Received: 14 February 2021 | Revised: 5 December 2021 | Accepted: 8 December 2021

DOI: 10.1002/sim.9297

RESEARCH ARTICLE

Statistics in Medicine WILEY

Sample size formula for a win ratio endpoint

Ron Xiaolong Yu[1] | Jitendra Ganju[2]

[1]Biostatistics, Gilead Sciences, Foster City, California, USA

[2]Ganju Clinical Trials, LLC,
San Francisco, California, USA

Correspondence
Jitendra Ganju, Ganju Clinical Trials, LLC, San Francisco, CA 94109, USA.
Email: ganjuclinicaltrials@gmail.com

Abstract

The win ratio composite endpoint, which organizes the components of the composite hierarchically, is becoming popular in late-stage clinical trials. The method involves comparing data in a pair-wise manner starting with the endpoint highest in priority (eg, cardiovascular death). If the comparison is a tie, the endpoint next highest in priority (eg, hospitalizations for heart failure) is compared, and so on. Its sample size is usually calculated through complex simulations because there does not exist in the literature a simple sample size formula. This article provides a formula that depends on the probability that a randomly selected patient from one group does better than a randomly selected patient from another group, and on the probability of a tie. We compare the published 95% confidence intervals, which require patient-level data, with that calculated from the formula, requiring only summary-level data, for 17 composite or single win ratio endpoints. The two sets of results are similar. Simulations show the sample size formula performs well. The formula provides important insights. It shows when adding an endpoint to the hierarchy can increase power even if the added endpoint has low power by itself. It provides relevant information to modify an on-going blinded trial if necessary. The formula allows a non-specialist to quickly determine the size of the trial with a win ratio endpoint whose use is expected to increase over time.

KEYWORDS

composite endpoints, phase 3 clinical trials, power, sample size

1 | INTRODUCTION

The primary endpoint in Phase 3 trials is often a composite of endpoints. For example, the components may include (all-cause or cardiovascular) mortality and a hospitalization related endpoint (hospitalization for heart failure or cardiovascular hospitalizations)[1–6] although other combinations are possible as well.[7] Conventionally, the endpoint has been defined as the time to the first occurrence of any component of the composite, but now attention has shifted to a hierarchical combination of endpoints.[5,8,9] The hierarchy arranges the endpoints according to priority such that an endpoint of lower priority contributes if the comparison on the endpoint of higher priority is indeterminate. Aside from evaluating endpoints hierarchically, a hierarchical combination has one another important advantage. The conventional approach combines endpoints of the same type—for example, both component endpoints are time to event. Hierarchically combining endpoints allows for combining different types of endpoints—for example, a time to event endpoint such as all-cause mortality with a recurrent event endpoint such as the *frequency* of cardiovascular-related hospitalizations.[5]

Of the three methods to combine endpoints hierarchically, the Finkelstein-Schoenfeld method,[10] Buyse's proportion in favor of treatment,[11] and the win ratio,[12] the win ratio is arguably the most popular method. A likely explanation for its popularity is that for a time to event endpoint it provides an estimate of treatment benefit that corresponds to the reciprocal of the hazard ratio.[13] The Finkelstein-Schoenfeld method does not provide an estimate and the proportion in favor of treatment is perhaps not as clinically interpretable as the win ratio. We show that the three methods are fundamentally similar which is one of the key results in deriving the sample size for the win ratio.

All three methods compare pairs of patient data (which is based on the Mann-Whitney idea[14]) at the minimum of follow-up times within each pair. If the comparison is indeterminate for any given pair for the endpoint highest in priority, the endpoint next highest in priority is compared, and so on. As there is no simple formula available for calculating the sample size of a win ratio endpoint, it is usually calculated through simulations. The simulation code is not straightforward because data have to be generated for multiple endpoints and drop-out patterns, pairs of patient data have to be hierarchically compared at the minimum of follow-up times, and within each run of the simulation, the data are resampled (via the bootstrap method) to calculate the variance. Trial planners interested in using the win ratio have to depend on a specialist to perform simulations to calculate the trial size. As is commonly the case, sample sizes are calculated under a range of different scenarios involving different numbers of endpoints, different treatment effects associated with each endpoint, different drop-out rates and so on. Evaluation of different scenarios is important not only from the perspective of statistical power but also for budgeting and planning timelines. Simulations evaluating several scenarios are complex and time-consuming.

We simplify the process by providing a formula for the variance and sample size of a (log-transformed) win ratio endpoint that is easy to use. The formula depends on the probability that a randomly selected patient from one group does better than a randomly selected patient from another group, and on the probability of tie. To test the reliability of the formula we: (a) Compare 95% confidence intervals (CI) calculated from the variance formula, which depends on summary-level data, with that reported in the literature, which depend on patient-level data, for 17 sets of CI. (b) Compare the power obtained from the formula with that obtained from bootstrap simulations. Both sets of comparisons demonstrate that the formula works well.

2 | AN OVERVIEW OF THE THREE METHODS

2.1 | Win ratio

Each patient in one treatment group is compared with each patient in the other treatment group. For each pair, if the comparison does not result in a tie, a winner or a loser is declared. If the comparison is indeterminate (ie, is a tie), the endpoint next highest in priority is compared in a similar manner, and so on.

Let $\#W$ denote the number of wins, $\#L$ denote the number of losses, and $\#T$ denote the number of ties. Further, let

- N denote the total sample size;
- k denote the proportion of patients allocated to one group (so $(1-k)$ is the proportion allocated to the other group);
- let $p_{\text{tie}} = \frac{\#T}{k(1-k)N^2}$ represent the proportion of ties.

Then, $k(1-k)N^2 = \#W + \#L + \#T$, where $k(1-k)N^2$ is the product of the sample sizes per group. The win ratio, which has been defined as the ratio of wins to losses, $\frac{\#W}{\#L}$, ranges between 0 and infinity. By this definition, the average value of the win ratio is the average of the ratios. This average is a function of the sample size, which is an undesirable characteristic. Therefore, we prefer defining it as the ratio of the probabilities of wins to losses[15]: $\frac{p_{\text{Win}}}{p_{\text{Loss}}}$, where p_{Win} and p_{Loss} are estimated, respectively, as $\frac{\#W}{k(1-k)N^2}$ and $\frac{\#L}{k(1-k)N^2}$. This average is the average of wins divided by the average of losses, and is invariant to the sample size.

