A win ratio approach for comparing crossing survival curves in clinical trials

Sirui Zheng[1], Duolao Wang[1*], Junshan Qiu[2],
Tao Chen[1], Margaret Gamalo[3]

[1] Global Health Trials Unit, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK

[2] Division of Biometrics I, Office of Biostatistics, Office of Translational Science, CDER, FDA, US

[3] Global Biometrics & Data Management, Pfizer Innovative Health, PA 19426, US

*** Corresponding author**

Duolao Wang, PhD, Professor of Biostatistics
Global Health Trials Unit, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK
E-mail: Duolao.Wang@lstmed.ac.uk
Phone: 44-151 705 3301

Abstract

Many clinical trials include time-to-event or survival data as an outcome. To compare two survival distributions, the log-rank test is often used to produce a *P*-value for a statistical test of the null hypothesis that the two survival curves are identical. However, such a *P*-value does not provide the magnitude of the difference between the curves regarding the treatment effect. As a result, the *P*-value is often accompanied by an estimate of the hazard ratio estimated from the proportional hazards model or Cox model as a measurement of treatment difference. However, one of the most important assumptions for Cox model is that the hazard functions for the two treatment groups are proportional. When the hazard curves cross, the Cox model could lead to misleading results and the log-rank test could also perform poorly. To address the problem of crossing curves in survival analysis, we propose the use of the win ratio method put forward by Pocock et al. as an estimand for analysing such data. The subjects in the test and control treatment groups are formed into all possible pairs. For each pair, the test treatment subject is labelled a winner or a loser if it is known who had the event of interest such as death. The win ratio is the total number of winners divided by the total number of losers and its standard error can be estimated using Bebu and Lachin method. Using real trial datasets and Monte Carlo simulations, this study investigates the power and type I error and compares the win ratio method with the log-rank test and Cox model under various scenarios of crossing survival curves with different censoring rates and distribution parameters. The results show that the win ratio method has similar power as the log-rank test and Cox model to detect the treatment difference when the assumption of proportional hazards holds true, and that the win ratio method outperforms log-rank test and Cox model in terms of power to detect the treatment difference when the survival curves cross.

Keywords: Win ratio; crossing survival curves; estimand; survival analysis; log-rank test; Cox model

1. Introduction

Many clinical trials would include time-to-event or survival data as an outcome. To compare the two survival distributions, the log-rank test is conducted and its associated *P*-value is used for evaluating whether the two survival curves are statistically identical or not. While the *P*-value provides evidence whether there is a difference, it does not measure its magnitude. Therefore, the *P*-value is often accompanied by an estimate of the hazard ratio with 95% confidence interval (CI) estimated from the Cox hazard model (Gregson et al. 2019). However, the Cox hazards model assumes proportionality of hazards and this assumption is violated when the hazard ratio changes over time, e.g., the two hazard/survival curves intersect. The log-rank test is also likely to lose statistical power in such cases (X. Lin and Xu 2010; P. Qiu and Sheng 2008; Liu, Qiu, and Sheng 2007; Mao 2019). This phenomenon often occurs in oncology clinical trials with delayed treatment effect, treatment switchers and treatment dilution (Luo et al. 2019). Consequently, the calculated results do not accurately reflect the treatment difference and the log-rank test loses its power as the differences in favour of one treatment are offset by the other after passing the crossing points (Klein and Moschberger 2003; X. Lin and Xu 2010). Based on a survey, the log-rank test was still applied in 70% of studies to detect the difference between two crossing survival curves, which clearly went against the assumption of proportional hazard rates (Li et al. 2015).

Several measures have been proposed to address the problem of intersecting curves in survival analysis. For instance, the Renyi test (weighted log-rank test) is based on the maximum absolute value of the difference between cumulative hazard rates. The general idea behind this test is similar to the Kolmogorov-Smirnov test with the exception that it considers censored data (Klein and Moschberger 2003). Lin and Wang introduced a modified log-rank test which measures squared differences at each time points (Bouliotis and Billingham 2011). Two-stage procedures, comprised of a conventional method, e.g., log-rank test and a proposed procedure for addressing the crossing hazard rates have also been used (P. Qiu and Sheng 2008). Kraus developed an adaptive Neyman's smooth method for testing homogeneity of two survival distribution with right censoring based on Neyman's embedding idea combined with Schwarz's selection rule (Kraus 2009). Li et al. (2015) demonstrated the robust power of the adaptive Neyman's smooth method dealing with crossing survival curves. While these methods address the hazard proportionality issue, they only provide a *P*-value for a statistical test of the null hypothesis that the two survival curves are identical and not an interpretable magnitude and precision of the difference between the curves corresponding to the treatment effect.

In this paper, we use the win ratio method proposed by Pocock et al. (2012) as an estimand to address the above problem not only by providing a *P*-value but also a point estimate and CI for the difference in two survival curves. The adjusted win ratio method was proposed to analyse hierarchical composite endpoints in clinical trials (Gasparyan et al. 2020). For all pairwise matches, Luo derived a closed-form variance estimator for the win ratio statistic by applying the *U*-statistics technique (Luo et al. 2015). Mao (2019) widened alternative hypotheses for win ratio from the hazard orders to upper quadrant stochastic order. Dong et al. (2018) have proposed a general form of stratified win ratio to reduce heterogeneity in pairwise comparison. The stratified win ratio approach performs in a similar manner to the conventional time-to-event analysis in EVOLVE trial (Abdalla et al. 2016). The win ratio approach has been recommended to measure the treatment difference in conjunction with rank-based analysis in ICH Harmonised Guideline (Ratitch et al. 2020). Dong et al. (2019) introduced the win odds to handle ties. In practice, the win ratio approach has been applied to design of clinical trials, such as cardiovascular trials (Redfors et al. 2020; Ferreira et al. 2020; Maurer et al. 2018), a randomized trial in kidney transplantation (Fergusson et al. 2018) and device-based hypertension trials (David et al. 2021). The event-specific win ratios were applied to design and analyse semi-competing risk data in clinical trials (Yang et al. 2022). The win ratio method was investigated to analyse recurrent events in composite endpoints (Mao, Kim, and Li 2022). In this paper, we apply the win ratio method to detect and measure the treatment effect differences in survival data, and compare its statistical properties with the conventional survival analysis methods.

Section 2 reviews the details of the proposed method and introduces inferential statistics. Then, the method is applied to two clinical studies are presented in Section 3. Section 4 performs some simulations to assess the statistical properties of the proposed method, and Section 5 summarises the key findings of this paper and discusses the limitations.

2. Methods

Win ratio statistic

Pocock et al. (2012) introduced a win ratio approach to analyse composite endpoints by considering the prioritised outcome. In cardiovascular trial studies, cardiovascular death is

more important than stroke or non-fatal myocardial infarction, for instance. In this paper, we apply the win ratio method to analyse the survival time. The details are as follows:

- Patients in treatment A (N_A) and B (N_B) are formed into all possible pairs ($N_A \times N_B$);
- For each pair, the treatment A patient is labelled a "winner" or a "loser" or "tied" according to their outcomes;
- Calculate the total number of winners (N_W), losers (N_L), and tied (N_T). $N_W + N_L + N_T = N_A \times N_B$;

- $R_W = N_W/N_L$ is the win ratio, the statistic for assessing the treatment effect for a survival time/time-to-event outcome in a clinical trial.

