**Stratified win ratio**

The win ratio was first proposed by Pocock et al. (2012) to analyze a composite endpoint while considering the clinical importance order and the relative timing of its components. It has attracted increasing attention since then, in applications as well as methodologically. It is not uncommon that some clinical trials require a stratified analysis. In this paper we propose a stratified win ratio statistic in a similar way as the Mantel-Haenszel stratified odds ratio (Mantel and Haenszel, 1959), derive a general form of its variance estimator with a plug-in of existing or potentially new variance/covariance estimators of the winners for the two treatment groups, and assess its statistical performance using simulation studies. Our simulations show that our proposed Mantel-Haenszel-type stratified win ratio performs similarly to the Mantel-Haenszel stratified odds ratio with asymptotic variance by Robins, Breslow and Greenland (1986) for the simplified situation when the win ratio reduces to the odds ratio, and our proposed Mantel-Haenszel-type stratified win ratio is preferred compared to the inverse-variance weighted win ratio and unweighted win ratio particularly when the data are sparse. We also formulate a homogeneity test following Cochran (1954) that assesses whether the stratum-specific win ratios are homogeneous across strata, as this method is used frequently in meta-analyses and a better test for the win ratio homogeneity is not available yet.

**Keywords**: Win ratio, stratified win ratio, odds ratio, Mantel-Haenszel stratified odds ratio

# Introduction

A composite endpoint is frequently used as the primary endpoint in randomized clinical trials in cardiovascular, cancer, transplant, diabetes and other disease areas, often in a time-to-event setting. One advantage of using a composite endpoint is that its event rate is higher than that of any of its components alone, resulting in a smaller sample size. However, the conventional statistical approach in this setting, which is a time-to-first-event analysis, suffers from the fact that all components are treated as equally important, while often the first event in a particular subject may not be the most important one. For example, in a solid organ transplantation trial, a common primary endpoint is the composite of treated biopsy-proven acute rejection (tBPAR), graft loss or death. A conventional primary analysis ignores any deaths of subjects that have already experienced tBPAR, although, arguably, death is clinically the most impactful event. To address this issue, several methods have been proposed that take into account the clinical importance of the different component events. These methods include the hazard ratio for multiple event times (Wei, Lin and Weissfeld, 1989; Wei and Glidden, 1997), a measure of total disease burden (Samson et al., 2010), the number of days alive and out of hospital (Ariti et al., 2011), the proportion in favor of treatment (Buyse, 2010), the net chance of a longer survival time (Péron et al., 2016), a weighted composite endpoint (Bakal, Westerhout and Armstrong, 2015; Duc and Wolbes, 2016), and the win ratio (Pocock et al., 2012; Luo et al., 2015; Bebu and Lachin, 2016; Wang and Pocock, 2016; Dong et al., 2016; Oakes, 2016; Luo et al., 2017).

 The win ratio in particular is appealing for its simplicity and straightforward interpretation. Pocock et al. (2012) introduced it in two versions: a matched pairs approach and an unmatched approach. In the matched pairs approach, patients in the treatment and the control group are matched on their risk profiles. For any pair, the time to the most important event (say, death) determines the “winner” (e.g. the one who died later). If the winner cannot be determined (e.g., both patients remain alive), the second-most important event is considered (say, graft loss), and so forth until the winner can be determined; otherwise the pair is tied. This comparison process based on the prioritized outcomes can be referred as *prioritized pairwise comparison* (or *layered comparison procedure* as Luo et al. (2017) described). The unmatched approach applies the same principle to all possible patient pairs between the treatment group and the control group (i.e. each patient in the treatment group is compared with every patient in the control group); it does not require a matching strategy and may therefore be more widely applicable in practice. The win ratio is the ratio of winners to losers for the treatment group (Pocock et al, 2012), or equivalently, the ratio of winners for the treatment group to winners for the control group (Dong et al. 2016). Hence, the win ratio is effectively a worst-event composite rather than a first-event composite. Values greater than 1 indicate a favour for treatment over control.

In recent years, the win ratio has enjoyed increasing attention, in applications as well as methodologically. It has been applied in a number of post-hoc clinical trial analyses (Rogers et al., 2014; Prakash et al., 2014; Bakal et al., 2015; Abdalla et al., 2016; Dong et al., 2016), and explored in a pneumonia trial in patients on mechanical ventilation (Montgomery, Abuan and Kollef, 2014). One of the largest cell therapy trials in heart failure patients to date prespecified the use of the win ratio as a secondary endpoint (Henry et al., 2016). Most methodological advances have been put forward for the unmatched approach. Luo et al. (2015) developed a closed-form variance estimator for the win ratio based on a two-component composite and a specific algorithm defining winners, losers, and ties. Bebu and Lachin (2016) and Dong et al. (2016) proposed inference methods for the win ratio that are valid irrespective to a rule (an algorithm) of how winners, losers and ties are determined, so that users are free to define this in any way based on their particular situation. In addition, Dong et al. (2016)’s simulations demonstrated that for the case when there are high event rates of less important outcomes and low event rates of more important outcomes, the conventional analysis (e.g. odds ratio) may show a benefit of a treatment group driven by the less important outcomes; however, the win ratio analysis considering the importance order among the prioritized outcomes may reveal that the other treatment group is actually better. Wang and Pocock (2016) introduced the win ratio for continuous non-normal outcomes. Oakes (2016) presented expressions for the win and loss probabilities for general bivariate survival models when follow-up of all patients is limited to a specified time horizon. Luo et al. (2017) discussed weighted win and loss for a non-stratified setting.

The rule (algorithm) defining winners, losers and ties is a key operation of the win ratio. However, there is no a general rule that can fit all disease indications. Therefore, interactions with clinical study stakeholders to define the rules are a must. As an example, for cardiovascular studies with the primary composite endpoint of cardiovascular (CV) death or hospitalization for chronic heart failure, Pocock et al. (2012) illustrated that the patient who had a longer time from randomization to CV death is a winner; if the patient without a CV death was censored before the CV death from the other patient within a pair, the winner cannot be determined based on the CV death outcome. Dong et al. (2016) provided an application with a phase III liver transplant study with the composite endpoint of treated biopsy-proven acute rejection (tBPAR) or graft loss or death. They argued that given the very low mortality rate (e.g., 3.8% vs. 2.6% in the two groups), if a patient dropped out before the death event in the other patient, it was assumed that the drop-out was alive beyond the death of that patient. As another example, Peron et al. (2016) discussed the net chance of a longer progression-free survival in an oncology setting by at least m months where m months is considered clinically worthwhile and relevant to the patients.

Often, clinical trials use stratified randomization to prevent an imbalance between treatment groups on known factors that influence prognosis or treatment outcome. The stratification variables are then typically considered in subsequent data analysis. Also, it makes sense to perform a stratified analysis with prognostic variables in a trial that was randomized without stratification. Cochran (1983), Gail, Wieand and Piantadosi (1984), and Lachin (2011) among others showed that analyzing stratified data with marginal unadjusted methods may lead to biased estimates; therefore, stratified analyses are often preferred. In this paper, we propose a stratified win ratio statistic, motivated by the correspondence to the Mantel-Haenszel stratified odds ratio (Mantel and Haenszel, 1959). We formulate a homogeneity test that assesses whether the stratum-specific win ratios are homogeneous across strata. We briefly illustrate the stratified win ratio with a case study, then explore its statistical properties in simulation studies.