Under the null hypothesis, the true win ratio equals 1. In the original win ratio paper,[12] the authors concluded that the variance of the win ratio was complex. They suggested to use a bootstrap resampling method. The bootstrap is a computer-based resampling procedure requiring patient-level data. Subsequently, a formula for the variance was derived[16,17]; but because it depends on patient-level data it cannot be used in its current form to calculate the sample size.

2.2 | Buyse's proportion in favor of treatment

As with the win ratio, each patient in one group is compared with each patient in the other group. The statistic proposed is the difference between number of wins and number of losses divided by the number of comparisons: $\frac{\#W-\#L}{k(1-k)N^2}$. The null hypothesis is that the proportion in favor of treatment is 0. Here, too, the original paper did not provide a formula for the variance. Buyse recommended using the permutation method, another computer-based resampling method, for drawing inferences.

2.3 | Finkelstein-Schoenfeld's method

This method compares *all* pairs of patients, including comparisons with patients belonging to the same treatment group.

For any pair of patients, let

$$u_{ij}\ (i \neq j) = \begin{cases} 1, \ \text{if patient } i \text{ does better (or wins) than patient } j \\ 0, \text{if the comparison is indeterminate} \\ -1, \text{if patient } i \text{ does worse (or loses) than patient } j \end{cases}.$$

If $u_{ij} = 1$, then it means that $u_{ji} = -1$. The rank for patient i is the sum of the u_{ij} over all pairwise comparisons of patient i with all other patients. The null hypothesis is that the sum of the ranks, called S, over any one treatment is 0.

3 | THE RELATIONSHIP AMONG THE THREE METHODS

The derivation of results in this section and the next are provided in the Appendix.

This article is based on the following pair of important expressions:

$\#W = \frac{k(1-k)N^2-\#T+S}{2}$ and that $\#L = \frac{k(1-k)N^2-\#T-S}{2}$. This result connects the number of wins and losses to Finkelstein-Schoenfeld's sum of ranks, S. Moreover, it shows that Buyse's proportion in favor of treatment can be written as $p_{\text{Win}}-p_{\text{Loss}}$. Thus, the three methods are fundamentally similar. It is through this connection that approximate formulas for the variance and sample size for the win ratio are derived. It also implies that the formula can be adapted to determining the sample size for Finkelstein-Schoenfeld's and Buyse's methods.

4 | SAMPLE SIZE

The sample size for any statistic depends on the variability of the statistic. We recommend log transforming the win ratio (recall that the Cox model also log transforms the ratio of hazard rates) because the distribution of the log transformed statistic is closer to a normal distribution than the untransformed win ratio.

The approximate variance of the log of the win ratio is

$$\text{Var}(\ln(WR)) \approx \frac{1}{N} \times \left\{ \frac{4\,(1+p_{\text{tie}})}{3k(1-k)\,(1-p_{\text{tie}})} \right\} \tag{1}$$

Using conventional notation, we write this as $\frac{1}{N} \times \sigma^2$, where σ^2 represents the term within {}.

The required sample size is

$$N \approx \frac{\sigma^2 \times \left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{\ln^2\left(WR_{\text{true}}\right)} \tag{2}$$

where WR_{true} refers to the assumed or true value of the win ratio, and $\ln\left(WR_{\text{true}}\right)$ refers to its natural log-transformed value. As usual, α and β refer, respectively, to type I and II error rates, with Z denoting the quantile value from the standard normal distribution. As an example, under 1:1 allocation, one-sided $\alpha = 2.5\%$, power = 90% or $\beta = 10\%$, a small proportion of ties, $p_{\text{tie}} = 0.1$, and 50% more wins on treatment than control, the required total sample size is 417.

The formula for power is

$$\text{Power} = 1 - \Phi\left(Z_\alpha - \ln\left(WR_{\text{true}}\right)\frac{\sqrt{N}}{\sigma}\right) \tag{3}$$

with Φ denoting the cumulative distribution function of the standard normal distribution. For weighted stratified win ratio, $WR_{\text{stratified}} = \frac{\sum_{i=1}^{M} w_i(\#W)_i}{\sum_{i=1}^{M} w_i(\#L)_i}$, where M is the total number of strata, and w_i, $(\#W)_i$ and $(\#L)_i$ are the fixed weight, number of wins and number of losses for the ith stratum respectively, after log transformation, its variance can be approximated by

$$\text{Var}\left(\ln\left(WR_{\text{stratified}}\right)\right) \approx \sigma^2 \cdot \frac{\sum_{i=1}^{M} w_i^2 N_i^3}{\left(\sum_{i=1}^{M} w_i N_i^2\right)^2} \tag{4}$$

where N_i is the number of patients within the ith stratum ($1 \leq i \leq M$). When all strata have the same weight, for example, $w_i = 1$ for all $1 \leq i \leq M$, it is called an unweighted stratified win ratio, and variance formula in (4) reduces to $\sigma^2 \cdot \frac{\sum_{i=1}^{M} N_i^3}{\left(\sum_{i=1}^{M} N_i^2\right)^2}$.

4.1 | Comparison with published results

We calculate the 95% CI as

$$\exp(\ln(WR) \mp 1.96\sqrt{\text{var}}) \tag{5}$$

where var is the variance calculated from (1) for an unstratified trial or from (4) for a stratified trial.

To test the reliability of (5) we compare it against published 95% CI. Equation (5) is based on summary-level data, whereas the published intervals are based on patient-level data. The comparisons are performed for composite endpoints (PARTNER,[2] ATTR-ACT,[5] COAPT,[18] DIG,[19] CHARM,[1] PARADIGM-HF,[4] EMPHASIS-HF,[3] and ATLAS ACS2-TIMI 51[20]) and for single endpoints (CHARM,[1] PLACIDE,[21] PARTNER,[2] and EMPHASIS-HF[3]) for which the win ratio results are published elsewhere as noted here: PARTNER,[12] ATTR-ACT,[22] COAPT,[23] DIG,[24] CHARM,[12] PARADIGM-HF,[24] EMPHASIS-HF,[12] ATLAS ACS2-TIMI 51,[17] PLACIDE.[25] Because the published result for ATTR-ACT is based on an unweighted stratified analysis, our result is based on the stratified variance formula (4). The CHARM program included three patient types which are compared: CHARM Added (ejection fraction <40% and on ACE-inhibitor), CHARM Alternative (ejection fraction <40% and intolerant to ACE-inhibitor), and CHARM Preserved (ejection fraction ≥40%). This gives a total of 17 pairs of 95% CI, of which 10 are for composite endpoints and 7 for single endpoints.