Given the outcome X_i $(i = 1,2, \dots n_t)$ for subject i in the treatment group and outcome Y_j $(j = 1,2, \dots n_c)$ for subject j in the control group (Dong et al. 2016). The general win loss statistic indicator is defined as a $n_t \times n_c$ matrix with elements:

$C_{ij} = C(X_i, Y_j)$

$$= \begin{cases} 1, if\ the\ i^{th}\ subject\ in\ treatment\ group\ wins\ over\ the\ j^{th}\ subject\ in\ control\ group \\ -1, if\ the\ i^{th}\ subject\ in\ treatment\ group\ loses\ against\ the\ j^{th}\ subject\ in\ control\ group \\ 0, if\ the\ i^{th}\ subject\ in\ treatment\ group\ ties\ with\ the\ j^{th}\ subject\ in\ control\ group \end{cases}$$

Then the total number of wins for the treatment group is:

$$TW = \sum_{i=1}^{n_t} \sum_{j=1}^{n_c} I\{C\left(X_i, Y_j\right) = 1\} \quad (1)$$

Where I is an indicator function.

Similarly, the total number of losses for

$$TL = \sum_{i=1}^{n_t} \sum_{j=1}^{n_c} I\{C\left(X_i, Y_j\right) = -1\} \quad (2)$$

The general win ratio can be defined as:

$$W_R = \frac{TW}{TL} \quad (3)$$

The win ratio statistic is able to assess the treatment effect in clinical trials. The value of win ratio >1 indicates the treatment effect is in favour of treatment A to B. Win ratio =1 corresponds there is no true treatment difference between treatment A and B.

Inferential statistics for the win ratio

The following are the hypothesis being tested when using the win ratio statistic:

H_0: win ratio=1. There is no difference in the number of "winners" between treatment and standard groups.

H_1: win ratio≠1. There is a true difference in the number of "winners" between treatment and standard groups.

The general win ratio statistic in equation (3) above is implemented in R via WWR package. WWR package was created by J. Qiu et al. (2017). The asymptotic variance of the win ratio can be obtained from WWR package, which is based on the idea of Bebu and Lachin (2016)

An alternative approach to measuring the difference between the compared treatment groups is called win difference (also named proportion in favour of treatment) which performs a generalised pairwise comparison between groups based on different clinical prioritised outcomes (Buyse 2010; Luo et al. 2017). The proportion in favour of treatment determines by the net difference between the proportion of favourable pairs and the proportion of unfavourable pairs over the total number of pairs.

3. Two real trial examples

In this section, two datasets are used to illustrate the win ratio approach. All calculation procedures were implemented in R.

Example 1: Survival time in gastric trials

The Gastrointestinal Tumour Study group reported the effects of chemotherapy against the combination of chemotherapy and radiotherapy on the survival time of locally unresectable gastric cancer patients (Stablein and Koutrouvelis 1985). This data is also analysed in Hsieh (2001), Bagdonavicius, Hafdi, and Nikulin (2004) and Klein and Moschberger (2003). Forty-five patients were randomly allocated to each of treatment groups and followed about eight

years. The censoring rates of chemotherapy and combined chemotherapy and radiotherapy are 4.4% and 13.3%, respectively. Kaplan-Meier curves of the two treatment groups are shown in Figure 1. The two estimated survival curves indicate that Chemotherapy plus radiotherapy treatment would initially be detrimental to a patient survival but become beneficial later on.

A test for treatment equality with log-rank statistic provides $\chi^2 = 0.2, P = 0.6$. This *P*-value indicates that the null hypothesis should not be rejected ($P > 0.05$), i.e., there is no sufficient evidence that chemotherapy combined with radiotherapy is better than chemotherapy alone. However, because the survival curves intersect (or are crossing), it is known that the log-rank test is not powerful. The result can be explained by the fact that the differences in favour of one treatment are offset by the other after passing the crossing points.

From the perspective of Cox model, it can be deduced that the survival distributions for the two treatments have no significant difference ($P = 0.60$). The hazard rate for chemotherapy is estimated as about 0.90 times that of chemotherapy combined with radiotherapy but could lie between about 0.58 and 1.39 times based on the 95% confidence interval. Nevertheless, the two survival functions, which cross at approximately $S(t) = 0.2$, and the corresponding test indicate a violation of the proportional hazards assumption ($\chi^2 = 13.13,\ P < 0.001$). The median survival is 499 days (95% CI (383, 748)) in the chemotherapy group and 254 days (95% CI (193, 542)) in chemotherapy combined with radiotherapy group. Chemotherapy improves survival in the initial 1000 days. However, chemotherapy combined with radiotherapy results in a better survival rate in the late follow-up time.

By using the win ratio approach, there are 2025 possible pairs in this case. The total number of wins and losses in the chemotherapy group is 1251 and 762, respectively. Therefore, the win ratio is $\frac{1251}{762} = 1.64$. The *P*-value is 0.06. The 95% CI for the win ratio is 0.98 to 2.74. The estimated win ratio greater than 1 suggests a favour for chemotherapy group over chemotherapy combined with radiotherapy group.

Example 2: Time to CV death or CHF hospitalisation in CHARM trial

The CHARM program is the largest trial program in chronic heart failure (CHF). The program evaluates the effectiveness of candesartan in reducing mortality and morbidity in heart failure.

Overall, 7599 subjects were allocated randomly to three different double-blind controlled trials:

- CHARM-Alternative trial: subjects were intolerant to angiotensin-converting enzyme (ACE) inhibitor and had left ventricular ejection fraction (LVEF)≤ 0.40;
- CHARM-Added trial: subjects were on ACE inhibitor and had LVEF ≤ 0.40;
- CHARM-Preserved trial: subjects had LVEF ≥ 0.40.

The CHARM program has a composite primary endpoint for each trial (cardiovascular death or hospitalisation for CHF), which was analysed as a survival time (the time from randomisation to the first occurrence of cardiovascular death or hospitalisation for CHF). Median follow-up time was 3.14 years (Pfeffer et al. 2003). The *P*-value for the proportional hazards assumption test in Cox model is 0.057, 0.096, 0.42 for CHARM-Alternative trial, CHARM-Added trial and CHARM-Preserved trial, respectively, suggesting that hazards proportionality assumption is not seriously violated for the tree studies. Results from win ratio method, log-rank test and Cox model are displayed in Table 1.

For CHARM-Alternative trial, 2028 patients were assigned to either the candesartan or placebo groups. Approximately 33% and 40% patients had primary outcome in the candesartan and placebo groups, respectively. The log-rank test shows a significant difference in the effect of treatment (P <0.0001). Similar results can also be drawn from the Cox model analysis. The inverse of unadjusted hazard rate for candesartan is estimated as about 1.30 times that of placebo but could lie between about 1.12 and 1.49 times. The inverse of hazard ratio together with 95% CI was calculated to compare with the win ratio. The win ratio method gives rise to the similar treatment effect. The *P*-value for win ratio test is < 0.0001, which indicates a significant difference in number of "winners" between the candesartan and placebo groups. Apart from this, the win ratio between candesartan and placebo is 1.33 with 95% CI (1.15, 1.55). For CHARM-Added study, the results of the analysis suggests that the effect of candesartan is significant different from that of placebo. However, CHARM-Preserved study indicates that treatment effect of candesartan is similar to placebo from log-rank test, Cox model and win ratio method.

