


## Article

# Exact Permutation and Bootstrap Distribution of Generalized Pairwise Comparisons Statistics

William N. Anderson<sup>1</sup> and Johan Verbeeck<sup>2,\*</sup> <sup>1</sup> Independent Researcher, Carpinteria, CA 93013, USA<sup>2</sup> Data Science Institute, I-Biostat, University of Hasselt, 3590 Diepenbeek, Belgium

\* Correspondence: johan.verbeeck@uhasselt.be; Tel.: +32-11-26-82-68

**Abstract:** To analyze multivariate outcomes in clinical trials, several authors have suggested generalizations of the univariate Mann–Whitney test. As the Mann–Whitney statistic compares the subjects' outcome pairwise, the multivariate generalizations are known as generalized pairwise comparisons (GPC) statistics. For GPC statistics such as the net treatment benefit, the win ratio, and the win odds, asymptotic based or re-sampling tests have been suggested in the literature. However, asymptotic methods require a sufficiently high sample size to be accurate, and re-sampling methods come with a high computational burden. We use graph theory notation to obtain closed-form formulas for the expectation and the variance of the permutation and bootstrap sampling distribution of the GPC statistics, which can be utilized to develop fast and accurate inferential tests for each of the GPC statistics. A simple example and a simulation study demonstrate the accuracy of the exact permutation and bootstrap methods, even in very small samples. As the time complexity is  $O(N^2)$ , where  $N$  is the total number of patients, the exact methods are fast. In situations where asymptotic methods have been used to obtain these variance matrices, the new methods will be more accurate and equally fast. In situations where bootstrap has been used, the new methods will be both more accurate and much faster.

**Keywords:** bootstrap test; generalized pairwise comparisons; graph theory; multivariate outcome; net treatment benefit; permutation test; win odds; win ratio

**MSC:** 62G99

**Citation:** Anderson, W.N.; Verbeeck, J. Exact Permutation and Bootstrap Distribution of Generalized Pairwise Comparisons Statistics. *Mathematics* **2023**, *11*, 1502. <https://doi.org/10.3390/math11061502>

Academic Editor: Christophe Chesneau

Received: 21 February 2023

Revised: 16 March 2023

Accepted: 17 March 2023

Published: 20 March 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In randomized clinical trials, often multiple outcome measures are selected to assess the effect of a treatment on two groups of subjects. In clinical areas, such as cardiology, oncology, psychology, and infectious disease, the combination of several clinical meaningful outcome measures is very common [1–4]. The standard method of analysis for multivariate survival outcomes is a time-to-first event analysis, which uses only the first event per subject [5]. However, this approach has been criticized, since it ignores subsequent events and the clinical relevance of the events. To remediate the shortcomings of the time-to-first event analysis, a new class of non-parametric methods has been proposed that allows for the analysis of multiple events per subject and allows for prioritizing the events by clinical relevance [6–8]. These methods are a generalization to multivariate prioritized outcomes of the classical univariate Mann–Whitney [9] and Gehan–Wilcoxon test [10,11]. Similar to the univariate outcome statistic, the generalized statistics compare the outcome per pair of subjects and hence are named generalized pairwise comparisons (GPC) statistics [7], while others use the term win statistics [12].

GPC is based on a prioritization of a number of  $k$  outcome measures, usually according to a decreasing level of clinical relevance. For each outcome measure, every subject is compared to every other subject in a pairwise manner. Per pair and per outcome, a score,

$u_{ijk}$ , is assigned, which is chosen to reflect whether subject  $i$  or subject  $j$  has the more favorable outcome  $k$ . The concept of more favorable can involve censoring, missing data, and can include a threshold of clinical relevance [7,13]. The score  $u_{ijk}$  that is most often used is the score proposed by Gehan [10] for censored observations:

$$u_{ijk} = \begin{cases} 1, & \text{if subject } i \text{ has a more favorable outcome } k \text{ than subject } j \\ 0, & \text{if the subjects cannot be compared on outcome } k \text{ or are tied} \\ -1, & \text{if subject } i \text{ has a less favorable outcome } k \text{ than subject } j \end{cases}$$

The scores can be computed separately for each outcome measure  $k$ , and the score  $u_{ij}$  for a pair in the prioritized GPC is the first non-zero score  $u_{ijk}$ , where the first is defined by the prespecified priority list of outcomes. It is quite possible that for many pairs of subjects, the overall comparison will remain 0. It is also possible to have non-transitive comparisons, such as three patients where  $u_{ij} = 1$ ,  $u_{jk} = 1$ , and  $u_{ki} = 1$  [13]. Alternatively, if prioritization of the outcome measures is not feasible or appropriate, a non-prioritized GPC sums the scores  $u_{ijk}$  over all outcome measures [13,14]. Additionally, the scores of the pairs that are uninformative due to censored observations, and thus cannot be compared, can be corrected by estimating the chance of a favorable outcome using Kaplan–Meier estimates [15–18]. The Kaplan–Meier corrected score matrix will consist of scores  $|u_{ij}| < 1$ . The score matrix resulting from the pairwise comparisons is a skew-symmetric  $N \times N$  matrix, with entries in the set  $\{-1, 0, +1\}$  for the prioritized GPC without the censoring corrections and entries in  $\mathbb{R}$  otherwise. Although scores in a non-prioritized GPC can be  $> 1$  per pair, most GPC statistics will adjust the scoring by dividing by  $k$  outcomes. Because of the summing, this adjustment can be performed at the level of the scores or after summing all scores. In the former, scores will be  $|u_{ij}| < 1$ , while in the latter, scores will be  $|u_{ij}| > 1$ .

