The New England Journal of Statistics in Data Science Volume 3, 188–199 (2025)
DOI: https://doi.org/10.51387/25-NEJSDS85

Utilizing Win Ratio Approaches and Two-Stage Enrichment Designs for Small-Sized Clinical Trials

Jialu WANG, Yeh-Fong CHEN*, and Thomas GWISE

Abstract

Conventional methods for analyzing composite endpoints in clinical trials often only focus on the time to the first occurrence of all events in the composite. Therefore, they have inherent limitations because the individual patients' first event can be the outcome of lesser clinical importance. To overcome this limitation, the concept of the win ratio (WR), which accounts for the relative priorities of the components and gives appropriate priority to the more clinically important event, was examined. For example, because mortality has a higher priority than hospitalization, it is reasonable to give a higher priority when obtaining the WR. In this paper, we evaluate three innovative WR methods (stratified matched, stratified unmatched, and unstratified unmatched) for two and multiple components under binary and survival composite endpoints. We compare these methods to traditional ones, including the Cox regression, O'Brien's rank-sum-type test, and the contingency table for controlling study Type I error rate. We also incorporate these approaches into two-stage enrichment designs with the possibility of sample size adaptations to gain efficiency for rare disease studies.

KEYWORDS AND PHRASES: Adaptive clinical trial, Composite endpoints, Enrichment strategy, Win ratio method.

1. INTRODUCTION

In the United States, according to the "Rare Diseases Act of 2002", there are more than 6,000 rare diseases [18, 8]. A rare disease is defined as a condition that affects fewer than 200,000 individuals, or 1 in 1,500 people. The development of efficient approaches to utilizing individual patient data, e.g. improved study designs and sound statistical methods, is instrumental in bringing breakthrough therapies to the market early [21, 9, 20]. Examples of treating rare diseases include but not limit to Gaucher disease and Neuronal ceroid lipofuscinosis, where trial sponsors had been recommended to use innovative designs, including umbrella designs and single-arm historical controlled designs [7, 17]. In the nonmalignant hematology disease area, there are also many rare disease clinical trials that require the careful identification of endpoints to assess the efficacy of drugs (e.g. WHIM syndrome and immune thrombocytopenia). In addition, it is not possible with many diseases to conduct well-controlled, adequately powered clinical trials for pediatric populations because of ethical concerns.

Given the concern over lacking adequate study power in conducting small-sized clinical trials, innovative designs utilizing different types of efficacy endpoints with proper statistical analyses and study-wise type I error control need to be considered. Patients are likely to be heterogeneous in rare disease clinical trials. When conducting such trials, composite endpoints can be created by combining multiple components, either requiring all components or a certain number of components or winning on multiple endpoints (e.g., 3 out of 5). Doing so can be beneficial and should be considered [14]. Furthermore, valid statistical methods are imperative to efficiently handle these types of endpoints to increase the chances of detecting treatment effect.

In this paper, we examine statistical methods utilizing win ratio methods (WR) based on both matched and unmatched pairs [5, 16]. We cover different types of endpoints (i.e., survival, binary, and continuous) as described in Section 3. A closed-form sample size formula is also provided. The sequential enriched design is introduced in Section 4.

To demonstrate the pros and cons of the WR methods, we consider different winning criteria, and results are illustrated by comparing WR methods with those via O'Brien's rank-sum-type test and the contingency table. We follow Section 5 to generate different types of data. Section 6 shows our simulation results and findings. Besides examining the WR methods mainly applied in the single parallel design, covariates stratification and innovative designs such as two-stage designs, including sequential parallel comparison designs and sequential enriched design, are used to provide further efficiency [4, 20, 22].

2. WIN RATIO METHODS AND NOTATIONS

For simplicity, we consider two treatment groups: one for the study drug and the other for the control, which can be a placebo. We are interested in assessing the treatment effect

*Corresponding author.

that can come from any component of a composite endpoint. In our evaluation, we examine the WR performance on the continuous or survival endpoint with multiple components. For example, the test hypotheses for a composite endpoint with two binary components of equal importance are $H_0 : p_{j,t} = p_{j,p}$ for $\forall j = 1, 2$, and $H_1 : p_{1,t} \neq p_{1,p}$ or $p_{2,t} \neq p_{2,p}$, where $p_{j,t}$ and $p_{j,p}$ are the survival probability of component j ($j = 1, 2$) in the treatment group and placebo group, respectively. Similarly, the test hypotheses for a composite endpoint with three equally important continuous components are: $H_0 : E_{p,j} = E_{t,j}$ for $\forall j = 1, 2, 3$, and $H_1 : E_{p,1} \neq E_{t,1}$ or $E_{p,2} \neq E_{t,2}$ or $E_{p,3} \neq E_{t,3}$, where $E_{p,j}$ and $E_{t,j}$ as the time to component j's improvement in the placebo group and the treatment group, respectively. Later, we also take the priority of the components' importance into consideration.

2.1 Motivation with Toy Example

The composite endpoints have been used in many clinical trials to increase the chances of collecting more data from many domains of a disease to increase the study power. Although this idea sounds feasible and can be useful, having a clear understanding of when a composite endpoint should be considered and how to use it properly is very important. We use Figure 1 as a toy example to illustrate that if a composite endpoint is not constructed wisely, the results can be misleading.

Figure 1 displays a composite endpoint with two components. We assume that all the eight patients in the drug group respond to event A but not B. For the eight placebo patients, we assume that half of them respond to both events A and B, and the other half don't respond to neither A nor B.

When we consider the composite endpoint by winning either A or B, results tell us that the drug response rate is 100% and the placebo response rate is 50%. However, if we further study the two individual events, we can see that this result is mainly driven by the event A, because the drug performs worse than the placebo on the event B. In particular, although for the composite endpoint A or B and the component A, the placebo response rate is 50% and drug response rate is 100%, for the component B the placebo response rate is still 50% but the drug response rate is 0%. In other words, if we do not consider any specific winning criteria, Event A and Event B should be equally important. Otherwise, results can be very misleading, and the study will not be powerful.

Drug Patient Number	Event A of Drug	Event B of Drug	Placebo Patient Number	Event A of Placebo	Event B of Placebo
1	Y	N	9	Y	Y
2	Y	N	10	Y	Y
3	Y	N	11	Y	Y
4	Y	N	12	Y	Y
5	Y	N	13	N	N
6	Y	N	14	N	N
7	Y	N	15	N	N
8	Y	N	16	N	N

Figure 1: Toy example of composite endpoint (A or B).

2.2 Literature Review for Two Types of Win Ratio Methods: Matched and Unmatched

The idea of WRs is not new and has been extensively studied. This type of endpoint has also been utilized in many large cardiovascular and renal clinical trials [15, 6]. The basic idea of constructing a WR is first to pair all patients in two treatment arms and compare their performance according to pre-defined criteria to determine their winning status. At the end, combine all pairs' winning status for making the final statistical inference. These pairs can be either coming from matched or unmatched samples [12, 19, 1, 13]. More details regarding how we applied the WR methods in either unmatched or matched pairs will be discussed and illustrated in Section 3. As noted in our toy example, how all the components are prioritized in the composite endpoint will affect the performance and interpretability of the WR results.

3. WIN RATIO WINNING CRITERIA AND SAMPLE SIZE CALCULATION

3.1 Composite Endpoint with Prioritized Components

3.1.1 Prioritized Binary Component

We begin the evaluation by considering the composite endpoint with two binary prioritized components. Suppose the two components we consider are death and hospitalization. We also assume that the death event is more clinically critical than hospitalization. We theoretically derive the test statistics and confidence interval under the null hypothesis and the analytical formula for sample size calculation.

