Adjusted win ratio using the inverse probability of treatment weighting

Duolao Wang [1*], Sirui Zheng [1], Ying Cui [2], Nengjie He [1], Tao Chen [3], Bo Huang [4]

[1] Biostatistics Unit, Department of Clinical Sciences, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK

[2] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA

[3] Centre for Health Economics, University of York, York YO10 5DD, UK

[4] Pfizer Inc, Groton, CT, USA.

[*] Corresponding author: E-mail: Duolao.Wang@lstmed.ac.uk

Tel: +44 (0) 151 705 3301, Fax: +44 (0) 151 705 3370

Abstract:

The win ratio method has been increasingly applied in the design and analysis of clinical trials. However, the win ratio method is a univariate approach that does not allow for adjusting for baseline imbalances in covariates, although a stratified win ratio can be calculated when the number of strata is small. This paper proposes an adjusted win ratio to control for such imbalances by the inverse probability of treatment weighting (IPTW). We derive the adjusted win ratio with its variance and suggest three IPTW adjustments: IPTW-average treatment effect (IPTW-ATE), stabilized IPTW-ATE (SIPTW-ATE), and IPTW-average treatment effect in the treated (IPTW-ATT). The proposed adjusted methods are applied to analyse a composite outcome in the CHARM trial. The statistical properties of the methods are assessed through simulations. Results show that adjusted win ratio methods can correct the win ratio for covariate imbalances at baseline. Simulation results show that the three proposed adjusted win ratios have about the similar power to detect the treatment difference and have slightly lower power than the corresponding adjusted Cox models when the assumption of proportional hazards holds true but have consistently higher power than adjusted Cox models when the proportional hazard assumption is violated.

Key words: adjusted win ratio, inverse probability of treatment weighting (IPTW), propensity score, baseline imbalance, proportional hazard assumption

1. Introduction

The win ratio method was proposed by Pocock et al. for the analyses of composite endpoints (Pocock et al. 2012), for example, cardiovascular (CV) death and heart failure hospitalization in CV trials (Redfors et al. 2020). This method overcomes the problem in the conventional practice of analysing the time to occurrence of the first event, whichever component of the composite, which is often of lesser clinical importance. The win ratio method can also be used to analyse non-normal outcomes and was shown to have about the same power as the Mann–Whitney method (D. Wang and Pocock 2016). Furthermore, this method was demonstrated to provide a good alternative to the traditional survival analysis methods such as the log-rank test and Cox model in terms of power to detect the treatment difference for the situation of nonproportional hazards (Zheng et al. 2023; Dong, Huang, et al. 2020).

Since its introduction in 2012, the win ratio has been applied in the design and analysis of some high-profile clinical trials, such as the EMPULSE study of the SGLT2 inhibitor empagliflozin in patients hospitalized for acute heart failure (Voors et al. 2022) and the ACTION trial of therapeutic versus prophylactic anticoagulation for patients admitted to hospital with COVID-19 and elevated D-dimer concentration (Lopes et al. 2021). Redfors et al. summarised the principles behind the win ratio and provided insights into how to implement the win ratio in CV trial design and reporting, including determining trial size (Redfors et al. 2020). However, the win ratio method has a major limitation. When the win ratio method is used to analyse an outcome in a randomized clinical trial, it is assumed that two treatment groups are balanced on patients' characteristics at baseline. When two treatment groups are not balanced, it would be desirable to control for possible confounding factors. This limitation has restricted the applications of the win ratio method. Dong et al. proposed the stratified win ratio in a similar way to the Mantel-Haenszel stratified odds ratio, and derived a general form of its variance estimator with a plug-in of existing or potentially new variance/covariance estimators of the number of wins for the two treatment groups (Dong et al. 2018). Dong et al. proposed the inverse probability of censoring weighting (IPCW) adjusted win ratio statistic (i.e., the IPCW-adjusted win ratio statistic) to address censoring issues in time-to-event composite outcome analysis by giving different weights calculated from the baseline covariates to censored cases (Dong, Mao, et al. 2020). Mao and Wang introduced a regression method for the win ratio to control for confounders in the model (Mao and Wang 2021).

Hill et al. applied the win ratio method to analyse a ranked composite of death, heart transplantation, or any of 13 major complications involving infants (<1 year of age) undergoing heart surgery with cardiopulmonary bypass (Hill et al. 2022). Furthermore, Gasparyan et al. provided a unified theory of win odds estimation in the presence of stratification and adjustment for one numeric variable (Gasparyan et al. 2021). The common problem with the above adjusted win ratio methods is that they only allow for a small number of strata or covariates, which limits their applications.

In this paper, we propose an adjusted win ratio to address the imbalanced distributions in patients' characteristics at baseline in clinical trials by means of the inverse probability of treatment weighting (IPTW). IPTW involves three main steps: (1) to calculate the probability or propensity of an individual being assigned to one treatment arm (e.g., active treatment), given an individual's characteristics. This probability is also referred to as the propensity score; (2) to calculate the weight for each individual as the inverse of the probability of receiving an actual treatment; and (3) to calculate the adjusted treatment effect by applying the above weights in the statistical analysis such as Cox regression model. IPTW is a methodology that can be used to adjust for confounding and selection bias in observational studies (Hernán, Brumback, and Robins 2000) and adjust for baseline characteristics to calculate the adjusted treatment effect in randomized controlled trials (Williamson, Forbes, and White 2014; Shen, Li, and Li 2014; Zeng et al. 2021). Even if there is no confounding in a randomized clinical trial, the IPTW approach could still yield a gain in statistical efficiency by reducing variance (Williamson, Forbes, and White 2014). The adjusted win ratio by IPTW has two major advantages over other approaches that have been used to adjust the win ratio: (1) it can allow for many covariates and (2) as a weighted analysis, it is easy to interpret and implement.

This paper is structured as follows. In Section 2, we describe the IPTW adjusted win ratio statistic and its variance. We illustrate the IPTW adjusted win ratio approach by analysis of a composite outcome from the CHARM trial in Section 3. Some simulation results are presented to assess the statistical performances of three adjustment methods in Section 4. In Section 5, we summarise the key findings and conclude this research work with a discussion.

2. The IPTW adjusted win ratio

2.1 Three weighting schemes for IPTW

The propensity score is frequently used to adjust for confounding and selection bias in observational studies (Hernán, Brumback, and Robins 2000) and calculate adjusted treatment effects in clinical trials (Williamson, Forbes, and White 2014). It is defined as the probability of being in a group which is often estimated using a logistic regression model and sometimes using other binomial regression models such as the logistic LASSO regression model, generalized additive logistic model, and generalized boosted models (McCaffrey et al. 2013). For the win ratio analysis, each subject in the Treatment group is compared with every subject in the Control group. The two subjects in some of such pairwise comparisons may not be comparable with respect to key baseline characteristics. Therefore, in this paper, we apply the well-establizhed propensity score approach to adjust the win ratio. We consider three weighting schemes.

2.1.1 IPTW-ATE

The first one is inverse probability of treatment weighting (IPTW). We use p_i and p_j as the propensity score for the ith subject in the Treatment group and jth subject in the Control group, respectively, and then the IPTW can be expressed as:

$$\omega_i = \frac{1}{p_i} \text{ for } i\text{th subject in the Treatment group}$$

$$\omega_j = \frac{1}{p_j} \text{ for } j\text{th subject in the Control group}$$

Weighting the sample using these weights generates a synthetic sample in which observed baseline covariates are expected not to be confounded with treatment assignment. The use of these weights in the covariate-adjusted analysis in clinical trials allows one to estimate the average treatment effect (ATE). These weights are referred to as the conventional IPTW-ATE weights (Austin and Stuart 2015; Lunceford and Davidian 2004).

2.1.2 SIPTW-ATE

Expected IPTW-ATE weights are given by the proportion of subjects in a group as follows:

$$\bar{\omega} = \begin{cases} \dfrac{N_t}{N_t + N_c} & \text{for all subjects in the Treatment group} \\ \dfrac{N_c}{N_t + N_c} & \text{for all subjects in the Control group} \end{cases}$$

where N_t and N_c are the total number of subjects in the Treatment and Control groups, respectively.

If a subject has a propensity score close to 0, the resulting IPTW-ATE weight can be large. If a few observations have very large weights, the resulting IPTW-ATE estimator can have a large variance and may not be approximately normally distributed (Robins, Hernán, and Brumback 2000). To address this issue of large variances, Robins, Hernan, and Brumback suggested the IPTW-ATE with stabilized weights (Hernán, Brumback, and Robins 2000):

$$\omega_i^a = \frac{N_t}{N_t+N_c}\omega_i \text{ for } i\text{th subject in the Treatment group}$$

$$\omega_j^a = \frac{N_c}{N_t+N_c}\omega_j \text{ for } j\text{th subject in the Control group}$$

The stabilized IPTW-ATE (SIPTW-ATE) weights are computed by multiplying the IPTW-ATE weights by the marginal probability of receiving the given treatment.