Inferential tests and confidence intervals for GPC are then constructed from the scoring matrix, utilizing statistics such as the Finkelstein–Schoenfeld statistic [6], the net treatment benefit [7], the win ratio [8] and win odds [19]. Suppose there are  $N$  subjects in a two-arm trial, with  $m$  subjects in the experimental group from a distribution  $F_1$  and  $n$  subjects in the control group from a distribution  $F_2$ . Moreover, let the indicator  $D_i = 1$  for subjects in the experimental group, and  $D_i = 0$  for patients in the control group. The Finkelstein–Schoenfeld statistic [6] is then the sum of the scores for the experimental group. If  $U_i = \sum_j u_{ij}$ , the Finkelstein–Schoenfeld statistic is:

$$FS = \sum_{i=1}^N D_i U_i, \quad (1)$$

which can be interpreted as the difference between the favorable outcomes in the experimental arm,  $W_T = \sum_{i=1}^N D_i U_i$  with  $u_{ij} > 0$ , and the favorable outcomes in the control arm,  $W_C = \sum_{i=1}^N D_i U_i$  with  $u_{ij} < 0$ . If the score entries are restricted to  $\{-1, 0, +1\}$ ,  $W_T$  and  $W_C$  merely count the number of wins per treatment arm, and  $FS = W_T - W_C$ .

It is easy to show (Appendix A) that  $\mathbb{E}(FS) = 0$ . The variance suggested by Finkelstein and Schoenfeld [6] is based on a permutation test

$$\text{Var}(FS) = \frac{mn}{N(N-1)} \sum_{i=1}^N U_i^2, \quad (2)$$

following [10,11,20]. Although Finkelstein and Schoenfeld presented the formula for  $\text{Var}(FS)$ , they did not give expectations and variances for  $W_T$  and  $W_C$  separately. For many of the GPC models considered below, those separate variances are needed.

The permutation test assumes exchangeability between observations and tests the null hypothesis:  $H_0 : F_1 = F_2$ . Gehan [10], Gilbert [11] and Mantel [20] derived a closed-form expression of the mean and variance of the exact null distribution, which means that no re-sampling permutation is actually required. Consequently, the asymptotic null distribution of the test statistic  $FS / \sqrt{\text{Var}(FS)}$  is shown to be standard normal [10,11], and the  $p$ -values

computed from this null distribution are valid, even for very small sample sizes of only 10 observations [10]. Alternatively, inference can be based on the proportion of re-sampling permutation samples with a test statistic greater than or equal to the observed statistic. Even though the permutation test is formulated under the  $H_0$  null hypothesis, the test statistic is only consistent under alternatives of the form  $FS \neq 0$  [10,11]. The permutation test is thus not consistent against all alternatives of the form  $F_1 \neq F_2$  [21,22].

Other GPC statistics include the net treatment benefit (NTB) [7], which is a U-statistic [23] transformation of the Finkelstein–Schoenfeld statistic, by dividing it by the total number of pairs possible between the two treatment arms,  $(W_T - W_C)/(nm)$ . It has been shown that the NTB and its transformations are unbiased and efficient estimators for univariate and uncensored observations [24]. The win ratio (WR) is defined as the ratio of the number of times a subject has a favorable outcome in the treatment arm and in the control arm,  $W_T/W_C$  [8]. Finally, the win odds adds half of the number of tied outcomes ( $W_0$ ) to both the numerator and denominator,  $(W_T + 1/2W_0)/(W_C + 1/2W_0)$  [25]. Since the win odds (WO) is a transformation of the net treatment benefit,  $WO = (1 + NTB)/(1 - NTB)$ , any test proposed for the net treatment benefit or Finkelstein–Schoenfeld statistic can also be applied to the win odds, and we will not focus further on the win odds.

For the inference of the GPC statistics, also re-sampling bootstrap [8], re-sampling re-randomization [7] or U-statistic asymptotic methods have been proposed [14,26,27]. The asymptotic properties are used in two different ways:

- Asymptotic formulas for the mean and variances of the numbers of wins,  $W_T$  and  $W_C$ , which are both U-statistics and thus asymptotically normal distributed [23]. Our method avoids the errors in these asymptotic formulas, since we give exact formulas.
- Delta method derivations of results for the win ratio based on the computed approximate means and variances.

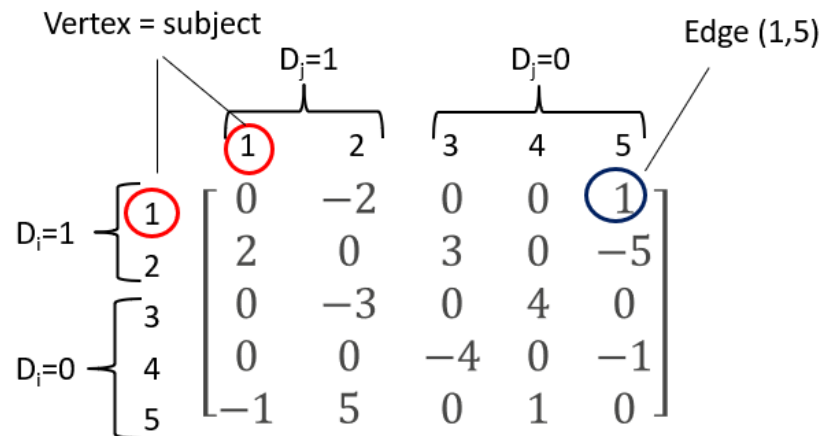
For larger trials, the asymptotic formulas may be satisfactory, but the re-sampling methods require a large number of replications to be accurate. In contrast, the permutation test (2) presented by Finkelstein and Schoenfeld [6] avoids the first step in the use of asymptotic properties and the associated error. This would be important in smaller clinical trials, such as in rare disease trials. Moreover, as the permutation test is an exact test in finite sampling, it is more accurate than the re-sampling methods and less time-consuming.

While the permutation test, as derived by Gehan [10], can be easily extended to the net treatment benefit and the win odds through their transformations, it is not obvious to extend it to the win ratio. Moreover, the permutation test determines the variability under the null hypotheses, making it unsuited for the determination of a confidence interval of the GPC statistics. Therefore, we re-derive the closed-form expression of the expectation and the variance of the permutation distribution using graph theory notation on the skew-symmetric score matrix, which allows extension of these results to a two-sample bootstrap distribution and to the win ratio. Additionally, we generalize the expression to allow scores in the  $\mathbb{R}$  field, generalizing the application to non-prioritized GPC algorithms and censoring correcting scores. It will be shown that the algorithm complexity of our expressions is  $O(N^2)$  in both time and space, and thus, the exact permutation and bootstrap method will be both faster and more accurate than re-sampling permutation and bootstrap tests.