4.1.1 | Composite endpoints

Table 1 and Figure 1 show the results for the composite endpoints (the horizontal axis is on the log scale as is appropriate for displays of hazard ratios). The two sets of 95% CI are shown in the last column of the table. The results are shown in increasing order of the proportion of ties. The published results for CHARM are based on a matched analysis yet our results, which are based on an unmatched analysis, are in excellent agreement. With the exception of PARTNER and EMPHASIS-HF, the 95% CI calculated by the two methods are similar. Even when the proportion of win ratio ties is very large (ATLAS, 0.94) or the sample size is not too large (ATTR-ACT, N = 441), the two sets of results are in good agreement.

For PARTNER, the published 95% CI is (1.35, 2.54) while the formula gives (1.43, 2.45). We speculate that the interval produced by the formula is more accurate for the following reason. The two methods give Z-scores (equivalently, P-values) that are nearly equal. The published Z-score is 4.58, and with the formula it equals 4.56. The formula-based Z-score is calculated as $\ln(WR)/\sqrt{\text{var}}$, where var is given by (1). Because the Z-scores are almost equal, one would expect the CI to

TABLE 1 95% CI comparisons for composite endpoints

Trial: Endpoints	$\frac{\#W}{\#L}$	Win ratio	p_{tie}	Top row: Published 95% CI Bottom row: Formula (1) 95% CI
PARTNER: all-cause mortality, hosp.[a] N = 358, $k = 0.5$	$\frac{18,445}{9,843}$	1.87	0.12	1.35, 2.54
				1.43, 2.45
ATTR-ACT: death, frequency of CVH N = 441, $k = 0.6$	$\frac{8,595}{5,071}$	1.69	0.21	1.26, 2.29
				1.25, 2.30
COAPT: death, HF hosp. N = 614, $k = 0.49$	$\frac{42,330}{26,277}$	1.61	0.27	1.29, 2.04
				1.27, 2.04
DIG: CV death, HF hosp. N = 6800, $k = 0.5$	$\frac{4,113,387}{3,644,017}$	1.13	0.33	1.05, 1.20
				1.04, 1.22
CHARM Added: CV death, HF hosp. N = 2548, $k = 0.5$	$\frac{421}{324}$	1.30	0.41	1.13, 1.50
				1.13, 1.49
CHARM Alternative: CV death, HF hosp. N = 2028, $k = 0.5$	$\frac{316}{222}$	1.42	0.47	1.20, 1.70
				1.20, 1.68
PARADIGM-HF: CV death, HF hosp. N = 8399, $k = 0.5$	$\frac{3,672,811}{2,918,490}$	1.27	0.63	1.15, 1.39
				1.14, 1.40
CHARM Preserved: CV death, hosp. N = 3023, $k = 0.5$	$\frac{294}{251}$	1.17	0.64	0.99, 1.39
				0.98, 1.40
EMPHASIS-HF: CV death, HF hosp. N = 2737, $k = 0.5$	$\frac{338,735}{210,952}$	1.61	0.71	1.37, 1.89
				1.30, 1.98
ATLAS ACS2-TIMI 51: CV death, MI N = 9525, $k = 0.5$	$\frac{772,505}{595,754}$	1.30	0.94	1.03, 1.63
				1.00, 1.69

Note: N is the total sample size, k is the fraction of N assigned to one treatment group, CVH is cardiovascular related hospitalizations, CV death is cardiovascular death, Hosp. is hospitalization, HF is heart failure, MI is myocardial infarction. Shown in order of increasing proportion of ties.

[a] Hospitalization due to valve or procedure-related deterioration.

also be very similar. For EMPHASIS-HF the reported 95% CI is (1.37, 1.89), whereas the formula gives a wider interval of (1.30, 1.98). The reason for the discrepancy is unclear to us.

4.1.2 | Single endpoints

Table 2 and Figure 2 show the win ratio 95% CI results for the single endpoints. For CHARM, the published 95% CI is (1.02, 1.08), and that based on the formula is (1.00, 1.11). The formula-based interval seems more accurate because the proportion of ties is nearly equal to 0 for which the formula would be expected to be a good approximation. Except for EMPHASIS-HF, the results are in good agreement.

5 | COMPARISON OF FORMULA (3) WITH SIMULATIONS

Although the comparison to published CI demonstrate the variance formula works well in practice, we now perform a more comprehensive assessment via simulations. The purpose is to determine if the power from (3) is similar to the power from the bootstrap procedure. Other methods for calculating power relying on asymptotic theory for the variance (but which require patient-level data) may also be used for comparison, and the results will be similar.[24] Simulations provide the bootstrap power, the average (over all simulations) probability of wins and losses, and p_{tie}. The power from the formula is obtained by plugging in the value of the win ratio and p_{tie} in (3). Two hierarchically combined endpoints