Notation Let Y_{ti} denote the death event for the ith patient who is assigned in the treatment group (i.e., patients take the assigned drug) T, and assume their death events are independent. Therefore, $Y_{ti} \xrightarrow{iid} Bernoulli(p_t)$, where $Y_{ti} = 1$ represents that the ith patient dead and $Y_{ti} = 0$ represents the patient living after the treatment. Similarly, we let Y_{ci} be the indicator of the death event for ith patient who is assigned in the control group C, and $Y_{ci} \xrightarrow{iid} Bernoulli(p_c)$. In addition, X_{ti} indicate the hospitalization event for the ith patient in treatment group T, and $X_{ti} \xrightarrow{iid} Bernoulli(q_t)$. That is $X_{ti} = 1$ if the ith patient in the treatment group

Paired Patient Type	Death		Hospitalization		Treatment win/lose/tie
	Treatment	Placebo	Treatment	Placebo	
1	1	1	0	1	
2	0	1	1	0	
3	0	1	1	1	win
4	0	1	0	0	
5	0	1	0	1	
6	0	0	0	1	
7	1	0	0	1	
8	1	0	0	0	
9	1	0	1	1	
10	1	0	1	0	lose
11	1	1	1	0	
12	0	0	1	0	
13	1	1	0	0	
14	1	1	1	1	
15	0	0	1	1	tie
16	0	0	0	0	

Figure 2: The comparison principle for composite endpoint with two prioritized binary components.

requires hospitalization, and $X_{ti} = 0$ if the ith patient does not. Similarly, let indicator X_{ci} denote the hospitalization event for the ith patient under the control group C, and $X_{ci} \xrightarrow{iid} Bernoulli(q_c)$. The principle for comparing a composite endpoint with two prioritized binary components, i.e., the winning rule of WR calculation, is specified in Figure 2. It emphasizes that treatment versus placebo's impact to death will be evaluated first; if no decision could be made at the first stage, their impact on hospitalization will be evaluated as the second step; if still no decision can be made, we say 'tie'.

Sample Size for Matched Win Ratio In the previous section, we introduced the way we pair patients; either coming from matched or unmatched samples will affect the performance and interpretability of the WR results. Here we derive the asymptotic properties of WR test statistics and the sample size formula for any given Type I and power requirement with details in Appendix A. We first analyze the matched win ratio method and then the unmatched method.

First, the probability of a treatment wins under all scenarios is derived as

$$p_w = p_t(1-q_t)p_cq_c + (1-p_t)q_tp_c + (1-p_t)(1-q_t)\big(1-(1-p_c)(1-q_c)\big).$$

The probability of a treatment losses under all scenarios is:

$$p_l = p_t(1-q_t)(1-p_c) + p_tq_t(1-p_cq_c) + (1-p_t)q_t(1-p_c)(1-q_c).$$

The probability that treatment and control tie under is

$$p_{tie} = 1 - p_w - p_l.$$

Next, we let the binary random variable X_i follow $Bernoulli(p)$, which denotes every win-loss comparison, where $X_i = 1$ if treatment wins; otherwise, $X_i = 0$, and

$$p = P(\text{treatment win}|\text{all non-tie pairs}) = \frac{p_w}{1-p_{tie}}.$$

Suppose a total number of N patients are randomized, and we let $n = N(1-p_{tie})$ denote the total number of non-tie units. Based on the Delta Method, we derive that

$$\sqrt{n}\left(\frac{\bar{X}}{1-\bar{X}} - \frac{p}{1-p}\right) \xrightarrow{D} N\left(0, \frac{p^2}{(1-p)^2}\right), \bar{X} = \sum_{i=1}^{n} X_i/n. \tag{3.1}$$

It is obvious that under the null hypothesis, $p = 0.5$ and $\sqrt{n}(\frac{\bar{X}}{1-\bar{X}} - 1) \xrightarrow{D} N(0,1)$. Besides, the minimum sample size N required for power β under Type I error α is

$$N = \frac{n}{1-p_{tie}} \quad \text{and} \quad n = \left(\frac{\frac{p}{1-p}Z_\alpha - \frac{p_a}{1-p_a}Z_\beta}{\frac{p}{1-p} - \frac{p_a}{1-p_a}}\right)^2, \tag{3.2}$$

where p_a is the proportion under alternative hypothesis.

Sample Size for Unmatched Win Ratio Similar to the matched WR, we first consider all the scenarios in which treatment wins and treatment losses.

For treatment and control pair (i, j), treatment wins when $Y_{ti} = 0$, $Y_{cj} = 1$, or $Y_{ti} = 1$, $Y_{cj} = 1$, $X_{ti} = 0$, $X_{cj} = 1$, or $Y_{ti} = 0$, $Y_{cj} = 0$, $X_{ti} = 0$, $X_{cj} = 1$. Similarly, control wins when $Y_{ti} = 1$, $Y_{cj} = 0$, or $Y_{ti} = 1$, $Y_{cj} = 1$, $X_{ti} = 1$, $X_{cj} = 0$, or $Y_{ti} = 0$, $Y_{cj} = 0$, $X_{ti} = 1$, $X_{cj} = 0$.

Therefore, we derive the test statistics for win ratio $g(\mathbf{X})$ by dividing the total number of treatment wins by the total number of control wins, where $\mathbf{X} = (\overline{Y_t}, \overline{X_t}, \overline{XY_t}, \overline{Yc}, \overline{Xc}, \overline{XY_c})$ and $\overline{Y_t} = \sum_{i=1}^{n_1} Y_{ti}$, $\overline{X_t} = \sum_{i=1}^{n_1} X_{ti}$, $\overline{XY_t} = \sum_{i=1}^{n_1} X_{ti}Y_{ti}$, $\overline{Yc} = \sum_{j=1}^{n_0} Y_{cj}$, $\overline{Xc} = \sum_{j=1}^{n_0} X_{cj}$, $\overline{XY_c} = \sum_{j=1}^{n_0} X_{cj}Y_{cj}$. The n_1 is the number of patients assigned to the treatment group, and n_0 is the number of patients assigned to the control group. $n_t = n_1 + n_0$.

Then by the Delta Method, we derive

$$\sqrt{n_t}\big(g(\mathbf{X}) - g(\boldsymbol{\theta})\big) \xrightarrow{D} N\big(0, C^2\big), \quad C^2 = \left(\frac{d}{d\boldsymbol{\theta}}g(\boldsymbol{\theta})\right)^T COV(\mathbf{X})\left(\frac{d}{d\boldsymbol{\theta}}g(\boldsymbol{\theta})\right), \tag{3.3}$$

where $\boldsymbol{\theta} = (p_t, q_t, p_tq_t, p_c, q_c, p_cq_c)$, $g(\boldsymbol{\theta}) = g(E(\mathbf{X}))$.

Therefore, under the null hypothesis

$$\sqrt{n_t}\big(g(\mathbf{X}) - 1\big) \xrightarrow{D} N\big(0, C_0^2\big), \quad C_0^2 = \left(\frac{d}{d\boldsymbol{\theta}}g(\boldsymbol{\theta})\right)^T COV(\mathbf{X})\left(\frac{d}{d\boldsymbol{\theta}}g(\boldsymbol{\theta})\right)_{|\boldsymbol{\theta}=\boldsymbol{\theta_0}}, \tag{3.4}$$

where $\boldsymbol{\theta_0} = (0.5, 0.5, 0.25, 0.5, 0.5, 0.25)$.

Similarly, under the alternative hypothesis

$$\sqrt{n_t}\big(g(\mathbf{X}) - g(\boldsymbol{\theta_1})\big) \xrightarrow{D} N\big(0, C_1^2\big),$$
$$C_1^2 = \left(\frac{d}{d\boldsymbol{\theta}} g(\boldsymbol{\theta})\right)^T COV(\mathbf{X}) \left(\frac{d}{d\boldsymbol{\theta}} g(\boldsymbol{\theta})\right)_{|\boldsymbol{\theta}=\boldsymbol{\theta_1}}, \tag{3.5}$$

where $\boldsymbol{\theta_1} = (p_{t1}, q_{t1}, p_{t1}q_{t1}, p_{c1}, q_{c1}, p_{c1}q_{c1})$.