This paper is organized as follows. In Section 2, the graphical model is presented that will be used to derive the expected values and variances of the GPC statistic under the permutation and bootstrap distribution. The exact permutation formulas for the mean and variance are derived in Section 3, while the exact bootstrap formulas follow in Section 4. The time complexity of the exact methods is evaluated in Section 5. Throughout the development of the theory, a small example is used to demonstrate the new methodologies. Additionally, a simulation shows the type I error and the 95% confidence interval coverage in Section 6, and the methodology is demonstrated in an example in Section 7. Both R<sup>®</sup> and SAS code for the exact permutation, and bootstrap tests are presented in the Supplementary Material as well as full detailed derivations and an extensive test of the algorithm.

## 2. Graphical Model

Any pairwise comparison analysis of an outcome measure in  $N = n + m$  subjects, that allows for non-prioritized analyses and censoring corrected scores to denote the more favorable outcome, will result in a skew-symmetry  $N \times N$  score matrix  $U$  with entries in the  $\mathbb{R}$  field. This will be the case when evaluating both a single outcome measure, by use of the Mann–Whitney test, or the Gehan–Wilcoxon test, and when evaluating multiple outcomes, by a GPC analysis. In the remainder of this manuscript, we consider that the score matrix  $U$  has been produced by some further unspecified mechanism; the only restriction is the skewness and symmetry. A small example (Figure 1) illustrates such a  $U$  matrix.



**Figure 1.** A small example of a score matrix  $U$ , resulting from a pairwise comparison. Here,  $D_{ij} = 1$  represents the treatment subjects,  $D_{ij} = 0$  represents the control subjects.

Since all GPC statistics can be constructed from  $W_T$  and  $W_C$ , it is our aim to compute the expectations and the variances of  $W_T$  and  $W_C$  as well as their covariance for both the permutation as the bootstrap distribution. These mean and (co)variances are computed over all possible permutations of the treatment arms or all possible bootstrap samples. For the net treatment benefit ( $\Delta$ ),

$$\mathbb{E}(\Delta) = \frac{\mathbb{E}(W_T) - \mathbb{E}(W_C)}{nm} \quad (3)$$

$$\text{Var}(\Delta) = \frac{\text{Var}(W_T) + \text{Var}(W_C) - 2\text{Cov}(W_T W_C)}{(nm)^2}. \quad (4)$$

For the win ratio ( $\Psi$ ), it is known that the logarithm of the win ratio is approximately normal distributed [24,26]. Its variance can then be approximated using the delta method:

$$\mathbb{E}(\Psi) \approx \frac{\mathbb{E}(W_T)}{\mathbb{E}(W_C)} \quad (5)$$

$$\text{Var}(\log \Psi) \approx \frac{\text{Var}(W_T)}{W_T^2} + \frac{\text{Var}(W_C)}{W_C^2} - 2 \frac{\text{Cov}(W_T W_C)}{W_T W_C}. \quad (6)$$

In order to derive the expectations, variances, and covariance for the  $W_T$  and  $W_C$ , it will be convenient to think of the score matrix  $U$  as the adjacency matrix of a directed graph  $\mathbb{G}$  [28] with  $N$  vertices, where a vertex represents a subject. A pair of subjects or vertices can be joined by an edge  $e = (i, j)$ . Since we are interested in the favorable outcomes, an edge  $e = (i, j)$  is defined when  $u_{ij} > 0$ . When  $D_i = 1$  and  $D_j = 0$ , such an edge is called a treatment edge and when  $D_i = 0$  and  $D_j = 1$ , it is called a control edge. The value  $u_{ij}$  is called the weight of the edge, and it is denoted by  $w_e$ . Note that the score of the comparison of subject  $i$  with subject  $j$  is exactly the opposite of the score of the comparison of subject  $j$  with subject  $i$ . Hence, the edge  $e = (i, j)$  is an ordered pair of distinct vertices, where the

head of  $e$  is vertex  $j$  and the tail is vertex  $i$ . In a pictorial representation, the edge  $e = (i, j)$  is represented by an arrow going from vertex  $i$  to vertex  $j$  (Figure 2). Let then  $E$  denote the total number of edges of  $\mathbb{G}$ . In our small example (Figures 1 and 2),  $E = 6$ .

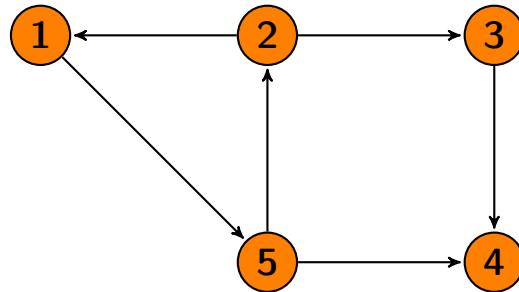


Figure 2. Small graph example.

Using the graph theory notation, the number of wins for the treatment,  $W_T$ , and control arm,  $W_C$ , can be redefined as follows. For a subject or vertex  $v$ , let  $N_v$  denote the number of times that the vertex  $v$  appears in the sample. In a permutation sample, each subject can only appear once and  $N_v = 1$  for all samples. In a bootstrap sample, subjects can appear more than once and  $N_v$  can be zero or higher. To count the number of wins for the treatment arm, the value  $T_e$  is defined as the number of times that edge  $e$  is a treatment edge in the sample, and similarly, the value  $C_e$  is the number of times that edge  $e$  is a control edge. If the treatment edge  $e$  has subjects  $(v, w)$ , the number of times this edge appears in the sample is then  $T_e = N_v N_w$ . For example, suppose that in a bootstrap sample of the five subjects in our small example (Figure 1), subjects 1 and 2 appear twice, subject 3 appears once, and subjects 4 and 5 do not appear at all; then,  $N_1 = N_2 = 2$ ,  $N_3 = 1$ , and  $N_4 = N_5 = 0$ . Furthermore, the pair or treatment edge  $(2,1)$  will appear  $T_{e(2,1)} = N_1 \times N_2 = 4$  times, the treatment edge  $T_{e(2,3)} = 2$  times and all other treatments edges  $T_e = 0$  times. The vectors  $\mathbf{T}$  and  $\mathbf{C}$  are the  $E \times 1$  column vectors composed of the various  $T_e$  and  $C_e$  and the vector  $\mathbf{W}$  is the column vector of the weights  $w_e$ .

