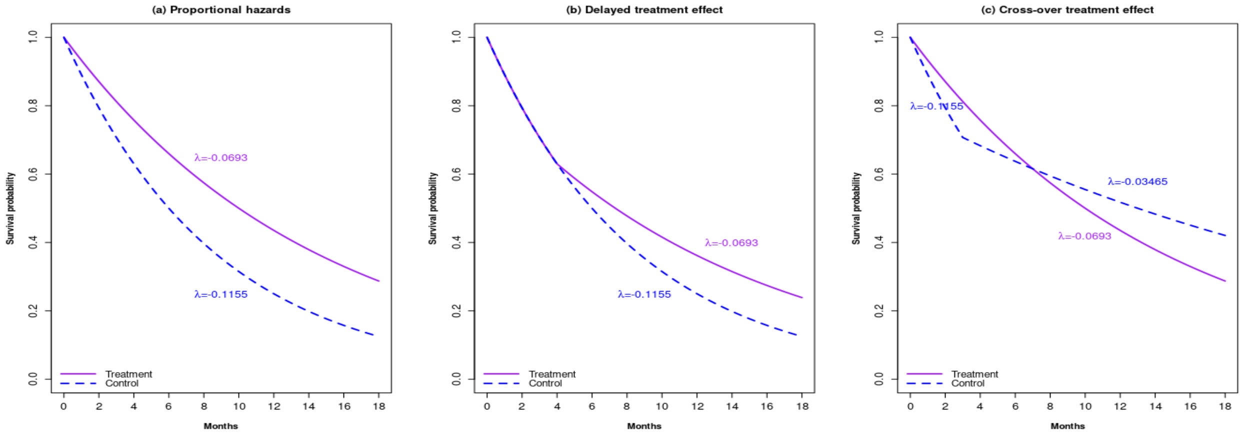
4. Simulation studies

We extend the simulation studies presented in Dong et al.[17] to investigate win ratio, win odds and net benefit in the setting of time-to-event outcomes.

4.1 Simulation study 1

In this simulation study, we analyze a single time-to-event outcome without censoring. We use three scenarios that arise in practice: (a) proportional hazards in the two groups, (b) delayed treatment effect in the experimental group, and (c) cross-over treatment effect between the two groups. For all three scenarios, following Huang and Kuan[33], we use exponential or piecewise exponential functions to generate 1000 simulated datasets with 200 patients per group. Figure 2 shows the hazard rates and survival curves for each scenario.

Figure 2 Assumed hazard rates (λ) and survival curves for simulations



For Scenario (a), proportional hazards in the two groups (Table 1a), the win ratio is the reciprocal of the hazard ratio[4,15,16,20]. Therefore, the estimated win ratio (the median over the 1000 simulated datasets) is close to the true value of 1.67 at all timepoints. The width of the estimated 95% confidence interval is 2.68 at Month 1, which is relatively wide because few events are observed and the evidence of the treatment effect is not yet strong. The estimated 95% confidence interval narrows to width 0.81 at Month 18, as most events have been observed and the evidence of the treatment effect has become very strong. In contrast, the point estimates of the win odds and the net benefit increase over time as the evidence of the treatment effect becomes stronger. The widths of their 95% confidence intervals also increase over time. For example, the estimated win odds is 1.08 at Month 1, reflecting that a considerable proportion of pairwise comparisons results in ties and as such provide little evidence for an effect of the treatment. Similarly, at Month 1, the 95% confidence interval is narrowest. As more events are observed and the evidence of the treatment effect becomes stronger over time, the point estimate of the win odds and the lower limit of the confidence interval move more away from the null value of 1.0. Although the 95% confidence interval becomes wider over time, the significance level (i.e., p-value, not shown in the tables) becomes smaller.

These results make sense. For this scenario of proportional hazards in the two groups, WR = 1/HR is constant and proportion of ties, P_{tie}, decreases as more events are observed over time. As expressed in (8a) and (8c), $NB = \frac{WR-1}{WR+1}(1 - P_{tie})$ increases with decreasing ties over time, and $WO = \frac{1+NB}{1-NB}$ also increases over time.

For Scenario (b), delayed treatment effect in the experimental group (Table 1b), because the survival curves start to separate at Month 4, with hazard ratio = 0.60, all three win statistics increase over time after Month 4 as the evidence of the treatment effect increases. The 95% confidence interval for the win ratio becomes narrower over time, and those for the win odds and the net benefit become wider.