Therefore, the minimum sample size required for power β under Type I error α is

$$n_t = \left(\frac{C_0 Z_\alpha - C_1 Z_\beta}{g(\boldsymbol{\theta_1}) - 1}\right)^2. \tag{3.6}$$

3.1.2 Prioritized Survival Component

In this section, we show the winning rules of matched and unmatched methods for the composite endpoint of two prioritized survival components. To further explore the pros and cons of the WR methods, traditional Cox regression in survival analysis and O'Brien's rank-sum-type test are considered and incorporated [2]. Point estimation and its corresponding confidence interval and power comparison are extensively explored via numerical studies in Section 6.2.

(Stratified) Matched Win Ratio We stratify patients into different strata based on their baseline covariates, and then form matched pairs on the study drug and the control. For each matched pair, according to the following criteria, we then compare each patient in the study drug group with the one matched in the placebo group is a winner or a loser and its asymptotic properties via Algorithm 1 [15]. We also note that [12] proposed a closed-form variance estimator and approximate $1-\alpha$ confidence interval, which could be utilized for testing the null hypothesis.

(Stratified) Unmatched Win Ratio We utilize the stratified Finkelstein and Schoenfeld (FS) test from [5] and [15] and derive the corresponding power by simulations. It proceeds as follows

1. Stratify patients into k strata and let A_k denote n_k patients in the kth strata.
2. Irrespective of treatment group, compare all possible pairs of patients i, j to determine whether patient i is a winner, loser, or tie.
3. Calculate N_w and N_L via the same way as in the matched method.
4. Define u_{ij} and assign $u_{ij} = +1, -1, 0$ according to winning status of patient i (i.e., winner, loser, or tie).
5. Within each stratum, calculate U_i where for $i \in A_k$, $U_i = \sum_{j \in A_k} u_{ij}$. It will be a positive integer if patient i wins more often than losses compared with all other patients.

We calculate the WR R_w and test statistics z as follows:

$$R_w = N_{\mathrm{w}}/N_{\mathrm{L}}, \quad z = T/V^{1/2}, \quad T = \sum_k \sum_{i \in A_k} D_i U_i,$$

Algorithm 1: (Stratified) Matched Winning Rule

if *stratified* **then**
  patients are stratified into k different strata based on their covariates, form matched pairs within each stratum for the new treatment and the control; all pairs are then collected.
else
  Form matched pairs based on the whole sample.
for *every matched pair* **do**
  if *one of the two patients die* **then**
    if *patient in the treatment group dies first* **then**
      Control wins (Treatment loses)
    if *patient in the control group dies first* **then**
      Treatment wins (Control loses)
  else
    ▷ Both patients die on the same day, or neither of them die
    if *patient in the treatment group has hospitalization first* **then**
      Control wins
    if *patient in the control group has hospitalization first,* **then**
      Treatment wins
    else
      Tie ▷ Both patients were hospitalized on the same day; no patient was hospitalized
Obtain:

1. The number of patients that fall into categories: (a) new treatment patient has death first N_a; (b) control patient has death first N_b; (c) new treatment patient has hospitalization first N_c; (d) control patient has hospitalization first N_d.
2. $N_w = N_b + N_d$, the number of "winners" for the new treatment. $N_{\mathrm{L}} = N_{\mathrm{a}} + N_{\mathrm{c}}$, the number of "losers" for the new treatment.
3. The proportion p_W: $p_W = \frac{N_{\mathrm{w}}}{N_{\mathrm{w}}+N_{\mathrm{L}}}$, $p_L, p_U = p_w \pm 1.96[\frac{p_w(1-p_w)}{(N_w+N_L)}]^{1/2}$
4. The "WR" $R_W = \frac{N_{\mathrm{w}}}{N_{\mathrm{L}}} = \frac{p_w}{1-p_w}$, $CI_{R_W,0.95} = (\frac{p_L}{1-p_L}, \frac{p_U}{1-p_U})$
5. The test statistics via a standardized normal assumption, for a significance hypothesis testing:

$$z = (p_w - 0.5)/[p_w(1-p_w)/(N_w+N_L)]^{\frac{1}{2}} \tag{3.7}$$

$$V = \sum_k \frac{m_k(n_k - m_k)}{n_k n_k - 1} \sum_{i \in A_k} U_i^2,$$

where $D_i = 1$ for subjects in the new group and $D_i = 0$ for patients in the standard group.

For hypothesis testing, we also utilize the standardized normal statistics z in the equation (3.7) of Algorithm 1. For the confidence interval (CI) and power, we first calculate

lnR_w and its approximate standard error $s = lnR_w/z$. Then we have $CI_{lnR_w,0.95} = (lnR_{w,L}, lnR_{w,U}) = (lnR_w - 1.96s, lnR_w + 1.96s)$, and thus $CI_{R_w,0.95} = (e^{lnR_{w,L}}, e^{lnR_{w,U}})$.

For the unstratified unmatched WR method, we follow the same step as the stratified unmatched WR method except for the stratification.

Cox Regression We use cox regression to analyze the time to the first event of the composite endpoint. For example, in a typical Cox regression equation

$$h(t) = h_0(t)\exp(\beta_t x_t + \beta_c x_{cov}) \tag{3.8}$$

The $h(t)$ is hazard rate at given time t, where $t = min(E_d, E_{hos})$. The x_t is an indicator representing whether the patient is in the treatment group, and x_{cov} are patients' baseline covariates. $h_0(t)$ is the baseline hazard, which does not depend on treatment indicator x_t and covariates x_{cov1}, x_{cov2}. Finally, β_t is the expected log hazard ratio (HR) that compares the risk of a patient in treatment to those in the control arm for both death and hospitalization events. We are interested in testing whether β_t is 0 or not under required Type I error.

O'Brien's Rank-Sum-Type Test Peter C. O'Brien proposed a rank-sum-type test in [2]. We incorporate it within the context of composite endpoint as follows:

1. Let Y_{ijk} represent the kth variable for the jth subject in group i, where $k = 1, \ldots, K$, $j = 1, \ldots, n_i$, $i = 1, \ldots, I$. Y_{ijk} is defined such that large values are better than small values for each k. (For example, k is death or hospitalization, i is treatment or control group, and j_i is the jth patient in group i.)
2. Let R_{ijk} represent the rank of Y_{ijk} among all values of variable k in the pooled set of I samples. Define S_{ij} as the sum of the ranks assigned to the jth person in sample i.
3. Perform a One-Way Analysis of Variance (ANOVA) on the S_{ij} values.