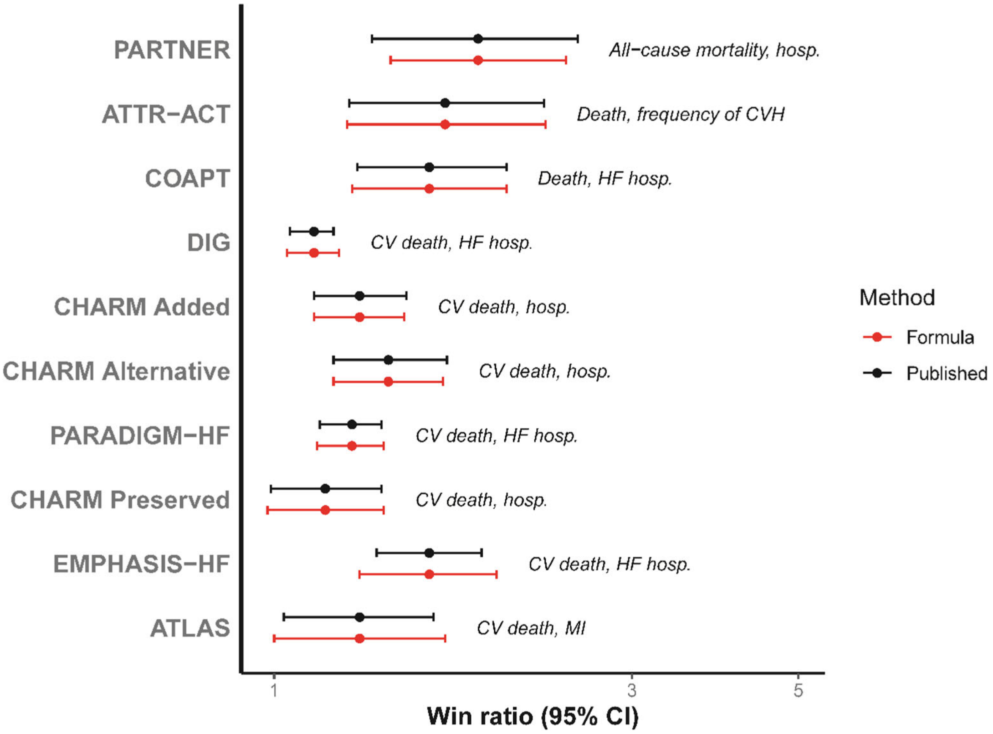


FIGURE 1 Comparison of published and formula-based 95% CI for trials with composite endpoints. Unabbreviated endpoints are noted in Table 1. The results show a close match between the formula-based and published intervals even when the proportion of ties is very large (ATLAS; 0.94) or the sample size is not too large (ATTR-ACT; N = 441)

are considered. Endpoint 1 is a time to event endpoint, and endpoint 2 is a recurrent event endpoint, as in ATTR-ACT. Power is evaluated at the conventional two-sided $\alpha = 5\%$.

5.1 | Details on the simulation

Each simulated trial is of fixed duration of 1 year for all patients. A total of between 100 and 700 patients (or between 50 and 350 patients per group) are enrolled in each simulated trial. The primary endpoint is mortality where the survival times are drawn from exponential distributions with a one-year mortality rate of 40% in the control arm and a hazard ratio (treatment vs control) that ranges between 0.50 and 0.80. The second endpoint is recurrent hospitalizations which are assumed to follow a negative binomial distribution. The dispersion parameter was varied to change the proportion of ties. Table 3 shows the parameters on which the simulation results are based.

Each patient's hospitalization event times are generated from a Poisson (λ) process (ie, time to first hospitalization and time between consecutive hospitalizations are identically and independently drawn from an exponential distribution with rate parameter λ), where λ is a patient-specific event rate and is itself a random variable drawn from a gamma distribution. The parameters of the gamma distribution are chosen such that the resulting gamma-Poisson mixture of event times is negative binomially distributed with dispersion parameters and the desired mean event rates shown in Table 3. The simulation assumes independence between survival times and hospitalization rates. Results in Table 4 are based on 5000 simulated trials for each scenario and 2000 bootstrap resamples for each simulated trial. The code for the simulations is available from ron.yu@gilead.com.

5.2 | Simulation results

Results comparing the power via formula (3) and the bootstrap procedure are shown in Table 4. To illustrate the results with an example, suppose it is believed that the proportion of ties is 0.16 and true win ratio is 1.43 (or equivalently, the

TABLE 2 95% CI comparisons for single endpoints

Trial: Endpoint	#*W* #*L*	Win ratio	p_{tie}	Top row: Published 95% CI Bottom row: Formula (1) 95% CI
CHARM: DAOH N = 7599, $k = 0.5$	7,402,980 7,011,257	1.06	~0	1.02, 1.08 1.00, 1.11
PLACIDE: Hospital stay N = 2939, $k = 0.5$	981,742 1,002,760	0.98	0.08	0.90, 1.08 0.89, 1.07
PARTNER: all-cause mortality N = 358, $k = 0.5$	14,466 8,498	1.70	0.28	1.23, 2.38 1.24, 2.34
CHARM Added: CV death N = 2548, $k = 0.5$	289 220	1.31	0.60	1.10, 1.57 1.10, 1.57
CHARM Alternative: CV death N = 2028, $k = 0.5$	202 148	1.37	0.65	1.11, 1.70 1.10, 1.70
CHARM Preserved: CV death N = 3023, $k = 0.5$	150 136	1.10	0.81	0.88, 1.39 0.86, 1.42
EMPHASIS-HF: CV death N = 2737, $k = 0.5$	163,129 124,825	1.31	0.85	1.04, 1.66 0.97, 1.76

Note: DAOH is days alive and out of hospital. CV death is cardiovascular death. Shown in order of increasing proportion of ties.

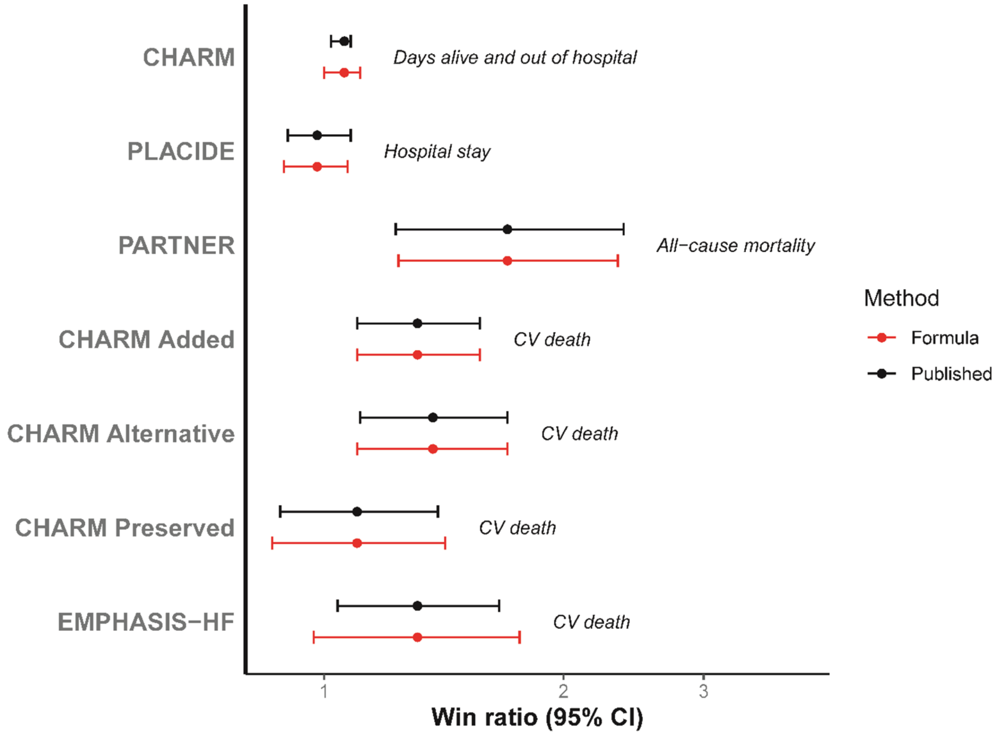


FIGURE 2 Comparison of published and formula-based 95% CI for trials with single endpoints. Unabbreviated endpoints are noted in Table 1. The results show a close match between the formula and published intervals