$W_T$  and  $W_C$  are then redefined as:

$$W_T = \mathbf{W}^t \mathbf{T} = \sum_e w_e T_e = [w_1, w_2, \dots, w_E] \mathbf{T}$$

$$W_C = \mathbf{W}^t \mathbf{C} = \sum_e w_e C_e = [w_1, w_2, \dots, w_E] \mathbf{C}$$

In our small example (Figure 1):

- Vector of edges =  $\{(1,5), (2,1), (2,3), (3,4), (5,2), (5,4)\}$ ;
- Vector  $\mathbf{W}^t = \{1, 2, 3, 4, 5, 1\}$ ;
- Vector  $\mathbf{T}^t = \{1, 0, 1, 0, 0, 0\}$  and  $W_T = 4$ ;
- Vector  $\mathbf{C}^t = \{0, 0, 0, 0, 1, 0\}$  and  $W_C = 5$ .

The development of the expectations and variances of  $W_T$  and  $W_C$ , and their covariance for the permutation and the bootstrap distribution, then follow the same general pattern.

The expectation of  $W_T$  is derived by:

$$\begin{aligned} \mathbb{E}(W_T) &= \mathbb{E}(\mathbf{W}^t \mathbf{T}) \\ &= [w_1, w_2, \dots, w_E] \mathbb{E}(\mathbf{T}) \\ &= \sum_e w_e \mathbb{E}(T_e). \end{aligned}$$

and similarly,  $\mathbb{E}(W_C) = \sum_e w_e \mathbb{E}(C_e)$ .

The variance is derived by

$$\begin{aligned}
 \text{Var}([W_T, W_C]) &= \text{Var}\left([\mathbf{W}^t, \mathbf{W}^t] \begin{bmatrix} \mathbf{T} \\ \mathbf{C} \end{bmatrix}\right) \\
 &= [\mathbf{W}^t, \mathbf{W}^t] \text{Var}\left(\begin{bmatrix} \mathbf{T} \\ \mathbf{C} \end{bmatrix}\right) \begin{bmatrix} \mathbf{W} \\ \mathbf{W} \end{bmatrix} \\
 &= [\mathbf{W}^t, \mathbf{W}^t] \left( \mathbb{E}\left(\begin{bmatrix} \mathbf{T} \\ \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{T} \\ \mathbf{C} \end{bmatrix}^t\right) - \left(\mathbb{E}\begin{bmatrix} \mathbf{T} \\ \mathbf{C} \end{bmatrix}\right) \left(\mathbb{E}\begin{bmatrix} \mathbf{T} \\ \mathbf{C} \end{bmatrix}^t\right) \right) \begin{bmatrix} \mathbf{W} \\ \mathbf{W} \end{bmatrix} \\
 &= [\mathbf{W}^t, \mathbf{W}^t] \left( \mathbb{E}\left(\begin{bmatrix} \mathbf{T}\mathbf{T}^t & \mathbf{T}\mathbf{C}^t \\ \mathbf{C}\mathbf{T}^t & \mathbf{C}\mathbf{C}^t \end{bmatrix}\right) - \begin{bmatrix} \mathbb{E}(\mathbf{T})\mathbb{E}(\mathbf{T}^t) & \mathbb{E}(\mathbf{T})\mathbb{E}(\mathbf{C}^t) \\ \mathbb{E}(\mathbf{C})\mathbb{E}(\mathbf{T}^t) & \mathbb{E}(\mathbf{C})\mathbb{E}(\mathbf{C}^t) \end{bmatrix} \right) \begin{bmatrix} \mathbf{W} \\ \mathbf{W} \end{bmatrix} \quad (7) \\
 &= \begin{bmatrix} V_{TT} & V_{TC} \\ V_{CT} & V_{CC} \end{bmatrix} \quad (8)
 \end{aligned}$$

The variance of the GPC statistics thus requires two counting steps. In the first step, the expected values  $\mathbb{E}(T_e)$ ,  $\mathbb{E}(C_e)$  and the expected value of an ordered pair of edges  $(e, f)$ , not necessarily distinct,  $\mathbb{E}(T_e T_f)$ ,  $\mathbb{E}(T_e C_f)$ ,  $\mathbb{E}(C_e T_f)$ , and  $\mathbb{E}(C_e C_f)$  are computed using elementary calculations involving binomial coefficients. These calculations differ between edge pairs, depending on the trial arm assignments and the geometric relationship of the edges. Note that because the variance matrix is symmetric, we do not explicitly need the individual terms  $\mathbb{E}(C_e T_f)$ . In the second step, the number of times that each of these geometric configurations of edges is present in the data set is counted.

### 3. The Permutation Distribution

In this section, it is shown that our derivation of the graph theory concepts lead to the exact same permutation test as proposed by Gehan [10], Gilbert [11], Mantel [20] and Finkelstein and Schoenfeld [6]. It allows, however, to develop a permutation test for the win ratio, and it can be extended to a bootstrap test.

In a permutation test, subjects are randomly re-sampled to the treatment groups without replacement. If all  $\binom{m+n}{m}$  possible permutations of  $m$  treatment assignments and  $n$  control assignments are considered, the  $W_T$  and  $W_C$  in each of these permutation samples will lead to their permutation distribution. The expectations, variances, and covariance of this permutation distribution of  $W_T$  and  $W_C$  can be calculated explicitly. An edge will thus always join the same subjects, but whether or not this edge contributes to the treatment wins or control wins depends on the treatment assignment in the permutation sample.

The expectations of  $W_T$  and  $W_C$  are the same in the permutation distribution (Supplementary Material S1 Section S4.1):

$$\mathbb{E}_P(W_T) = \mathbb{E}_P(W_C) = \frac{mn}{N(N-1)} \sum_e w_e.$$

These provide the second half of the Equation (7). For our small example (Figure 1), the expectations of  $W_T$  and  $W_C$  equal  $\frac{6}{5 \times 4} (1 + 2 + 3 + 4 + 5 + 1) = 4.8$ .