2.1.3 IPTW-ATT

The third popular weighting scheme is defined as follow:

$$\omega_i^b = 1 \text{ for } i\text{th subject in the Treatment group}$$

$$\omega_j^b = \frac{p_i}{p_j} \text{ for } j\text{th subject in the Control group}$$

These weights can be used in the analysis of an outcome to estimate the average treatment effect by weighting the Control group only and allow one to estimate the average treatment effect in the treated (ATT) (Morgan and Todd 2008). We refer to these weights as IPTW-ATT weights. With these weights, subjects in the Treatment arm receive a weight of one, while those in the Control arm receive a weight of the odds of receiving the active treatment. Thus, the population in the Treatment arm serves as the reference population to which the treated and control populations are standardised.

2.2 The IPTW adjusted win ratio and its variance

Dong et al. proposed an IPCW approach to adjust win statistics (win ratio, win odds and the net benefit) for dependent censoring that can be predicted by baseline covariates and/or time-dependent covariates (producing the CovIPCW-adjusted win statistics) (Dong et al. 2021). They also showed that the CovIPCW-adjusted win statistics were unbiased estimators of treatment effect in the presence of dependent censoring. We use similar notations for IPTW adjusted win ratio.

Denote $\{\omega_i\}_{i=1}^{N_t}$ and $\{\omega_j\}_{j=1}^{N_c}$ as the weight for each subject. Let $N_t^T = \sum_{i=1}^{N_t} \omega_i$ and $N_c^T = \sum_{j=1}^{N_c} \omega_j$ denote the number of subjects in a pseudo-population defined based on the IPTW.

The IPTW adjusted win ratio is defined as

$$WR^T = \frac{P_t^T}{P_c^T}$$

where $P_t^T = \frac{n_t^T}{N_t^T N_c^T}$ and $P_c^T = \frac{n_c^T}{N_t^T N_c^T}$ are win proportions in the Treatment and Control groups, respectively with

$$n_t^T = \sum_{i=1}^{N_t} \sum_{j=1}^{N_c} K_{ij}^T \text{ and } n_c^T = \sum_{i=1}^{N_t} \sum_{j=1}^{N_c} L_{ij}^T.$$

The kernel functions are defined as

$$K_{ij}^T = \omega_i \omega_j K_{ij} \text{ and } L_{ij}^T = \omega_i \omega_j L_{ij}$$

respectively for K_{ij} and L_{ij} as defined in Dong et al. (Dong et al. 2021). The kernel function $K_{ij} = 1$ if the ith subject from the Treatment group wins against the jth subject from the Control group, and = 0 otherwise. Similarly, the kernel function $L_{ij} = 1$ if the jth subject from the Control group wins against the ith subject from the Treatment group, and = 0 otherwise. The proportion of ties is $P_{tie}^T = 1 - P_t^T - P_c^T$.

The logarithm of the win ratio is asymptotically normally distributed with the variance under the null hypothesis of equal win probabilities between two groups respectively as

$$\hat{\sigma}^2_{log(WR^T)} = \frac{\hat{\sigma}_t^2}{\hat{\theta}_t{}^2} + \frac{\hat{\sigma}_c^2}{\hat{\theta}_c{}^2} - \frac{2\hat{\sigma}_{tc}}{\hat{\theta}_t \hat{\theta}_c},$$

where $\hat{\theta}_t = \hat{\theta}_c = \frac{n_t^T + n_c^T}{2}$; $\hat{\sigma}_t^2$ and $\hat{\sigma}_c^2$ are the variances estimator for n_t^T and n_c^T, respectively; and $\hat{\sigma}_{tc}$ is their covariance estimator defined as

$$\hat{\sigma}_t^2 = \frac{N_c^T}{(N_c^T-1)}\sum_{i=1}^{N_t}\sum_{j=1}^{N_c}\sum_{j'=1,j'\neq j}^{N_c}\left[K_{ij}^T-\hat{\theta}_{K^T}\right]\left[K_{ij'}^T-\hat{\theta}_{K^T}\right] + \frac{N_t^T}{(N_t^T-1)}\sum_{j=1}^{N_c}\sum_{i=1}^{N_t}\sum_{i'=1,i'\neq i}^{N_t}\left[K_{ij}^T-\hat{\theta}_{K^T}\right]\left[K_{i'j}^T-\hat{\theta}_{K^T}\right],$$

$$\hat{\sigma}_c^2 = \frac{N_t^T}{(N_t^T-1)}\sum_{j=1}^{N_c}\sum_{i=1}^{N_t}\sum_{i'=1,i'\neq i}^{N_t}\left[L_{ij}^T-\hat{\theta}_{L^T}\right]\left[L_{i'j}^T-\hat{\theta}_{L^T}\right] + \frac{N_c^T}{(N_c^T-1)}\sum_{i=1}^{N_t}\sum_{j=1}^{N_c}\sum_{j'=1,j'\neq j}^{N_c}\left[L_{ij}^T-\hat{\theta}_{L^T}\right]\left[L_{ij'}^T-\hat{\theta}_{L^T}\right],$$

$$\hat{\sigma}_{tc} = \frac{N_c^T}{(N_c^T-1)}\sum_{i=1}^{N_t}\sum_{j=1}^{N_c}\sum_{j'=1,j'\neq j}^{N_c}\left[K_{ij}^T-\hat{\theta}_{K^T}\right]\left[L_{ij'}^T-\hat{\theta}_{L^T}\right] + \frac{N_t^T}{(N_t^T-1)}\sum_{j=1}^{N_c}\sum_{i=1}^{N_t}\sum_{i'=1,i'\neq i}^{N_t}\left[K_{ij}^T-\hat{\theta}_{K^T}\right]\left[L_{i'j}^T-\hat{\theta}_{L^T}\right].$$

In the above formula, $\hat{\theta}_{K^T} = \hat{\theta}_{L^T} = \frac{n_t^T+n_c^T}{2N_t^T N_c^T}$.

In this section, ω_i and ω_j represent any weights, and therefore apply to all three IPTW approaches.

3. Example

The CV trial of Candesartan in Heart Failure Assessment of Reduction in Mortality and Morbidity (CHARM) programme was a randomized, double-blind, controlled trial comparing candesartan with placebo in patients with chronic heart failure (Pfeffer et al. 2003). The primary outcome was a composite outcome of CV death or hospitalizations due to chronic heart failure. It is obvious that CV death is clinically more important than hospitalizations due to chronic heart failure and this order of clinical importance was used in our win ratio analysis of this composite outcome in this study. The traditional survival analysis methods such as the Cox proportional hazards model, analyse the composite outcome as the time to the first occurrence of CV death or hospitalization due to chronic heart failure, in which the more frequent outcome of hospitalizations dominates the

result. The win ratio method, on the other hand, takes into account the clinical importance of the composite elements and orders them accordingly.

In the adjusted win ratio analysis, we assessed the impact of all important predictors of CV death or hospitalizations due to chronic heart failure identified in a previous investigation on the treatment effect in the CHARM program (Pocock et al. 2006). These variables, in the order of strength of association with CV death or hospitalizations due to chronic heart failure, are listed in Table 1.

Table 1 presents the summary statistics of the baseline characteristics of patients by pre- and post-IPTW adjustment. To assess the homogeneity of treatment groups at baseline before and after IPTW adjustment, we calculated the standardised mean difference (SMD) in each of the baseline covariates as well as the sum of absolute SMD.

The SMD is commonly used as a measure of the balance in characteristics before and after propensity score (i.e., IPTW) matching (Austin 2011). It can be interpreted as a measurement of the mean difference between Treatment and Control groups in terms of units in the pooled deviation. It is a unit-free measurement and can therefore be compared across different groups or covariates (Andrade 2020).

For a continuous variable, let $\bar{X}_t$ and $\bar{X}_c$ be the sample mean of X in the Treatment and Control subjects, respectively, and let S_t^2 and S_c^2 be the sample variance of X in the Treatment and Control subjects.

$$SMD = \frac{(\bar{X}_t - \bar{X}_c)}{\sqrt{\frac{S_t^2 + S_c^2}{2}}}$$

Similarly for a dichotomous variable:

$$SMD = \frac{(\bar{P}_t - \bar{P}_c)}{\sqrt{\frac{\bar{P}_t(1-\bar{P}_t) + \bar{P}_c(1-\bar{P}_c)}{2}}}$$

Where $\bar{P}_t$ and $\bar{P}_c$ denote the proportion of the dichotomous variable in Treatment and Control subjects, respectively.

The sum of the absolute SMD is 0.34726, 0.00215, 0.00215 and 0.01413 for unadjusted, IPTW-ATE, SIPTW-ATE, and IPTW-ATT, respectively, suggesting that the three IPTW methods could effectively reduce the imbalances between treatment groups in the baseline characteristics and that both IPTW-ATE and SIPTW-ATE produced the smallest and almost identical total SMDs.