4. Simulation studies based on hypothetical scenarios

A general framework for simulation of survival data

We will perform Monte Carlo simulation studies to assess the performance of the win ratio method against the log-rank test and Cox model in terms of type I error and power to detect treatment difference in two survival distributions with various censoring mechanisms. Four scenarios under various random censoring rates (0%, 20%, 40% and 75%) for a survival time are considered:

1) Two groups with identical survival distributions;
2) Two groups which have proportional hazard rates;
3) Two survival curves cross at $S(t) = 0.0 \sim 0.5$;
4) Two survival curves cross at $S(t) = 0.5 \sim 1.0$.

The equal sample sizes ($N_A = N_B = 10, 20, 30, 40, 50, 100, 150, 200, 400$) are considered in each treatment group. For each sample size, the descriptive statistics of the win ratio and the proportion of tests with $P < 0.05$ are obtained after running 1000 times simulations. The descriptive statistics include mean, geometric mean, median and standard deviation. The proportion of significant results ($P < 0.05$) give an estimate of the type I error when the two underlying survival distributions are identical but gives an estimate of power when two underlying survival distributions are different. The four scenarios mentioned above are shown in Figure 2. The simulated survival time were implemented using PWEALL in R developed by Luo et al. (2019).

Simulation 1: Two groups have identical survival distributions

The first scenario is conducted to investigate the type I error of three methods (Win ratio, log-rank test and Cox model) (Figure 2: a). The summary statistics of the win ratio and the estimated type I error rates for win ratio method, log-rank test, and Cox model are presented in Table 2. Under various censoring rates, the mean, geometric mean and median of win ratio are approximately 1. The values of the central tendency of the win ratio fluctuate around 1 but stabilises eventually to 1 when the sample size is increased. The type I error of the three methods (win ratio method, log-rank test and Cox model) is approximately at the nominal level of 5%. Under the same percentage of censoring rate, the type I error rates in all three tests are slightly fluctuating with different sample sizes. The type I error is not sensitive to the different censoring rates.

Simulation 2: Two survival distributions have proportional rates

The second scenario is to assess the performance of the win ratio when the two survival distributions have proportional hazard rates (Figure 2: b). The descriptive statistics of the win ratio and the empirical power of the three tests are summarised in Table 3. The geometric mean of the win ratio is less than 1 for various sample sizes and censoring rates, which means that treatment effect is in favour of treatment B than A, i.e., treatment response to B is better than A. At 0% of censoring rate, a sample size of 800 gives 68.9% power for the win ratio, 69.3% for the log-rank test and 69.3% for the Cox model. Under the same amount censoring, the power in all three tests increases proportionally with increasing sample size. However, with the same sample size, the power of the tests decreases when the amount of censoring rate is increased. When censoring rate is 75% and the sample size is 800, the power of the three tests reduces to 22.8%, 23.7% and 23.7% respectively. The results indicate that the win ratio method has the similar power as log-rank test and Cox model to detect the treatment difference when the hazard proportional assumption holds true.

Simulation 3: Two survival curves cross at $S(t) = 0.0 \sim 0.5$

Table 4 presents the results for two survival curves that intersect at $S(t) = 0.0 \sim 0.5$ (Figure 2: c). The geometric mean of win ratio is greater than 1 for different sample sizes and censoring rates. The win ratio method has the best performance among the three tests when the survival curves are intersecting at later time points. At 0% censoring rate, a sample size of 800 the win ratio method has the highest statistical power of 76.2% compared to19.8% for log-rank test and 19.8% for Cox model. The result indicates that the win ratio method trends to have higher power than log-rank test and Cox model when two survival curves cross at later time. However, increasing censoring rates reduce the power for all three methods. When the censoring rate reaches 75% with the sample size of 800, the power decreases to 27.1% for win ratio, 10.5% for log-rank test and 10.3% for Cox model.

Simulation 4: Two survival curves cross at $S(t) = 0.5 \sim 1.0$

Table 5 displays the simulation results for two survival curves crossing at $S(t) = 0.5 \sim 1$ (Figure 2: d). The geometric mean of win ratio being larger than 1 implies that there is a survival difference between treatment A and B. Table 5 shows a consistently larger power for win ratio method than log-rank test and Cox model. When there is no censoring and sample size is 800, the power to detect the treatment difference is 42.2%, 7.8% and 7.8% for win ratio method,

log-rank test and Cox model, respectively. When the censoring rate reaches 75% and sample size is 800, the power is 15.7%, 7.4% and 7.1% for win ratio method, log-rank test and Cox model, respectively. Of note, the power for win ratio method appears sensitive to the censoring rate, but the log-rank and Cox model are not when the survival curves cross at early time S(t) = 0.5 ~ 1.0.

5. Discussion

The log-rank test and Cox model are widely used methods for the comparison of two survival distributions (Koziol and Jia 2014). The log-rank test and Cox model have sufficient power to detect the difference in treatment effect between the two groups under the proportional hazards assumption. However, the log-rank test has poor power to detect such a difference when the hazard proportionality assumption is violated. This is because the log-rank test provides an overall comparison of the entire survival experience, therefore, the differences in favour of one treatment are offset by the other after passing the crossing points (X. Lin and Xu 2010; Klein and Moschberger 2003). On the other hand, the Cox model yields a biased estimation of the treatment effect when the hazard ratios change over time (R.S. Lin and León 2017). The objective of this paper was to assess the statistical properties of the win ratio method as an estimand for measuring the treatment effect when hazard proportionality is violated and to compare the win ratio method with log-rank method and Cox model in terms of the type I error when two underlying survival distributions are identical, and power when two underlying survival distributions are different, and in particular when the proportionality assumption is violated.

The results from the two real clinical trials and simulations under four scenarios show that the win ratio method has similar power as the log-rank test and Cox model to detect the treatment difference when the assumption of proportional hazards holds true, and that the win ratio method outperforms log-rank test and Cox model in terms of power to detect the treatment difference when survival curves cross.

In this study, we simulated survival distributions via piecewise constant hazard for four complex scenarios with different assumptions of proportional hazards. Simulation results in Scenario 1 indicates that the win ratio method, log-rank test and Cox model have the same type I error of 5% under various censoring rates and sample sizes when the two survival

distributions are identical. This is expected since the simulations are set up as so. The results from Scenario 2 demonstrate that when the hazard rates are proportional, the win ratio method, log-rank test and Cox model have approximately the same power. Simulation results under Scenarios 3 and 4 suggest that when two survival curves cross at early or late time points, the win ratio method outperforms log-rank test and Cox model under various censoring rates and sample sizes. The favourable differences in estimated power between win ratio and the other two methods increase with increase in the sample size.