For Scenario (c), cross-over treatment effect between the two groups (Table 1c), before Month 3, the experimental group performs better, with hazard ratio = 0.60; after Month 3, the control group performs better, with hazard ratio = 2.0. Therefore, the point estimates of the win ratio, the win odds and the net benefit first increase and then decrease from Month 3. The 95% confidence intervals for the win ratio become narrower over time, whereas those for the net benefit become wider. Interestingly, the width of the 95% confidence intervals for the win odds increase first, then decrease from Month 9.

For all three scenarios, at Month 18, the win ratio and the win odds get closer to each other since majority of events have been observed and there are few ties; consequently, their confidence intervals also get closer to each other.

Table 1a Win statistics and 95% confidence intervals for Scenario (a)

Time	Win proportion (%)		WR		WO		Net benefit (%)	
	Treatment	Control	Median (95% CI)	Width of 95% CI	Median (95% CI)	Width of 95% CI	Median (95% CI)	Width of 95% CI
Month 1	10.2	6.1	1.67 (0.86, 3.54)	2.68	1.08 (0.97, 1.22)	0.25	4.1 (-1.3, 9.9)	11.2
Month 3	25.9	15.7	1.68 (1.11, 2.50)	1.39	1.22 (1.04, 1.45)	0.41	10.0 (2.2, 18.3)	16.1
Month 6	41.3	24.9	1.66 (1.21, 2.32)	1.11	1.40 (1.13, 1.71)	0.58	16.5 (6.2, 26.2)	20.0
Month 9	50.2	30.3	1.66 (1.27, 2.19)	0.92	1.50 (1.21, 1.85)	0.64	20.0 (9.6, 29.7)	20.1
Month 12	55.3	33.3	1.67 (1.28, 2.16)	0.88	1.58 (1.25, 1.98)	0.73	22.2 (11.0, 32.9)	21.9

Month 18	60.0	36.1	1.66 (1.31, 2.12)	0.81	1.63 (1.30, 2.06)	0.76	23.9 (12.9, 34.7)	21.9

95% CI (confidence interval) is constructed as the 95% percentile interval (2.5th percentile, 97.5th percentile) from 1000 simulations.

Table 1b Win statistics and 95% confidence intervals for Scenario (b)

Time	Win proportion (%)		WR		WO		Net benefit (%)	
	Treatment	Control	Median (95% CI)	Width of 95% CI	Median (95% CI)	Width of 95% CI	Median (95% CI)	Width of 95% CI
Month 1	10.0	10.0	1.00 (0.63, 1.55)	0.91	1.00 (0.92, 1.09)	0.17	0.0 (-4.4, 4.3)	8.7
Month 3	24.6	24.7	1.00 (0.78, 1.29)	0.52	1.00 (0.88, 1.14)	0.25	0.0 (-6.2, 6.4)	12.6
Month 6	37.4	34.5	1.08 (0.88, 1.33)	0.45	1.06 (0.91, 1.23)	0.32	2.9 (-4.5, 10.4)	14.9
Month 9	44.8	39.0	1.14 (0.96, 1.38)	0.42	1.12 (0.97, 1.31)	0.34	5.7 (-1.6, 13.5)	15.1
Month 12	49.2	41.6	1.18 (1.00, 1.40)	0.41	1.16 (1.00, 1.36)	0.36	7.5 (0.0, 15.4)	15.4
Month 18	52.8	43.9	1.21 (1.02, 1.42)	0.40	1.20 (1.02, 1.40)	0.38	9.0 (1.0, 16.8)	15.8

95% CI (confidence interval) is constructed as the 95% percentile interval (2.5th percentile, 97.5th percentile) from 1000 simulations.