Table 2 displays the unadjusted and adjusted win ratio statistics from analyses of the composite outcome of CV death or hospitalizations due to chronic heart failure, the primary outcome in the CHARM trial. The IPTW-adjusted win ratios from the IPTW-ATE, SIPTW-ATE, and IPTW-ATT analyses of the primary outcome produced very similar estimates to the unadjusted win ratio in terms of win proportion, point and interval estimates of win ratio, and *P*-value. This is because two treatment arms were well balanced in the patient characteristics at baseline in this large, randomized study with a total of 7599 patients as shown in Table 1. Table 2 also presents the results of the Cox regression model analysis of the time from randomization to the first occurrence of the CV death or hospitalization due to chronic heart failure. Win ratio statistics from the unadjusted and adjusted methods are almost identical to those from unadjusted and adjusted Cox models in the point estimate, interval estimate of 1/HR, and *P*-value.

4. Simulation studies

To assess the performance of three adjusted IPTW methods to estimate the adjusted win ratios, we performed Monte Carlo simulation studies of survival time mimicking the CHARM dataset. We also applied the adjusted methods to analyse the simulated survival data using the Cox model and compared unadjusted and adjusted win ratio methods with the corresponding unadjusted and adjusted Cox models in terms of type I error and power to detect treatment differences in two survival distributions with various censoring mechanisms.

The simulations were conducted in the following steps:

Step 1. Simulate data for a setting with three baseline covariates (X_1, X_2, X_3). These covariates were simulated from independent standard normal distributions. Of these three covariates, X_1 and X_2 were related to treatment allocation, while X_2 and X_3 were related to the outcome. X_1, X_2 and X_3 were structured this way so that only X_2 was a confounding factor of treatment effect (it was related to both treatment allocation and the outcome of interest) whereas X_1 and X_3 were not confounding factors.

For the i^{th} subject, the probability of being in the Treatment group was determined from the following logistic model: $logit(p_i) = \alpha_{0A} + \alpha_1 x_{1i} + \alpha_2 x_{2i}$. The intercept of the treatment-allocation model (α_{0A}) was selected so that the proportion of subjects in the simulated sample that were treated was fixed at a desired proportion. We allowed the proportion of subjects treated in

the active arm to be 50%, corresponding to the ratio of sample size between the Treatment and Control groups being 1:1, and thus the value of α_{0A} was $logit(.50)$.

The other parameters in the simulations were estimated using the CHARM dataset. Of all baseline covariates listed in Table 1, the estimated largest and most significant effect on being allocated to the active treatment group was log(0.88) in log odds ratio per standard deviation for the covariate "bundle branch block". Therefore, the regression coefficient (the log odds of being allocated to the active treatment group) α_1 was set to log(0.88). For each subject, treatment status (1=Treatment, 0=Control) was generated from a Bernoulli distribution with subject-specific parameter p_i: $Z_i \sim B(1, p_i)$.

Step 2. For each subject, the linear predictor was defined as $LP_i = \alpha_2 x_{2i} + \alpha_3 x_{3i}$. α_2 and α_3 were estimated to be log(1.11) (from the effect of bundle branch block abnormal, which was related to both treatment allocation and the study outcome) and log(1.34) (the effect of age on the study outcome, the largest effect of all covariates), respectively.

Although different values of α_2 could be used in Step 1 and 2, we used the same α_2 value estimated in Step 2 in the simulations for the convenience of expression and modelling.

We then simulated survival time T without censoring from two Weibull distributions depending on Z_i for each arm.

$$T_A \sim Weibull(\eta_A e^{LP_i}, \theta_A), if\ Z_i = 1$$
$$T_B \sim Weibull(\eta_B e^{LP_i}, \theta_B), if\ Z_i = 0$$

Where η_A and η_B are the scale parameters, θ_A and θ_B are the shape parameters for Weibull distributions in arm A (Treatment) and B (Control), respectively. The event indicator δ was set as 1, indicating that every subject experienced an event.

To simulate survival time T_C with an expected proportion of censoring (pc), we generated the censoring time C from an exponential distribution with rate $\frac{1}{\eta}\frac{pc}{1-pc}$, where η is the scale parameter of Weibull distribution:

$$\eta = \begin{cases} \eta_A e^{LP_i} & if\ Z_i = 1 \\ \eta_B e^{LP_i} & if\ Z_i = 0 \end{cases}$$

Each individual subject was assigned to have a survival time T_C =min (T, C) and the event indicator $\delta = I[T \leq C]$, where $\delta = 1$ for the occurrence of event and $\delta = 0$ otherwise.

We simulated survival time from two Weibull distributions based on three scenarios as illustrated in Figure 1:

(1) Scenario 1: Two identical Weibull distributions with the same scale and shape parameters (scale η = 10.405, shape θ =0.786 in both treatment arms A and B) (Scenario 1 in Figure 1). The two parameters were estimated by fitting a Weibull distribution to the primary endpoint (time from randomization to CV death or hospitalization due to chronic heart failure) in the CHARM trial for the combined treatment A (candesartan) and B (placebo). This means that the hazard ratio (HR) between A and B is 1.

(2) Scenario 2: Two different Weibull distributions with different scales but the same shape: scale η_A = 10.678, shape θ_A =0.838 in arm A; scale η_B = 8.677, shape θ_B =0.838 in arm B (Scenario 2 in Figure 1). Those values were also estimated by fitting a Weibull distribution to the primary endpoint in each arm in the CHARM trial, with a restriction that the HR between treatment A and B is 0.84. This is a typical situation of hazard proportionality, a basic requirement for the Cox proportional hazard model.

(3) Scenario 3: Survival time in this scenario was generated from two different Weibull distributions; the following parameters were estimated from the CHARM trial by fitting the distributions to data from each of the arms: scale η_A = 10.678, shape θ_A =0.838 in arm A; scale η_B = 9.935, shape θ_B =0.747 in arm B (Scenario 3 in Figure 1).

For each of the above three scenarios, we set four proportions of censoring (0, 25%, 50% and 75%) and two sets of alpha values ($\alpha_1 = \alpha_2 = \alpha_3 = 0$ for two homogeneous treatment groups and α_1, α_2 and α_3 estimated from the CHARM database for two heterogeneous treatment groups).

Each simulated data set consisted of samples with the given sample size *N* roughly equally divided into two groups: N_A and N_B, where $N_A \approx N_B$, representing sample size in treatment A and B, respectively. Simulations are repeated 1000 times with sample size *N*=400. Within each of these simulated datasets, we examine the effects of the population heterogeneity and the proportion of censoring on the performance of the adjusted win ratios using the following measures over 1000 simulations: median of the sum of absolute values of standardised mean differences (SSMD) in simulated covariates (X_1, X_2, X_3), the median of proportion of wins by treatment arm, the median of unadjusted and adjusted win ratios and their 95%CIs, and *P*-value, the proportion of simulations with significant tests at 5% significance level for null hypothesis that there is no difference in the win proportion between the Treatment arm and Control arm (H_0: Win Ratio=1). In addition, we

also calculated the following statistics from the Cox model analysis of simulated survival data: the median of unadjusted and adjusted HRs between the Treatment and Control group (1/HR between Control and Treatment) and their 95%CIs, and *P*-value, the proportion of significant tests at 5% significance level for the null hypothesis that there is no difference in the hazard between Treatment and Control arm (H_0: HR=1).

The proportion of significant results ($P < 0.05$) gives an estimate of the type I error when there is no difference in treatment effect between two simulated treatment groups but gives an estimate of power when there is a difference.

The simulation results for the win ratio analyses and Cox model analyses of simulated survival times for two-arm randomized control trials are displayed in Table 3, 4 and 5 for Scenarios 1, 2 and 3, respectively. To further assess the impact of sample size on the power of win ratio analysis and Cox model analysis of survival time, we set the sample size as 100, 200, 300, 400, 500, 600, 700, and 800 and the censoring rate as 67.6% as estimated from the CHARM trial for Scenario 2 and Scenario 3. Figure 2 displays the power of the win ratio method and the Cox model under Scenario 2 and Scenario 3 by five analysis methods: unadjusted analysis for homogeneous populations, unadjusted, IPTW-ATE, SIPTW-ATE and IPTW-ATT analyses for heterogeneous populations.

The following observations can be made from the simulation results:

(1) The three IPTW methods could effectively reduce the imbalance between treatment groups in the baseline characteristics and both IPTW-ATE and SIPTW-ATE generate the smallest total SMDs (Table 3, 4 and 5).

(2) When there is no difference in treatment effect corresponding to $\eta_A = \eta_B$ for the scale parameters, $\theta_A = \theta_B$ for the shape parameters for Weibull distributions in the Scenario 1, the estimated unadjusted and adjusted win proportions are almost identical (Table 3). For example, in the two homogeneous treatment arms ($\alpha_1 = \alpha_2 = \alpha_3 = 0$), the win proportions in the Treatment and Control arms are about 50% for simulations with no censoring, about 39% for simulations with 25% of censoring, about 27% for simulations with 50% of censoring, and 14% for simulations with 75% of censoring. As a result, unadjusted and adjusted win ratios are close to 1, and the proportions of simulations with significant tests at 5% significance level (α error) are close to 5% for the unadjusted and adjusted win ratio methods.