TABLE 3 Parameters for the two endpoints on which the results in Table 4 are based

	N = 100 $p_{\text{tie}} = 0.04$	N = 100 $p_{\text{tie}} = 0.10$	N = 100 $p_{\text{tie}} = 0.18$	N = 500 $p_{\text{tie}} = 0.16$	N = 700 $p_{\text{tie}} = 0.14$
Hazard ratio	0.60	0.60	0.60	0.70	0.80
Rate ratio	0.50	0.50	0.50	0.70	0.60
Control rate/year	5	2	1.0	1.0	1.0
Treat. rate/year	2.5	1	0.5	0.7	0.8
Dispersion	0.2	0.5	1.0	1.0	1.0

Note: p_{tie} is the proportion of win ratio ties.

TABLE 4 Simulation power vs formula power for a two-component (time to event and recurrent event) win ratio endpoint for varying proportion of ties and treatment effects

	N = 100 $p_{\text{tie}} = 0.04$	N = 100 $p_{\text{tie}} = 0.10$	N = 100 $p_{\text{tie}} = 0.18$	N = 500 $p_{\text{tie}} = 0.16$	N = 700 $p_{\text{tie}} = 0.14$
Endpoint 1 HR	0.60	0.60	0.60	0.70	0.80
Endpoint 2 RR	0.50	0.50	0.50	0.70	0.60
Win ratio	2.21	1.89	1.77	1.43	1.36
Power, %					
Bootstrap	88.8	69.6	54.9	83.9	86.2
Formula (3)	90.7	70.2	54.5	83.8	85.5

Note: HR is the hazard ratio for endpoint 1 (time to event), RR is the rate ratio for endpoint 2 (recurrent event), p_{tie} is the proportion of win ratio ties. Given the HR and RR, the simulation calculated p_{tie}, the win ratio, and power with the bootstrap procedure. The values for power via (3) are based on values of the win ratio and p_{tie} prior to rounding. The table shows the rounded values. As a result, there is a very slight difference in power calculated before and after rounding of win ratio and p_{tie}.

probability of a win equals 0.494, the probability of a loss equals 0.346). Then with 250 patients per group for a total N = 500, the power from formula (3) is 83.8%. From the bootstrap procedure, power is 83.9% (Table 4, second to last column). Both methods give similar power for a range of scenarios including when the sample size is as small as 50 patients per group (or N = 100). A small sample size was selected because formulas based on approximations tend to work better as the sample size increases. The first three columns of the table show that the formula works well even when the sample size is small. For all scenarios, the difference in the power between the two methods is within 2%.

6 | DISCUSSION

Having a formula for sample size has the obvious advantage that given the win ratio and the probability of a tie (or equivalently, given the probability of a win and the probability of a loss), no simulations are needed. A reviewer pointed us to a recent publication on the win ratio sample size.[26] The authors provide a more precise solution (ie, it is not an approximation like ours), but the formula involves a parameter called standard rank deviation (SRD), which in general requires numerical evaluation (eg, via standard numerical routines in R or Monte Carlo integration) and sometimes the evaluation can be computationally prohibitive. Our formula provides three insights. The first is that adding an endpoint to break ties may increase the power even if the win ratio is reduced by the inclusion of the endpoint. For example, suppose with two endpoints the win ratio is 1.41 and that the proportion of ties is 0.30. Then for, say, a sample size of 600 patients, allocated 1:1, power is 76% at one-sided $\alpha = 2.5\%$. Now suppose adding a third endpoint with a smaller treatment effect reduces the proportion of ties to 0, and the resultant win ratio is smaller, say 1.32. When this is the case, power increases to 84%. Thus, power may increase even if inclusion of an endpoint dilutes the win ratio. To illustrate this point further, take a single time to event endpoint analyzed the conventional way (eg, via the logrank statistic). Suppose increasing the trial duration brings the hazard ratio closer to the null because the treatment effect is time-dependent and decreases after some period, but the additional events reduce the variance sufficiently so that power increases. As a point of clarity, we note that the expectation of an improvement in power of a win ratio endpoint depends on a positive effect on all component endpoints. Power will decrease if treatment results in a worsening of outcomes in the added endpoint.

A second insight concerns the proportion of ties. If it is larger than expected, power is adversely affected. Calculation of the win ratio ties requires unblinded data because the method involves comparisons of pairs of patients between treatment groups. However, Finkelstein-Schoenfeld ties can be calculated from blinded data because the method involves comparisons between all pairs of patients. Furthermore, the proportion of win ratio ties should nearly equal to the proportion of Finkelstein-Schoenfeld ties. Thus, the proportion of Finkelstein-Schoenfeld ties calculated from blinded data (when the data are sufficiently mature) can be used as a proxy for the win ratio ties. If the proportion of ties is substantially higher than expected such that it puts the success of the trial at high risk, some consideration may be given to modifying certain aspects of the study design (eg, increasing the sample size), while the trial is still blinded. This is analogous to monitoring the event rate in a conventional time to event trial to determine if the trial is adequately powered.

The third is that the distribution of the data do not materially affect the power of the win ratio endpoint so long as the win ratio and the proportion of ties are unchanged. That is, the power will not change much if there is a lag or attenuation of effect in the benefit with treatment so long as the win ratio and the proportion of ties are unchanged. Thus, under one scenario where there is no delay in benefit, and another scenario where there is a delay in benefit, the power (for fixed sample size) for both cases will be similar if the win ratio and proportion of ties stays the same.

The proposed sample size formula has a simple form because of the assumptions made in its derivation that there is no intransitivity and that tied ranks occur in patient subgroups of equal size (see Appendix for details). Thus, the formula works well when the intransitivity rate is low, and when the win ratio is not very large because the variance is derived under the null. Extensive simulations[27] conducted on a univariate win ratio endpoint show that the variance does not increase much for win ratios <1.8 and sample sizes $\geq$ 100. Conversely, the null variance will tend to underestimate the true variance when the sample size is small or if the win ratio is very large. On both points of intransitivity and the null variance, Tables 1 and 2 show that the formula performs well when tested against data from real trials. If intransitivity is present, then the proposed formula will underestimate the true power. However, in practice, component endpoints tend to be positively correlated, and when this holds, the rate of intransitivity is expected to be low.