3.2 Composite Endpoint with Equally Important Continuous Components

To generalize the use of the WR method in a composite endpoint with more than two components, we consider the situation in which a composite endpoint has multiple equally important components. For example, a composite endpoint with three equally important continuous components has notations described as follows

Suppose $y_{p,j,i}$ is the ith patient's time to its jth component improvement in the placebo group, $y_{t,j,i}$ is the ith patient's time to its jth component improvement in the treatment group, and y_{base} is a baseline. We identify the indicators of successful improvement for patients in the placebo group via the following indicators:

$$\mathcal{I}_{p,j,i} = \begin{cases} 1 & y_{p,j,i}/y_{base,i} < c_t, \\ 0 & y_{p,j.i}/y_{base,i} \geq c_t, \end{cases}$$

$$\mathcal{I}_{p,i} = \begin{cases} 1 & \sum_{j=1}^{3} \mathcal{I}_{p,j,i} \geq 1, \\ 0 & \sum_{j=1}^{3} \mathcal{I}_{p,j,i} = 0, \end{cases}$$

where $\mathcal{I}_{p,j,i}$ is an indicator that implies whether the ith patient in placebo group successfully improves on the jth component with cutoff c_t, and $\mathcal{I}_{p,i}$ is an indicator that implies whether the ith patient in placebo group successfully improves on at least one component. Similarly, we identify the indicators of successful improvement $\mathcal{I}_{t,j,i}$ and $\mathcal{I}_{t,i}$ for patients in the treatment group via the following indicators:

$$\mathcal{I}_{t,j,i} = \begin{cases} 1 & y_{t,j,i}/y_{base,i} < c_t, \\ 0 & y_{t,j,i}/y_{base,i} \geq c_t, \end{cases}$$

$$\mathcal{I}_{t,i} = \begin{cases} 1 & \sum_{j=1}^{3} \mathcal{I}_{t,j,i} \geq 1, \\ 0 & \sum_{j=1}^{3} \mathcal{I}_{t,j,i} = 0. \end{cases}$$

(Stratified) Matched Win Ratio The logic here is similar to the Algorithm 1 except for some modification, especially the way to define the winner in every matched pair comparison. We stratified patients into different strata based on their baseline covariates, and then form matched pairs on the study drug and the control. For each matched pair, we determine that the patient in the study drug is a winner or a loser by the following rule:

1. Calculate the total number of successful improvements for each patient in placebo, i.e., calculate $\sum_{j=1}^{3} \mathcal{I}_{p,j,i}$, $i = 1, \ldots, n_0$.
2. Calculate the total number of successful improvements for each patient in treatment, i.e., calculate $\sum_{j=1}^{3} \mathcal{I}_{t,j,i}$, $i = 1, \ldots, n_1$.
3. Within each pair, if the total number of successful improvements for the patient in treatment is greater than that for the patient in placebo, treatment wins.
4. Within each pair, if the total number of successful improvements for the patient in treatment is less than that for the patient in placebo, control wins.
5. Otherwise, tie.

Calculate N_w, the number of winners, and N_L, the number of losers for the study drug. The test statistics is the same as the one in Algorithm 1.

(Stratified) Unmatched Win Ratio The procedure here is the same as the unmatched WR method for the composite endpoint with the prioritized survival components. However, like the above matched WR for continuous components, the rule to define the winner in every matched pair comparison is completely different and should follow the winning rule in the new matched WR.

Contingency Table For evaluating the advantage of WR methods, we construct a conventional contingency table as in Table 1. We let $n_{11} = \sum_{i=1}^{n_{1\cdot}} \mathcal{I}_{t,i}$, $n_{10} = \sum_{i=1}^{n_{1\cdot}} (1 - \mathcal{I}_{t,i})$, $n_{01} = \sum_{i=1}^{n_{0\cdot}} \mathcal{I}_{p,i}$, and $n_{00} = \sum_{i=1}^{n_{0\cdot}} (1 - \mathcal{I}_{p,i})$.

Table 1. Contingency table.

	Success	Failure	Total
Treatment	n_{11}	n_{10}	$n_{1.}$
Placebo	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	N

Then we perform hypothesis test via odds ratio. The idea is, instead of calculating the total number of improvements in the treatment (placebo) group for ith patient, a success of treatment (placebo) is counted if the patient has at least one improved component after being allocated to the treatment (placebo) group. Therefore, the test statistic and its distribution is

$$\hat{OR} = \frac{n_{11}n_{00}}{n_{10}n_{01}}, \quad \log(\hat{OR}) \sim N(0, \hat{se}),$$
$$\hat{se} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}.$$

4. SEQUENTIAL ENRICHED DESIGN

To further enhance trial efficacy, two-stage designs can be considered for rare disease clinical trials. In our illustration, we considered sequential enriched design (SED). As seen in Figure 3, SED has two stages. However, before patients are randomized to the first main stage, a placebo lead-in phase is built in to determine their placebo response status. The first major stage of SED is a traditional parallel design, and at the end of the first stage, only patients in the drug group of Stage 1 and are also responders will be further rerandomized to the second stage. The goal of SED is to only study patients who are both placebo non-responders and drug responders.

We use c_{s0} to denote the cutoff for determining placebo nonresponders, i.e., if $y_{pj,i}/y_{base,i} > c_{s0}$ for $\forall j = 1, 2, 3$, then the ith patient is a placebo nonresponder. Let c_{s1} be the cutoff for determining drug nonresponders, i.e., if $y_{j,i}/y_{base,i} > c_{s1}$ for all $j = 1, 2, 3$, the ith patient is drug nonresponder.

As shown in Table 2, the overall patient population is composed of four subpopulations according to the treatments patients receive, and whether they respond to the

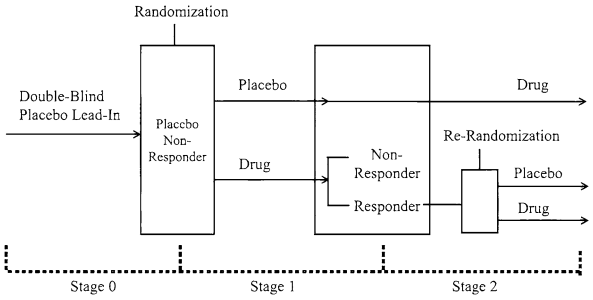

Figure 3: SED procedure [3].

Table 2. Distribution of overall patient population.

Proportion	Drug responder	Drug non-responder
Placebo responder	p_1	p_2
Placebo non-responder	p_3	p_4

treatments or not. The four categories are drug responders and placebo responder p_1, drug non-responders and placebo responders p_2, drug responders and placebo non-responders p_3, and drug non-responders and placebo non-responders p_4. Note that in SED, the target patient population is the type of patients with p_3 probability.

5. DATA GENERATION

5.1 Composite Endpoint with Two Survival Components

We utilize 'coxed' package in R statistical software to generate survival time response [10, 11]. For simplicity, we illustrate our idea by only considering two components, death and hospitalization.

Time to the Component Improvements with Less Clinical Importance

$$E_{hos} = H_0^{-1}\big[-\log(u)\exp(-X\beta_{hos})\big],$$

where $X = (x_t, x_{cov1}, x_{cov2})$, $\beta_{hos} = (\beta_t, \beta_{cov1}, \beta_{cov2})$. The x_t is an indicator of whether the patient is in the treatment group. β_t is the expected log hazard ratio (HR) that compares the risk of a patient in treatment to that in control for hospitalization. The drug is effective if $\beta_t > 0$. β_{cov1} and β_{cov2} are coefficients of covariate x_1 and x_2, respectively. The u is randomly drawn from a standard uniform distribution $\mathcal{U}[0, 1]$. $H_0 = \int_0^t h_0(s)ds$ is cumulative baseline hazard function, where $h_0(t)$ represents baseline hazard;

Time to the Component Improvements with More Clinical Importance

$$E_d = H_0^{-1}\big[-\log(u)\exp(-X\beta_d)\big],$$

where $X = (x_t, x_{dhraito}, x_{cov1}, x_{cov2})$, $\beta_d = (\beta_t + \beta_{in}, \beta_{dhratio}, \beta_{cov1}, \beta_{cov2})$. β_{in} is expected log HR that describes the difference between the risk of a patient for death and hospitalization in treatment group. Therefore, $\beta_t' = \beta_t + \beta_{in}$ is the expected log HR that compares the risk of a patient in the treatment to that in control for the death event. The $x_{dhraito}$ is a standardized random variable that describes the strength of the relationship between risk of death and hospitalization for each patient without treatment effect. The $\beta_{dhraito}$ describes the strength of the relationship between E_d and $x_{dhratio}$. $\beta_{dhraito} = 0$ indicates that the patient's risk of hospitalization is equal to their risk of death in the control group.