(3) When there is a treatment effect corresponding to a proportional HR of 0.84 from two different Weibull distributions with different scales but the same shape (scale η_A=10.678, shape θ_A =0.838 in arm A; scale η_B=8.677, shape θ_B=0.838 in arm B) in the Scenario 2, the estimated unadjusted and adjusted win ratios and HRs are quite stable and close to the true value of 1.19 (1/HR=1/0.84) regardless of unadjusted and adjusted methods and censoring mechanisms (Table 4). For example, the median of win ratio is from 1.17 to 1.21.

(4) When there is a nonproportional hazard of treatment effect from two Weibull distributions with different scales and shapes (scale η_A=10.405, shape θ_A=0.786 in arm A; scale η_B=8.973, shape θ_B=0.786 in arm B) in the Scenario 3, the differences in the estimated win ratios and 1/HRs become larger and larger with the increase in the proportion of censoring (Table 5). For example, when the proportion of censoring is 75%, the median win ratio is 1.28, 1.28, 1.26, 1.26, and 1.26, compared with the median 1/HR of 1.20, 1.19, 1.18, 1.18, and 1.18 for unadjusted analyses for homogeneous and heterogeneous populations, IPTW-ATE, SIPTW-ATE, and IPTW-ATT, respectively.

(5) When there is a proportional hazard of treatment effect between two arms, unadjusted and adjusted win ratio methods have lower power than unadjusted and adjusted Cox models, particularly, if the proportion of censoring is low (Table 4, Figure 2-a and Figure 2-b). For example, power of win ratio vs Cox model at a 50% censoring rate is 18.6% vs 23.7%, 20.8% vs 24.7%, 18.2% vs 22.5%, 18.8% vs 22.5%, and 18.9% vs 21.9% for unadjusted analyses of homogenous and heterogeneous populations, IPTW-ATE, SIPTW-ATE, and IPTW-ATT, respectively. Figure 2-a shows that unadjusted win ratio for heterogeneous population seems to have the largest power whereas the other methods seem to have about the similar power. The results from Cox models seem insensitive to the three IPTW adjustment methods (Figure 2-b).

(6) When there is a nonproportional hazard of treatment effect, unadjusted and adjusted win ratio methods have considerably higher power than unadjusted and adjusted Cox models (Table 5, Figure 2-c and Figure 2-d). For example, power of win ratio vs Cox model at a 50% censoring rate is 18.4% vs 13.4%, 18.1% vs 10.8%, 14.9% vs 10.2%, 15.3% vs 10.1%, and 15.2% vs 9.1% for unadjusted analyses of homogenous and heterogeneous populations, IPTW-ATE, SIPTW-ATE, and IPTW-ATT, respectively.

5. Discussion

The win ratio method has been used for analysing different types of composite endpoints in clinical trials, including time-to-event outcomes, continuous measurements, and categorical data, to account for relative clinical importance of components (e.g., CV death, hospitalization, stroke, myocardial infarction). Examples of win ratio applications included the EMPULSE study of the SGLT2 inhibitor empagliflozin in patients hospitalized for acute heart failure (Voors et al. 2022) and the ACTION trial of therapeutic versus prophylactic anticoagulation for patients admitted to hospital with COVID-19 and elevated D-dimer concentration (Lopes et al. 2021), and STRESS trial on methylprednisolone for heart surgery in infants (Hill et al. 2022). Win ratio method has also been used to analyse non-normal continuous outcome. For example, in DAPA-HF trial on dapagliflozin in patients with heart failure and reduced ejection fraction, the win ratio method was used in the analysis of change in KCCQ total symptom score (McMurray et al. 2019).

In this study, we proposed an adjusted win ratio approach which advances the method of Dong et al. (2021) by adjusting for possible imbalances at baseline in terms of the characteristics of patients by the IPTW propensity score analysis. All covariates are factored into a single measurement (propensity score), which is then used to generate different weighting schemes for calculating the weight-adjusted win ratios. We propose three IPTW-adjusted win ratios to estimate: IPTW-average treatment effect (IPTW-ATE), stabilized IPTW-ATE (SIPTW-ATE), and IPTW-average treatment effect in the treated (IPTW-ATT). We provide the mathematical expressions for the weighted-adjusted win ratio statistic together with its variance. IPTW-adjusted methods based on propensity score can provide alternative methods to regression models to calculate the adjusted treatment effects in clinical trials, can increase the precision of the treatment effect estimate, and are particularly useful when non-convergence of covariate adjusted regression model occurs such as generalized linear models for calculating adjusted risk difference and ratio (Williamson, Forbes, and White 2014).

We applied the proposed adjusted win ratio methods to analyse the composite primary outcome in the CHARM trial (CV death or hospitalizations due to chronic heart failure). The CHARM results show that the three adjusted methods (IPTW-ATE, SIPTW-ATE, and IPTW-ATT) could considerably reduce the baseline imbalance in patient characteristics and produce win ratio statistics that are almost identical to those of 1/HR from unadjusted and adjusted Cox models.

We then conducted extensive simulations to assess the statistical properties of the proposed IPTW-adjusted win ratio methods by fitting different Weibull distributions for two treatment arms based on the parameters estimated from the primary outcome in the CHARM trial. We compared the unadjusted and adjusted win ratio methods with the unadjusted and adjusted Cox proportional hazard models in terms of point and interval estimates of treatment effect, Type I error, and power.

The most important finding is that the three adjusted methods could balance the covariate distributions in the simulation studies and adjust for baseline imbalances in covariates. It is observed that, when the two survival distributions are identical, α error is close to 5% for the unadjusted and adjusted win ratio methods. It is also observed that if there is a nonproportional hazard treatment effect, the statistical power is consistently higher for the unadjusted win ratio method than the unadjusted Cox model. This result is consistent with previous studies (Zheng et al. 2023; Dong, Huang, et al. 2020). The same is true for adjusted win ratio methods and Cox models. The three adjusted win ratio methods have about the similar power.

It is noted that the win ratio is very close to the reciprocal of the hazard ratio when the proportional hazard assumption is true. Such a reciprocal relationship between the two has been found in previous studies (Oakes 2016; Verbeeck et al. 2019; Finkelstein and Schoenfeld 2019).

To address the limitation of the win ratio method that it does not allow directly to adjust for covariates at baseline in clinical trials, some stratified approaches have been suggested but they can only allow for a limited number of strata (Dong et al. 2018; T. Wang and Mao 2022). Our proposed propensity score adjustment methods have some advantages over stratified methods. First, the adjusted win ratio methods can allow for as many as possible covariates at baseline so long as the propensity score (i.e., IPTW) can be estimated. Second, it is conceptually easy to understand and implement using the R-package WINS, publicly available on the Comprehensive R Archive Network (CRAN) and the computing time is less demanding. Third, the adjusted win ratio methods have a larger statistical power than the adjusted Cox proportional hazard models when the proportional hazard assumption does not hold.

The current study has some limitations. First, the simulation parameters (e.g., parameters for measuring population heterogeneity and for fitting Weibull distributions for the three different scenarios) were estimated using the CHARM trial which has a small extent of heterogeneity due to the well-conducted randomization and the very large sample size of 7599 patients. Second, for the convenience of simulations, we simulated a simple survival data instead of two different

survival data for two different components of a composite outcome, following a strategy used in previous simulation studies of a composite outcome (Dong, Mao, et al. 2020; Dong et al. 2018). Third, we assessed the performances of win ratio adjusted methods for a composite of time-to-event components. The statistical properties of other types of composite outcomes such as continuous outcomes, ordinal outcomes, categorical outcomes, or combinations of different outcomes need investigation too.

In summary, the adjusted win ratio approach can correct the win ratio from the baseline imbalances in covariates. The three adjusted win ratio methods (IPTW-ATE, SIPTW-ATE and IPTW-ATT) seem to have about the similar power to detect the treatment difference, have slightly smaller power than the corresponding adjusted Cox proportional hazard models for comparing composite outcomes in clinical trials when the proportional hazard assumption is true, but have much larger power than adjusted Cox proportional hazard models when the assumption of hazard proportionality is violated.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work featured in this article was initiated by the Trials Methodology Research Partnership funded by UK Medical Research Council (Project Reference: MR/S014357/1)

Data availability statement

The data that support the findings in this article are available on request from the corresponding author. The original data are not publicly available due to privacy or ethical restrictions.

References

Andrade, Chittaranjan. 2020. "Mean Difference, Standardized Mean Difference (SMD), and Their Use in Meta-Analysis: As Simple as It Gets." *The Journal of clinical psychiatry* 81. https://doi.org/10.4088/JCP.20f13681.

Austin, P. C. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behav Res* 46 (3): 399-424. https://doi.org/10.1080/00273171.2011.568786.

Austin, P. C., and E. A. Stuart. 2015. "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies." *Stat Med* 34 (28): 3661-79. https://doi.org/10.1002/sim.6607.

Dong, G., B. Huang, Y. W. Chang, Y. Seifu, J. Song, and D. C. Hoaglin. 2020. "The win ratio: Impact of censoring and follow-up time and use with nonproportional hazards." *Pharm Stat* 19 (3): 168-177. https://doi.org/10.1002/pst.1977.