Some new methods have been recently developed for analysing survival data with non-proportional hazards. The combo test is such a method which combines several weighted log-rank tests and increases the power towards the violated assumption of proportional hazards (Luo et al. 2019). The maximum-type combination test has a strong power towards to improper weights of weighting schemes under non-proportional hazards scenarios (Ristl et al. 2021). We were unable to compare win ratio method with those methods since the primary objective of this study was to compare the win ratio method with the conventional survival analysis methods. While these methods may increase the power compared with the traditional survival analysis methods, their main limitation lies that they only provide a *P*-value for a statistical test of the null hypothesis that there is no difference between two survival curves but do not provide an interpretable magnitude and precision of treatment difference.

The win ratio method is based on the counts of "winners" and "losers" in each treatment group for a survival time outcome among all possible pairwise comparisons. Under the minimal assumptions, a close-form variance for win ratio can be derived. In this paper, we used the method proposed by Bebu and Lachin (2016) for calculating the variance for win ratio statistic, which can be implemented via WWR package in R (J. Qiu et al. 2017). The win ratio approach has some advantages over the traditional methods for survival data analyses. First, the win ratio is calculated by the total number of winners divided by the total numbers of losers, and its 95% confidence interval and *P*-value for the win ratio are obtained using asymptotic theory based on the number of winners, losers, and ties. The counting of those numbers does not require the assumption of the hazard proportionality, which is a fundamental requirement for the log-rank test and Cox model. It is therefore an assumption-free method but has larger power than log-rank test and Cox model to detect the difference when survival curves intersect. Second, the win ratio approach deals with a composite endpoint (multiple survival times with each component contributing on separate survival time) by giving ranking priorities to its

components, in contrast to the traditional survival analysis methods which treat each component of a composite endpoint equally and analyse only the first occurrence of any component events. Third, empirically, estimates of win ratio are very close to the reciprocal of hazard ratios in terms of direction and magnitude, and similar strength of evidence although an exact mathematical relationship has not been established (Gregson et al. 2019). Therefore, win ratio could be used a new estimand for assessing the treatment difference in survival analysis.

The study has some limitations. First, we have not discussed in this paper the sample size calculation for the win ratio method in design of a clinical trial with time-to-event data as the primary outcome. Two recently published papers dealt with sample size calculation based on win ratio statistic. Yu and Ganju (2022) provided an approximation formula based on probability of win ratio endpoint. Mao, Kim, and Miao (2022) provided more precise formula to calculate sample size for win ratio analysis of different types of outcomes. Second, the win ratio method for analysis of survival data suggested in this paper is a univariate method, which generates a crude (unadjusted) win ratio between two treatment arms. Estimation of the adjusted treatment effect is often needed, particularly when randomisation imbalance occurs for some important prognostic factors at baseline. Inverse probability weighting based on propensity score could be employed to calculate the adjusted win ratio as was done by Dong et al. (2020) for dealing with survival data in the presence of independent censoring. Adjusted win ratio with stratification proposed by Gasparyan et al. (2020) could also be employed to control for possible imblances in baseline characteristics of patients. Finally, in this paper, we adopted the definition of the win ratio given Pocock et al. (2012) and excluded tied observations when calculating the win ratio statistic in the setting of independent censoring. The resulting win ratio statistic may be not reliable when censoring is dependent (eg, baseline covariate related). Under such a circumstance, inverse-probability-of-censoring weighting adjusted win ratio (IPCW-adjusted win ratio) (Dong et al. 2020) and inverse-probability-of-censoring weighting adjusted win ratio for baseline covariates and/or time-dependent covariates (CovIPCW-adjusted win ratio) (Dong et al. 2021) may be recommended. Further work is needed to assess the properties of those adjusted win ratios when the survival curves cross.

Disclosures

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abdalla, S., M. E. Montez-Rath, P. S. Parfrey, and G. M. Chertow. 2016. "The win ratio approach to analyzing composite outcomes: An application to the EVOLVE trial." *Contemp Clin Trials* 48: 119-24. https://doi.org/10.1016/j.cct.2016.04.001. https://www.ncbi.nlm.nih.gov/pubmed/27080930.

Bagdonavicius, V., M. A. Hafdi, and M. Nikulin. 2004. "Analysis of survival data with cross-effects of survival functions." *Biostatistics* 5 (3): 415-25. https://doi.org/10.1093/biostatistics/5.3.415. https://www.ncbi.nlm.nih.gov/pubmed/15208203.

Bebu, I., and J. M. Lachin. 2016. "Large sample inference for a win ratio analysis of a composite outcome based on prioritized components." *Biostatistics* 17 (1): 178-87. https://doi.org/10.1093/biostatistics/kxv032. https://www.ncbi.nlm.nih.gov/pubmed/26353896.

Bouliotis, G., and L. Billingham. 2011. "Crossing survival curves: alternatives to the log-rank test." *Trials* 12 (Suppl 1): A137-A137. https://doi.org/10.1186/1745-6215-12-S1-A137. https://lstmed.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=cookie,ip,uid&db=edsdoj&AN=edsdoj.855011f10bcc4f48b08872fc5170878e&site=eds-live&scope=site.

Buyse, M. 2010. "Generalized pairwise comparisons of prioritized outcomes in the two-sample problem." *Stat Med* 29 (30): 3245-57. https://doi.org/10.1002/sim.3923. https://www.ncbi.nlm.nih.gov/pubmed/21170918.

David, E. Kandzari, L. Hickey Graeme, J. Pocock Stuart, A. Weber Michael, Böhm Michael, A. Cohen Sidney, Fahy Martin, Lamberti Giuseppina, and Mahfoud Felix. 2021. "Prioritised endpoints for device-based hypertension trials: the win ratio methodology." *EuroIntervention* 16 (18): e1496-e1502. https://eurointervention.pcronline.com/article/prioritized-endpoints-for-device-based-hypertension-trials-the-win-ratio-methodology.

Dong, G., D. C. Hoaglin, J. Qiu, R. A. Matsouaka, Y. Chang, J. Wang, and M. Vandemeulebroecke. 2019. "The Win Ratio: On Interpretation and Handling of Ties." *Statistics in Biopharmaceutical Research* 12 (1): 99-106. https://doi.org/10.1080/19466315.2019.1575279. https://lstmed.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=cookie,ip,uid&db=asx&AN=141626801&site=eds-live&scope=site.

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. https://www.ncbi.nlm.nih.gov/pubmed/33247544.

Dong, G., D. Li, S. Ballerstedt, and M. Vandemeulebroecke. 2016. "A generalized analytic solution to the win ratio to analyze a composite endpoint considering the clinical importance order among components." *Pharm Stat* 15 (5): 430-7. https://doi.org/10.1002/pst.1763. https://www.ncbi.nlm.nih.gov/pubmed/27485522.

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. https://www.ncbi.nlm.nih.gov/pubmed/32552451.

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. https://www.ncbi.nlm.nih.gov/pubmed/29172988.