# The stratified win ratio

Consider a clinical trial with patients randomized into two groups with *M* strata. Let Nt(m) and Nc(m) denote the number of patients in the treatment group and the control group, respectively, and N(m) = Nt(m) + Nc(m) the total sample size, in the mth stratum (m = 1, 2, …, *M*). Let Xi(m) be the prioritized outcome information for the ith patient in the treatment group, and Yj(m) be the prioritized outcome information for the jth patient in the control group, in the mth stratum (i = 1, 2, … Nt(m);j = 1, 2, … Nc(m);m = 1, 2, …, *M*). Let K be a kernel function such that K(Xi(m), Yj(m)) = 1 if Xi(m) “wins” over Yj(m) and 0 otherwise, following *prioritized pairwise comparison* process, and similarly L be a kernel function such that L(Xi(m), Yj(m)) = 1 if Yj(m) “wins” over Xi(m) and 0 otherwise. Using the partial ordering notation by Bebu and Lachin (2016),  and. K and L should be compatible in the sense that K(Xi(m), Yj(m)) + L(Xi(m), Yj(m)) is either 1 or 0 (in case of a tie), but it is never 2. Then, in the mth stratum, the total number of wins for the treatment group is, while the total number of wins for the control group is.

In principle, the idea of the stratified win ratio analysis is to adjust or control the prognostic variables (i.e. strata). The data are divided into strata first that are relatively homogeneous with respect to the prognostic variables. Then the winners are counted separately within each stratum (i.e. stratum-specific winners) and the potential confounding effect of the prognostic variables is eliminated from the analysis. Subsequently the stratum-specific winners are combined to estimate the stratified win ratio. We define the (weighted) *stratified win ratio* as

 . (1)

 In the unstratified situation (*M* = 1), this definition reduces to nt(1)/nc(1), the original win ratio proposed by Pocock et al. (2012). Luo et al. (2015) developed a closed-form variance estimator for this unstratified win ratio based on a two-component composite and a specific algorithm defining winners, losers, and ties. Bebu and Lachin (2016) and Dong et al. (2016) derived an asymptotic distribution of this unstratified win ratio that are valid irrespective to a rule (algorithm) of how winners, losers and ties are determined, so that users are free to define this in any way based on their particular situation.

How should the weights *w(m)* be chosen? The simplest possible choice is *w(m)*=1 for all *m*, which gives the same weight to all strata. This was used by Abdalla et al. (2016); we call it the unweighted (stratified) win ratio. We propose *w(m)* = 1/N(m), for its straightforward interpretation as the reciprocal of the stratum size, as well as for its correspondence to the Mantel-Haenszel stratified odds ratio estimator, as we explain below. Other choices are possible; some will be briefly covered in the Discussion.

To motivate our choice, consider a simplified situation with only one stratum (*M*=1), one endpoint (an undesired event), and ignoring any time-to-event information. To compute the win ratio, each patient in the treatment group is compared with every patient in the control group. We only consider whether a patient experienced an event or not, as in Table I (ignoring the stratum superscript). A treatment-control patient pair counts as a “win” for the treatment group if the control patient had an event but the treatment patient did not, and similarly vice versa. If both had an event, or both had not, the pair is tied. In this simplified situation, the number of wins for the treatment group is *nt = n11 n22*, and the number of wins for the control group is *nc = n21 n12*. Hence, the win ratio reduces to the odds ratio in favor of treatment:

. (2)

 The choice of *w(m)* = 1/N (m) extends this correspondence to the stratified situation. With *M* > 1, a single endpoint without the time-to-event information considered, and the same definition of winning pairs as above (now within stratum, as in Table I), the stratified win ratio is

 (3)

This is the same as the stratified odds ratio estimator by Mantel and Haenszel (1959); it gives more weight to bigger strata (e.g. Agresti, 2002 and 2013). In general, if the importance order among the components of the composite endpoint and time-to-event information are ignored, the win ratio is reduced to the odds ratio.

# Distribution of the stratified win ratio

To derive an asymptotic distribution of the stratified win ratio, we assume fixed *M* and fixed weights *w(m)* for all strata m = 1, 2, …, M. Bebu and Lachin (2016) (with little reparameterization from the winners over the number of  pairs) and Dong et al. (2016) showed that the numbers of wins for the treatment and control groups are asymptotically normally distributed. This can be applied to a stratum, namely  and  are asymptotically normally distributed as follows:

, (4)

where *AN* denotes an asymptotic normal distribution. For a convenience to readers as well as to supplement this paper, we include the mean, variance and covariance estimators (i.e. ,  , , and ) by Dong et al. (2016) in Appendix A.1.

Across strata, we use the fact that each stratum is an independent sample for clinical trials with a stratified randomization. For any two different strata m and n (m = 1, 2, …, M; n = 1, 2, …, M; m ≠ n), nt(m) and nt(n) are independent, nc(m) and nc(n) are independent, and nt(m) and nc(n) are independent. This allows the derivation of the asymptotic distributions of the numerator and the denominator of the stratified win ratio, as well as their covariance, by elementary operations. As shown in Appendix A.2, the asymptotic joint distribution of the numerator and the denominator of the stratified win ratio is

. (5)

Following the delta method, we can obtain an asymptotic joint distribution of the logarithm of the numerator and the denominator:

, (6)

where and . Therefore, the logarithm of the stratified win ratio WR is asymptotically normally distributed as:

, (7a)

where , (7b)

. (7c)

For the variance estimator, the users can plug in those (i.e., , and ) proposed by Dong et al (2016), Bebu and Lachin (2016) with little reparameterization or new estimators that may be developed in the future. When w(m) = 1/N(m) for our proposed Mantel-Haenszel-type stratified win ratio,

, (8)

and when w(m) = 1 for the unweighted stratified win ratio,

. (9)

For this logarithm of the stratified win ratio, the point estimate is . For its variance estimate , in addition to plugging in , , and ,  and can be estimated by  under the null hypothesis (H0) of the same treatment effect in the two groups, for which the subscript ‘0’ denotes the estimate under H0 (also see a discussion on the estimate under the alternative hypothesis (H1) in Section 7).

# A homogeneity test

The Breslow-Day test (1980) is typically used to assess whether the odds ratios across strata are homogeneous. However, for the win ratio, we have not seen a way to derive the correspondence of the Breslow-Day homogeneity test. One way to assess the homogeneity of the win ratios across strata could be to apply the following inverse-variance weighted approach by Cochran (1954):

, (10)

where , is the *inverse-variance weighted (log) win ratio*, and $w\_{IV}^{\left(m\right)}={1}/{σ\_{log\left(WR^{\left(m\right)}\right)}^{2}}$ is the inverse-variance weight for the mth stratum (see Appendix A.1 for $σ\_{log\left(WR^{\left(m\right)}\right)}^{2}$). Again, plug-in estimators can be used. With large stratum sizes, the Q statistic approximately follows a distribution with *M-1* degrees of freedom under the null hypothesis that all stratum-specific win ratios WR(m) (m = 1, 2, …, M) are equal.