The variance for the treatment wins,  $W_T$ , is then (Supplementary Material S1 Section S4.3)

$$\begin{aligned}
 \text{Var}_P(W_T) &= \frac{mn}{N(N-1)} \sum_v I_v^s + \frac{mn(m-1)}{N(N-1)(N-2)} \sum_v (I_v^2 - I_v^s) + \frac{mn(n-1)}{N(N-1)(N-2)} \sum_v (O_v^2 - O_v^s) \\
 &\quad + \frac{mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)} \left\{ \left( \sum_v I_v \right)^2 - \sum_v \left[ I_v^s + (I_v^2 - I_v^s) + (O_v^2 - O_v^s) + 2I_v O_v \right] \right\} \\
 &\quad - \left[ \frac{mn}{N(N-1)} \sum_v I_v \right]^2,
 \end{aligned}$$



with

$$I_v = \sum_i w_e \quad \text{the column sums of } w_e \text{ for vertex } v.$$

$$I_v^s = \sum_i w_e^2 \quad \text{the column sums of } w_e^2 \text{ for vertex } v.$$

$$O_v = \sum_j w_e \quad \text{the row sums of } w_e \text{ for vertex } v.$$

$$O_v^s = \sum_j w_e^2 \quad \text{the row sums of } w_e^2 \text{ for vertex } v.$$

Note that  $\sum_v I_v = \sum_v O_v = \sum_e w_e$  and that, due to the symmetry of the  $U$ -matrix, the column sums of  $w_e$  (or entries with  $u_{ij} > 0$ ) are equal to the row sums of  $u_{ij} < 0$ .

The variance for  $W_C$  is similar, but with the roles of  $I_v$  and  $O_v$  reversed. The covariance of  $W_T$  and  $W_C$  equals (Supplementary Material S1 Section S4.3):

$$\begin{aligned} \text{Cov}_P(W_T W_C) &= \frac{mn}{N(N-1)} \sum_v I_v O_v + \frac{mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)} \left\{ \left( \sum_v I_v \right)^2 - \sum_v \left[ I_v^s + (I_v^2 - I_v^s) \right. \right. \\ &\quad \left. \left. + (O_v^2 - O_v^s) + 2I_v O_v \right] \right\} - \left[ \frac{mn}{N(N-1)} \sum_v I_v \right]^2 \end{aligned}$$

Finally, the variance of the Finkelstein–Schoenfeld statistic, the net treatment benefit and win ratio statistic can be calculated from (3)–(6). For the Finkelstein–Schoenfeld statistic (and the net treatment benefit), the exact permutation variance simplifies to simple row sums of  $u_{ij}$  (Figure 3):

$$\begin{aligned} \text{Var}_P(W_T - W_C) &= \text{Var}_P(W_T) + \text{Var}_P(W_C) - 2\text{Cov}_P(W_T W_C) \\ &= \frac{mn}{N(N-1)} \left( \sum_v O_v^2 - 2 \sum_v I_v O_v + \sum_v I_v^2 \right) \\ &= \frac{mn}{N(N-1)} \sum_v (O_v - I_v)^2 \end{aligned}$$

It is easy to see that this is equal to (2), as proposed by Gehan [10], Gilbert [11], Mantel [20] and Finkelstein and Schoenfeld [6].

$v$		$(O_v - I_v)^2$
1	$\begin{bmatrix} 0 & -2 & 0 & 0 & 1 \end{bmatrix}$	1
2	$\begin{bmatrix} 2 & 0 & 3 & 0 & -5 \end{bmatrix}$	0
3	$\begin{bmatrix} 0 & -3 & 0 & 4 & 0 \end{bmatrix}$	1
4	$\begin{bmatrix} 0 & 0 & -4 & 0 & -1 \end{bmatrix}$	25
5	$\begin{bmatrix} -1 & 5 & 0 & 1 & 0 \end{bmatrix}$	25
	$\Sigma$	52

**Figure 3.** Graphical representation of counts needed of a score matrix  $U$ , with entries  $u_{ij}$ , to calculate the exact permutation variance for the win difference and net treatment benefit. With  $O_v$ , the row sums of  $u_{ij} > 0$ , and with  $I_v$ , the columns sums of  $u_{ij} > 0$  of a vertex  $v$ . Due to the symmetry of the matrix, the row sums of  $u_{ij} < 0$  are equal to  $-I_v$  and  $O_v - I_v$  equals the row sums of  $u_{ij}$ .

For our small example (Figure 1), it is possible to take every permutation sample, calculate the win difference for each sample and determine the variance of the win differences. If we do this, the variance coincides with what is calculated using the exact permutation formulas (see Supplementary Material S2, Section S1.1):

$$\text{Var}_P(W_T - W_C) = \frac{6}{5 \times 4} \times 52 = 15.6.$$

Other examples can be found in Supplementary Material S2, Section S1.2.

#### 4. The Bootstrap Distribution

Using a similar approach, we can now also develop a closed-form formula for the bootstrap distribution for both the net treatment benefit and the win ratio rather than the originally proposed re-sampling bootstrap test [8] for testing the hypothesis  $H_0 : \Delta = 0$  or  $H_0 : \Psi = 1$  or for the confidence interval construction.

In a bootstrap, subjects are randomly re-sampled with replacement within their treatment group. If all possible bootstrap samples,  $m^m n^n$  in total, are considered, the number of treatment wins,  $W_T$  and control wins,  $W_C$ , in each of these samples will lead to their bootstrap distribution. The expectations, variances, and covariance of this bootstrap distribution of  $W_T$  and  $W_C$  can also be calculated explicitly, following a similar reasoning as in the permutation. This method will be more accurate than actually re-sampling via bootstrap, since the randomization error is eliminated.

Since the bootstrap uses the same observed information as the permutation, the graph  $\mathbb{G}$  described in Section 2 remains useful. The edges and vertices remain the same as previously. However, since an edge  $e$  can now be repeated in a sample, be present once, or not be present at all, the number of times an edge contributes to a win in **T** or **C** can now be between 0 and  $n$ , respectively,  $m$ . Additionally, since subjects or vertices remain in their treatment arm over the bootstrap samples, edges joining vertices within the same treatment arm ( $D_i = D_j$ ) will never be a win in any bootstrap sample, and they will also not contribute to the variance. In other words, the variance of wins in a bootstrap sample does not depend on the within-arm comparisons. This is in contrast to the permutation variance, where within-arm comparisons contribute to the final variance.