Table 1c Win statistics and 95% confidence intervals for Scenario (c)

Time	Win proportion (%)		WR		WO		Net benefit (%)	
	Treatment	Control	Median (95% CI)	Width of 95% CI	Median (95% CI)	Width of 95% CI	Median (95% CI)	Width of 95% CI
Month 1	10.2	6.1	1.67 (0.86, 3.54)	2.68	1.08 (0.97, 1.22)	0.25	4.1 (-1.3, 9.9)	11.2
Month 3	25.9	15.7	1.68 (1.11, 2.50)	1.39	1.22 (1.04, 1.45)	0.41	10.0 (2.2, 18.3)	16.1
Month 6	31.4	26.0	1.21 (0.88, 1.69)	0.81	1.11 (0.93, 1.35)	0.42	5.5 (-3.7, 14.8)	18.5
Month 9	35.3	33.6	1.04 (0.79, 1.45)	0.66	1.03 (0.85, 1.29)	0.44	1.5 (-8.2, 12.5)	20.7
Month 12	38.0	39.2	0.96 (0.74, 1.30)	0.56	0.97 (0.80, 1.22)	0.43	-1.5 (-11.4, 10.2)	21.5
Month 18	41.5	46.1	0.90 (0.71, 1.18)	0.47	0.91 (0.74, 1.16)	0.41	-4.8 (-14.6, 7.3)	22.0

95% CI (confidence interval) is constructed as the 95% percentile interval (2.5th percentile, 97.5th percentile) from 1000 simulations.

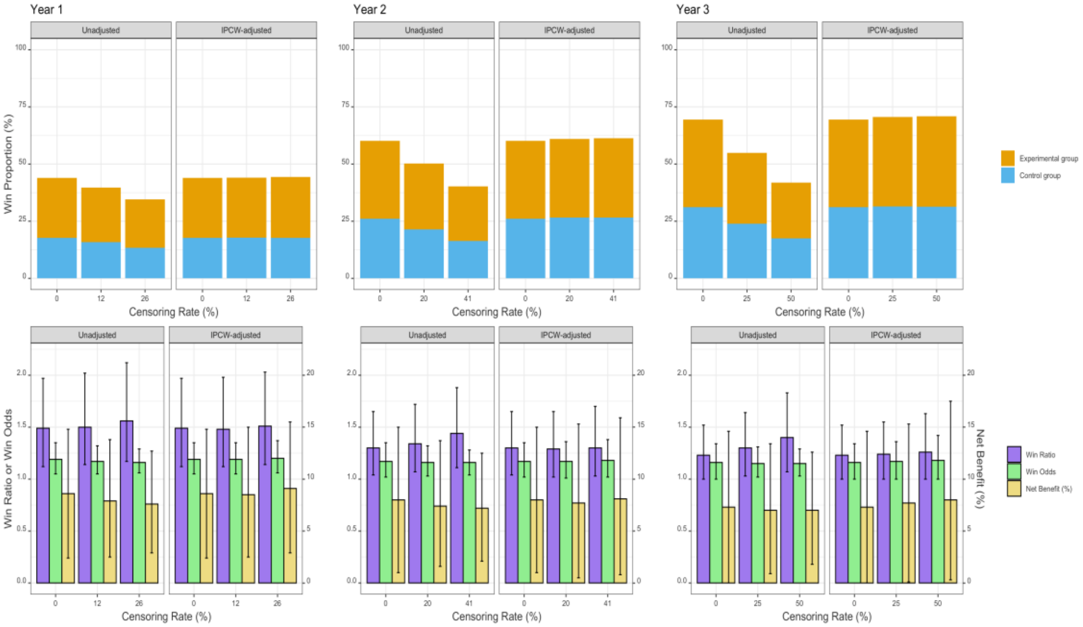
In summary, as follow-up time increases, more events are observed, and evidence of the treatment effect becomes stronger. For all three scenarios the win proportions in both groups increase over time, and the 95% confidence interval for the win ratio becomes narrower. With respect to the net benefit, its 95% confidence interval becomes wider over time. This is not surprising since increase in variance when the number of ties decreases is a well-known property of the Mann-Whitney statistic[31,32], and the net benefit is a direct transformation of the Mann-Whitney U statistic as explained in Section 2.3. For Scenario (c), however, the width of the 95% confidence intervals for the win odds may first increase and then decrease, corresponding to the cross-over pattern of the treatment effect.

Nevertheless, from this simulation without censoring, it looks that the three win statistics complement one another to show the strength of the treatment effect. Therefore, it may be helpful to present three win statistics together when there is no (or little) censoring (i.e., early dropout). This makes sense as explained in Section 3 because the win probabilities at time *x* can be calculated by $\pi_t(x) = -\int_0^x S_t dS_c$ and $\pi_c = -\int_0^x S_c dS_t$ when ideally there is no censoring for a single time-to-event outcome. In practice, S_t and S_c are replaced with their corresponding Kaplan-Meier estimators to estimate the win statistics[4], or equivalently the U-statistics approach described in Section 2.1 can be used. The latter (i.e., the U-statistics approach) can be applied regardless of whether there is censoring.