While the Q statistic provides an opportunity to perform the homogeneity test for the win ratio, it should be used with caution. In general, homogeneity tests tend to have low power. This holds for the Breslow-Day test as well as for Cochran’s inverse-variance weighted approach (Sutton et al., 2000; Shuster, Guo and Skyler, 2012). Using a significance level of 0.10 rather than 0.05 is therefore often recommended (e.g., Sutton et al., 2000). Conversely, even if a homogeneity test rejects its null hypothesis, a summary measure such as an estimated stratified odds ratio can still be useful when the stratum-specific odds ratios don’t vary much (e.g. Agresti, 2002 and 2013). This can also be applied to the proposed stratified win ratio as well due to its similarity to the Mantel-Haenszel stratified odds ratio. That is, if the win ratios do not vary dramatically across strata, the stratified win ratio is still a useful summary of the stratified data. We refer to Hoaglin (2016) for further cautionary notes on the use of “Cochran’s Q”. One should bear in mind that it may not always be necessary to use a formal test for heterogeneity. We suggest the inverse-variance weighted approach (Cochran, 1954) to test the homogeneity of win ratios across strata if a homogeneity test is desired. This is because the Cochran approach is used frequently in meta-analyses (e.g. Sutton et al., 2000) and a better statistical test for the win ratio homogeneity is not available yet.

# An application

The CHARM program (Pfeffer et al., 2003; Pocock et al. 2005) was designed as 3 separate randomized trials comparing candesartan with placebo in patients with chronic heart failure (CHF) who (1) were intolerant to angiotensin-converting enzyme (ACE) inhibitor and had left ventricular ejection fraction (LVEF) ≤ 0.40 (CHARM-Alternative trial); (2) were on ACE inhibitor and had LVEF ≤ 0.40 (CHARM-Added trial); or (3) had LVEF > 0.40 (CHARM-Preserved trial). The primary composite efficacy endpoint for each trial was cardiovascular (CV) death or hospitalization for CHF. The CHARM program eventually included 7599 subjects with a median follow-up time of 3.14 years.

Table II shows the primary composite efficacy endpoint and its components from the CHARM program, following Pocock et al. (2012). The number of CV deaths was overall similar to the number of CHF hospitalizations, despite some imbalance in CHARM-Preserved. However, in the conventional time-to-first-event analysis, only about half of the CV deaths (54%) contributed to the primary composite endpoint – the rest were ignored because they occurred after a CHF hospitalization.

Pocock et al. (2012) pointed out that a win ratio analysis can make greater use of CV deaths if these are considered clinically more important than CHF hospitalizations. They presented results from the matched pairs approach, with patients matched on a risk score derived from 32 baseline covariates. Table II shows the results that we obtained with the unmatched approach, by study and overall. The stratified win ratio (95% CI) is 1.179 (1.087, 1.280), 1.190 (1.098, 1.290) and 1.189 (1.096, 1.288) for the unweighted, inverse-variance weighted and the proposed stratified win ratio, respectively. We also include the win ratio on the “naïve pool” (treating the three studies as one), which yielded win ratio = 1.184 with 95% CI (1.093, 1.284). Interestingly, the proposed stratified win ratio using Mantel-Haenszel-type weights is practically identical to the inverse-variance weighted win ratio, while the unweighted stratified win ratio and the win ratio on the “naïve pool” are very slightly different.

# Simulation studies

## Simulation Study 1

In Section 2, we have theoretically demonstrated that the point estimate of our proposed stratified win ratio is same as that of the Mantel-Haenszel stratified odds ratio for the simplified case when the win ratio reduces to the odds ratio. To assess the asymptotic variance of our proposed stratified win ratio vs the variance of the Mantel-Haenszel stratified odds ratio by Robins, Breslow and Greenland (1986) for the simplified case, as well as to assess the performance of homogeneity test by the inverse-variance weighted approach (Cochran, 1954) vs the Breslow-Day method (1980), we conduct a Monte Carlo simulation study to only consider a single study endpoint and ignore the time-to-event information (i.e. a binary outcome). We assume that (1) there are 4 strata and (2) the true event rate is 50% in any stratum in the control group, and the true event rate is 40%, 30%, 20% and 10% in the Stratum 1, 2, 3 and 4 in the treatment group. To evaluate the impact of stratum sizes, we choose the sample size = 25, 50, 100 and 500 for the Stratum 1, 2, 3 and 4 respectively in each group.

 Table III presents the results from one simulated data. The point estimate of the win ratio is identical to that of the odds ratio within each stratum, and the point estimate of the proposed stratified win ratio is identical to the Mantel-Haenszel stratified odds ratio, as also demonstrated in Section 2. The 95% asymptotic confidence intervals of the stratum-specific win ratios and the proposed stratified win ratio are almost identical to those of the corresponding stratum-specific odds ratios and the Mantel-Haenszel stratified odds ratio. For the homogeneity test, the inverse-variance weighted approach by Cochran (1954) and the Breslow-Day method (1980) show almost identical results (e.g. Q = 3.74 vs 3.61, P-value = 0.291 vs 0.307).

We perform Monte Carlo simulations to systematically compare the win ratio and the odds ratio. Based on the assumed event rates as aforementioned, we simulate 5000 datasets. As shown in Figure 1, the upper and lower limits of the asymptotic 95% confidence intervals for the proposed stratified win ratio are almost identical to those for the Mantel-Haenszel stratified odds ratio. Therefore, for the situation when the win ratio reduces to the odds ratio, the proposed stratified win ratio can provide the result very similar to the Mantel-Haenszel stratified odds ratio (also see a further discussion in Section 7).

 With respect to the homogeneity of win ratios across strata, the Q value and p-value are shown in Figure 2. In general, the two Q values (per the inverse-variance weighted approach vs. per the Breslow-Day method) are very similar when the homogeneity null hypothesis is not rejected. Only when the heterogeneity is apparent (e.g. Q > 10 corresponding to P-value < 0.018), the difference between the two Q values becomes noticeable. With respect to the two corresponding p-values (per the inverse-variance weighted approach vs. per the Breslow-Day method), they are similar for most cases, however, the inverse-variance weighted approach may provide a smaller p-value.

## Simulation Study 2

To assess the performance of our proposed stratified win ratio, we conduct another Monte Carlo simulation study to compare the stratified win ratio with the inverse-variance weighted win ratio and the unweighted win ratio. For illustration purpose, we consider a clinical study with a composite endpoint of three components: endpoint 1, endpoint 2 and endpoint 3 in the clinical importance order from the least important to the most important.

### A large sample

We first consider a relatively large sample, for which we set stratum size to 100, 150 and 75 for the Stratum 1, Stratum 2 and Stratum 3, respectively for each group. The assumed true event rates are listed in Table IV. For a simplicity, we simulate the times to events by the absolute value of 40 + 40\*R, 50 + 50\*R and 80 + 80\*R for endpoint 1, endpoint 2 and endpoint 3 respectively, and censoring time by the absolute value of 100 + 100\*R for all strata in both groups. Here R is a random variable following the standard normal distribution.