Dong, G., B. Huang, D. Wang, J. Verbeeck, J. Wang, and D. C. Hoaglin. 2021. "Adjusting win statistics for dependent censoring." *Pharm Stat* 20 (3): 440-450. https://doi.org/10.1002/pst.2086.

Dong, G., L. Mao, B. Huang, M. Gamalo-Siebers, J. Wang, G. Yu, and D. C. Hoaglin. 2020. "The inverse-probability-of-censoring weighting (IPCW) adjusted win ratio statistic: an unbiased estimator in the presence of independent censoring." *J Biopharm Stat* 30 (5): 882-899. https://doi.org/10.1080/10543406.2020.1757692.

Dong, G., J. Qiu, D. Wang, and M. Vandemeulebroecke. 2018. "The stratified win ratio." *J Biopharm Stat* 28 (4): 778-796. https://doi.org/10.1080/10543406.2017.1397007.

Finkelstein, D. M., and D. A. Schoenfeld. 2019. "Graphing the Win Ratio and its components over time." *Stat Med* 38 (1): 53-61. https://doi.org/10.1002/sim.7895.

Gasparyan, S. B., F. Folkvaljon, O. Bengtsson, J. Buenconsejo, and G. G. Koch. 2021. "Adjusted win ratio with stratification: Calculation methods and interpretation." *Stat Methods Med Res* 30 (2): 580-611. https://doi.org/10.1177/0962280220942558.

Hernán, M. A., B. Brumback, and J. M. Robins. 2000. "Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men." *Epidemiology* 11 (5): 561-70. https://doi.org/10.1097/00001648-200009000-00012.

Hill, K. D., P. J. Kannankeril, J. P. Jacobs, H. S. Baldwin, M. L. Jacobs, S. M. O'Brien, D. P. Bichel, E. M. Graham, B. Blasiole, A. Resheidat, A. S. Husain, S. R. Kumar, J. L. Kirchner, D. S. Gallup, J. W. Turek, M. Bleiweis, B. Mettler, A. Benscoter, E. Wald, T. Karamlou, A. H. Van Bergen, D. Overman, P. Eghtesady, R. Butts, J. S. Kim, J. P. Scott, B. R. Anderson, M. F. Swartz, P. I. McConnell, D. F. Vener, and J. S. Li. 2022. "Methylprednisolone for Heart Surgery in Infants - A Randomized, Controlled Trial." *N Engl J Med* 387 (23): 2138-2149. https://doi.org/10.1056/NEJMoa2212667.

Lopes, R. D., E. Silva P. G. M. de Barros, R. H. M. Furtado, A. V. S. Macedo, B. Bronhara, L. P. Damiani, L. M. Barbosa, J. de Aveiro Morata, E. Ramacciotti, P. de Aquino Martins, A. L. de Oliveira, V. S. Nunes, L. E. F. Ritt, A. T. Rocha, L. Tramujas, S. V. Santos, D. R. A. Diaz, L. S. Viana, L. M. G. Melro, M. S. de Alcântara Chaud, E. L. Figueiredo, F. C. Neuenschwander, M. D. A. Dracoulakis, Rgsd Lima, V. C. de Souza Dantas, A. C. S. Fernandes, O. C. E. Gebara, M. E. Hernandes, D. A. R. Queiroz, V. C. Veiga, M. F. Canesin, L. M. de Faria, G. S. Feitosa-Filho, M. B. Gazzana, I. L. Liporace, A. de Oliveira Twardowsky, L. N. Maia, F. R. Machado, A. de Matos Soeiro, G. E. Conceição-Souza, L. Armaganijan, P. O. Guimarães, R. G. Rosa, L. C. P. Azevedo, J. H. Alexander, A. Avezum, A. B. Cavalcanti, and O. Berwanger. 2021. "Therapeutic versus prophylactic anticoagulation for patients admitted to hospital with COVID-19 and elevated D-dimer concentration (ACTION): an open-label, multicentre, randomised, controlled trial." *Lancet* 397 (10291): 2253-2263. https://doi.org/10.1016/s0140-6736(21)01203-4.

Lunceford, Jared K., and Marie Davidian. 2004. "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study." *Statistics in Medicine* 23 (19): 2937-2960. https://doi.org/10.1002/sim.1903. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1903.

Mao, L., and T. Wang. 2021. "A class of proportional win-fractions regression models for composite outcomes." *Biometrics* 77 (4): 1265-1275. https://doi.org/10.1111/biom.13382.

McCaffrey, D. F., B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette. 2013. "A tutorial on propensity score estimation for multiple treatments using generalized boosted models." *Stat Med* 32 (19): 3388-414. https://doi.org/10.1002/sim.5753. https://www.ncbi.nlm.nih.gov/pubmed/23508673.

McMurray, J. J. V., S. D. Solomon, S. E. Inzucchi, L. Køber, M. N. Kosiborod, F. A. Martinez, P. Ponikowski, M. S. Sabatine, I. S. Anand, J. Bělohlávek, M. Böhm, C. E. Chiang, V. K. Chopra, R. A. de Boer, A. S. Desai, M. Diez, J. Drozdz, A. Dukát, J. Ge, J. G. Howlett, T. Katova, M. Kitakaze, C. E. A. Ljungman, B. Merkely, J. C. Nicolau, E. O'Meara, M. C. Petrie, P. N. Vinh, M. Schou, S. Tereshchenko, S. Verma, C. Held, D. L. DeMets, K. F. Docherty, P. S. Jhund, O. Bengtsson, M. Sjöstrand, and A. M. Langkilde. 2019. "Dapagliflozin in Patients with Heart Failure and Reduced Ejection Fraction." *N Engl J Med* 381 (21): 1995-2008. https://doi.org/10.1056/NEJMoa1911303.

Morgan, SL, and JL. Todd. 2008. "A diagnostic routine for the detection of consequential heterogeneity of causal effects." *Sociological Methodology* 38: 231–281.

Oakes, D. 2016. "On the win-ratio statistic in clinical trials with multiple types of event." *Biometrika* 103 (3): 742-745.

Pfeffer, M. A., K. Swedberg, C. B. Granger, P. Held, J. J. McMurray, E. L. Michelson, B. Olofsson, J. Ostergren, S. Yusuf, and S. Pocock. 2003. "Effects of candesartan on mortality and morbidity in patients with chronic heart failure: the CHARM-Overall programme." *Lancet* 362 (9386): 759-66. https://doi.org/10.1016/s0140-6736(03)14282-1.

Pocock, S. J., C. A. Ariti, T. J. Collier, and D. Wang. 2012. "The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities." *Eur Heart J* 33 (2): 176-82. https://doi.org/10.1093/eurheartj/ehr352.

Pocock, S. J., D. Wang, M. A. Pfeffer, S. Yusuf, J. J. McMurray, K. B. Swedberg, J. Ostergren, E. L. Michelson, K. S. Pieper, and C. B. Granger. 2006. "Predictors of mortality and morbidity in patients with chronic heart failure." *Eur Heart J* 27 (1): 65-75. https://doi.org/10.1093/eurheartj/ehi555.

Redfors, B., J. Gregson, A. Crowley, T. McAndrew, O. Ben-Yehuda, G. W. Stone, and S. J. Pocock. 2020. "The win ratio approach for composite endpoints: practical guidance based on previous experience." *Eur Heart J* 41 (46): 4391-4399. https://doi.org/10.1093/eurheartj/ehaa665.

Robins, J. M., M. A. Hernán, and B. Brumback. 2000. "Marginal structural models and causal inference in epidemiology." *Epidemiology* 11 (5): 550-60. https://doi.org/10.1097/00001648-200009000-00011.

Shen, C., X. Li, and L. Li. 2014. "Inverse probability weighting for covariate adjustment in randomized studies." *Stat Med* 33 (4): 555-68. https://doi.org/10.1002/sim.5969. https://www.ncbi.nlm.nih.gov/pubmed/24038458.

Verbeeck, J., E. Spitzer, T. de Vries, G. A. van Es, W. N. Anderson, N. M. Van Mieghem, M. B. Leon, G. Molenberghs, and J. Tijssen. 2019. "Generalized pairwise comparison methods to analyze (non)prioritized composite endpoints." *Stat Med* 38 (30): 5641-5656. https://doi.org/10.1002/sim.8388.

Voors, A. A., C. E. Angermann, J. R. Teerlink, S. P. Collins, M. Kosiborod, J. Biegus, J. P. Ferreira, M. E. Nassif, M. A. Psotka, J. Tromp, C. J. W. Borleffs, C. Ma, J. Comin-Colet, M. Fu, S. P. Janssens, R. G. Kiss, R. J. Mentz, Y. Sakata, H. Schirmer, M. Schou, P. C. Schulze, L. Spinarova, M. Volterrani, J. K. Wranicz, U. Zeymer, S. Zieroth, M. Brueckmann, J. P. Blatchford, A. Salsali, and P. Ponikowski. 2022. "The SGLT2 inhibitor empagliflozin in patients hospitalized for acute heart failure: a multinational randomized trial." *Nat Med* 28 (3): 568-574. https://doi.org/10.1038/s41591-021-01659-1.