4.2 Simulation study 2

As described in Dong at al. [17], we selected 800 patients from clinical trials in cardiovascular (CV) disease with the composite of death and hospitalization as the primary endpoint. We used the data up to 3 years, and excluded patients who dropped out prior to Year 3, so that we could estimate the win statistics without bias from censoring (i.e., we considered these estimated values as true win statistics) up to Year 3. Then we applied independent exponentially distributed censoring, Exp(0.0004) and Exp(0.001), corresponding to 25% and 50% censoring, respectively, at Year 3. As discussed in Dong at al.[17], the experimental group performs better over time than the control group, and the hazards in the two groups are nonproportional without a particular pattern. We apply the inverse-probability-of-censoring weighting approach to adjust win statistics for independent censoring (i.e., IPCW-adjusted win statistics).

Figure 3 Unadjusted vs IPCW-adjusted win proportions and win statistics



This simulation study produces the same findings as scenarios (a) and (b) in Simulation study 1. Figure 3 presents unadjusted vs IPCW-adjusted win proportions and win statistics. As also shown in Table 2a and Table 2b, regardless of censoring and adjustment (unadjusted vs IPCW-adjusted), the 95% confidence interval for the win ratio becomes narrower over time, and those for the win odds and the net benefit become slightly wider over time. The win statistics decrease slightly from Year 1 to Year 3. This pattern means that the evidence for a slightly larger treatment effect is stronger early in the study.

In the presence of censoring, the number of events becomes less and the variability in the unadjusted win ratio becomes larger (i.e., the 95% confidence interval becomes wider, Table 2a). However, the 95% confidence intervals for the unadjusted win odds and net benefit become narrow as the amount of censoring increases. For example, at Year 3, the width of the 95% confidence interval for the unadjusted win odds is 0.34, 0.29 and 0.26 corresponding to 0%, 25% and 50% censoring. This indicates that, in presence of censoring particularly when the proportion of censoring is not small, the win odds and the net benefit may have an advantage to interpret treatment effect compared to the win ratio, as also reported in Brunner, Vandemeulebroecke and Mütze[20].

Table 2a Unadjusted win statistics and 95% confidence intervals for CV example

Time	Censor-ing (%)	Win proportion (%)		WR		WO		Net benefit (%)	
		Treat-ment	Control	Median (95% CI)	Width of 95% CI	Median (95% CI)	Width of 95% CI	Median (95% CI)	Width of 95% CI
Year 1	0%	26.3	17.7	1.49 (1.12, 1.97)	0.85	1.19 (1.05, 1.35)	0.30	8.6 (2.4, 14.8)	12.4
	12%	23.8	15.9	1.50 (1.14, 2.02)	0.88	1.17 (1.05, 1.32)	0.27	7.9 (2.5, 13.8)	11.3
	26%	21.1	13.4	1.56 (1.17, 2.12)	0.95	1.16 (1.06, 1.29)	0.23	7.6 (2.9, 12.7)	9.8
Year 2	0%	34.1	26.1	1.30 (1.04, 1.65)	0.61	1.17 (1.02, 1.35)	0.33	8.0 (1.0, 15.0)	14.0
	20%	28.8	21.4	1.34 (1.07, 1.72)	0.65	1.16 (1.03, 1.32)	0.29	7.4 (1.6, 13.7)	12.1
	41%	23.8	16.4	1.44 (1.11, 1.88)	0.77	1.16 (1.04, 1.28)	0.24	7.2 (2.1, 12.5)	10.4
Year 3	0%	38.4	31.1	1.23 (1.00, 1.52)	0.52	1.16 (1.00, 1.34)	0.34	7.3 (0.0, 14.6)	14.6
	25%	31.0	23.9	1.30 (1.03, 1.64)	0.61	1.15 (1.02, 1.31)	0.29	7.0 (0.9, 13.4)	12.5
	50%	24.4	17.5	1.40 (1.07, 1.83)	0.76	1.15 (1.03, 1.29)	0.26	7.0 (1.8, 12.6)	10.7

95% CI (confidence interval) is constructed as the 95% percentile interval (2.5th percentile, 97.5th percentile) from 1000 simulations.