There are thus  $(W_T + W_C)^2$  ordered pairs of edges corresponding to wins and contributing to the variance. Let the indicator  $D_v = 1$  when  $D_i = 1$  and  $D_j = 0$ , and  $D_v = 0$  when  $D_i = 0$  and  $D_j = 1$ . An edge corresponding to a treatment win ( $u_{ij} > 0$  for  $D_v = 1$  and  $u_{ij} < 0$  for  $D_v = 0$ ) or a control win ( $u_{ij} < 0$  for  $D_v = 1$  and  $u_{ij} > 0$  for  $D_v = 0$ ) in the observed data will be called a treatment edge or a control edge.

One could also consider bootstrap sampling from the entire population to test the same null hypothesis of the permutation test,  $H_0 : F_1 = F_2$ . The formulas of the one-sample bootstrap are similar in spirit to those given here, and they are detailed in the Supplementary Material S1, Section S6.

Using elementary sums of multinomial coefficients, the expectations of  $W_T$  and  $W_C$  are shown to be  $\mathbb{E}_B(W_T) = \sum_e w_e T_e = W_T$  and similarly,  $\mathbb{E}_B(W_C) = \sum_e w_e C_e = W_C$  (Supplementary Material S1, Section S5.1). Consequently, the expectation of the win difference  $\mathbb{E}_B(W_T - W_C) = W_T - W_C$  and the approximation of the expectation for the win ratio  $\mathbb{E}_B(W_T)/\mathbb{E}_B(W_C) = W_T/W_C$ .

The variance for the treatment wins,  $W_T$ , is then (Supplementary Material S1 Section S5.3)

$$\text{Var}_B(W_T) = \sum_{D_v=1} \#T_v^s + \frac{n-1}{n} \sum_{D_v=1} \#T_v^2 + \frac{m-1}{m} \sum_{D_v=0} \#T_v^2 - \frac{m+n-1}{nm} W_T^2 \quad (9)$$

with

$$\begin{aligned} \#T_v &= \sum w_e \text{ the row } (D_{v=1}) \text{ and column } (D_{v=0}) \text{ sums of } w_e > 0 \text{ for vertex } v. \\ \#T_v^s &= \sum w_e^2 \text{ the row } (D_{v=1}) \text{ and column } (D_{v=0}) \text{ sums of } w_e^2 \text{ for } w_e > 0 \text{ vertex } v. \end{aligned}$$

The variance of  $W_C$  is obtained similarly by replacing all treatment with control parameters and defining  $\#C_v$  as the row ( $D_{v=1}$ ) and column ( $D_{v=0}$ ) sums of  $w_e < 0$  for



vertex  $v$  and  $\#C_v^s$  as the row ( $D_{v=1}$ ) and column ( $D_{v=0}$ ) sums of  $w_e^2$  for  $w_e < 0$  vertex  $v$ . The covariance equals (Supplementary Material S1 Section S5.3):

$$\text{Cov}_B(W_T W_C) = \frac{n-1}{n} \sum_{D_v=1} \#T_v \#C_v + \frac{m-1}{m} \sum_{D_v=0} \#T_v \#C_v - \frac{m+n-1}{nm} W_T W_C \quad (10)$$

Finally, the variance of the Finkelstein–Schoenfeld statistic, the net treatment benefit and win ratio statistic can be calculated from (3)–(6). For the Finkelstein–Schoenfeld statistic and the net treatment benefit, the exact bootstrap variance simplifies to row and column sums of  $\#T$  and  $\#C$  counts (Figure 4).

$v$						For $D_v=1$				
	$\#T_v$	$\#C_v$	$\#T_v^s$	$\#C_v^s$	$(\#T_v - \#C_v)^2$	$\#T_v$	$\#C_v$	$\#T_v^s$	$\#C_v^s$	$(\#T_v - \#C_v)^2$
1	0	-2	0	0	1	1	0	1	0	1
2	2	0	3	0	-5	3	5	9	25	4
3	0	-3	0	4	0	$\Sigma$				
4	0	0	-4	0	-1					
5	-1	5	0	1	0	4	5	10	25	5
$(\#T_v - \#C_v)^2$					9	0	16	$\Sigma$		
					For $D_v=0$					

**Figure 4.** Graphical representation of counts needed of a  $U$  matrix to calculate the exact bootstrap variance for the win difference and net treatment benefit. Notice that the  $\sum_{D_v=1} \#T_v = W_T$  and  $\sum_{D_v=1} \#C_v = W_C$ .

For our small example (Figure 1), it is possible to take every bootstrap sample, calculate the win difference for each sample and determine the variance of the win differences. If we do this, the variance coincides with what is calculated using the exact bootstrap formulas (see Supplementary Material S2, Section S2.1):

$$\begin{aligned} \text{Var}_B(W_T - W_C) &= \text{Var}_B(W_T) + \text{Var}_B(W_C) - 2\text{Cov}_B(W_T W_C) \\ &= \frac{n-1}{n} \left[ \sum_{D_v=1} (\#T_v - \#C_v)^2 \right] + \frac{m-1}{m} \left[ \sum_{D_v=0} (\#T_v - \#C_v)^2 \right] \\ &\quad + \sum_{D_v=1} \#T_v^s + \sum_{D_v=1} \#C_v^s - \frac{m+n-1}{nm} (W_T - W_C)^2 \\ &= \frac{2}{3}(5) + \frac{1}{2}(25) + 10 + 25 - \frac{4}{6}(4-5)^2 = 50.17 \end{aligned}$$

Other examples can be found in Supplementary Material S2, Section S2.2.

## 5. Complexity

In any GPC analysis, a bounded number of computations for each pair of patients is performed to obtain the score matrix  $U$ , and the number of pairs is  $O(N^2)$ . The complexity of computing  $U$  is therefore  $O(N^2)$ . It is conceivable that one could reduce the complexity for special cases of GPC, however, our analysis works for arbitrary skew matrices, and in such cases, every pair of patients must be examined at least once.