Wang, D., and S. Pocock. 2016. "A win ratio approach to comparing continuous non-normal outcomes in clinical trials." *Pharm Stat* 15 (3): 238-45. https://doi.org/10.1002/pst.1743.

Wang, T., and L. Mao. 2022. "Stratified proportional win-fractions regression analysis." *Stat Med* 41 (26): 5305-5318. https://doi.org/10.1002/sim.9570.

Williamson, E. J., A. Forbes, and I. R. White. 2014. "Variance reduction in randomised trials by inverse probability weighting using the propensity score." *Stat Med* 33 (5): 721-37. https://doi.org/10.1002/sim.5991. https://www.ncbi.nlm.nih.gov/pubmed/24114884.

Zeng, S., F. Li, R. Wang, and F. Li. 2021. "Propensity score weighting for covariate adjustment in randomized clinical trials." *Stat Med* 40 (4): 842-858. https://doi.org/10.1002/sim.8805. https://www.ncbi.nlm.nih.gov/pubmed/33174296.

Zheng, S., D. Wang, J. Qiu, T. Chen, and M. Gamalo. 2023. "A win ratio approach for comparing crossing survival curves in clinical trials." *J Biopharm Stat*: 1-14. https://doi.org/10.1080/10543406.2023.2170393. https://www.ncbi.nlm.nih.gov/pubmed/36749067.

Table 1: Baseline characteristics of patients by pre- and post-IPTW adjustment

Characteristics	Pre-IPTW adjustment*			Post-IPTW adjustment SMD		
	Candesartan (n=3803)	Placebo (n=3796)	SMD	IPTW-ATE	SIPTW-ATE	IPTW-ATT
Age (per 10 years over age 60)	6.78 (0.75)	6.79 (0.77)	-0.01240	-0.00009	-0.00009	0.00009
Diabetes mellitus						
Insulin treated	353 (9%)	354 (9%)	-0.00150	0.00001	0.00001	-0.00063
Other	735 (19%)	721 (19%)	0.00847	0.00014	0.00014	0.00001
Ejection fraction (per 5% decrease below 45)	-7.04 (1.93)	-7.04 (1.91)	0.00018	0.00003	0.00003	0.00012
Prior Hospitalization						
Prior CHF hosp within 6 months	1398 (37%)	1397 (37%)	-0.00086	0.00004	0.00004	-0.00085
Prior CHF hosp but not within 6 months	1327 (35%)	1304 (34%)	0.01140	-0.00002	-0.00002	0.00060
Cardiomegaly	814 (21%)	853 (22%)	-0.02580	0.00002	0.00002	0.00002
Diagnosis of CHF over 2 years ago	1978 (52%)	1903 (50%)	0.03760	0.00003	0.00003	-0.00005
NYHA						
III	1977 (52%)	2008 (53%)	-0.01830	-0.00016	-0.00016	-0.00102
IV	96 (3%)	102 (3%)	-0.01020	-0.00020	-0.00020	0.00042
DBP (per 10 mmHg decrease)	-7.66 (1.09)	-7.67 (1.06)	0.01470	-0.00004	-0.00004	-0.00018
Bundle branch block	865 (23%)	946 (25%)	-0.05110	-0.00012	-0.00012	0.00081
Heart rate (per 10 beats/min)	7.3 (1.33)	7.28 (1.28)	0.01520	0.00010	0.00010	-0.00040
Dependent edema	932 (25%)	922 (24%)	0.00508	-0.00011	-0.00011	-0.00085
Pulmonary crackles	606 (16%)	626 (16%)	-0.01510	-0.00014	-0.00014	-0.00066
Rest dyspnoea	1096 (29%)	1099 (29%)	-0.00292	0.00006	0.00006	-0.00106
Female	1186 (31%)	1214 (32%)	-0.01710	-0.00016	-0.00016	0.00095
Atrial fibrillation	1039 (27%)	1044 (28%)	-0.00408	-0.00009	-0.00009	-0.00145
BMI (per 1kg/m2 decrease below 27.5)	-25.83 (2.36)	-25.84 (2.39)	0.00537	-0.00002	-0.00002	0.00028
Mitral regurgitation	42 (1%)	39 (1%)	0.00750	-0.00028	-0.00028	-0.00052
Previous myocardial infarction	2024 (53%)	1980 (52%)	0.02130	-0.00013	-0.00013	0.00226
Pulmonary edema	90 (2%)	121 (3%)	-0.05000	-0.00017	-0.00017	0.00041
Current smoker	565 (15%)	549 (14%)	0.01110	0.00000	0.00000	-0.00050
Sum of absolute SMDs			0.34726	0.00215	0.00215	0.01413

SMD: standardised mean difference. IPTW: inverse-probability-of-treatment weighting. SIPTW: stabilized inverse-probability-of-treatment weighting. ATE: average treatment effect. ATT: average treatment effect in the treated. CHF: congestive heart failure. NYHA: New York Heart Association. DBP: diastolic blood pressure. BMI: Body Mass Index.

* Mean (standard deviation) is presented for continuous variables and number (%) for categorical variables.

Table 2: Win ratio statistics from analyses of composite outcome of cardiovascular death or hospitalizations due to chronic heart failure in the CHARM trial

Method	Win ratio analysis				Cox model analysis	
	Win proportion (%)					
	Candesartan	Placebo	WR (95% CI)	*P* Value	1/HR (95% CI) *	*P* Value
Unadjusted	28.1	23.6	1.19 (1.10, 1.29)	< 0.001	1.19 (1.10, 1.29)	< 0.001
IPTW-ATE	27.9	23.6	1.18 (1.09, 1.28)	< 0.001	1.18 (1.09, 1.28)	< 0.001
SIPTW-ATE	27.9	23.6	1.18 (1.09, 1.28)	< 0.001	1.18 (1.09, 1.28)	< 0.001
IPTW-ATT	27.8	23.6	1.18 (1.09, 1.28)	< 0.001	1.18 (1.09, 1.28)	< 0.001

HR: hazard ratio. WR: win ratio. GMR: geometric mean ratio. IPTW: inverse-probability-of-treatment weighting. SIPTW: stabilized inverse-probability-of-treatment weighting. ATE: average treatment effect. ATT: average treatment effect in the treated.
*HR was estimated from Cox model.

Table 3: The summary statistics of win ratio and the type I errors when there is no treatment difference in the hazard of time-to-event outcome. Scenario 1: Two identical Weibull distributions with same scale and shape (scale η=10.405, shape θ =0.786 in both treatment arms A and B). N=400.

% of Censoring	Population	Method	SSMD	Median win proportion (%)		Median of point estimate (95%CI)		Proportion of simulations with $P < 0.05$	
				Active	Control	Win ratio	Cox model (1/HR)	Win ratio	Cox model
0	Homogeneous	Unadjusted	0	50.0%	50.0%	1.00 (0.80, 1.26)	1.00 (0.82, 1.22)	4.7%	4.7%
	Heterogeneous	Unadjusted	0.32125	50.2%	49.8%	1.01 (0.81, 1.26)	1.00 (0.82, 1.22)	4.6%	5.8%
		IPTW-ATE	0.00516	50.1%	49.9%	1.00 (0.80, 1.26)	1.00 (0.82, 1.22)	4.3%	5.0%
		SIPTW-ATE	0.00516	50.1%	49.9%	1.00 (0.80, 1.26)	1.00 (0.82, 1.22)	4.1%	5.0%
		IPTW-ATT	0.02798	50.1%	49.9%	1.00 (0.80, 1.26)	1.00 (0.82, 1.22)	4.3%	5.0%
25	Homogeneous	Unadjusted	0	38.6%	38.9%	0.99 (0.77, 1.29)	0.99 (0.79, 1.24)	6.1%	5.2%
	Heterogeneous	Unadjusted	0.33045	38.9%	38.5%	1.01 (0.77, 1.30)	1.01 (0.80, 1.27)	5.1%	5.3%
		IPTW-ATE	0.00517	38.7%	38.7%	1.00 (0.77, 1.29)	1.00 (0.80, 1.26)	5.3%	4.9%
		SIPTW-ATE	0.00517	38.7%	38.7%	1.00 (0.77, 1.30)	1.00 (0.80, 1.26)	5.2%	4.9%
		IPTW-ATT	0.02848	38.7%	38.6%	1.00 (0.77, 1.30)	1.00 (0.80, 1.26)	5.1%	5.7%
50	Homogeneous	Unadjusted	0	27.0%	26.8%	1.01 (0.74, 1.38)	1.01 (0.76, 1.33)	3.4%	3.8%
	Heterogeneous	Unadjusted	0.32205	27.2%	27.2%	1.00 (0.73, 1.36)	1.00 (0.76, 1.32)	3.8%	3.7%
		IPTW-ATE	0.00472	27.1%	27.3%	0.99 (0.72, 1.36)	0.99 (0.75, 1.32)	3.9%	3.3%
		SIPTW-ATE	0.00472	27.1%	27.3%	0.99 (0.72, 1.36)	0.99 (0.75, 1.31)	3.7%	3.3%
		IPTW-ATT	0.02689	27.1%	27.2%	0.99 (0.72, 1.36)	0.99 (0.75, 1.31)	3.8%	3.5%
75	Homogeneous	Unadjusted	0	14.0%	13.8%	1.01 (0.65, 1.57)	1.00 (0.67, 1.49)	5.0%	5.1%
	Heterogeneous	Unadjusted	0.33428	14.3%	14.1%	1.01 (0.65, 1.56)	1.01 (0.68, 1.49)	6.3%	5.6%
		IPTW-ATE	0.00552	14.2%	14.2%	1.00 (0.64, 1.56)	1.00 (0.67, 1.48)	6.6%	5.6%
		SIPTW-ATE	0.00552	14.2%	14.2%	1.00 (0.64, 1.56)	1.00 (0.67, 1.48)	6.8%	5.6%
		IPTW-ATT	0.02846	14.2%	14.2%	1.00 (0.64, 1.56)	0.99 (0.67, 1.48)	6.0%	5.6%