Table 2b IPCW-adjusted win statistics and 95% confidence intervals for CV example

Time	Censor-ing (%)	Win proportion (%)		WR		WO		Net benefit (%)	
		Treat-ment	Control	Median (95% CI)	Width of 95% CI	Median (95% CI)	Width of 95% CI	Median (95% CI)	Width of 95% CI
Year 1	0%	26.3	17.7	1.49 (1.12, 1.97)	0.85	1.19 (1.05, 1.35)	0.30	8.6 (2.4, 14.8)	12.4
	12%	26.3	17.8	1.48 (1.12, 1.98)	0.86	1.19 (1.05, 1.35)	0.30	8.5 (2.5, 15.0)	12.5
	26%	26.7	17.7	1.51 (1.14, 2.03)	0.89	1.20 (1.06, 1.37)	0.31	9.1 (2.9, 15.5)	12.6
Year 2	0%	34.1	26.1	1.30 (1.04, 1.65)	0.61	1.17 (1.02, 1.35)	0.33	8.0 (1.0, 15.0)	14.0
	20%	34.4	26.6	1.29 (1.02, 1.65)	0.64	1.17 (1.01, 1.36)	0.35	7.7 (0.5, 15.3)	14.8
	41%	34.7	26.6	1.30 (1.03, 1.70)	0.65	1.18 (1.02, 1.38)	0.36	8.1 (0.8, 15.9)	15.1
Year 3	0%	38.4	31.1	1.23 (1.00, 1.52)	0.52	1.16 (1.00, 1.34)	0.34	7.3 (0.0, 14.6)	14.6
	25%	39.1	31.5	1.24 (1.00, 1.55)	0.55	1.17 (1.00, 1.36)	0.36	7.7 (0.1, 15.3)	15.2
	50%	39.6	31.3	1.26 (1.00, 1.63)	0.63	1.18 (1.00, 1.42)	0.42	8.0 (0.3, 17.5)	17.2

95% CI (confidence interval) is constructed as the 95% percentile interval (2.5th percentile, 97.5th percentile) from 1000 simulations.

With IPCW-adjustment (Table 2b), both the point estimates and the width of the 95% confidence intervals for the win odds and the net benefit are generally more ~~strikingly~~ stable over time than with the unadjusted win statistics as shown by the range of both the point estimates and interval widths. This indicates that the IPCW adjustment may be more effective at correcting bias due to censoring. The correction is especially evident for the more variable win ratio (Table 2b): after the IPCW adjustment the width of the 95% confidence interval at Year 3 is 0.55 with 25% censoring and 0.63 with 50% censoring, much closer to 0.52 with no censoring than the 0.61, 0.76

without the adjustment. Therefore, the IPCW adjustment corrects for bias, and it also aligns the confidence interval width to the width under no censoring. This indicates that, with an adjustment (e.g., IPCW adjustment) for censoring, the three win statistics may complement one another to show the strength of the treatment effect.

5. Application to CHARM studies

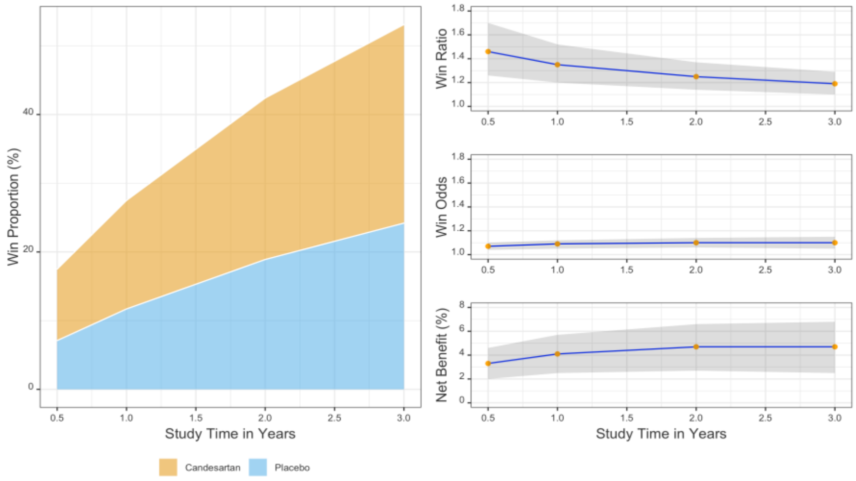
The CHARM trial[28] was a randomized, double-blind, placebo-controlled study comparing candesartan with placebo in patients with chronic heart failure. The primary endpoint was a composite of cardiovascular death or hospitalizations due to chronic heart failure. A total of 7599 patients were randomized to the two groups.