Fergusson, N. A., T. Ramsay, M. Chasse, S. W. English, and G. A. Knoll. 2018. "The win ratio approach did not alter study conclusions and may mitigate concerns regarding unequal composite end points in kidney transplant trials." *J Clin Epidemiol* 98: 9-15. https://doi.org/10.1016/j.jclinepi.2018.02.001. https://www.ncbi.nlm.nih.gov/pubmed/29428872.

Ferreira, J. P., P. S. Jhund, K. Duarte, B. L. Claggett, S. D. Solomon, S. Pocock, M. C. Petrie, F. Zannad, and J. J. V. McMurray. 2020. "Use of the win ratio in cardiovascular trials." *JACC Heart Fail* 8 (6): 441-450. https://doi.org/10.1016/j.jchf.2020.02.010. https://www.ncbi.nlm.nih.gov/pubmed/32466836.

Gasparyan, S. B., F. Folkvaljon, O. Bengtsson, J. Buenconsejo, and G. G. Koch. 2020. "Adjusted win ratio with stratification: Calculation methods and interpretation." *Stat Methods Med Res*: 962280220942558. https://doi.org/10.1177/0962280220942558. https://www.ncbi.nlm.nih.gov/pubmed/32726191.

Gregson, J., L. Sharples, G. W. Stone, C. F. Burman, F. Ohrn, and S. Pocock. 2019. "Nonproportional hazards for time-to-event outcomes in clinical trials: JACC review topic of the week." *J Am Coll Cardiol* 74 (16): 2102-2112. https://doi.org/10.1016/j.jacc.2019.08.1034. https://www.ncbi.nlm.nih.gov/pubmed/31623769.

Hsieh, F. 2001. "On heteroscedastic hazards regression models: theory and application." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (1): 63-79. https://doi.org/10.1111/1467-9868.00276. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00276.

Klein, J. P. , and M. L. Moschberger. 2003. *Survival analysis: techniques for censored and truncated data*. Book.*Statistics for biology and health*. New York: Springer.

Koziol, J. A., and Z. Jia. 2014. "Weighted Lin-Wang tests for crossing hazards." *Computational and Mathematical Methods in Medicine* 2014: 643457. https://doi.org/10.1155/2014/643457. https://doi.org/10.1155/2014/643457.

Kraus, D. 2009. "Adaptive Neyman's smooth tests of homogeneity of two samples of survival data." *Journal of Statistical Planning and Inference* 139 (10): 3559-3569. https://doi.org/https://doi.org/10.1016/j.jspi.2009.04.009. http://www.sciencedirect.com/science/article/pii/S0378375809000998.

Li, H., D. Han, Y. Hou, H. Chen, and Z. Chen. 2015. "Statistical inference methods for two crossing survival curves: a comparison of methods." *PLoS One* 10 (1): e0116774. https://doi.org/10.1371/journal.pone.0116774. https://www.ncbi.nlm.nih.gov/pubmed/25615624.

Lin, Ray S., and Larry F. León. 2017. "Estimation of treatment effects in weighted log-rank tests." *Contemporary Clinical Trials Communications* 8: 147-155. https://doi.org/https://doi.org/10.1016/j.conctc.2017.09.004. http://www.sciencedirect.com/science/article/pii/S2451865417300881.

Lin, X., and Q. Xu. 2010. "A new method for the comparison of survival distributions." *Pharmaceutical Statistics* 9 (1): 67-76. https://doi.org/https://doi.org/10.1002/pst.376. https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.376.

Liu, K., P. Qiu, and J. Sheng. 2007. "Comparing two crossing hazard rates by Cox proportional hazards modelling." *Stat Med* 26 (2): 375-91. https://doi.org/10.1002/sim.2544. https://www.ncbi.nlm.nih.gov/pubmed/16538703.

Luo, X., X. Mao, X. Chen, J. Qiu, S. Bai, and H. Quan. 2019. "Design and monitoring of survival trials in complex scenarios." *Stat Med* 38 (2): 192-209. https://doi.org/10.1002/sim.7975. https://www.ncbi.nlm.nih.gov/pubmed/30281165.

Luo, X., J. Qiu, S. Bai, and H. Tian. 2017. "Weighted win loss approach for analyzing prioritized outcomes." *Stat Med* 36 (15): 2452-2465. https://doi.org/10.1002/sim.7284. https://www.ncbi.nlm.nih.gov/pubmed/28343373.

Luo, X., H. Tian, S. Mohanty, and W. Y. Tsai. 2015. "An alternative approach to confidence interval estimation for the win ratio statistic." *Biometrics* 71 (1): 139-145. https://doi.org/10.1111/biom.12225. https://www.ncbi.nlm.nih.gov/pubmed/25156540.

Mao, L. 2019. "On the alternative hypotheses for the win ratio." *Biometrics* 75 (1): 347-351. https://doi.org/10.1111/biom.12954. https://www.ncbi.nlm.nih.gov/pubmed/30096729.

Mao, L., K. Kim, and Y. Li. 2022. "On recurrent-event win ratio." *Stat Methods Med Res* 31 (6): 1120-1134. https://doi.org/10.1177/09622802221084134.

Mao, L., KyungMann Kim, and Xinran Miao. 2022. "Sample size formula for general win ratio analysis." *Biometrics* 78 (3): 1257-1268. https://doi.org/https://doi.org/10.1111/biom.13501. https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13501.

Maurer, M. S., J. H. Schwartz, B. Gundapaneni, P. M. Elliott, G. Merlini, M. Waddington-Cruz, A. V. Kristen, M. Grogan, R. Witteles, T. Damy, B. M. Drachman, S. J. Shah, M. Hanna, D. P. Judge, A. I. Barsdorf, P. Huber, T. A. Patterson, S. Riley, J. Schumacher, M. Stewart, M. B. Sultan, C. Rapezzi, and Attr-Act Study Investigators. 2018. "Tafamidis treatment for patients with transthyretin amyloid cardiomyopathy." *N Engl J Med* 379 (11): 1007-1016. https://doi.org/10.1056/NEJMoa1805689. https://www.ncbi.nlm.nih.gov/pubmed/30145929.

Pfeffer, M. A., K. Swedberg, C. B. Granger, P. Held, J. J. McMurray, E. L. Michelson, B. Olofsson, J. Ostergren, S. Yusuf, S. Pocock, Charm Investigators, and Committees. 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. https://www.ncbi.nlm.nih.gov/pubmed/13678868.

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. https://www.ncbi.nlm.nih.gov/pubmed/21900289.

Qiu, J., X. Luo, S. Bai, H. Tian, and M. Mikailov. 2017. "WWR: An R package for analyzing prioritized outcomes." *J Med Stat Inform* 5: 4. https://doi.org/10.7243/2053-7662-5-4.

Qiu, P., and J. Sheng. 2008. "A two-stage procedure for comparing hazard rate functions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70 (1): 191-208. https://doi.org/https://doi.org/10.1111/j.1467-9868.2007.00622.x. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00622.x.