For the variance of a GPC statistic, one could in principle consider all possible permutations or bootstrap samples, compute the numbers of treatment and control wins for each sample, and then compute the mean and variances. Such a computation would be exponential in  $N$  and hence completely unsatisfactory for a real example. We have performed such computations in some very small examples as a test of our algorithm (Supplementary Material S2).

For a specific vertex  $v$ , in order to compute the various vertex-dependent terms in the variance formulas,  $(I_v, I_v^s, O_v, O_v^s, \#T_v, \#C_v, \#T_v^s, \text{ and } \#C_v^s)$ , each other vertex must be examined once. Thus, the complexity of the computation at each vertex is  $O(N)$ , and computing at all vertices is thus  $O(N^2)$ . The total number of edges and the wins  $W_T$  and  $W_C$  are computed from these numbers in an additional  $O(N)$  steps, and the final computations are  $O(1)$ . Accordingly, the time complexity of both the permutation and bootstrap algorithms is  $O(N^2)$ .

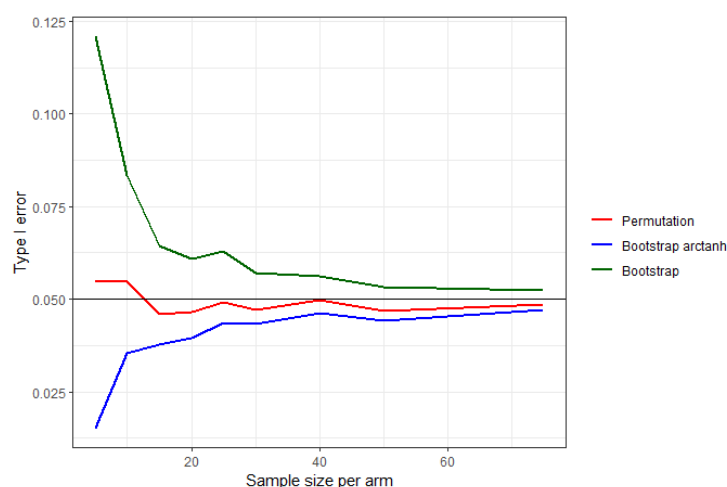
The exact computations will be faster than evaluating a large number of any permutation or bootstrap re-sampling, because evaluating only one sample is already  $O(N^2)$ .

## 6. Simulations

Similarly as for the Finkelstein-Schoenfeld test, inference for the permutation and bootstrap test is based on the standard normal assumption for the net treatment benefit and lognormal assumption for the win ratio of the asymptotic distribution of the permutation and bootstrap test statistic. While both the permutation and the bootstrap serve for hypothesis testing, albeit testing slightly different null hypotheses (see Introduction), only the bootstrap is additionally useful for confidence interval construction. As the permutation is estimating the variance under the null hypothesis, it is expected that the estimation bias of the variance under the alternative hypothesis will increase with increasing effect size. Since the bootstrap is estimating the variance under the alternative hypothesis, the bootstrap asymptotic distribution may be more likely to deviate from normality close to the boundary of potential values, which is 1 (or  $-1$ ) for the NTB. Therefore, we will additionally evaluate an inverse hyperbolic tangent transformation of the test statistic, which limits the confidence interval within the boundaries of the NTB and under which the normality assumption may hold for smaller sample sizes [29].

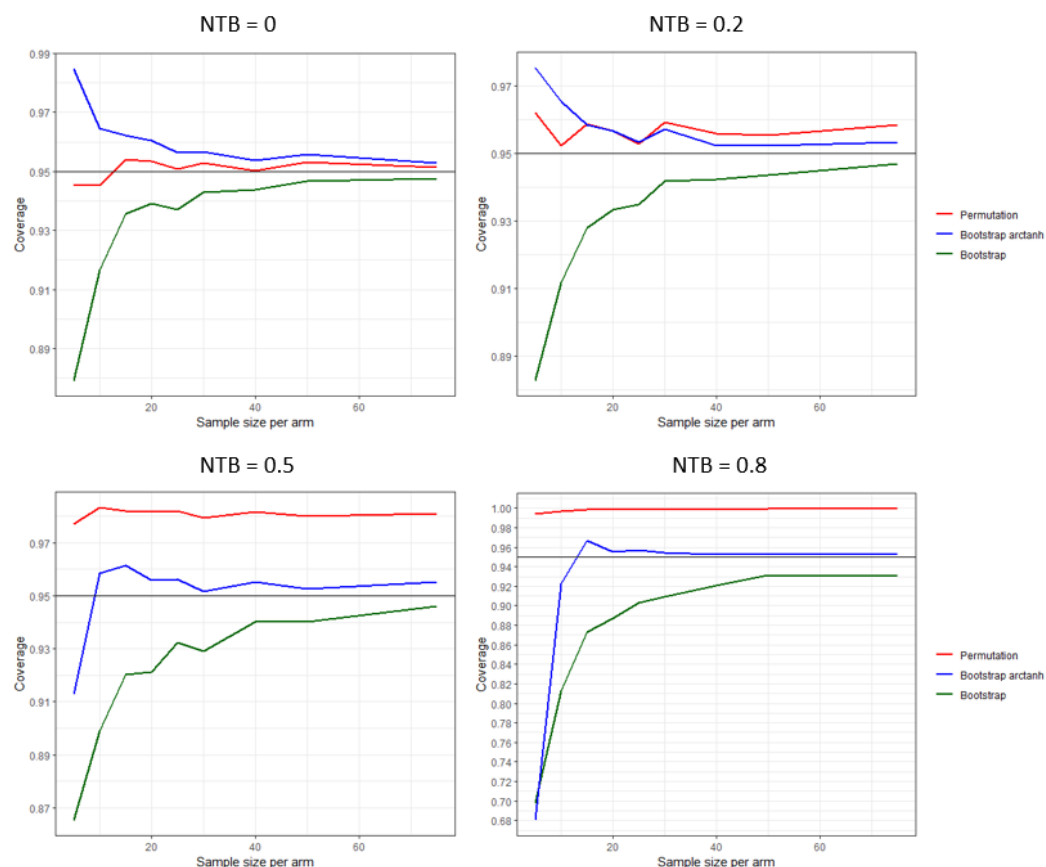
By means of a simulation study, the appropriateness of the normality assumption is evaluated by means of the nominal type I error and the confidence interval coverage for small samples. The simulated samples contain 5, 10, 15, 20, 25, 30, 40, 50 or 75 observations from a normal distribution  $N(0.3, 0.1)$  for the experimental arm and an equal number of observations from a  $N(\mu, 0.1)$  distribution with  $\mu = 0.3, 0.264, 0.205, 0.119$  for the control arm. These correspond to a net treatment benefit of 0, 0.2, 0.5 and 0.8. Each simulation is repeated 10,000 times.