Because only a small number of patients dropped out prior to Year 3, the unadjusted and IPCW-adjusted win statistics are very similar. Table 3 and Figure 4 present the IPCW-adjusted win proportions and win statistics. As in scenarios (a) and (b) of Simulation study 1, the width of 95% confidence intervals for the win ratio becomes narrower over time, and those for the net benefit become wider. Very interestingly, for this study as a large clinical trial, the point estimate of the win ratio declines with the follow up-time, whereas the point estimates of the win odds and the net benefit vary little with the follow-up time. Moreover, the 95% confidence interval for the win odds is very narrow; its width increases from 0.06 to 0.10, whereas the width for the win ratio decreases from 0.44 to 0.20. This may indicate that, compared to the win ratio, the win odds may have an advantage of narrow confidence interval when the proportion of ties is not small.

Table 3 IPCW-adjusted win statistics and 95% confidence intervals for the CHARM program

Time	Win proportion (%)		WR		WO		Net benefit (%)	
	Treat-ment	Control	WR (95% CI)	Width of 95% CI	WO (95% CI)	Width of 95% CI	NB (95% CI)	Width of 95% CI
Month 6	10.4	7.1	1.46 (1.26, 1.70)	0.44	1.07 (1.04, 1.10)	0.06	3.3 (2.0, 4.6)	2.6
Year 1	15.8	11.7	1.35 (1.20, 1.52)	0.32	1.09 (1.05, 1.12)	0.07	4.1 (2.5, 5.7)	3.2
Year 2	23.5	18.9	1.25 (1.14, 1.37)	0.23	1.10 (1.06, 1.14)	0.09	4.7 (2.7, 6.6)	3.9
Year 3	28.9	24.2	1.19 (1.10, 1.29)	0.20	1.10 (1.05, 1.15)	0.10	4.7 (2.5, 6.8)	4.4

Figure 4 IPCW-adjusted win proportions and win statistics over time



6. Conclusions and Discussions

The generalized pairwise comparisons and win statistics (win ratio, win odds and net benefit), in particular, win ratio and net benefit, have received increasing attention in methodological research. They also have been applied in the design and analysis of Phase III clinical trials and in supporting drug approval by health authorities. However, win ratio, win odds and net benefit have been typically used individually.

The three win statistics test the same null hypothesis of equal win probabilities in the experimental and control groups, and they provide similar p-values and statistical powers since the Z-values of the their corresponding statistical tests are approximately equal as proved in Section 2.5. Therefore, in this article, we target to show whether the three win statistics complement one another for analyzing the strength of the treatment effect. In the setting of time-to-event outcomes, we use simulation studies and data from a clinical trial to explain their behavior in relation to proportionality of hazards and in relation to the Mann-Whitney test. In our view, in the absence of censoring or when the amount of censoring is small, presenting win proportions, win ratio, win odds and net benefit together can give a more detailed picture of an analysis. Specifically, the win ratio and win odds are relative quantitative measures (similar to the hazard ratio) to evaluate the relative strength of one treatment group versus the other, while the net benefit is an absolute quantitative measure (similar to difference in response rates) that is bounded by -1.0 and 1.0 to evaluate the absolute strength of one treatment group versus the other. In the presence of a positive treatment effect (i.e., win proportion for the treatment group is higher than that for the control group), the win ratio is always greater than the win odds (they are equal in the absence of ties).