Ratitch, B., N. Goel, C. Mallinckrodt, J. Bell, J. W. Bartlett, G. Molenberghs, P. Singh, I. Lipkovich, and M. O'Kelly. 2020. "Defining efficacy estimands in clinical trials: examples illustrating ICH E9(R1) guidelines." *Ther Innov Regul Sci* 54 (2): 370-384. https://doi.org/10.1007/s43441-019-00065-7. https://www.ncbi.nlm.nih.gov/pubmed/32072586.

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*. https://doi.org/10.1093/eurheartj/ehaa665. https://www.ncbi.nlm.nih.gov/pubmed/32901285.

Ristl, R., NM. Ballarini, H. Götte, A. Schüler, M. Posch, and F. König. 2021. "Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze RCTs in oncology." *Pharmaceutical Statistics* 20 (1): 129-145. https://doi.org/https://doi.org/10.1002/pst.2062. https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.2062.

Stablein, D. M., and I. A. Koutrouvelis. 1985. "A two-sample test sensitive to crossing hazards in uncensored and singly censored data." *Biometrics* 41 (3): 643-52. https://www.ncbi.nlm.nih.gov/pubmed/4074816.

Yang, Song, James Troendle, Daewoo Pak, and Eric Leifer. 2022. "Event-specific win ratios for inference with terminal and non-terminal events." *Statistics in Medicine* 41 (7): 1225-1241. https://doi.org/https://doi.org/10.1002/sim.9266. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9266.

Yu, Ron Xiaolong, and Jitendra Ganju. 2022. "Sample size formula for a win ratio endpoint." *Statistics in Medicine* 41 (6): 950-963. https://doi.org/https://doi.org/10.1002/sim.9297. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9297.

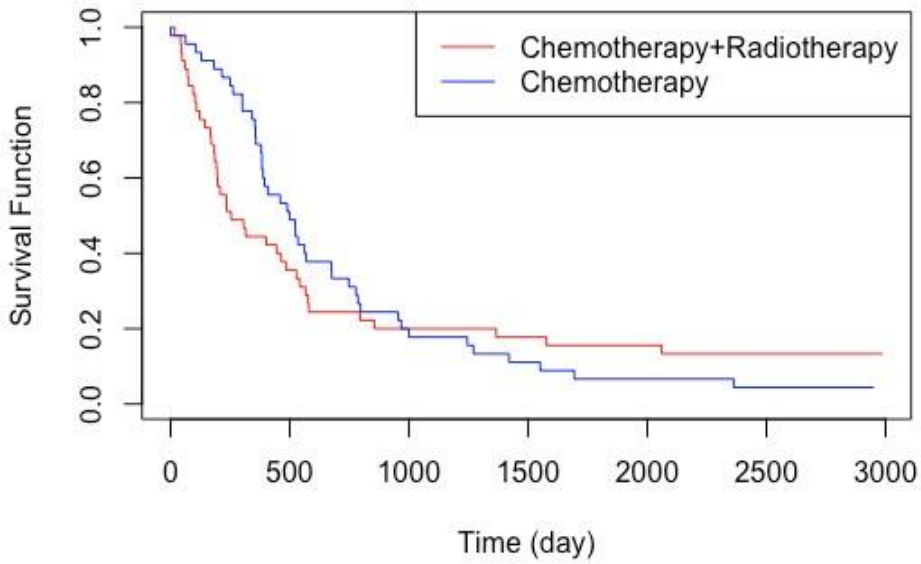
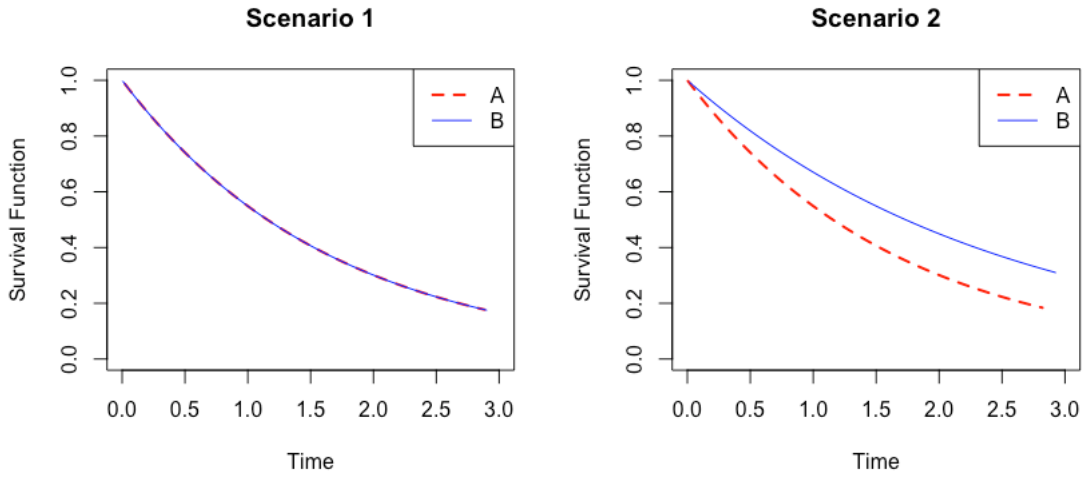
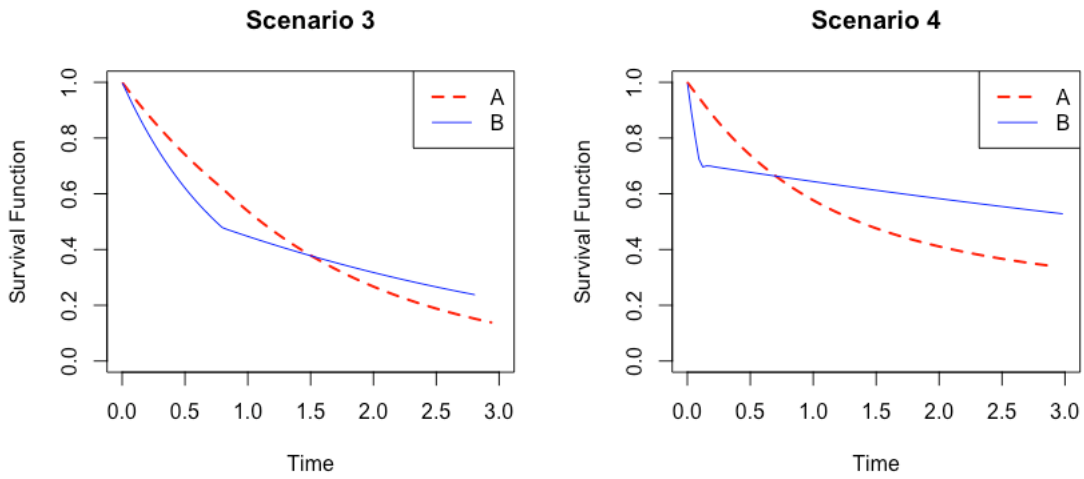


Figure 1: Kaplan-Meier estimates of survival for chemotherapy against chemotherapy combined with radiotherapy



(a) (b)



(c) (d)

Figure 2: Survival curves for four scenarios

Table 1: Summary results from log-rank, Cox model and win ratio analyses of three sub-studies in the CHARM programme