5.2 Composite Endpoint with Three Equally Important Continuous Components and Repeated Measurements

Time to Patient's Three Component Improvements in the Placebo Group

$$y_{base} = \beta_{cov1} x_1 + \beta_{cov2} x_2,$$
$$y_{pj} = \beta_{pj}(1 - x_k) + y_{base} + \epsilon_{pj}, \quad j = 1, 2, 3.$$

The y_{base} is a baseline vector and β_{cov1} and β_{cov2} are coefficients of covariate vectors x_1 and x_2, respectively. In addition, y_{pj} is a vector that stores the time (or any continuous measurements) to the jth component improvement of patients who are in the placebo group. The x_k is an indicator vector that shows whether patients are in the placebo group ($x_k = \mathbf{0}$) or treatment group ($x_k = \mathbf{1}$). The β_{pj} is the placebo effect that may reduce a patient's time to the jth component improvement in placebo to that in baseline. The placebo is effective if $\beta_{pj} < 0$. The ϵ_{pj} is the randomness that corresponds to the jth placebo response.

Time to Patient's Three Component Improvements in the Treatment Group

$$y_1 = \beta_{t1} x_k + y_{base} + \epsilon_{t1},$$
$$y_2 = (\beta_{t1} + \beta_{in2}) x_k + y_{base} + \epsilon_{t2},$$
$$y_3 = (\beta_{t1} + \beta_{in3}) x_k + y_{base} + \epsilon_{t3}.$$

β_{t1} is drug effect that reduces a patient's time (or any continuous measurements) to the first component improvement in treatment to that in baseline, which is effective if $\beta_t < 0$. The β_{in2} describes the difference of drug efficacy between the first and second components in treatment group, i.e., $\beta_{t2} = \beta_{t1} + \beta_{in2}$ is the drug effect that reduces a patient's time (or any continuous measurement) to the second component improvement in treatment to that in baseline. In addition, β_{in3} has a similar definition to β_{in2}, and ϵ_{tj} for $j = 1, 2, 3$ is the randomness that corresponds to the ith treatment response.

6. NUMERICAL STUDY

We evaluate WR methods on different type of composite endpoints, and compare it with conventional estimation methods under different experimental designs. In Section 6.1, we perform simulations to examine the close-form sample size formula for binary composite endpoints. We consider two scenarios, the WR can help save sample sizes and it does not have power advantage, respectively. In Section 6.2, we evaluate the utility of WR method for survival endpoints, comparing different estimation analyses, Type I error and study power under complete randomization (CR). In Section 6.3, we extend to the two-stage sequential enrichment design (SED) and show its benefit in further improving study efficiency using continuous endpoints, especially for small-size studies.

6.1 Toy Example: Sample Size Requirement for Prioritized Composite Endpoint with Two Binary Components

We use a toy example here to show how the matched win-ratio method in Section 3.1.1 saves samples for the composite endpoint with two prioritized binary components. In our simulation, we set Type I error $\alpha = 0.05$ and power $\beta = 95\%$. We use the same notation as in Section 3.1.1 and apply the closed-form sample size calculation formula (3.2). We let p_t, the probability of death in the treatment group, vary among $(0, 0.3)$ and keep other probabilities of an event fixed.

In Figure 4, we set $p_c = 0.3$, $q_t = q_c = 0.5$. It mimics the scenario that compared to a placebo, a drug does not improve the component of less importance. That is the drug is effective to death only and has no effect on hospitalization.

The blue line is always below the red line, showing a clear difference between the WR method and the conventional method which does not consider clinical importance and treats the two components equally. This smaller minimum sample size of WR method also matches Table 9, where WR has larger power than the conventional method (i.e. cox regression) when the treatment has effect on death only.

It can also be observed that the difference is small at the beginning, as it represents the true difference between p_c and p_t (i.e., the x-axis value) is large, and the more p_t approach the $p_c = 0.3$ the greater WR method can save the samples. This further demonstrates the advantage of WR method in detecting small treatment effect for prioritized composite endpoints, and its potential for small-size studies.

In Figure 5, we set $p_c = 0.3$, $q_t = 0.45$, $q_c = 0.5$, a scenario that a drug is effective to both two components. In contrast, the WR doesn't provide much benefit in power improvement, which aligns the Table 7. It can be observed that (1) although the blue line is below the red line when

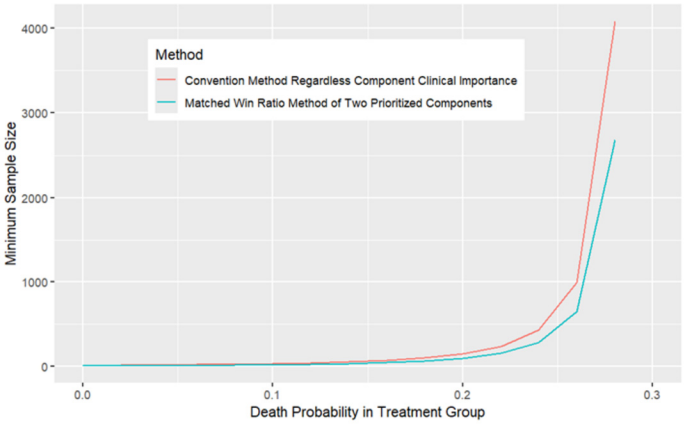

Figure 4: Sample size requirement for binary composite endpoint of two components when treatment has effect on death only.

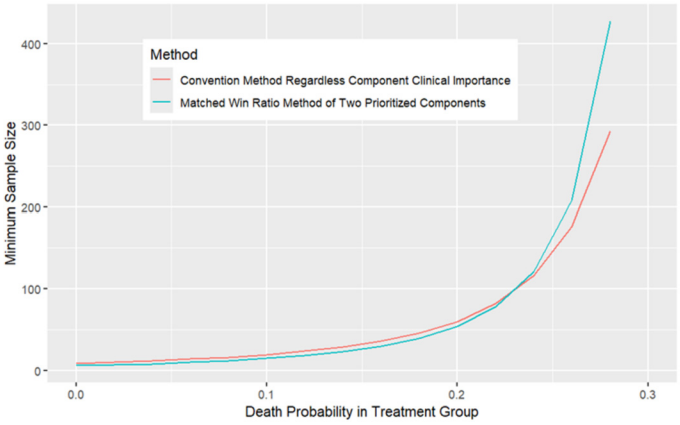

Figure 5: Sample size requirement for binary composite endpoint of two components when treatment has effect on both components.

$p_t < 0.24$, the minimum sample size differences between the two lines are very small; (2) the WR does not have advantage when $p_t >= 0.24$. That is the benefit of utilizing a prioritized composite endpoint decreases as the p_t approaches the p_c.

6.2 Survival Composite Endpoint with Two Components under Parallel Design

We let $\beta_{cov1} = -0.5$, $\beta_{cov2} = 0.5$, $x_{cov1}, x_{cov2} \sim Bernoulli(0.5)$, $x_{dhratio} \sim Uniform(0,1)$. Table 3 shows the distribution of patients in four generated strata.

Table 3. Distribution of patients.

Stratum	1	2	3	4
Percentage of patients (%)	24.5	23.6	26.5	25.4

We estimate the HR based on Cox regression and calculate the WR for our proposed SED and analyses. In addition, we calculate the corresponding confidence intervals and Type I error as well as power via the exact methods.

Type I Error When under the null hypothesis, a drug has no effect such that every patient is equally likely to have hospitalization/death in the treatment and control groups time. We show that either HR or win ratios are close to 1 and the Type I errors are controlled for all examined methods. Our results are displayed in Table 4 and Table 5.

Power for the Same Effects on Both Components Next, we examine the performance of WR methods by comparing it with other commonly used analyses for cases with either both two components have a similar effect or only one having an effect. Our results are shown below.

As seen in Table 7, it can be observed that the powers order is Cox regression > O'Brien's > stratified unmatched ∼ unstratified unmatched > stratified matched when assuming the same effects on both components.