In the case of continuous, ordinal and binary outcomes, the win odds may be preferable to the win ratio to handle the ties more appropriately[20], because a tie implies that the two patients in a pair had the same value of an outcome. For time-to-event outcomes, however, the censoring-induced ties do not necessarily mean that the two patients in a pair have the same value of such outcome (see details in Section 3). As noted in Oakes[4] and Dong et al.[16], unless the proportional hazards assumption holds, the win ratio can be impacted by censoring and follow-up time. The same issue also applies to the win odds and the net benefit which are sensitive to censoring-induced ties. Handling of ties caused by censoring is quite complex because of various censoring mechanisms (non-informative administrative censoring, informative censoring owing to drop-out or confounding intercurrent event). For example, when the censoring is primarily due to administrative censoring (i.e. follow-up time), the win ratio may have more meaningful clinical interpretation compared to the win odds. This is because the win ratio can be viewed as a special version of the win odds, by imputing $100 * \pi_t/(\pi_t + \pi_c)$% and $100 * \pi_c/(\pi_t + \pi_c)$% of the ties as win proportions for the experimental and control groups, respectively, instead of 50% of ties for each group as used in the win odds. This imputation approach assuming missing-at-random is analogous to the conditional power approach based on observed data at the interim analysis in group sequential designs. However, when censoring is primarily due to dropout or informative intercurrent events, without an adjustment for censoring, the win odds and the net benefit may have an advantage to interpret treatment effect as this type of censoring may cause the observed win ratio greatly away from the true value in either positive or negative direction; with an adjustment (e.g., IPCW adjustment) for censoring, the three win statistics may complement one another to show the strength of the treatment effect.

In general, for time-to-event outcomes, comparisons among the three win statistics is subtle. Nevertheless, the Z-values of the statistical tests for the three win statistics are approximately equal and their test provide similar p-values. Therefore, the three win statistics may complement one another to show the strength of the treatment effect, and presenting win proportions, win ratio, win odds and net benefit together can give a more detailed picture of an analysis. On the other hand, one may just use and present one statistical measure (win ratio, win odds, or net benefit) for clinical trial design and analysis because the three win statistics are complementary. It is also advisable to graphically display the win statistics over follow-up time following Finkelstein and Schoenfeld[15] to assess the variability and robustness of the win statistics since censoring-induced ties decrease over time.

Data availability statement

The simulated data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine.* 2010;29(30):3245-3257.
2. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal.*

2012;33(2):176-182.

3. Dong G, Hoaglin DC, Qiu J, Matsouaka RA, Chang Y, Wang J, Vandemeulebroecke M. The win ratio: on interpretation and handling of ties. *Statistics in Biopharmaceutical Research.* 2020;12(1):99-106.
4. Oakes D. On the win-ratio statistic in clinical trials with multiple types of event. *Biometrika.* 2016;103(3):742–745.
5. Wang D., Pocock S. A win ratio approach to comparing continuous non-normal outcomes in clinical trials. *Pharmaceutical Statistics.* 2016;15(3):238-245.
6. Luo X, Tian H, Mohanty S, Tsai WY. An alternative approach to confidence interval estimation for the win ratio statistic. *Biometrics.* 2015;71(1):139-145.
7. Bebu I, Lachin JM. Large sample inference for a win ratio analysis of a composite outcome based on prioritized components. *Biostatistics.* 2016;17(1):178-187.
8. Dong G, Li D, Ballerstedt S, Vandemeulebroecke M. A generalized analytic solution to the win ratio to analyze a composite endpoint considering the clinical importance order among components. *Pharmaceutical Statistics.* 2016;15(5):430-437.
9. Dong G, Qiu J, Wang D, Vandemeulebroecke M. The stratified win ratio. *Journal of Biopharmaceutical Statistics.* 2018;28(4):778-796.
10. Luo X, Qiu J, Bai S, Tian H. Weighted win loss approach for analyzing prioritized outcomes. *Statistics in Medicine.* 2017;36(15):2452-2465.
11. Gasparyan SB, Folkvaljon F, Bengtsson O, Buenconsejo J, Koch GG. Adjusted win ratio with stratification: Calculation methods and interpretation. *Statistical Methods in Medical Research.* 2021;30(2):580-611.
12. Verbeeck J, Ozenne B, Anderson WN. Evaluation of inferential methods for the net benefit and win ratio statistics. *Journal of Biopharmaceutical Statistics.* 2020;30(5):765-782.

13. Péron J, Buyse M, Ozenne B, Roche L, Roy P. An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. *Statistical Methods in Medical Research.* 2018;27(4):1230-1239.