For trial planning purposes it is useful to have a readily available formula to determine the approximate number of patients needed for given power for a range of scenarios which can be quickly accomplished with the formula. For example, one may wish to calculate the sample size for a modest effect of treatment (eg, win ratio = 1.20-1.35), a moderate effect (eg, win ratio 1.35-1.55) or a large effect (eg, win ratio > 1.55) with guesstimates for the proportion of ties (eg, between 0.10 and 0.30). The proportion of ties will often depend on the type of endpoint lowest in priority; in particular, a recurrent event endpoint will be expected to yield a greater proportion of ties than a continuous measure.

The win ratio prioritizes composite outcomes that allows for handling of competing risks such as fatal events. As a result, use of the win ratio as a key endpoint (primary or secondary endpoint) in large, long-term trials is likely to increase. As clinical trialists develop more experience with the win ratio, they will gain a better understanding of the inputs required to calculate the sample size. To help build experience, we urge trialists to publish the win ratio even for trials for which the endpoints are analyzed by conventional methods (specifically, we recommend the excellent presentation style in the original win ratio publication[12]). The proposed formula provides important insights and allows for a quick calculation of the sample size even by the non-specialist.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Jitendra Ganju https://orcid.org/0000-0003-4247-2682

REFERENCES

1. Pfeffer MA, Swedberg K, Granger CB, et al. Effects of candesartan on mortality and morbidity in patients with chronic heart failure: the CHARM overall programme. *Lancet*. 2003;362:759-766.
2. Leon MB, Smith CR, Mack M, et al. Transcatheter aortic-valve implantation for aortic stenosis in patients who cannot undergo surgery. *N Engl J Med*. 2010;363:1597-1607.
3. Zannad F, McMurray JJV, Krum H, et al. Eplerenone in patients with systolic heart failure and mild symptoms. *N Engl J Med*. 2011;364:11-21.

4. McMurray JJV, Packer M, Desai AS, et al. Angiotensin-Neprilysin inhibition versus enalapril in heart failure. *N Engl J Med*. 2014;371:993-1004.
5. Maurer MS, Schwartz JS, Gundpaneni B, et al. Tafamidis treatment for patients with transthyretin amyloid cardiomyopathy. *N Engl J Med*. 2018;379:1007-1016.
6. Teerlink JR, Diaz R, Felker GM, et al. Omecamtic mecarbil in chronic heart failure with reduced ejection fraction: rationale and design of GALATIC-HF. *JACC Heart Fail*. 2020;8:329-340.
7. Kuck K-H, Brugada J, Furnkranz A, et al. Cryoballoon or radiofrequency ablation for paroxysmal atrial fibrillation. *N Engl J Med*. 2016;374:2235-2245.
8. https://clinicaltrials.gov/ct2/show/NCT04157751. Accessed December 10, 2020.
9. https://clinicaltrials.gov/ct2/show/NCT03860935. Accessed December 10, 2020.
10. Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. *Stat Med*. 1999;18:1341-1354.
11. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med*. 2010;29: 3245-3257.
12. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J*. 2012;33:176-182.
13. Oakes D. On the win-ratio statistic in clinical trials with multiple types of event. *Biometrika*. 2012;99(1):5.
14. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat*. 1947;18:50-60.
15. Dong G, Hoaglin DC, Qiu J, et al. The win ratio: on interpretation and handling of ties. *Stat Biopharm Res*. 2020;12:99-106.
16. Dong G, Li D, Ballerstedt S, Vandemeulebroecke M. A generalized analytic solution to the win ratio to analyze a composite endpoint considering the clinical importance order among components. *Pharm Stat*. 2016;15(5):430-437.
17. Luo X, Tian H, Mohanty S, Tsai WY. An alternative approach to confidence interval estimation for the win ratio statistic. *Biometrics*. 2015;71:139-145.
18. Stone GW, Lindenfeld J, Abraham WT, et al. Transcatheter mitral-valve repair in patients with heart failure. *N Engl J Med*. 2018;379:2307-2318.
19. The Digitalis Investigation Group. The effect of digoxin on mortality and morbidity in patients with heart failure. *N Engl J Med*. 1997;336:525-533.
20. Mega JL, Braunwald E, Wiviott SD, et al. Rivaroxaban in patients with a recent acute coronary syndrome. *N Engl J Med*. 2012;366:9-19.
21. Allen SJ, Wareham K, Wang D, et al. Lactobacilli and bifidobacteria in the prevention of antibiotic-associated diarrhoea and Clostridium difficile diarrhoea in older inpatients (PLACIDE): a randomised, double-blind, placebo-controlled, multicentre trial. *Lancet*. 2013;382:1249-1257.
22. Pocock SJ, Collier TJ. Statistical appraisal of 6 recent clinical trials in cardiology. *J M Coll Cardiol*. 2019;73:2740-2755.
23. Redfors B, Gregson J, Crowley A, et al. The win ratio approach for composite endpoints: practical guidance based on previous experience. *Eur Heart J*. 2020;00:1-9.
24. Ferreira JP, Jhund PS, Duarte K, et al. Use of the win ratio in cardiovascular clinical trials. *J M Coll Cardiol*. 2020;8:441-450.
25. Wang D, Pocock SJ. A win ratio approach to comparing continuous non-normal outcomes in clinical trials. *Pharmceut Stat*. 2016;15:238-245.
26. Mao L, Kim K, Miao X. Sample size formula for general win ratio analysis. *Biometrics*. 2021.
27. Verbeeck J, Ozenne B, Anderson WN. Evaluation of inferential methods for the net benefit and win ratio statistics. *J Biopharm Stat*. 2020;30:765-782.