SSMD: sum of absolute standardised mean difference. IPTW: inverse-probability-of-treatment weighting. SIPTW: stabilized inverse-probability-of-treatment weighting. ATE: average treatment effect. ATT: average treatment effect in the treated.

Homogeneous population (α_1= α_2= α_3= 0); Heterogeneous population (α_1= log(0.88), α_2= log(1.11), α_3= log(1.34)).

Table 4: The summary statistics of win ratio and powers when there is a treatment difference in the hazard of time-to-event outcome. Scenario 2: Two different Weibull distributions with the same shape parameter (scale η_A =10.678, shape θ_A =0.838 in arm A; scale η_B =8.677, shape θ_B=0.838 in arm B). Proportional hazard ratio=0.84. N=400.

% of Censoring	Population	Method	SSMD	Median win proportion (%)		Median of point estimate (95%CI)		Proportion of simulations with $P < 0.05$	
				Active	Control	Win ratio	Cox model (1/HR)	Win ratio	Cox model
0	Homogeneous	Unadjusted	0	54.5%	45.5%	1.20 (0.96, 1.51)	1.20 (0.98, 1.46)	34.1%	42.9%
	Heterogeneous	Unadjusted	0.32520	54.5%	45.5%	1.20 (0.95, 1.50)	1.19 (0.98, 1.45)	33.6%	40.6%
		IPTW-ATE	0.00521	54.2%	45.8%	1.19 (0.94, 1.50)	1.18 (0.97, 1.44)	30.6%	38.5%
		SIPTW-ATE	0.00521	54.2%	45.8%	1.19 (0.94, 1.49)	1.18 (0.97, 1.44)	31.7%	38.5%
		IPTW-ATT	0.02717	54.2%	45.8%	1.19 (0.94, 1.49)	1.18 (0.96, 1.44)	30.6%	37.2%
25	Homogeneous	Unadjusted	0	41.7%	35.2%	1.18 (0.91, 1.54)	1.18 (0.94, 1.48)	21.8%	28.9%
	Heterogeneous	Unadjusted	0.33950	42.0%	34.8%	1.21 (0.93, 1.56)	1.20 (0.95, 1.51)	28.8%	33.5%
		IPTW-ATE	0.00512	41.7%	35.0%	1.19 (0.91, 1.56)	1.19 (0.94, 1.50	25.2%	31.5%
		SIPTW-ATE	0.00512	41.7%	35.0%	1.19 (0.92, 1.55)	1.19 (0.94, 1.51	26.5%	31.5%
		IPTW-ATT	0.02829	41.7%	35.0%	1.19 (0.92, 1.56)	1.19 (0.94, 1.52	25.3%	30.8%
50	Homogeneous	Unadjusted	0	28.5%	23.9%	1.19 (0.86, 1.63)	1.19 (0.90, 1.58)	18.6%	23.7%
	Heterogeneous	Unadjusted	0.32160	28.7%	24.0%	1.20 (0.87, 1.65)	1.19 (0.90, 1.58)	20.8%	24.7%
		IPTW-ATE	0.00525	28.7%	24.2%	1.19 (0.86, 1.64)	1.18 (0.89, 1.57)	18.2%	22.5%
		SIPTW-ATE	0.00525	28.7%	24.2%	1.19 (0.86, 1.64)	1.18 (0.89, 1.57)	18.8%	22.5%
		IPTW-ATT	0.02833	28.6%	24.1%	1.19 (0.86, 1.64)	1.19 (0.89, 1.58)	18.9%	21.9%
75	Homogeneous	Unadjusted	0	13.8%	11.8%	1.17 (0.74, 1.84)	1.18 (0.78, 1.78)	12.7%	12.8%
	Heterogeneous	Unadjusted	0.32565	14.2%	11.8%	1.21 (0.77, 1.90)	1.19 (0.80, 1.79)	11.5%	12.7%
		IPTW-ATE	0.00507	14.2%	11.9%	1.19 (0.75, 1.89)	1.19 (0.79, 1.78)	10.8%	12.0%
		SIPTW-ATE	0.00507	14.2%	11.9%	1.19 (0.75, 1.89)	1.18 (0.79, 1.78)	11.1%	12.0%
		IPTW-ATT	0.02757	14.2%	11.8%	1.20 (0.75, 1.90)	1.18 (0.79, 1.77)	10.8%	11.6%

SSMD: sum of absolute standardised mean difference. IPTW: inverse-probability-of-treatment weighting. SIPTW: stabilized inverse-probability-of-treatment weighting. ATE: average treatment effect. ATT: average treatment effect in the treated.

Homogeneous population (α_1= α_2= α_3= 0); Heterogeneous population (α_1=log(0.88), α_2=log(1.11), α_3= log(1.34)).

Table 5: The summary statistics of win ratio and powers when there is a treatment difference in the hazard of time-to-event outcome. Scenario 3: Nonproportional hazard ratio (Two different Weibull distributions (scale η_A =10.678, shape θ_A =0.838 in arm A; scale η_B =9.935, shape θ_B=0.747 in arm B). N=400.

% of Censoring	Population	Method	SSMD	Median win proportion (%)		Median of point estimate (95%CI)		Proportion of simulations with $P < 0.05$	
				Active	Control	Win ratio	Cox model (1/HR)	Win ratio	Cox model
0	Homogeneous	Unadjusted	0	52.0%	48.0%	1.08 (0.86, 1.36)	1.01 (0.83, 1.23)	11.0%	5.8%
	Heterogeneous	Unadjusted	0.33120	52.4%	47.6%	1.10 (0.88, 1.38)	1.02 (0.84, 1.24)	12.9%	5.8%
		IPTW-ATE	0.00535	52.2%	47.8%	1.09 (0.87, 1.38)	1.01 (0.83, 1.23)	10.8%	5.1%
		SIPTW-ATE	0.00535	52.2%	47.8%	1.09 (0.87, 1.37)	1.01 (0.83, 1.23)	11.6%	5.2%
		IPTW-ATT	0.02801	52.2%	47.8%	1.09 (0.87, 1.38)	1.01 (0.83, 1.24)	11.9%	5.6%
25	Homogeneous	Unadjusted	0	41.0%	36.5%	1.13 (0.87, 1.46)	1.05 (0.84, 1.32)	14.6%	8.4%
	Heterogeneous	Unadjusted	0.33040	41.3%	36.1%	1.15 (0.89, 1.49)	1.07 (0.85, 1.34)	15.2%	7.8%
		IPTW-ATE	0.00532	41.1%	36.3%	1.14 (0.87, 1.48)	1.06 (0.84, 1.33)	13.2%	6.8%
		SIPTW-ATE	0.00532	41.1%	36.3%	1.14 (0.87, 1.48)	1.06 (0.84, 1.33)	14.0%	6.8%
		IPTW-ATT	0.02836	41.1%	36.2%	1.13 (0.87, 1.48)	1.06 (0.84, 1.33)	13.4%	6.1%
50	Homogeneous	Unadjusted	0	29.3%	24.7%	1.18 (0.86, 1.62)	1.11 (0.84, 1.47)	18.4%	13.4%
	Heterogeneous	Unadjusted	0.32347	29.3%	24.9%	1.19 (0.86, 1.62)	1.12 (0.84, 1.47)	18.1%	10.8%
		IPTW-ATE	0.00509	29.2%	24.9%	1.17 (0.85, 1.62)	1.10 (0.83, 1.46)	14.9%	10.2%
		SIPTW-ATE	0.00509	29.2%	24.9%	1.17 (0.86, 1.61)	1.10 (0.83, 1.46)	15.3%	10.1%
		IPTW-ATT	0.02838	29.1%	24.9%	1.18 (0.86, 1.62)	1.10 (0.83, 1.47)	15.2%	9.1%
75	Homogeneous	Unadjusted	0	15.5%	12.1%	1.28 (0.82, 2.00)	1.20 (0.80, 1.78)	21.9%	15.4%
	Heterogeneous	Unadjusted	0.33525	15.6%	12.2%	1.28 (0.82, 2.00)	1.19 (0.80, 1.78)	19.6%	14.7%
		IPTW-ATE	0.00517	15.6%	12.2%	1.26 (0.81, 1.98)	1.18 (0.79, 1.76)	17.9%	14.3%
		SIPTW-ATE	0.00517	15.6%	12.2%	1.26 (0.81, 1.97)	1.18 (0.79, 1.76)	18.2%	14.3%
		IPTW-ATT	0.02906	15.5%	12.2%	1.26 (0.80, 1.99)	1.18 (0.79, 1.77)	17.7%	14.0%

SSMD: sum of absolute standardised mean difference. IPTW: inverse-probability-of-treatment weighting. SIPTW: stabilized inverse-probability-of-treatment weighting. ATE: average treatment effect. ATT: average treatment effect in the treated.