	CHARM-Alternative		CHARM-Added		CHARM-Preserved	
	Candesartan	Placebo	Candesartan	Placebo	Candesartan	Placebo
Descriptive statistics						
Number of patients	1013	1015	1276	1272	1514	1509
Primary composite event	334	406	483	538	333	366
Log-rank test						
P-value	0.0004		0.01		0.118	
Cox model analysis						
1/HR (95% CI)	1.30 (1.12,1.49)		1.18 (1.04,1.33)		1.12 (0.97,1.03)	
P-value	0.0001		0.01		0.118	
Win ratio analysis						
Win ratio (95% CI)	1.33 (1.15, 1.55)		1.20 (1.06, 1.36)		1.13 (0.97, 1.32)	
P-value	0.0001		0.004		0.107	

HR=Hazard ratio

Table 2: The summary statistics of win ratio and the type I errors of three methods when two groups have identical survival distributions: Simulation 1

% of censoring	Summary statistics of win ratio					Proportion of tests with $P<0.05$		
	sample size	mean	geometric mean	median	SD	win ratio	log-rank	Cox-model
0	20	1.19	1.01	1.04	0.77	4.6%	6.2%	5.4%
	40	1.08	1.00	1.02	0.44	4.2%	5.4%	5.0%
	60	1.05	1.00	1.01	0.32	4.4%	5.7%	5.7%
	80	1.05	1.01	1.01	0.27	4.0%	4.4%	4.1%
	100	1.03	1.00	1.01	0.25	5.1%	5.3%	5.3%
	200	1.01	1.00	1.00	0.17	5.3%	4.9%	4.9%
	300	1.01	1.00	1.00	0.13	4.2%	4.6%	4.6%
	400	1.01	1.00	1.00	0.11	5.1%	4.8%	4.8%
	800	1.00	1.00	1.00	0.08	4.3%	4.6%	4.6%
25	20	1.33	1.00	1.00	1.96	4.0%	5.8%	4.4%
	40	1.12	1.01	1.01	0.58	5.0%	5.8%	5.2%
	60	1.05	0.98	0.99	0.38	4.7%	4.6%	4.3%
	80	1.05	1.00	0.99	0.33	4.5%	5.3%	4.8%
	100	1.03	0.99	1.00	0.28	4.4%	5.0%	4.7%
	200	1.02	1.00	0.99	0.20	4.6%	5.2%	5.2%
	300	1.01	1.00	0.99	0.16	5.2%	4.7%	4.7%
	400	1.01	1.00	1.00	0.13	5.3%	4.9%	4.9%
	800	1.01	1.00	1.00	0.09	5.3%	4.5%	4.5%
50	20	1.46	1.00	1.03	1.64	4.7%	5.9%	3.5%
	40	1.13	0.97	0.96	0.66	4.3%	5.1%	4.0%
	60	1.12	1.01	1.01	0.54	4.6%	5.3%	4.8%
	80	1.08	1.00	1.00	0.42	4.1%	5.7%	4.8%
	100	1.07	1.01	1.00	0.38	5.1%	4.9%	4.2%
	200	1.04	1.01	1.02	0.24	5.0%	5.2%	5.0%
	300	1.01	0.99	0.99	0.19	4.8%	4.1%	4.0%
	400	1.01	1.00	1.00	0.17	5.6%	5.1%	5.1%
	800	1.01	1.00	1.00	0.12	4.4%	4.0%	4.0%
75	20	1.90	1.03	1.00	3.46	3.3%	5.4%	0.5%
	40	1.53	0.99	1.00	3.11	4.6%	4.8%	2.7%
	60	1.32	1.00	1.01	2.22	5.0%	4.2%	3.1%
	80	1.18	0.99	1.00	1.01	4.9%	4.5%	3.9%
	100	1.17	1.02	1.02	0.71	4.9%	4.6%	4.1%
	200	1.05	0.99	0.99	0.37	4.9%	4.8%	4.5%
	300	1.04	1.00	1.00	0.28	4.6%	4.6%	4.4%
	400	1.04	1.01	1.02	0.25	5.0%	5.2%	4.9%
	800	1.02	1.01	1.01	0.17	5.1%	5.6%	5.6%

SD: Standard deviation

Table 3: The summary statistics of win ratio and powers of three methods when two survival curves have proportional hazard rates: Simulation 2

0% of censoring	Summary statistics of win ratio					Proportion of tests with $P<0.05$		
	sample size	mean	geometric mean	median	SD	win ratio	log-rank	Cox-model
0	20	1.01	0.84	0.85	0.70	5.3%	8.7%	7.8%
	40	0.88	0.82	0.82	0.35	7.3%	7.9%	7.7%
	60	0.85	0.81	0.83	0.26	9.6%	12.4%	11.7%
	80	0.84	0.81	0.81	0.22	11.9%	12.1%	11.6%
	100	0.83	0.81	0.81	0.20	14.4%	16.2%	16.1%
	200	0.82	0.81	0.82	0.14	24.7%	25.8%	25.7%
	300	0.82	0.81	0.81	0.11	34.0%	33.5%	33.4%
	400	0.82	0.82	0.81	0.10	42.1%	42.4%	42.2%
	800	0.82	0.82	0.82	0.07	68.9%	69.3%	69.3%
25	20	1.00	0.78	0.79	0.84	6.6%	8.7%	6.8%
	40	0.88	0.79	0.80	0.44	7.5%	9.4%	8.6%
	60	0.86	0.81	0.82	0.31	8.5%	9.8%	9.1%
	80	0.86	0.82	0.81	0.28	10.4%	11.6%	11.0%
	100	0.84	0.81	0.81	0.23	12.0%	14.1%	13.9%
	200	0.83	0.81	0.81	0.17	20.6%	20.3%	20.3%
	300	0.83	0.82	0.82	0.13	24.4%	26.2%	26.1%
	400	0.82	0.81	0.81	0.11	32.8%	34.1%	33.8%
	800	0.82	0.82	0.82	0.08	57.6%	58.6%	58.6%
50	20	1.24	0.80	0.82	1.80	5.2%	6.5%	3.4%
	40	0.98	0.84	0.84	0.60	4.9%	6.3%	5.3%
	60	0.93	0.84	0.84	0.45	6.2%	7.5%	6.8%
	80	0.87	0.81	0.82	0.36	9.2%	9.3%	8.7%
	100	0.88	0.83	0.83	0.30	8.5%	8.4%	8.0%
	200	0.83	0.81	0.82	0.19	15.0%	14.6%	14.3%
	300	0.83	0.81	0.81	0.16	18.6%	18.9%	18.7%
	400	0.83	0.82	0.81	0.14	23.0%	25.6%	25.2%
	800	0.82	0.82	0.82	0.10	42.6%	44.3%	44.0%
75	20	1.39	0.80	1.00	2.24	3.2%	5.4%	0.8%
	40	1.10	0.77	0.80	1.20	5.9%	5.4%	2.5%
	60	1.02	0.80	0.83	0.82	7.4%	8.3%	6.5%
	80	0.92	0.79	0.80	0.57	6.4%	7.1%	6.2%
	100	0.90	0.80	0.82	0.45	5.6%	7.4%	6.8%
	200	0.86	0.81	0.80	0.32	9.9%	10.0%	9.3%
	300	0.83	0.81	0.81	0.22	11.8%	11.3%	10.9%
	400	0.83	0.81	0.82	0.20	15.1%	14.8%	14.4%
	800	0.82	0.81	0.81	0.13	22.8%	23.7%	23.7%