How to cite this article: Yu RX, Ganju J. Sample size formula for a win ratio endpoint. *Statistics in Medicine*. 2022;41(6):950-963. doi: 10.1002/sim.9297

APPENDIX A

A.1. Derivations of results

Note: the convention is to use uppercase and lowercase letters respectively for random variables and their realized values. For simplicity in notation, we do not make that distinction

A.1.1. Derivation of the connection between the number of wins and sum of ranks

$$\#W = \frac{k(1-k)N^2 - \#T + S}{2}$$

Denote the two treatment groups as A and B. The rank for patient i is $U_i = \sum_{j\neq i=1}^{N} u_{ij}, i = 1, 2, \dots , N$. Let $S = \sum_{i\in A} U_i$ denote the sum of ranks for patients in treatment group A. Consider the rank of, say, patient 1, U_1. U_1 is the sum of all pairwise comparisons of patient 1 with every other patient, for a total of $N - 1$ comparisons. Of these $N - 1$ pairs, $kN - 1$ pairs are with patients in group A and $(1 - k)N$ comparisons are with patients in group B. U_1 is therefore $\sum_{p\in A} u_{1p}+ \sum_{q\in B} u_{1q}$. Similarly, for the rank for patient 2 $U_2 = \sum_{p\in A} u_{2p}+ \sum_{q\in B} u_{2q}$. Suppose patients 1 and 2 belong to the same treatment group, say, A. Note that in the sum $U_1 + U_2$, the components u_{12} and u_{21} cancel each other out because by construction if $u_{12} = +1$, then $u_{21} = -1$, and vice versa. In this way, the sum of all u_{ij} and u_{ji} equal 0 within the same group, where $i \neq j$.

We write $\sum_{i\in A} U_i = \sum_{i\in A} \sum_{p\in A} u_{ip} + \sum_{i\in A} \sum_{q\in B} u_{iq}$, and because $\sum_{i\in A} \sum_{p\in A} u_{ip} = 0$, $\sum_{i\in A} U_i = \sum_{i\in A} \sum_{q\in B} u_{iq}$. This represents the comparison of each patient in group A with each patient in group B for a total of $k(1 - k)N^2$ pairs of comparisons. Thus, we have

$$S = \sum_{i\in A} \sum_{q\in B} u_{iq} = \#W - \#L, \tag{A1}$$

and

$$k(1 - k)N^2 = \#W + \#L + \#T \tag{A2}$$

From (A1) and (A2), we have $S + k(1 - k)N^2 = (\#W - \#L) + (\#W + \#L + \#T) = 2(\#W) + \#T$. Therefore, $\#W = \frac{k(1-k)N^2-\#T+S}{2}$.

A.1.2. Derivation of the sample size

By the delta method, $\ln(\#W) \approx \ln(\mu_W) + \frac{\#W-\mu_W}{\mu_W}$ and $\ln(\#L) \approx \ln(\mu_L) + \frac{\#L-\mu_L}{\mu_L}$, where μ_W and μ_L denote the expected values of wins and losses respectively, so

$$\ln(WR) = \ln(\#W) - \ln(\#L) \approx \ln\left(\frac{\mu_W}{\mu_L}\right) + \frac{\#W}{\mu_W} - \frac{\#L}{\mu_L}.$$

Under the null hypothesis, because

$$\mu_W = \mu_L = \mu = \frac{1}{2}k(1 - k)N^2 (1 - p_{\text{tie}}) \tag{A3}$$

therefore

$$\ln(WR) \approx \frac{\#W - \#L}{\mu} \quad \text{and} \quad \text{Var}(\ln(WR)) \approx \frac{1}{\mu^2}\text{Var}(\#W - \#L).$$

Because $\#W - \#L = \sum_{i\in A} U_i$ in the Finkelstein-Schoenfeld method, it has permutation variance[10] $\approx k(1 - k) \sum_{All\ i} U_i^2$, therefore

$$\text{Var}(\ln(WR)) \approx \frac{k(1 - k)}{\mu^2} \sum_{All\ i} U_i^2. \tag{A4}$$

In the absence of any ties and intransitivity (see below for an explanation), each patient will have a unique rank in the set $\{-(N - 1), -(N - 3), \dots , (N - 3), (N - 1)\}$,

therefore

$$\begin{aligned}\sum_{All\ i} U_i^2 &= (-(N - 1))^2 + (-(N - 3))^2 + \cdots + (N - 3)^2 + (N - 1)^2\\ &= \sum_{j=1}^{N}(N + 1 - 2j)^2 = \frac{N\left(N^2 - 1\right)}{3}.\end{aligned}$$

When there are ties and intransitivity exists, the value of $\sum_{All\ i} U_i^2$ will depend on the pattern of ties and intransitivity. Assuming no intransitivity (but allowing ties), $\sum_{All\ i} U_i^2$ can be well approximated by $\frac{N^3\left(1-p_{\text{tie}}^2\right)}{3}$, that is,

$$\sum_{All\ i} U_i^2 \approx \frac{N^3\left(1-p_{\text{tie}}^2\right)}{3} \tag{A5}$$

The right-hand side of (A5) can be derived by assuming that the N patients can be divided into $l\left(=\frac{N}{m}\right)$ groups of m patients each, and that all patients within each group have tied ranks. Given this assumption, the proportion of Finkelstein-Schoenfeld ties, $p_{FS,\text{tie}}$, is given by $\frac{lm(m-1)}{N(N-1)}$ or $\frac{m-1}{N-1}$. And because $p_{FS,\text{tie}} \approx p_{\text{tie}}$, we have

$$\frac{m-1}{N-1} = p_{FS,\text{tie}} \approx p_{\text{tie}}, \text{or } m \approx Np_{\text{tie}}.$$

Then

$$\begin{aligned}\sum_{All\ i} U_i^2 &= m\left[(-(l-1)m)^2 + (-(l-3)m)^2 + \cdots + ((l-3)m)^2 + ((l-1)m)^2\right]\\ &= m^3\sum_{j=1}^{K}(l+1-2j)^2 = \frac{m^3 l\left(l^2-1\right)}{3} = \frac{N\left(N^2-m^2\right)}{3} \approx \frac{N^3\left(1-p_{\text{tie}}^2\right)}{3}.\end{aligned}$$

Combining Equations (A4) and (A5), we have

$$\text{Var}(\ln(WR)) \approx \frac{k(1-k)}{\mu^2}\sum_{All\ i} U_i^2 \approx \frac{k(1-k)}{\mu^2}\cdot\frac{N^3\left(1-p_{\text{tie}}^2\right)}{3} \approx \frac{4\left(1+p_{\text{tie}}\right)}{3k(1-k)\left(1-p_{\text{tie}}\right)N},$$

which is obtained from plugging in the value for μ in (A3) and simplifying.