14. Verbeeck J, Spitzer E, de Vries T, van Es GA, Anderson WN, Van Mieghem NM, Leon MB, Molenberghs G, Tijssen J. Generalized pairwise comparison methods to analyze (non)prioritized composite endpoints. *Statistics in Medicine.* 2019;38(30):5641-5656

15. Finkelstein DM, Schoenfeld DA. Graphing the win ratio and its components over time. *Statistics in Medicine.* 2019;38(1):53-61.

16. Dong G, Huang B, Chang Y-W, Seifu Y, Song J, Hoaglin DC. The win ratio: Impact of censoring and follow-up time and use with nonproportional hazards. Pharmaceutical Statistics. 2020;19(3):168–177

17. Dong G, Mao L, Huang B, Gamalo-Siebers M, Wang J, Yu G, Hoaglin DC. The inverse-probability-of-censoring weighting (IPCW) adjusted win ratio statistic: an unbiased estimator in the presence of independent censoring. *Journal of Biopharmaceutical Statistics.* 2020;30(5):882-899.

18. Dong G, Huang B, Wang D, Verbeeck J, Wang J, Hoaglin DC. Adjusting win statistics for dependent censoring. *Pharmaceutical Statistics.* 2021;20(3):440-450.

19. Peng L. The use of the win odds in the design of non-inferiority clinical trials. *Journal of Biopharmaceutical Statistics.* 2020;30(5):941-946.

20. Brunner E, Vandemeulebroecke M, Mütze T. Win odds: An adaptation of the win ratio to include ties. *Statistics in Medicine.* 2021;40(14):3367-3384.

21. Gasparyan SB, Kowalewski EK, Folkvaljon F, Bengtsson O, Buenconsejo J, Adler J, Koch GG. Power and sample size calculation for the win odds test: application to an ordinal endpoint in COVID-19 trials. *Journal of Biopharmaceutical Statistics.* 2021. doi:

10.1080/10543406.2021.1968893.

22. Song J, Verbeeck J, Hoaglin DC, Gamalo-Siebers M, Seifu Y, Huang B, Wang D, Cooner F, Dong G. The Win Odds: Statistical Inference and Regression. *Journal of Biopharmaceutical Statistics* (under revision). 2022.
23. Mao L, Wang T. A class of proportional win-fractions regression models for composite outcomes. *Biometrics.* 2021;77(4):1265-1275.
24. Mao L, Kim K, Miao X. Sample size formula for general win ratio analysis. *Biometrics.* 2021; doi: 10.1111/biom.13501.
25. Mao L, Kim K. Statistical Models for Composite Endpoints of Death and Nonfatal Events: A Review. Statistics in Biopharmaceutical Research. 2021;13(3): 260-269.
26. Yang S, Troendle J, Pak D, Leifer E. Event-specific win ratios for inference with terminal and non-terminal events. *Statistics in Medicine.* 2021; doi: 10.1002/sim.9266.
27. Voors AA, Angermann CE, Teerlink JR, Collins SP, Kosiborod M, Biegus J, Ferreira JP, Nassif ME, Psotka MA, Tromp J, Borleffs CJW, Ma C, Comin-Colet J, Fu M, Janssens SP, Kiss RG, Mentz RJ, Sakata Y, Schirmer H, Schou M, Schulze PC, Spinarova L, Volterrani M, Wranicz JK, Zeymer U, Zieroth S, Brueckmann M, Blatchford JP, Salsali A, Ponikowski P. The SGLT2 inhibitor empagliflozin in patients hospitalized for acute heart failure: a multinational randomized trial. Nature Medicine. 2022 Mar; 28(3):568-574. doi: 10.1038/s41591-021-01659-1
28. Pfeffer MA, Swedberg K, Granger CB, et al. Effects of candesartan on mortality and morbidity in patients with chronic heart failure: the CHARM-Overall programme. *Lancet.* 2003;362:759-766.
29. Cui Y, Huang B. WINS: The R WINS Package. 2022. https://cran.r-project.org/web/packages/WINS/index.html.

30. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*. 1947;18(1):50-60.

31. Lehmann, E. L. Elements of Large-Sample Theory. New York: Springer; 1999.

32. Verbeeck J, Deltuvaite-Thomas V, Berckmoes B, et al. Unbiasedness and efficiency of non-parametric and UMVUE estimators of the probabilistic index and related statistics. *Statistical Methods in Medical Research*. 2021;30(3):747-768.

33. Huang B, Kuan PF. Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point. *Pharmaceutical Statistics.* 2018;17(3):202-213.