Based on the assumed true event rates and the times to events, we simulated 5000 datasets. As shown in Figure 3 for the histograms and the fitted normal distribution density curves, the logarithms of the proposed stratified win ratio, the inverse-variance weighted win ratio and the unweighted win ratio are all reasonably normally distributed. The summaries of the win ratios are presented in Table V. For the stratum-specific win ratio, the median value (95% CI) is 2.64 (1.63, 4.71), 1.25 (0.88, 1.78) and 2.64 (1.51, 5.14) in Stratum 1, Stratum 2 and Stratum 3, respectively. Although the win ratio in the Stratum 2 shows a numerical benefit of the treatment group vs the control group, the difference between the two groups is not statistically significant. For the stratified win ratios, the median value (95% CI) is 1.78 (1.38, 2.29), 1.76 (1.37, 2.28) and 1.62 (1.25, 2.12) for the proposed stratified win ratio, the inverse-variance weighted win ratio and the unweighted win ratio, respectively. Therefore, all three stratified win ratios show that the treatment group is statistically better than the control group.

For this relatively large sample, our proposed stratified win ratio and the inverse-variance weighted win ratio perform very similar (Figure 4) as their distributions are very similar as shown in Figure 3. This is due to the fact that the weighting of the two approaches are similar when the sample size is relatively large and the data are not sparse. The same finding has been reported for the Mantel-Haenszel stratified odds ratio and the inverse-variance weighted odds ratio (e.g. Sutton et al., 2000). However, overall the unweighted win ratio is lower compared to the other two methods (Figure 4). This can be explained from the results from one simulated sample as an example presented in Table VI. The stratum-specific win ratio is 2.47, 1.48 and 3.95 in Stratum 1, Stratum 2 and Stratum 3, respectively. To calculate the proposed stratified win ratio, the weights of 26.5%, 60.8% and 12.6% are applied to the 3 stratum-specific win ratios. As the most weight (60.8%) is applied to Stratum 2, the largest stratum, the proposed stratified win ratio (2.05) is closer to the win ratio in Stratum 2. However, for the unweighted win ratio, an even higher weight (71.7%) is applied to Stratum 2. Therefore, the treatment effect in Stratum 2, which is the lowest, is over-contributed to the unweighted win ratio, thus the unweighted win ratio is lower than the proposed stratified win ratio (1.87 vs 2.05). On the other hand, our simulations reveal that the treatment effect in a large stratum can be under-contributed to the unweighted win ratio in some other cases. Cochran (1983), Gail, Wieand and Piantadosi (1984), and Lachin (2011) among others have demonstrated that the use of marginal unadjusted methods in the analysis of stratified data can lead to a biased estimate.

### A sparse data sample

Similar to the sparseness of contingency tables for odds ratio analyses, sparse data in the win ratio setting can be referred to as relatively few winners for the treatment or control group in a stratum. It may occur typically when there is a large number of strata relative to the total sample size in a clinical trial. To show the performance of our proposed stratified win ratio for sparse data, we simulate datasets with 5 strata based on the assumed true event rates listed in Table VII. For a simplicity, we simulate the times to events by the absolute value of 40 + 40\*R, 50 + 50\*R and 80 + 80\*R for endpoint 1, endpoint 2 and endpoint 3 respectively, and censoring time by the absolute value of 100 + 100\*R for all strata in both groups. In order to generate sparse data samples, we use small stratum sample sizes ranging from 15 to 30 (Table VII).

 Our simulations reveal that the proposed stratified win ratio, the inverse-variance weighted win ratio and the unweighted win ratio can be similar when the data are sparse and homogeneous. Table VIII presents results of one simulated dataset as an example. The calculated stratum-specific win ratio is 2.49, 1.32, 1.68, 2.50 and 1.42 for the Stratum 1, 2, 3, 4 and 5 respectively, which result in Cochran Q = 1.03 with corresponding p-value = 0.905. Therefore, the win ratios across strata are homogeneous. For this homogeneous sparse data, the proposed stratified win ratio is 1.75 (95% CI: 1.05, 2.92), which is almost same as the inverse-variance weighted win ratio 1.73 (95% CI: 1.04, 2.88) and the unweighted win ratio 1.79 (95% CI: 1.05, 3.04). For a conventional analysis, the Mantel-Haenszel stratified odds ratio 2.00 (95% CI: 1.09, 3.67) is also very similar to the inverse-variance weighted odds ratio 1.99 (95% CI: 1.07, 3.67).

 Table IX presents results of one simulated dataset as an example of sparse and heterogeneous data. The calculated win ratio is 3.86, 25.43, 2.27, 2.69 and 1.04 for the Stratum 1, 2, 3, 4 and 5, respectively. Obviously, the win ratios across strata are heterogeneous (Cochran Q = 9.66 with corresponding p-value = 0.047). For this data, the proposed stratified win ratio is 3.08 (95% CI: 1.72, 5.52), which is lower than the inverse-variance weighted win ratio 3.34 (95% CI: 1.88, 5.93) and the unweighted win ratio 3.33 (95% CI: 1.81, 6.11). For a conventional analysis, the Mantel-Haenszel stratified odds ratio is also very different from the inverse-variance weighted odds ratio. In general, our simulations reveal that the proposed stratified win ratio, the inverse-variance weighted win ratio and the unweighted win ratio may greatly differ when the data are sparse and heterogeneous. It is well known that the Mantel-Haenszel method (e.g. for stratified odds ratio) is robust and generally preferred compared to the inverse-variance weighted method when data are sparse. Given the similarity between our proposed stratified win ratio and the Mantel-Haenszel stratified odds ratio, our proposed stratified win ratio is preferred particularly when the data are sparse.

# Discussion

In this paper, we have discussed the stratified win ratio, which is a useful concept that enables inference for a composite endpoint accounting for the clinical importance order of its components and for stratification of the study population. We identified a simplified situation such that the win ratio reduces to the odds ratio when the importance order among the components of a composite endpoint and time-to-event information are ignored. It is well-known that the Mantel-Haenszel stratified odds ratio is robust compared to other stratified odds ratio estimators such as logit and maximum likelihood estimators when the number of strata is large or the data are sparse (e.g. Agresti, 2002 and 2013). Therefore, we propose a stratified win ratio in a similar way as the Mantel-Haenszel stratified odds ratio. Simulation Study 1 shows that the proposed Mantel-Haenszel-type stratified win ratio performs very similarly or almost identically to the Mantel-Haenszel stratified odds ratio when the win ratio reduces to the odds ratio. However, in terms of the win ratio versus the odds ratio within a stratum, we noticed that if a cell count in a 2×2 table is around 5 or below in a small stratum (e.g. 25 patient per group), the confidence interval of the win ratio can be slightly narrower than that for the odds ratio, and the difference between the two confidence intervals becomes noticeable when the win ratio is > 4.0. As shown from Simulation Study 2, our proposed stratified win ratio and the inverse-variance weighted win ratio are similar when the sample is large; the unweighted win ratio could overestimate or underestimate the treatment effect with an over-contribution or a under-contribution of the treatment effect from a large stratum; and the proposed stratified win ratio can perform better when the data are sparse and heterogeneous. Per our simulations, the data can be considered sparse when the number of winners for a stratum is less than or equal to 25 (vs. cell count ≤ 5 in the odds ratio setting). In addition, similar to the Mantel-Haenszel stratified odds ratio regarding zero events (e.g. Dong (2017)), the strata with 0 wins for both groups do not have a contribution to our proposed Mantel-Haenszel-type stratified win ratio. Certainly, further research on how to handle rare events in the win ratio setting is needed.