To achieve a type I error rate of α (one-sided) and a study power of $1-\beta$, the required sample size, N, will need to satisfy the following equation:

$$\text{Var}(\ln(WR))\left(Z_{1-\alpha}+Z_{1-\beta}\right)^2 = \ln^2\left(WR_{\text{true}}\right) \text{ or}$$

$$N = \frac{4(1+p_{\text{tie}})\left(Z_{1-\alpha}+Z_{1-\beta}\right)^2}{3k(1-k)\left(1-p_{\text{tie}}\right)\ln^2\left(WR_{\text{true}}\right)}.$$

And the formula for power is given by.

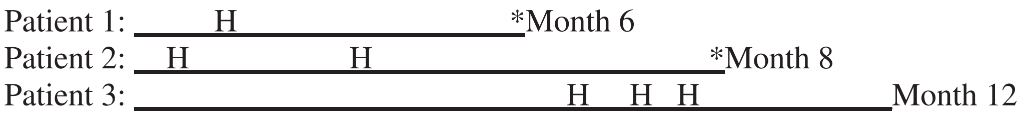
$$\text{Power} = 1-\Phi\left(Z_{1-\alpha}-\ln\left(WR_{\text{true}}\right)\sqrt{\frac{3Nk(1-k)\left(1-p_{\text{tie}}\right)}{4\left(1+p_{\text{tie.}}\right)}}\right)$$

with Φ denoting the cumulative distribution function of the standard normal distribution.

A.1.3. Intransitivity and its relationship to variance and sample size

Intransitivity refers to an outcome for one pairwise comparison that does not logically follow from the other two pairwise comparisons in a triplet of patients. Intransitive outcomes are possible when patients have unequal follow-up times. For any pairwise comparison, let ">" denote "better than" or equivalently a "win," let "<" denote "worse than" or equivalently a "loss," and let "=" denote a tie.

Here we illustrate how intransitivity can occur for a single recurrent event endpoint, frequency of hospitalizations. The figure below shows outcomes for patients 1-3. Time is shown on the horizontal axis, H denotes hospitalizations, * denotes a censored event (eg, loss to follow-up).



The pairwise comparisons of the triplet of patients 1-3 is intransitive, because patient 1 > patient 2, patient 2 > patient 3, but patient 1 < patient 3. Using the Finkelstein-Schoenfeld notation, $u_{12}=1$, $u_{23}=1$, $u_{13}=-1$, and ranks $U_1 = u_{12} +$

$u_{13} = 0$, $U_2 = u_{21} + u_{23} = 0$, and $U_3 = u_{31} + u_{32} = 0$. The ranks for the patients being the same, no patient fared better or worse than the other. The sum of squares of the ranks, on which the variance is based, equals 0. Recall that the variance is based on the sum of squares of the ranks.

Suppose instead there is no loss to follow-up, so all three patients are followed for 12 months with events as shown above. The relationship between the three pairs of patients is now transitive because patient 1 > patient 2 > patient 3. Moreover, the sum of squares of the ranks equals 8. When there is no loss to follow-up, the sum of squares of the ranks is larger, hence the variance and the sample size is larger.

A.1.4. Derivation of the stratified variance

Suppose the (weighted) stratified win ratio, $WR_{\text{stratified}}$, is computed as

$$WR_{\text{stratified}} = \frac{\sum_{i=1}^{M} w_i (\#W)_i}{\sum_{i=1}^{M} w_i (\#L)_i},$$

where w_i is the fixed weight, and $(\#W)_i$ and $(\#L)_i$ are the total number of wins and the total number of losses in the ith stratum, respectively ($1 \leq i \leq M$).

Then the variance formula for the stratified win ratio is

$$\text{Var}\left(\ln\left(WR_{\text{stratified}}\right)\right) \approx \frac{4\left(1 + p_{\text{tie}}\right)}{3k(1-k)\left(1 - p_{\text{tie}}\right)} \cdot \frac{\sum_{i=1}^{M} w_i^2 N_i^3}{\left(\sum_{i=1}^{M} w_i N_i^2\right)^2},$$

where N_i is the total number of patients in the ith stratum.

A.2. Derivation

By the delta method,

$$\ln\left(\sum_{i=1}^{M} (\#W)_i\right) \approx \ln\left(\mu_W\right) + \frac{\sum_{i=1}^{M} w_i\ (\#W)_i - \mu_W}{\mu_W},$$
$$\ln\left(\sum_{i=1}^{M} (\#L)_i\right) \approx \ln\left(\mu_L\right) + \frac{\sum_{i=1}^{M} w_i\ (\#L)_i - \mu_L}{\mu_L},$$

where $\mu_W = E\left(\sum_{i=1}^{M} w_i\ (\#W)_i\right)$, and $\mu_L = E\left(\sum_{i=1}^{M} w_i\ (\#L)_i\right)$.

Under the null hypothesis, $\mu_W = \mu_L$, which we will denote by μ,

$$\ln\left(WR_{\text{stratified}}\right) \approx \frac{\sum_{i=1}^{M} w_i\ \left[(\#W)_i - (\#L)_i\right]}{\mu}.$$

Therefore,

$$\begin{aligned} \text{Var}\left(\ln\left(WR_{\text{stratified}}\right)\right) &\approx \frac{1}{\mu^2}\text{Var}\left(\sum_{i=1}^{M} w_i\ \left[(\#W)_i - (\#L)_i\right]\right) \\ &= \frac{1}{\mu^2}\sum_{i=1}^{M} w_i^2 \text{Var}\left((\#W)_i - (\#L)_i\right) \end{aligned}$$

This follows, because $(\#W)_i - (\#L)_i$ and $(\#W)_j - (\#L)_j$ are independent whenever $i \neq j$, and by repeating the same argument used to derive (1) for each of the strata and assuming that the proportion of ties are homogeneous across the strata, we get

$$\approx \frac{1}{\mu^2}\sum_{i=1}^{M} \frac{w_i^2 k(1-k) N_i^3 \left(1 - p_{\text{tie}}^2\right)}{3} \tag{A6}$$

Plugging $\mu = \frac{1}{2}\sum_{i=1}^{M} w_i k(1-k) N_i^2 \left(1-p_{\text{tie}}\right)$into (A6), and simplifying, we have

$$\text{Var}\left(\ln\left(WR_{\text{stratified}}\right)\right) \approx \frac{4\left(1+p_{\text{tie}}\right)}{3k(1-k)\left(1-p_{\text{tie}}\right)} \cdot \frac{\sum_{i=1}^{M} w_i^2 N_i^3}{\left(\sum_{i=1}^{M} w_i N_i^2\right)^2}$$