Homogeneous population (α_1= α_2= α_3= 0); Heterogeneous population (α_1=log(0.88), α_2=log(1.11), α_3=log(1.34)).

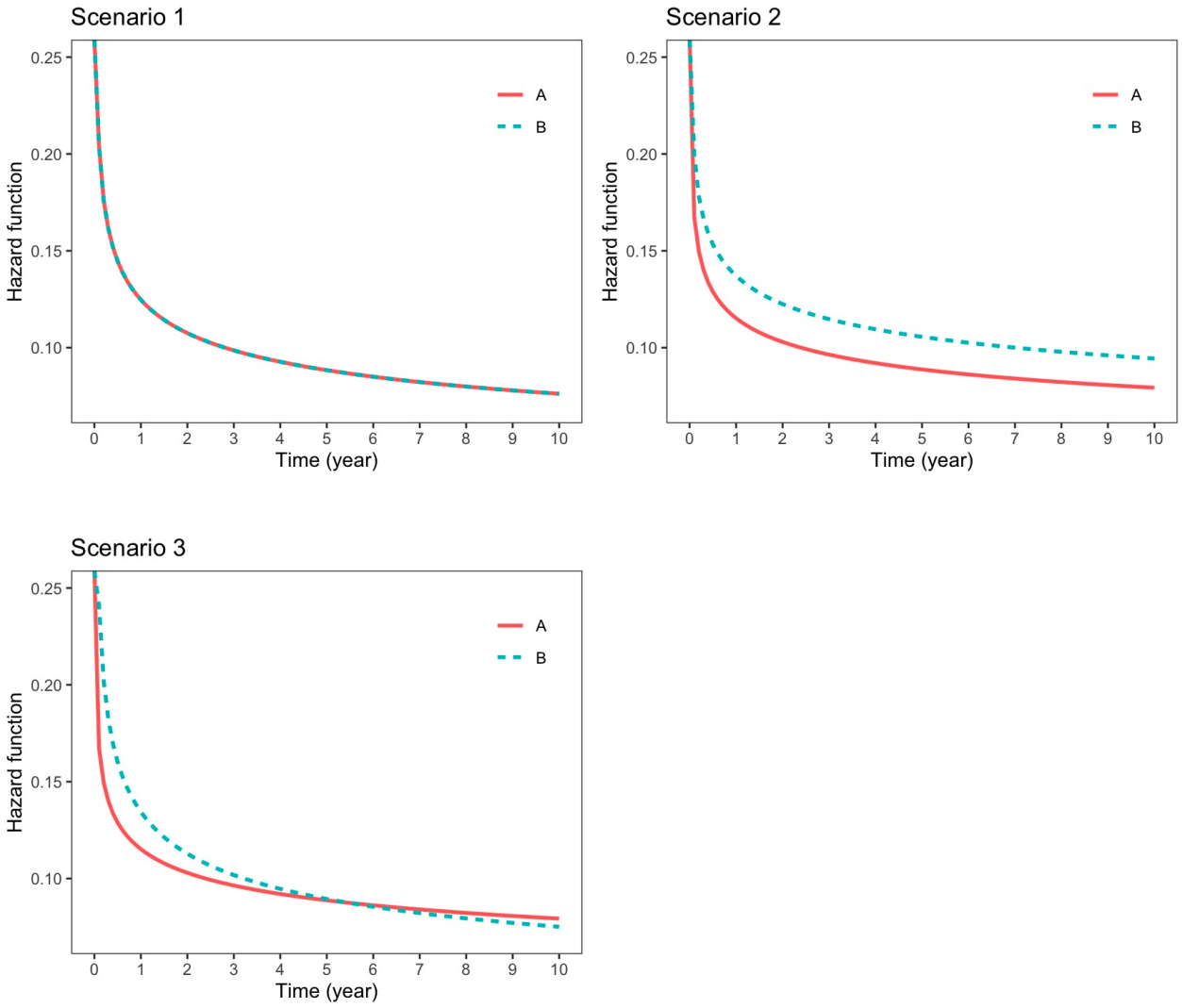

Figure 1: Hazard functions for three simulation scenarios.

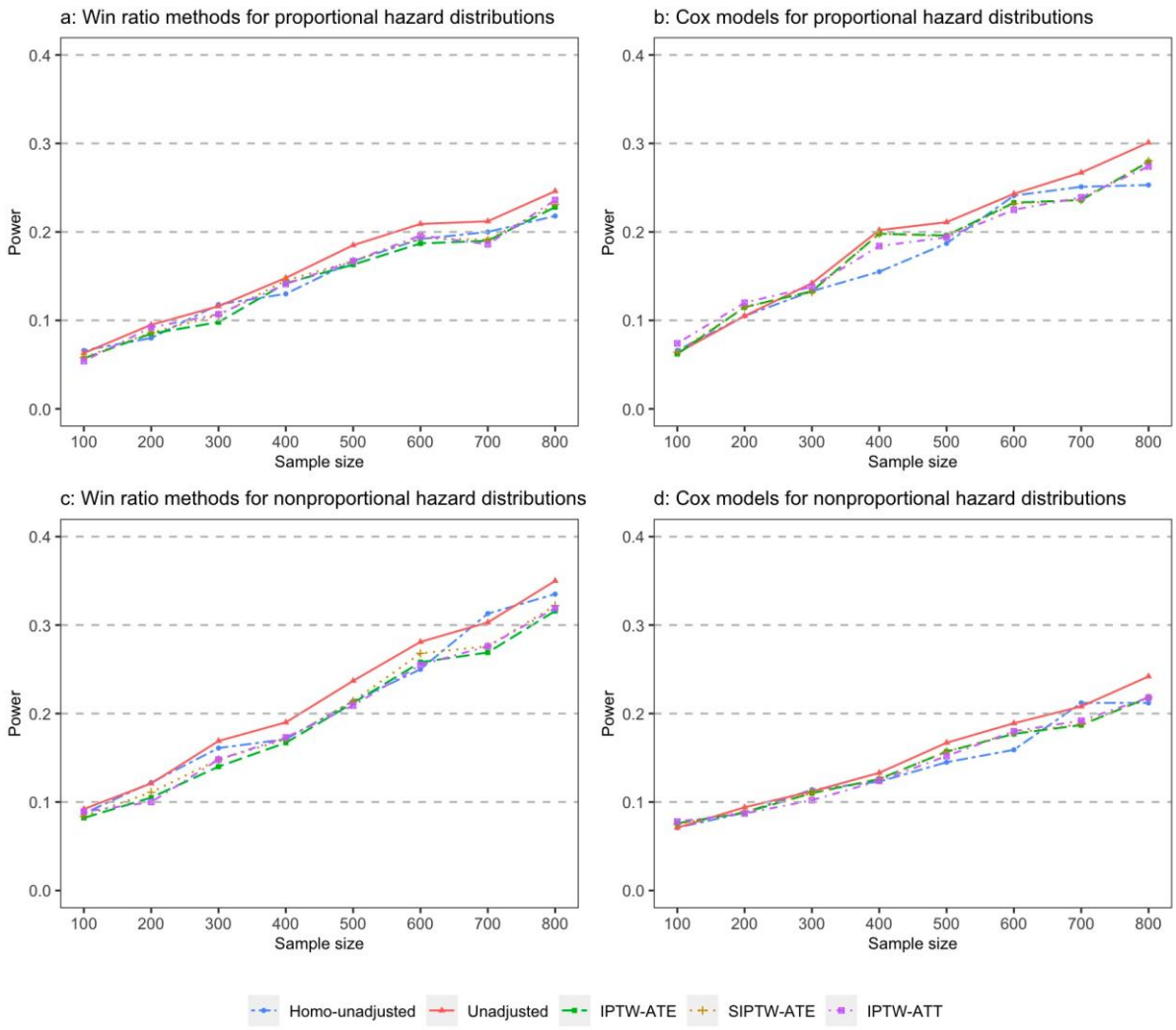

Figure 2: Comparison of powers by different methods for two simulation scenarios (proportional vs nonproportional hazard distributions).