We do not see a way to perform simulations with a true win ratio rather than with a true event assumed. Therefore, the bias or mean square error of the proposed stratified win ratio, inverse-variance weighted win ratio and unweighted win ratio vs the true win ratio cannot be assessed. However, from a meta-analysis point of view, there have been extensive discussions on weighted vs unweighted approaches for point estimates as well as for homogeneity tests (e.g. an early discussion by DerSimonian and Laird (1986)). It looks clear that the weighted approach is more recommended and widely used. Given the similarity between our proposed stratified win ratio and the Mantel-Haenszel stratified odds ratio, our proposed stratified win ratio is preferred particularly when the data are sparse.

For the variance estimatorof the stratified win ratio, the users can plug in variance and covariance estimators (i.e., , and ) proposed by Dong et al (2016), Bebu and Lachin (2016) with little reparameterization or new estimators that may be developed in the future, and these plug-ins can be under the null hypothesis (H0) or under the alternative hypothesis (H1) as a referee pointed out. Under H1, a natural choice is maximum likelihood estimators such as , , , ,  and . Dong et al (2016)’s simulations compared variance estimates per their estimator under H0 with those per bootstrap under H1, they reported that there is no a meaningful difference between the two methods unless the sample size is very small (like 25 subjects/group), for which the estimates under H0 could be anticonservative. We re-assessed the two types of estimates (under H0 vs under H1) by varying the win ratio from 0.25 to 4.0 from a practice point of view, we still do not see a meaningful difference. However, when data are sparse, the estimates under H1 may not be stable. Therefore we suggest the variance estimator under H0. The details of this re-assessment can be found in Dong et al. (2017).

The general form of the stratified win ratio defined in (1) also depends on a selection of weights *w(m)*, corresponding to the different strata m = 1, 2, …, *M*. We have proposed to use the Mantel-Haenszel-type weight *w(m)* = 1/N(m), for its straightforward interpretation and correspondence to the Mantel-Haenszel stratified odds ratio (Mantel and Haenszel, 1959). Other choices are possible. We already mentioned *w(m)*=1 (for all *m*), which gives the same weight to all strata and had been used by Abdalla et al. (2016). Alternatively, one may argue that the win ratio is constructed based on the number of possible treatment-control pairs Nt(m)×Nc(m) in the mth stratum, and hence, *w(m)* = 1/(Nt(m)×Nc(m)) appears as an interesting choice. Another set of weights can be *w(m)* = 1/(nt(m)+nc(m)) based on the total number of wins from both groups; however, these weights are data-dependent, and nt(m) andnc(m) themselves are random variables. Therefore, an asymptotic distribution of the stratified win ratio with these weights would be difficult to derive. A comprehensive investigation of different weights may be of interest for future research, but the proposed Mantel-Haenszel-type weights appear reasonable, given the well-known robustness of the Mantel-Haenszel stratified odds ratio (e.g. Agresti, 2002 and 2013) and the similarity between the win ratio and the odds ratio.

To derive an asymptotic distribution of the stratified win ratio, we have assumed that the weights *w(m)* (m = 1, 2, … , M) are fixed, the strata are pre-specified, and each stratum is an independent sample. For an observational study or a clinical trial without a stratified randomization, one needs to assess implications first to see if the stratified win ratio can be applied. Maximum likelihood estimation of the common odds ratio assumes that the margins of treatment group × strata are fixed, so that the likelihood is the product of the independent binomials (i.e. the total number of binomials = treatment group × strata). The asymptotic variance of the Mantel-Haenszel stratified odds ratio Robins, Breslow and Greenland (1986) derived assumes that the number of strata *M* is fixed. This assumption on the number of strata is also applicable to the stratified win ratio. To have a good statistical performance, the asymptotic variance by Robins, Breslow and Greenland (1986) requires that the total number of strata is large if the data are sparse. This requirement may also apply to the proposed stratified win ratio due to its similarity to the Mantel-Haenszel stratified odds ratio.

We have also formulated a homogeneity test that assesses whether the stratum-specific win ratios are homogeneous across strata. While this is typically done by the Breslow-Day test (Breslow and Day, 1980) for the odds ratio, a corresponding test for the win ratio is more difficult to derive. We have adopted the inverse-variance weighted approach by Cochran (1954), which is often used in meta-analysis. As Hoaglin (2016) pointed out for the odds ratio, this may partly be due to using an incorrect null distribution; Kulinskaya and Dollinger (2015) eventually wind up recommending the Breslow-Day test instead. Conversely, the stratified win ratio can still be useful irrespective of any homogeneity test if the stratum-specific win ratios don’t vary much. Finally, since the power of the Cochran approach is usually low anyway (e.g. Sutton et al., 2000; Shuster, Guo and Skyler, 2012), one referee pointed out that one may instead consider a consistency test, testing the null hypothesis that the smallest stratum-specific win ratio is at least as large as C times the largest, with C in (0,1) often set to 0.5.

In addition to stratified analyses, the Mantel-Haenszel method is often used for meta-analyses on binary outcomes under the framework of the fixed effect model. Similarly, our proposed stratified win ratio can be utilized for meta-analyses, for which study can be considered as a stratification variable.