SD: Standard deviation

Table 4: The summary statistics of win ratio and powers of three methods when two survival curves cross at S(t) = 0.0 ~ 0.5: Simulation 3

% of censoring	Summary statistics of win ratio					Proportion of tests with $P<0.05$		
	sample size	mean	geometric mean	median	SD	win ratio	log-rank	Cox-model
0	20	1.58	1.28	1.27	1.48	6.9%	8.9%	8.2%
	40	1.33	1.24	1.23	0.53	7.0%	7.1%	6.7%
	60	1.31	1.25	1.25	0.42	10.1%	7.1%	6.9%
	80	1.29	1.25	1.25	0.36	13.6%	6.9%	6.8%
	100	1.29	1.25	1.26	0.31	15.0%	8.9%	8.7%
	200	1.27	1.25	1.25	0.21	27.2%	10.6%	10.5%
	300	1.26	1.25	1.25	0.17	36.7%	11.5%	11.4%
	400	1.26	1.25	1.25	0.15	46.9%	14.8%	14.8%
	800	1.25	1.24	1.24	0.10	76.2%	19.8%	19.8%
25	20	1.66	1.29	1.27	1.67	4.2%	7.2%	5.6%
	40	1.37	1.24	1.23	0.65	7.0%	6.7%	6.1%
	60	1.35	1.26	1.27	0.51	8.3%	6.7%	6.2%
	80	1.29	1.23	1.23	0.43	10.3%	8.2%	7.8%
	100	1.29	1.24	1.24	0.36	12.2%	7.8%	7.5%
	200	1.27	1.25	1.25	0.25	21.0%	9.7%	9.7%
	300	1.27	1.25	1.24	0.20	30.4%	12.5%	12.2%
	400	1.25	1.24	1.24	0.17	35.8%	9.4%	9.4%
	800	1.25	1.25	1.25	0.12	63.5%	18.9%	18.8%
50	20	2.17	1.35	1.25	3.30	6.3%	7.8%	4.4%
	40	1.46	1.25	1.27	0.97	4.8%	5.6%	4.6%
	60	1.41	1.28	1.28	0.66	8.7%	6.2%	5.7%
	80	1.35	1.26	1.28	0.53	8.4%	5.5%	5.3%
	100	1.33	1.26	1.25	0.47	9.5%	7.0%	6.6%
	200	1.31	1.27	1.27	0.33	17.9%	8.7%	8.6%
	300	1.24	1.22	1.22	0.24	17.8%	7.2%	7.1%
	400	1.27	1.25	1.25	0.21	26.8%	9.7%	9.6%
	800	1.25	1.25	1.24	0.15	46.8%	12.8%	12.8%
75	20	2.17	1.18	1.00	3.62	3.9%	5.8%	0.7%
	40	2.03	1.25	1.25	4.21	7.6%	5.9%	3.3%
	60	1.61	1.27	1.27	1.37	5.5%	5.7%	4.0%
	80	1.51	1.27	1.25	0.98	9.3%	6.1%	5.5%
	100	1.40	1.24	1.25	0.73	6.3%	4.7%	4.0%
	200	1.33	1.25	1.25	0.46	11.6%	6.3%	6.1%
	300	1.27	1.22	1.22	0.35	11.3%	6.9%	6.4%
	400	1.28	1.24	1.23	0.31	16.3%	6.8%	6.6%
	800	1.26	1.25	1.25	0.21	27.1%	10.5%	10.3%

SD: Standard deviation

Table 5: The summary statistics of win ratio and powers of three methods when two survival curves cross at S(t) = 0.5 ~ 1.0: Simulation 4

% of censoring	Summary statistics of win ratio					Proportion of tests with *P*<0.05		
	sample size	mean	geometric mean	median	SD	win ratio	log-rank	Cox-model
0	20	1.36	1.16	1.17	0.89	4.5%	5.8%	4.7%
	40	1.25	1.15	1.15	0.53	6.0%	7.1%	6.6%
	60	1.21	1.16	1.13	0.39	6.8%	4.5%	4.2%
	80	1.19	1.15	1.15	0.33	8.5%	4.7%	4.3%
	100	1.21	1.17	1.17	0.29	9.6%	5.2%	4.9%
	200	1.17	1.16	1.16	0.20	13.6%	6.1%	6.1%
	300	1.17	1.16	1.17	0.16	17.7%	5.7%	5.6%
	400	1.17	1.16	1.16	0.14	22.2%	6.3%	6.3%
	800	1.16	1.16	1.16	0.10	42.2%	7.8%	7.8%
25	20	1.45	1.14	1.14	1.22	4.8%	6.3%	5.4%
	40	1.33	1.19	1.19	0.67	7.4%	5.7%	5.4%
	60	1.24	1.16	1.16	0.48	6.7%	5.0%	4.8%
	80	1.20	1.14	1.14	0.38	6.6%	5.7%	5.4%
	100	1.20	1.16	1.15	0.32	6.8%	5.3%	5.2%
	200	1.19	1.16	1.17	0.25	11.1%	7.0%	7.0%
	300	1.17	1.15	1.16	0.18	15.0%	5.0%	4.9%
	400	1.18	1.17	1.17	0.16	19.9%	4.6%	4.4%
	800	1.16	1.15	1.15	0.11	31.2%	8.2%	8.0%
50	20	1.80	1.17	1.19	2.61	5.2%	5.2%	2.9%
	40	1.37	1.16	1.17	0.88	5.6%	4.7%	4.1%
	60	1.32	1.19	1.21	0.62	6.0%	5.7%	5.3%
	80	1.23	1.14	1.16	0.48	5.7%	5.3%	5.0%
	100	1.23	1.16	1.16	0.41	6.9%	4.4%	4.0%
	200	1.20	1.16	1.17	0.29	10.5%	6.5%	6.3%
	300	1.18	1.16	1.16	0.23	11.8%	4.9%	4.8%
	400	1.17	1.16	1.16	0.19	12.7%	4.7%	4.6%
	800	1.17	1.16	1.16	0.14	24.5%	7.7%	7.7%
75	20	2.07	1.12	1.00	3.47	3.0%	5.1%	0.5%
	40	1.70	1.09	1.09	3.28	5.6%	5.8%	2.9%
	60	1.50	1.17	1.17	1.63	5.8%	5.6%	4.3%
	80	1.38	1.15	1.14	1.08	6.3%	5.1%	4.5%
	100	1.32	1.17	1.18	0.70	6.2%	5.6%	4.7%
	200	1.24	1.17	1.16	0.44	7.7%	6.3%	6.2%
	300	1.19	1.15	1.14	0.34	8.9%	5.0%	4.7%
	400	1.18	1.15	1.15	0.28	10.8%	4.8%	4.6%
	800	1.17	1.16	1.15	0.20	15.7%	7.4%	7.1%

SD: Standard deviation