Power for Having Effect on Death Only (No Effect on Hospitalization) As seen in Table 9, it can be observed that the powers order is stratified unmatched > unstratified unmatched ∼ stratified matched > O'Brien's > Cox regression.

Table 9 thus demonstrates that WR methods can more greatly increase trial efficiency than traditional methods when treatment is effective on a prioritized component that occurs after a prioritized component, where the traditional methods that measure the first event cannot be detected thus. Specifically, in Table 9, when $N = 60$, the two unmatched WR methods increases around 30% more power than 'Cox regression' and 'O'Brien's rank sum-type test' (i.e. the two traditional methods); when $N = 200$, the improvement is 20% for 'Cox regression' and is 10% for 'O'Brien's rank sum-type test'. The 'stratified matched WR' also shows the same trend. This shows the advantage of WR in trial efficiency enhancement: for small sized studies, considering a composite endpoint with win ratio can help increase study power.

Power for Having Effect on Death Only but Assuming Wrong Winning Criteria As seen in Table 11, it can be observed that the powers order is O'Brien's > Cox regression > stratified unmatched ∼ unstratified unmatched ∼ stratified matched when assuming that only effect exists on the death event, not the hospitalization event.

Table 5. Type I error comparison with $\beta_t = \beta_{in} = \beta_{dhratio} = 0$.

Type I error	$N = 60$	$N = 100$	$N = 200$
Cox regression	0.05	0.05	0.05
Stratified matched WR	0.06	0.06	0.06
Stratified unmatched WR	0.04	0.05	0.05
Unstratified unmatched WR	0.04	0.05	0.05
O'Brien's rank-sum-type test	0.05	0.05	0.05

Table 4. The estimation of treatment effect for different sample sizes.

Total Sample Size	$N = 60$		$N = 100$		$N = 200$	
Estimation	Beta (SE)	CI	Beta (SE)	CI	Beta (SE)	CI
HR	1.05 (0.40)	(0.60, 1.83)	1.02 (0.27)	(0.67, 1.54)	1.02 (0.18)	(0.76, 1.37)
Stratified, matched WR	1.01 (0.67)	(0.44, 2.33)	1.01 (0.45)	(0.54, 1.89)	1.00 (0.27)	(0.65, 1.53)
Stratified, unmatched WR	1.05 (0.49)	(0.54, 2.02)	1.03 (0.33)	(0.63, 1.68)	1.00 (0.21)	(0.72, 1.41)
Unstratified, unmatched WR	1.04 (0.44)	(0.56, 1.90)	1.03 (0.32)	(0.65, 1.65)	1.01 (0.21)	(0.73, 1.40)

Table 6. The estimation of treatment effect of different sample sizes.

Total Sample Size	$N = 60$		$N = 100$		$N = 200$	
Estimation	Beta (SE)	CI	Beta (SE)	CI	Beta (SE)	CI
HR	0.62 (0.25)	(0.36, 1.10)	0.61 (0.16)	(0.40, 0.93)	0.60 (0.11)	(0.45, 0.82)
Stratified matched WR	1.51 (1.20)	(0.69, 3.87)	1.49 (0.75)	(0.82, 2.95)	1.49 (0.43)	(0.98, 2.34)
Stratified unmatched WR	1.59 (0.77)	(0.81, 3.12)	1.54 (0.50)	(0.95, 2.54)	1.52 (0.32)	(1.08, 2.14)
Unstratified unmatched WR	1.55 (0.68)	(0.84, 2.89)	1.51 (0.47)	(0.94, 2.43)	1.49 (0.30)	(1.07, 2.08)

Table 7. Power comparison with setting $\beta_t = \log(0.6)$ and let $\beta_{in} = 0$ ($\beta' = \beta_t + \beta_{in} = \log(0.6)$) to make HR = 0.6.

Power	$N = 60$	$N = 100$	$N = 200$
Cox regression	0.44	0.66	0.92
Stratified matched WR	0.17	0.26	0.47
Stratified unmatched WR	0.19	0.36	0.65
Unstratified unmatched WR	0.21	0.35	0.61
O'Brien's rank-sum-type test	0.32	0.51	0.82

6.3 Continuous Composite Endpoint with Three Components and Repeated Measurements under SED

As highlighted in the introduction, two-stage enrichment designs such as sequential parallel comparison design, SED and sequential multiple assignment randomized trial have been proposed and used in clinical trials. After learning that the use of WR can increase the study power, we are interested in assessing whether the idea of WR can be implemented in two-stage design to further increase trial efficiency for rare disease clinical trials. We consider the SED and compare it with complete randomization (CR) in our evaluation in the followings.

Check Type I Error Drug and placebo are equally effective in all the three components.

All Type I errors in Table 12 are preserved when sample size N is big. In addition, The Type I error under stratified matched WR is preserved more slowly than others.

Power Comparison

Scenario 1 The drug is equally effective in improving all three components, and it's more effective than placebo in all the three components. The results are in Table 13.

When $\sum_{j=1}^{3} |\beta_{pj} - \beta_{tj}| = 1.5$, SED always outperforms CR. The WR methods for composite components under both designs achieve higher power than other tests when sample size N is large. Stratified methods have higher power than nonstratified methods.

Scenario 2 The drug is much more effective than placebo in the first component, but it's equally effective as placebo in the 2nd and 3rd components. We decrease the drug's overall efficacy to the three components. The results are in Table 14.

When $\sum_{j=1}^{3} |\beta_{pj} - \beta_{tj}| = 0.5$, although powers decrease, SED still outperforms CR.

Scenario 3 We keep assuming that a drug is equally effective in improving the three components and more effective than placebo. However, we adjust the distribution of patients by decreasing the proportion of the target patient p_3. The results are in Table 15. When $\sum_{j=1}^{3} |\beta_{pj} - \beta_{tj}| = 1.5$ and target population is low, the SED even more outperforms the CR than the scenario when $p_3 = 0.8$ when the sample size N is small.

Table 9. Power comparison with setting $\beta_t = 0$ and $\beta_{in} = \log(0.18)$ ($\beta' = \beta_t + \beta_{in} = \log(0.18)$) such that HR = 0.6 under cox regression.

Power	$N = 60$	$N = 100$	$N = 200$
Cox regression	0.51	0.65	0.81
Stratified matched WR	0.78	0.94	0.99
Stratified unmatched WR	0.90	0.99	1
Unstratified unmatched WR	0.89	0.99	1
O'Brien's rank-sum-type test	0.50	0.74	0.93

Table 8. The estimation of treatment effect of different sample sizes.

Total Sample Size	$N = 60$		$N = 100$		$N = 200$	
Estimation	Beta (SE)	CI	Beta (SE)	CI	Beta (SE)	CI
HR	0.61 (0.25)	(0.35, 1.09)	0.59 (0.16)	(0.39, 0.91)	0.60 (0.11)	(0.45, 0.81)
Stratified matched WR	3.02 (4.48)	(1.38, 11.8)	3.06 (2.31)	(1.66, 7.58)	2.98 (1.13)	(1.92, 5.20)
Stratified unmatched WR	3.29 (1.80)	(1.58, 6.84)	3.24 (1.20)	(1.87, 5.59)	3.05 (0.70)	(2.10, 4.43)
Unstratified unmatched WR	3.14 (1.58)	(1.59, 6.23)	3.09 (1.10)	(1.84, 5.21)	2.96 (0.67)	(2.06, 4.25)

Table 10. The estimation of treatment effect of different sample sizes.