# References

1. Abdalla S, Montez-Rath ME, Parfrey PS, Chertow GM. (2016). The win ratio approach to analyzing composite outcomes: An application to the EVOLVE trial. *Contemp Clin Trials* 48:119-124.
2. Agresti A. (2002). Categorical Data Analysis. 2nd ed. New York: John Wiley & Sons.
3. Agresti A. (2013). Categorical Data Analysis. 3rd ed. New York: John Wiley & Sons.
4. Ariti CA, Cleland JG, Pocock SJ, Pfeffer MA, Swedberg K, Granger CB, McMurray JJ, Michelson EL, Ostergren J, Yusuf S. (2011). Days alive and out of hospital and the patient journey in patients with heart failure: insights from the candesartan in heart failure: assessment of reduction in mortality and morbidity (CHARM) program. *Am Heart J.* 162(5):900-906.
5. Bakal JA, Roe MT, Ohman EM, Goodman SG, Fox KA, Zheng Y, Westerhout CM, Hochman JS, Lokhnygina Y, Brown EB and Armstrong PW. (2015). Applying novel methods to assess clinical outcomes: insights from the TRILOGY ACS trial. Eur Heart J. 36(6):385-392a.
6. Bakal JA, Westerhout CM, Armstrong PW. (2015). Impact of weighted composite compared to traditional composite endpoints for the design of randomized controlled trials. *Stat Methods in Medical Research* 24(6):980-988.
7. Bebu I, Lachin JM. (2016). Large sample inference for a win ratio analysis of a composite outcome based on prioritized components. *Biostatistics* 17(1):178-187.
8. Breslow N. E, Day N. E. (1980). Statistical Methods in Cancer Research, Volume I: The Analysis of Case-Control Studies*,* IARC Scientific Publications, No. 32, Lyon, France: International Agency for Research on Cancer.
9. Buyse M.(2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Stat Med. 29(30):3245-3257.
10. Claggett B, Wei LJ, Pfeffer MA. (2013). Moving beyond our comfort zone. Eur Heart J. 34(12):869-871.
11. Cochran W. G. (1954). The combination of estimates from different experiments. *Biometrics* 10:101-129.
12. Cochran W. G. (1983). Planning and Analysis of Observational Studies. New York, Wiley.
13. DerSimonian R, Laird N.M. (1986). Meta-analysis in clinical trials, *Control. Clin. Trials* 7(3):177–188.
14. Dong G, Li D, Ballerstedt S, Vandemeulebroecke M. (2016). A generalized analytic solution to the win ratio to analyze a composite endpoint considering the clinical importance order among components. *Pharmaceutical Statistics*  15(5): 430–437.
15. Dong G. (2017). Meta-analysis for rare events as binary outcomes - a Chapter for New Advances in Statistics and Data Science (2016 ICSA Symposium Book)*,* Springer, 2017 (in publishing, editors: Chen D, Jin Z, Li G, Li Y and Zhao Y).
16. Dong G, Guo M, Hoaglin C. D, Vandemeulebroecke M. (2017). Is the asymptotic variance estimator of the logarithm of the win ratio under the null hypothesis sound? – A note on the generalized analytic solution to the win ratio by Dong et al. (2016). Post-publishing supportive information. *Pharmaceutical Statistics*  15(5): 430–437.
17. Duc AN, Wolbers M.(2017).Weighted analysis of composite endpoints with simultaneous inference for flexible weight constraints. *Stat Med.* 10;36(3):442-454.
18. Gail M.H, Wieand W.Y, Piantadosi S. (1984). Tests for no treatment effect in randomized clinical trials. B*iometrika* 75: 57-64.
19. Henry TD, Schaer GL, DeMaria A, Recker D, Remmers AE, Goodrich J, Patel A. (2016). The iXCELL-DCM Trial: Rationale and Design. *Cell Transplant* 25(9):1689-1699.
20. Hoaglin DC. (2016). Misunderstandings about Q and 'Cochran's Q test' in meta-analysis (with discussions). Stat Med. 35(4):485-495.
21. Kulinskaya E, Dollinger MB (2015). An accurate test for homogeneity of odds ratios based on Cochran’s Q-statistic. *BMC Medical Research Methodology* **15**:49.
22. Lachin J. (2011). Biostatistical Methods: The Assessment of Relative Risks. 2nd edition. Hoboken, NJ, Wiley.
23. Luo X, Tian H, Mohanty S, Tsai WY. (2015). An alternative approach to confidence interval estimation for the win ratio statistic. *Biometrics* 71(1):139-145.
24. Luo X, Qiu J, Bai S, Tian H. (2017). Weighted win loss approach for analyzing prioritized outcomes. Stat Med. 36(15):2452-2465.
25. Mantel, N., and Haenszel, W. (1959). Statistical Aspects of Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute* 22:719–748.
26. Montgomery A, Abuan T, and Kollef M. (2014). The win ratio method: a novel hierarchical endpoint for pneumonia trials in patients on mechanical ventilation. 34th International Symposium on Intensive Care and Emergency Medicine. 18-21 March 2014, Brussels, Belgium.
27. Oakes D. (2016). On the win-ratio statistic in clinical trials with multiple types of event. *Biometrika*  103**(**3):742–745.
28. Pfeffer MA, Swedberg K, Granger CB, Held P, McMurray JJV, Michelson EL, Olofsson B, Ostergren J, Yusuf S for the CHARM Investigators and Committees. (2003). Effects of candesartan on mortality and morbidity in patients with chronic heart failure: the CHARM-Overall programme. Lancet 362: 759–766.
29. Péron J, Roy P, Ozenne B, Roche L, Buyse M. (2016). The Net Chance of a Longer Survival as a Patient-Oriented Measure of Treatment Benefit in Randomized Clinical Trials. *JAMA Oncology*  2(7):901-905.
30. Pocock SJ, Ariti CA, Collier TJ, Wang D. (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.
31. Pocock S, Wang D, Wilhelmsen L, Hennekens CH. (2005). The data monitoring experience in the Candesartan in Heart Failure Assessment of Reduction in Mortality and morbidity (CHARM) program. Am Heart J. 149(5):939-943.
32. Prakash R, Horsfall M, Markwick A, Pumar M, Lee L, Sinhal A, Joseph MX and Chew DP. (2014). Prognostic impact of moderate or severe mitral regurgitation (MR) irrespective of concomitant comorbidities: a retrospective matched cohort study. BMJ Open 4(7):e004984.
33. Robins J. M, Breslow N, Greenland S. (1986). Estimators of Mantel-Haenszel Variance Consistent in Both Sparse Data and Large-Strata Limiting Models. Biometrics 42: 311–323.
34. Rogers JK, Pocock SJ, McMurray JJ, Granger CB, Michelson EL, Östergren J, Pfeffer MA, Solomon SD, Swedberg K, Yusuf S. (2014). Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to CHARM-Preserved. *Eur J Heart Fail*  16(1):33-40.
35. Sampson UK, Metcalfe C, Pfeffer MA, Solomon SD, Zou KH. (2010). Composite outcomes: weighting component events according to severity assisted interpretation but reduced statistical power. *J Clin Epidemiol* 63:1156–1158.
36. Shuster JJ, Guo JD, Skyler JS. (2012). Meta-analysis of safety for low event-rate binomial trials. Res Synth Methods 3(1):30-50.
37. Sutton A, Abrams K, Jones D, Sheldon T, and Song F. (2000). Methods for meta-analysis in medical research. West Sussex, England, Wiley.
38. Wang D, Pocock S. (2016). A win ratio approach to comparing continuous non-normal outcomes in clinical trials. *Pharmaceutical Statistics*  15(3):238-45.
39. Wei LJ, Lin DY, Weissfeld L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. J Am Stat Assoc. 84:1064–1072.
40. Wei LJ, Glidden DV. (1997). An overview of statistical methods for multiple failure time data in clinical trials. Stat Med. 16(8):833–839.

# Appendix

# A.1 Log(win ratio) in the mth stratum

Bebu and Lachin (2016) (with little reparameterization from the winners over the number of  pairs) and Dong et al. (2016) showed that the numbers of wins for the treatment and control groups are asymptotically normally distributed. This can be applied to a stratum, namely  and  are asymptotically normally distributed as follows:

.

To use this distribution, the users can plug-in variance and covariance estimators (i.e. $σ^{2}\_{t}^{\left(m\right)}$, $σ^{2}\_{c}^{\left(m\right)}$ and $σ\_{tc}^{\left(m\right)}$) such as those proposed by Bebu and Lachin (2016), Dong et al. (2016) or new estimators that may be developed in the future. The followings are the estimators derived by Dong et al. (2016).