Total Sample Size	$N = 60$		$N = 100$		$N = 200$	
Estimation	Beta (SE)	CI	Beta (SE)	CI	Beta (SE)	CI
HR	0.60 (0.24)	(0.34, 1.06)	0.59 (0.16)	(0.39, 0.91)	0.60 (0.11)	(0.44, 0.81)
Stratified matched WR	1.12 (0.78)	(0.49, 2.65)	1.17 (0.65)	(0.63, 2.22)	1.12 (0.31)	(0.74, 1.73)
Stratified unmatched WR	1.19 (0.54)	(0.62, 2.29)	1.19 (0.39)	(0.73, 1.96)	1.15 (0.23)	(0.82, 1.61)
Unstratified unmatched WR	1.18 (0.52)	(0.64, 2.19)	1.17 (0.37)	(0.73, 1.88)	1.14 (0.22)	(0.82, 1.58)

Table 11. Power comparison with setting $\beta_t = 0$ and $\beta_{in} = \log(0.18)$ ($\beta' = \beta_t + \beta_{in} = \log(0.18)$) such that HR $= 0.6$ under cox regression.

Power	$N = 60$	$N = 100$	$N = 200$
Cox regression	0.50	0.66	0.82
Stratified matched WR	0.07	0.07	0.10
Stratified unmatched WR	0.06	0.09	0.12
Unstratified unmatched WR	0.09	0.07	0.11
O'Brien's rank-sum-type test	0.51	0.72	0.91

In summary, given the same sample size N, the power of SED is at least approximately equal to or greater than the one under CR, especially for smaller N. That is, two-stage enrichment designs can further enhance trial efficiency, especially for a small-size clinical trial. Let us take the 'stratified unmatched WR' as an example. In Table 14 (scenario 2), when $N = 100$ the 'stratified unmatched WR' under SED increases 14% power than the 'Contingency Table' (i.e. the traditional method) but increases 4% under CR; when $N = 500$ the 'stratified unmatched WR' under SED continues to increase 14% power and increases 12% under CR. The 'unstratified unmatched WR' has the same trend. Table 15 (scenario 3) further confirms the benefit of SED in improving power for win-ratio methods.

APPENDIX A. APPENDIX

A.1 Derivation of $p_.$ under Matched Win Ratio

We consider all the scenarios that treatment wins and the corresponding probability p_w.

$$\begin{aligned} p_w &= P(Y_T = 1, X_t = 0, Y_c = 1, X_c = 1) \\ &\quad + P(Y_T = 0, X_t = 1, Y_c = 1, X_c = 0) \\ &\quad + P(Y_T = 0, X_t = 1, Y_c = 1, X_c = 1) \\ &\quad + P(Y_T = 0, X_t = 0, Y_c = 1, X_c = 0) \\ &\quad + P(Y_T = 0, X_t = 0, Y_c = 1, X_c = 1) \\ &\quad + P(Y_T = 0, X_t = 0, Y_c = 0, X_c = 1) \\ &= p_t(1 - q_t)p_c q_c + (1 - p_t)q_t p_c \\ &\quad + (1 - p_t)(1 - q_t)\big(1 - (1 - p_c)(1 - q_c)\big). \end{aligned}$$

Also, we consider all the scenarios that control wins and the corresponding probability p_l.

$$\begin{aligned} p_l &= P(Y_T = 1, X_t = 0, Y_c = 0, X_c = 1) \\ &\quad + P(Y_T = 1, X_t = 0, Y_c = 0, X_c = 0) \\ &\quad + P(Y_T = 1, X_t = 1, Y_c = 0, X_c = 1) \\ &\quad + P(Y_T = 1, X_t = 1, Y_c = 0, X_c = 0) \\ &\quad + P(Y_T = 1, X_t = 1, Y_c = 1, X_c = 0) \end{aligned}$$

Table 12. Type I error comparison with setting $(p_1, p_2, p_3, p_4) = (0.05, 0.05, 0.8, 0.1)$, $\epsilon \sim N(0, 1)$, $\beta_{pj} = \beta_{t1} = -1.5$, $\beta_{in2} = \beta_{in3} = 0$, $\beta_{cov1} = \beta_{cov2} = 5$, $c_t = 0.8$, $c_{s0} = 0.8$, $c_{s1} = 0.9$.

Type I error	$N = 100$		$N = 200$		$N = 500$	
Design	CR	SED	CR	SED	CR	SED
Contingency table	0.05	0.05	0.05	0.05	0.05	0.05
Stratified matched WR	0.08	0.13	0.07	0.07	0.06	0.06
Stratified unmatched WR	0.05	0.05	0.06	0.04	0.05	0.05
Unstratified unmatched WR	0.05	0.05	0.06	0.04	0.05	0.05

Table 13. Power comparison with setting $(p_1, p_2, p_3, p_4) = (0.05, 0.05, 0.8, 0.1)$, $\epsilon \sim N(0, 1)$, $\beta_{pj} = -1.5$, $\beta_{t1} = -2$, $\beta_{in2} = \beta_{in3} = 0$, $\beta_{cov1} = \beta_{cov2} = 5$, $c_t = 0.8$, $c_{s0} = 0.8$, $c_{s1} = 0.9$.

Power	$N = 100$		$N = 200$		$N = 500$	
Design	SED	CR	SED	CR	SED	CR
Contingency table	0.30	0.30	0.58	0.45	0.92	0.90
Stratified matched WR	0.48	0.46	0.77	0.69	0.99	0.99
Stratified unmatched WR	0.49	0.47	0.81	0.74	0.99	0.99
Unstratified unmatched WR	0.33	0.32	0.59	0.51	0.92	0.93

Table 14. Power comparison with setting $(p_1, p_2, p_3, p_4) = (0.05, 0.05, 0.8, 0.1)$, $\varepsilon \sim N(0,1)$, $\beta_{pj} = -1.5$, $\beta_{t1} = -2$, $\beta_{in2} = \beta_{in3} = 0.5$, $\beta_{cov1} = \beta_{cov2} = 5$, $c_t = 0.8$, $c_{s0} = 0.8$, $c_{s1} = 0.9$.

Power	$N = 100$		$N = 200$		$N = 500$	
Design	SED	CR	SED	CR	SED	CR
Contingency table	0.09	0.07	0.16	0.13	0.27	0.20
Stratified matched WR	0.15	0.14	0.23	0.20	0.40	0.31
Stratified unmatched WR	0.23	0.11	0.27	0.17	0.41	0.32
Unstratified unmatched WR	0.22	0.07	0.24	0.14	0.32	0.22

Table 15. Power comparison with setting $(p_1, p_2, p_3, p_4) = (0.6, 0.05, 0.3, 0.05)$, $\epsilon \sim N(0,1)$, $\beta_{pj} = -1.5$, $\beta_{t1} = -2$, $\beta_{in2} = \beta_{in3} = 0$, $\beta_{cov1} = \beta_{cov2} = 5$, $c_t = 0.8$, $c_{s0} = 0.8$, $c_{s1} = 0.9$.

Power	$N = 100$		$N = 200$		$N = 500$	
Design	SED	CR	SED	CR	SED	CR
Contingency table	0.07	0.06	0.10	0.10	0.20	0.17
Stratified matched WR	0.12	0.07	0.13	0.11	0.23	0.23
Stratified unmatched WR	0.23	0.07	0.25	0.15	0.33	0.26
Unstratified unmatched WR	0.20	0.06	0.24	0.10	0.29	0.19

$$
\begin{aligned}
&+ P(Y_T = 0, X_t = 1, Y_c = 0, X_c = 0) \\
&= p_t(1-q_t)(1-p_c) \\
&\quad + p_t q_t (1 - p_c q_c) + (1-p_t) q_t (1-p_c)(1-q_c).
\end{aligned}
$$

Then, we consider all the scenarios that treatment and control tie and the corresponding probability p_{tie}.

$$
\begin{aligned}
p_{tie} &= P(Y_T = 1, X_t = 0, Y_c = 1, X_c = 0) \\
&\quad + P(Y_T = 1, X_t = 1, Y_c = 1, X_c = 1) \\
&\quad + P(Y_T = 0, X_t = 1, Y_c = 0, X_c = 1) \\
&\quad + P(Y_T = 0, X_t = 0, Y_c = 0, X_c = 0) \\
&= 1 - p_w - p_l.
\end{aligned}
$$

Suppose a total of N units are randomized, and we let $n = N(1 - p_{tie})$ denote the total number of non-tie units. Also, we let the binary random variable X_i follow $Bernoulli(p)$, where

$$
\begin{aligned}
p &= P(\text{treatment win} | \text{all non-tie pairs}) \\
&= \frac{P(\text{treatment wins in all pairs})}{P(\text{non-tie pairs})} \\
&= \frac{p_w}{1 - p_{tie}}.
\end{aligned}
$$

A.2 Derivation of $g(\mathbf{X})$ under Unmatched Win Ratio

Here we derive the $g(\mathbf{X})$ in equation (3.3)

$$
g(\mathbf{X}) = \frac{\bar{Y}_t - \bar{Y}_t\bar{Y}_c - 2X\bar{Y}_tX\bar{Y}_c + \bar{X}_tX\bar{Y}_c + 2X\bar{Y}_t\bar{Y}_c - \bar{X}_t\bar{Y}_c + \bar{X}_cX\bar{Y}_t - X\bar{Y}_t - \bar{X}_cX\bar{X}_t + \bar{X}_t}{2X\bar{Y}_c\bar{Y}_t - \bar{Y}_c\bar{Y}_t - \bar{X}_c\bar{Y}_t - 2X\bar{Y}_tX\bar{Y}_c + \bar{X}_tX\bar{Y}_c - X\bar{Y}_c + \bar{Y}_c + \bar{X}_cX\bar{Y}_t - \bar{X}_cX\bar{X}_t + \bar{X}_c}
$$

$$
\begin{aligned}
&g(\boldsymbol{\theta}) \\
&= g(E(\mathbf{X})) \\
&= \frac{p_t - p_tp_c - 2p_tq_tp_cq_c + q_tq_tp_cq_c + 2p_tq_tp_c - q_tp_c + q_cp_tq_t - p_tq_t - q_cq_t + q_t}{2p_cq_cp_t - p_cp_t - q_cp_t - 2p_tq_tp_cq_c + q_tp_cq_c - p_cq_c + p_c + q_cp_tq_t - q_cq_t + q_c}
\end{aligned}
$$

ACKNOWLEDGEMENTS

The authors express their gratitude to editorial support that greatly enhanced the presentation of this manuscript. Disclaimer: The contents, views or opinions expressed in this publication or presentation are those of the authors and do not necessarily reflect official policy or position of the U.S. Food and Drug Administration.

FUNDING

This work was supported by the ORISE Research Program of the U.S. Food and Drug Administration.

Accepted 12 March 2025

REFERENCES

[1] Bebu, I. and Lachin, J. M. (2016). Large sample inference for a win ratio analysis of a composite outcome based on prioritized components. *Biostatistics* **17** 178–187. https://doi.org/10.1093/biostatistics/kxv032. MR3449859

[2] O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40** 1079–1087. https://doi.org/10.2307/2531158. MR0786180

[3] Chen, Y. F., Zhang, X., Tamura, R. N. and Chen, C. M. (2014). A sequential enriched design for target patient population in psychiatric clinical trials. *Statistics in Medicine* **33** 2953–2967. https://doi.org/10.1002/sim.6116. MR3260515

[4] Fava, M., Evins, A. E., Dorer, D. J. and Schoenfeld, D. A. (2003). The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics* **72** 115–127.

[5] Finkelstein, D. M. and Schoenfeld, D. A. (1999). Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine* **18** 1341–1354.

[6] Finkelstein, D. M. and Schoenfeld, D. A. (2019). Graphing the win ratio and its components over time. *Statistics in Medicine* **38** 53–61. https://doi.org/10.1002/sim.7895. MR3887266

[7] Food and Drug Administration (2017). BRINEURA (Cerliponase Alfa) Injection.
[8] Food and Drug Administration, Center for Drug Evaluation and Research (2017 Dec). Pediatric Rare Diseases–A Collaborative Approach for Drug Development Using Gaucher Disease as a Model. Draft Guidance for Industry.
[9] Guo, M., Ma, Y., Eworuke, E., Khashei, M., Song, J., Zhao, Y. and Jin, F. (2023). Identifying COVID-19 cases and extracting patient reported symptoms from Reddit using natural language processing. *Scientific Reports* **13(1)** 13721.
[10] Harden, J. J. and Kropko, J. (2019). Simulating duration data for the Cox model. *Political Science Research and Methods* **7**(4) 921–928. https://doi.org/10.1017/psrm.2018.19.
[11] Kropko, J. and Harden, J. J. (2020). Beyond the hazard ratio: generating expected durations from the Cox proportional hazards model. *British Journal of Political Science* **50**(1) 303–320. https://doi.org/10.1017/S000712341700045X.
[12] Luo, X., Tian, H., Hong, M., Surya, T. and Wei, Y. (2015). An alternative approach to confidence interval estimation for the win ratio statistics. *Biometrics* **71** 139–145. https://doi.org/10.1111/biom.12225. MR3335358
[13] Mao, L., Kim, K.-M. and Miao, X. (2022). Sample size formula for general win ratio analysis. *Biometrics* **78** 1257–1268. https://doi.org/10.1111/biom.13501. MR4493522
[14] Mielke, J., Jones, B., Posch, M. and König, F. (2021). Testing procedures for claiming success. *Biopharmaceutical Research* **13** 106–112.
[15] Pocock, S. J., Ariti, C. A., Collier, T. J. and Wang, D. (2012). The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal* **33** 176–182. https://doi.org/10.1002/sim.6205. MR3274506
[16] Redfors, B., Gregson, J., Crowley, A., McAndrew, T., Ben-Yehuda, O., Stone, G. W. and Pocock, S. J. (2020). The win ratio approach for composite endpoints: practical guidance based on previous experience. *European Heart Journal* **41** 4391–4399.
[17] Sharma, A., Frangoul, H., Locatelli, F., Kuo, K., Bhatia, M., Mapara, M., Eckrich, M., Imren, S., Li, N., Rubin, J., Zhang, S., Liu, T., Hobbs, W. and Grupp, S. A. (2024). Health-related quality-of-life improvements after Exagamglogene autotemcel in patients with severe sickle cell disease. *Blood* **144** 7453.
[18] U.S. Congress (2002). Rare Disease Act of 2002. Public Law No. 107-280.
[19] Wang, D. and Pocock, S. (2016). A win ratio approach to comparing continuous non-normal outcomes in clinical trials. *Pharmaceutical Statistics* **15** 238–245.
[20] Wang, J., Li, P. and Hu, F. (2023). A/B testing in network data with covariate-adaptive randomization. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR. https://doi.org/10.1002/sam.70003. MR4853591
[21] Yin, X., Hamasaki, T. and Evans, S. (2021). Sequential multiple assignment randomized trials for COMparing Personalized Antibiotic StrategieS (SMART COMPASS): design considerations. *Statistics in Biopharmaceutical Research* **13(2)** 181–191.
[22] Zhang, X., Chen, Y.-F. and Tamura, R. (2018). The plan of enrichment designs for dealing with high placebo response. *Pharmaceutical Statistics* **17(1)** 25–37.

Jialu Wang. Department of Statistics, George Washington University, USA.
E-mail address: jialu@gwmail.gwu.com

Yeh-Fong Chen. Division of Biometrics IX, OB/OTS/CDER, FDA, Silver Spring, MD, USA.
E-mail address: yehfong.chen@fda.hhs.gov

Thomas Gwise. Division of Biometrics IX, OB/OTS/CDER, FDA, Silver Spring, MD, USA.
E-mail address: thomas.gwise@fda.hhs.gov