$n\_{t}^{\left(m\right)}\~AN\left(ϑ\_{t}^{\left(m\right)},σ^{2}\_{t}^{\left(m\right)}\right)$

with $ϑ\_{t}^{\left(m\right)}=N\_{t}^{\left(m\right)}N\_{c}^{\left(m\right)}ϑ\_{K}^{\left(m\right)}$,

$ϑ\_{K}^{\left(m\right)}=E\left[K\left(X\_{i}^{\left(m\right)},Y\_{j}^{\left(m\right)}\right)\right]$,

$σ^{2}\_{t}^{\left(m\right)}=\frac{1}{N\_{t}^{\left(m\right)}}σ^{2}\_{1}^{\left(m\right)}+\frac{1}{N\_{c}^{\left(m\right)}}σ^{2}\_{2}^{\left(m\right)}$,

$σ^{2}\_{1}^{\left(m\right)}=\frac{N\_{t}^{\left(m\right)}N\_{c}^{\left(m\right)}}{\left(N\_{c}^{\left(m\right)}-1\right)}\sum\_{i=1}^{N\_{t}^{\left(m\right)}}\sum\_{j=1}^{N\_{c}^{\left(m\right)}}\sum\_{\begin{matrix}j'=1\\j'\ne j\end{matrix}}^{N\_{c}^{\left(m\right)}}\left[K\left(X\_{i}^{\left(m\right)},Y\_{j}^{\left(m\right)}\right)-ϑ\_{K}^{\left(m\right)}\right]\left[K\left(X\_{i}^{\left(m\right)},Y\_{j'}^{\left(m\right)}\right)-ϑ\_{K}^{\left(m\right)}\right]$,

and $σ^{2}\_{2}^{\left(m\right)}=\frac{N\_{t}^{\left(m\right)}N\_{c}^{\left(m\right)}}{\left(N\_{t}^{\left(m\right)}-1\right)}\sum\_{j=1}^{N\_{c}^{\left(m\right)}}\sum\_{i=1}^{N\_{t}^{\left(m\right)}}\sum\_{\begin{matrix}i'=1\\i'\ne i\end{matrix}}^{N\_{t}^{\left(m\right)}}\left[K\left(X\_{i}^{\left(m\right)},Y\_{j}^{\left(m\right)}\right)-ϑ\_{K}^{\left(m\right)}\right]\left[K\left(X\_{i'}^{\left(m\right)},Y\_{j}^{\left(m\right)}\right)-ϑ\_{K}^{\left(m\right)}\right]$.

Similarly,

$n\_{c}^{\left(m\right)}\~AN\left(ϑ\_{c}^{\left(m\right)},σ^{2}\_{c}^{\left(m\right)}\right)$.

The covariance of $n\_{t}^{\left(m\right)}$and $n\_{c}^{\left(m\right)}$ is

$σ^{2}\_{tc}^{\left(m\right)}=\frac{1}{N\_{t}^{\left(m\right)}}σ^{2}\_{tc1}^{\left(m\right)}+\frac{1}{N\_{c}^{\left(m\right)}}σ^{2}\_{tc2}^{\left(m\right)}$,

with $σ^{2}\_{tc1}^{\left(m\right)}=\frac{N\_{t}^{\left(m\right)}N\_{c}^{\left(m\right)}}{\left(N\_{c}^{\left(m\right)}-1\right)}\sum\_{i=1}^{N\_{t}^{\left(m\right)}}\sum\_{j=1}^{N\_{c}^{\left(m\right)}}\sum\_{\begin{matrix}j'=1\\j'\ne j\end{matrix}}^{N\_{c}^{\left(m\right)}}\left[K\left(X\_{i}^{\left(m\right)},Y\_{j}^{\left(m\right)}\right)-ϑ\_{K}^{\left(m\right)}\right]\left[L\left(X\_{i}^{\left(m\right)},Y\_{j'}^{\left(m\right)}\right)-ϑ\_{L}^{\left(m\right)}\right]$

and $σ^{2}\_{tc2}^{\left(m\right)}=\frac{N\_{t}^{\left(m\right)}N\_{c}^{\left(m\right)}}{\left(N\_{t}^{\left(m\right)}-1\right)}\sum\_{j=1}^{N\_{c}^{\left(m\right)}}\sum\_{i=1}^{N\_{t}^{\left(m\right)}}\sum\_{\begin{matrix}i'=1\\i'\ne i\end{matrix}}^{N\_{t}^{\left(m\right)}}\left[K\left(X\_{i}^{\left(m\right)},Y\_{j}^{\left(m\right)}\right)-ϑ\_{K}^{\left(m\right)}\right]\left[L\left(X\_{i'}^{\left(m\right)},Y\_{j}^{\left(m\right)}\right)-ϑ\_{L}^{\left(m\right)}\right]$.

To estimate $σ^{2}\_{t}^{\left(m\right)}$, $σ^{2}\_{c}^{\left(m\right)}$ and $σ^{2}\_{tc}^{\left(m\right)}$, $ϑ\_{K}^{\left(m\right)}$ and $ϑ\_{L}^{\left(m\right)}$need to be estimated first. Under the null hypothesis (H0) of the same treatment effect between the treatment and the control groups, $ϑ\_{K}^{\left(m\right)}$ and $ϑ\_{L}^{\left(m\right)}$can be estimated by:

,

where the subscript ‘0’ denotes the estimator under H0. With $ϑ\_{K}^{\left(m\right)}$ and $ϑ\_{L}^{\left(m\right)}$ replaced with  and  respectively in the above variance and covariance estimators, the estimates,  and can be obtained.

Following the delta method, the win ratio in the mth stratum, is asymptotic normally distributed as:

,

where .

Under H0, and can be estimated by:

.

With  and applied to to replace and respectively, the variance estimate of the win ratio in the mth stratum can be obtained.

# A.2 Derivation details of (5)

We define the stratified win ratio as follows per (1):

.

The numerator of the stratified win ratio WR asymptotically follows a normal distribution as follows:

 



Similarly for the denominator of the stratified win ratio, its asymptotic distribution is

.

Their covariance can be derived as:

 





+ …





 

 (since nt(v) and nc(u) are independent, v=1, 2, …M; u=1, 2, …,M; and v ≠ u)

 .

 

Therefore, the asymptotic joint distribution of the numerator and the denominator of the stratified win ratio WR is

.

# Supplemental material

Our SAS and R programs calculating our proposed stratified win ratio and its asymptotic confidence interval can be found in the online version of this paper at the Journal's web site.

**Table I 2×2 table in the mth stratum**

|  |  |
| --- | --- |
| Group | Number of patients |
| Without event | With event | Total |
| Treatment |  |  |  |
| Control |  |  |  |
| Total |  |  |  |

**Table II Win ratio and stratified win ratios in the CHARM program (unmatched approach)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | CHARM-Added | CHARM-Alternative | CHARM-Preserved |
| Candesartan | Placebo | Candesartan | Placebo | Candesartan | Placebo |
| Number of patients | 1276 | 1272 | 1013 | 1015 | 1514 | 1509 |
| Primary composite event | 483 | 538 | 334 | 406 | 333 | 366 |
| Hospitalization for CHF | 309 | 356 | 207 | 286 | 241 | 276 |
| CV death | 302 | 347 | 219 | 252 | 170 | 170 |
|  CV death counted in the primary composite event | 174 | 182 | 127 | 120 | 92 | 90 |
| Win ratio (95% CI) | 1.187 (1.048, 1.345) | 1.274 (1.098, 1.478) | 1.114 (0.958, 1.295) |
| Win ratio (95% CI) on pooled studies |  1.184 (1.093, 1.284) |
| Stratified win ratio (95%) | Unweighted |  1.179 (1.087, 1.280) |
| Inverse-variance weighted |  1.190 (1.098, 1.290) |
| Proposed |  1.189 (1.096, 1.288) |

**Table III Stratum-specific and stratified win ratios and odds ratios, and homogeneity test (results from one simulated data from Simulation Study 1)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stratum | Group | Number of patients | Win ratio(95% CI) | Odds ratio(95% CI) |
| Without event | With event | Total |
| 1 | Treatment | 11 | 14 | 25 | 1.40 (0.46, 4.25) | 1.40 (0.45, 4.35) |
| Control | 9 | 16 | 25 |
| 2 | Treatment | 37 | 13 | 50 | 2.85 (1.25, 6.47) | 2.85 (1.23, 6.60) |
| Control | 25 | 25 | 50 |
| 3 | Treatment | 74 | 26 | 100 | 2.85 (1.59, 5.10) | 2.85 (1.57, 5.16) |
| Control | 50 | 50 | 100 |
| 4 | Treatment | 401 | 99 | 500 | 3.83 (2.92, 5.02) | 3.83 (2.89, 5.07) |
| Control | 257 | 243 | 500 |
| Stratified | 3.41 (2.71, 4.30) | 3.41 (2.69, 4.33) |
| Homogeneity test (inverse-variance weighted approach [Cochran, 1954]): Q = 3.74, P-value =0.291Homogeneity test (Breslow-Day method): Q = 3.61, P-value =0.307 |

**Table IV Assumed true event rates (Simulation Study 2: large sample)**

|  |  |  |
| --- | --- | --- |
| Endpoint | Group 1 (n = 325) | Group 2 (n = 325) |
| Stratum 1(n = 100) | Stratum 2(n = 150) | Stratum 3(n = 75) | Stratum 1(n = 100) | Stratum 2(n = 150) | Stratum 3(n = 75) |
| Endpoint 1 | 10% | 20% | 10% | 30% | 30% | 30% |
| Endpoint 2 | 5% | 10% | 5% | 10% | 10% | 10% |
| Endpoint 3 | 5% | 10% | 5% | 10% | 10% | 10% |

**Table V Summaries of the win ratios (Simulation Study 2: large sample)**

| Stratum | Method | Summary of point estimates |
| --- | --- | --- |
|   | Median | 95% CI |
| 1 |  | 2.64 | 1.63, 4.71 |
| 2 | Stratum-specific win ratio | 1.25 | 0.88, 1,78 |
| 3 |  | 2.64 | 1.51, 5.14 |
|  | Proposed stratified win ratio | 1.78 | 1.38, 2.29 |
| Stratified | Inverse-variance weighted win ratio | 1.76 | 1.37, 2.28 |
|  | Unweighted win ratio | 1.62 | 1.25, 2.12 |

**Table VI Weights for the proposed stratified win ratio and the ‘unweighted’ win ratio (Simulation Study 2: One simulated large sample)**

| Stratum (m) | Stratum size /group | Winners in Treatment | Winners in Control | Stratum-specific win ratio | Weight for the proposed Stratified win ratio | Weight for the ‘unweighted’ win ratio |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 100 | 4104 | 1661 | 2.47 | 26.5% | 20.8% |
| 2 | 150 | 8452 | 5720 | 1.48 | 60.8% | 71.7% |
| 3 | 75 | 2345 | 594 | 3.95 | 12.6% | 7.4% |

The proposed stratified win ratio = 2.05; the unweighted win ratio = 1.87.

**Table VII Assumed true event rates (Simulation Study 2: sparse data sample)**

|  |  |  |
| --- | --- | --- |
| Endpoint | Group 1 (n = 110) | Group 2 (n = 110) |
| Stratum1 (n=30) | Stratum 2 (n=25) | Stratum 3 (n=25) | Stratum 4 (n=15) | Stratum 5(n=15) | Stratum 1 (n=30) | Stratum 2 (n=25) | Stratum 3 (n=25) | Stratum 4 (n=15) | Stratum 5(n=15) |
| Endpoint 1 | 5% | 10% | 15% | 10% | 15% | 30% | 30% | 30% | 30% | 30% |
| Endpoint 2 | 5% | 5% | 5% | 10% | 10% | 10% | 10% | 10% | 10% | 10% |
| Endpoint 3 | 5% | 5% | 5% | 10% | 10% | 10% | 10% | 10% | 10% | 10% |

**Table VIII Results of Simulation Study 2: One simulated sparse data sample resulting in homogeneous win ratios across strata**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stratum | Group | Number of patients | Number of winners | Win ratio (95% CI) |
| Without event | With event | Total |
| 1 | Treatment | 25 | 5 | 30 | 289 | 2.49 (0.93, 6.71) |
| Control | 17 | 13 | 30 | 116 |
| 2 | Treatment | 17 | 8 | 25 | 177 | 1.32 (0.52, 3.37) |
| Control | 14 | 11 | 25 | 134 |
| 3 | Treatment | 19 | 6 | 25 | 168 | 1.68 (0.61, 4.64) |
| Control | 16 | 9 | 25 | 100 |
| 4 | Treatment | 14 | 1 | 15 | 25 | 2.50 (0.27, 22.83) |
| Control | 13 | 2 | 15 | 10 |
| 5 | Treatment | 11 | 4 | 15 | 51 | 1.42 (0.36, 5.62) |
| Control | 11 | 4 | 15 | 36 |
| Proposed stratified win ratio (95% CI ): 1.75 (1.05, 2.92)Inverse-variance weighted win ratio (95% CI ): 1.73 (1.04, 2.88)Unweighted win ratio (95% CI ): 1.79 (1.05, 3.04) |
| Homogeneity test (inverse-variance weighted approach [Cochran, 1954]): Q = 1.03, P-value = 0.905 |

**Table IX Results of Simulation Study 2: One simulated sparse data sample resulting in heterogeneous win ratios across strata**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stratum | Group | Number of patients | Number of winners | Win ratio (95% CI) |
| Without event | With event | Total |
| 1 | Treatment | 27 | 3 | 30 | 189 | 3.86 (1.06, 14.00) |
| Control | 22 | 8 | 30 | 49 |
| 2 | Treatment | 24 | 1 | 25 | 178 | 25.43 (5.60, 115.55) |
| Control | 17 | 8 | 25 | 7 |
| 3 | Treatment | 18 | 7 | 25 | 220 | 2.27 (0.85, 6.07) |
| Control | 14 | 11 | 25 | 97 |
| 4 | Treatment | 12 | 3 | 15 | 70 | 2.69 (0.73, 9.89) |
| Control | 9 | 6 | 15 | 26 |
| 5 | Treatment | 12 | 3 | 15 | 28 | 1.04 (0.21, 15.24) |
| Control | 12 | 3 | 15 | 27 |
| Proposed stratified win ratio (95% CI ): 3.08 (1.72, 5.52)Inverse-variance weighted win ratio (95% CI ): 3.34 (1.88, 5.93)Unweighted win ratio (95% CI ): 3.33 (1.81, 6.11) |
| Homogeneity test (inverse-variance weighted approach [Cochran, 1954]): Q = 9.66, P-value = 0.047 |