The permutation test controls the nominal alpha level well, even for very small sample sizes of five observations per treatment arm (Figure 5). The bootstrap test and its inverse hyperbolic tangent transformation on the other hand require at least 30–40 observations per treatment arm.



**Figure 5.** Type I error of the permutation, bootstrap and the inverse hyperbolic tangent (arctanh) transformation of the bootstrap test for varying sample sizes.

As anticipated, the confidence interval coverage based on the permutation distribution is good under the null hypothesis, but it increasingly deteriorates with increasing effect size (Figure 6). The coverage for the transformed bootstrap is for all effect sizes better in range than the untransformed bootstrap, and depending on the effect size requires 20–30 observations per treatment arm.



**Figure 6.** Confidence interval coverage of the permutation, bootstrap and the inverse hyperbolic tangent (arctanh) transformation of the bootstrap test for varying treatment effects and sample sizes. NTB = net treatment benefit.

Further evaluation of the permutation and bootstrap Type I error control and confidence interval coverage, in comparison with other GPC methods for inference and with other type of data, is available in Verbeek et al. [30].

## 7. Illustration

The rare, genetic skin disease epidermolysis bullosa simplex (EBS) is characterized by the formation of blisters under low mechanical stress [31]. An innovative immunomodulatory 1% diacerein cream was postulated to reduce the number of blisters compared to placebo and evaluated in a randomized, placebo-controlled, double-blind, two-period cross-over phase II/III trial in 16 pediatric patients [32]. After daily treatment during 4 weeks and monitoring for an additional 3 months, patients crossed over to the opposite treatment after a washout period. In each treatment period, both the number of blisters in a treated body surface area were counted and the quality of life (QoL) was assessed. The primary endpoint, the proportion of patients with more than 40% reduction in blisters at week 4 compared to baseline, however, leads to inconclusive results [32]. The primary analysis with the Barnard [33] test requires separate analyses per treatment period, which showed that during the first treatment period, there was a treatment effect in favor of the diacerein

cream ( $p = 0.007$ ) in contrast to the second treatment period ( $p = 0.32$ ). Wally et al. [32] discuss reasons why the second period effect might be smaller.

Since the Barnard test is ignoring the cross-over design and the QoL, it only uses a fraction of the available information in the EBS trial. It is well known that the QoL of EBS patients is poor due to the hindrance of daily activities by the blisters [32]. While the reduction in the number of blisters is important, the healing but not yet disappearance of a blister may affect positively the QoL outcome, which is relevant for both the patient and clinician. The ability to add the QoL outcome with the blister outcome is clinically very relevant and very straightforward with GPC. We will re-analyze the EBS trial with a GPC both when prioritizing and not prioritizing the outcomes. Although GPC variants exist for matched designs [34], such as cross-over trials, under certain circumstances, which are applicable in the general GPC test, the matching can be ignored [35]. This means that we can consider the contribution of each subject to both treatment periods as a contribution of two independent subjects. Hence, the re-analysis will include all available information in a single analysis, which evades difficulties in interpreting conflicting results from separate analyses per treatment period.

The GPC permutation hypothesis test of the binary 40% reduction in blisters outcome and the continuous change in QoL outcome separately does not show evidence of a treatment effect of diacerein on the reduction in blisters ( $p = 0.0701$ ), while there is evidence for improvement in QoL ( $p = 0.0019$ ) (Table 1). In a prioritized GPC permutation test, where the blisters are evaluated first in the pairwise comparisons, there is evidence for a positive treatment effect of diacerein ( $p = 0.0051$ ) (Table 1). In addition, when evaluating all pairs for both outcomes in a non-prioritized GPC, the permutation test shows evidence for a positive treatment effect ( $p = 0.0022$ ) (Table 1). In addition, the confidence intervals around the net treatment benefit, obtained using the inverse hyperbolic tangent bootstrap, show that there is evidence for a treatment effect by diacerein, 59% (95% CI: 19–82%) with the prioritized GPC and 48% (95% CI: 21–68%) for the non-prioritized GPC (Table 1).

**Table 1.** EBS trial data analysis of the composite blister and QoL outcomes with the prioritized and non-prioritized GPC.  $W_T$  = number of wins for the diacerein arm,  $W_C$  = number of wins for the placebo arm,  $W_0$  = number of ties, NTB = net treatment benefit, CI = confidence interval.

	$W_T$	$W_C$	$W_0$	NTB (95%CI)	$p$ -Value Two-Sided
<b>Prioritized GPC</b>					
Blister	99 (44%)	24 (11%)		0.33	
QoL	72 (32%)	14 (6%)		0.26	
Total	171 (76%)	38 (17%)	16 (7%)	0.59 (0.19;0.82)	0.0051
<b>Non-Prioritized GPC</b>					
Blister	99 (44%)	24 (11%)	102 (45%)	0.33	0.0701
QoL	162 (72%)	22 (10%)	41 (18%)	0.62	0.0019
Total				0.48 (0.21;0.68)	0.0022

## 8. Discussion

Efficient closed-form formulas to compute the expectation and the variance of the permutation and bootstrap distribution of GPC statistics are developed using graph theory notation. These methodologies are shown to give the exact means and variances and are not subject to sampling errors, which do result from any randomized permutation or bootstrap test. Additionally, it is shown that the time complexity is  $O(N^2)$ , which is faster than any randomization permutation or bootstrap test. Since in most of the applications of the GPC statistics, only the means and variances are used [36–45], the proposed exact methods eliminate the need for randomization tests. In order to construct a hypothesis test for the GPC statistics, any use of the means and variances would require the assumption of asymptotic normality. This normality assumption is either explicit or implicit in the various

references. The normality assumption does seem satisfactory in simulations [30], even for sample sizes as small as 10 observations for the null permutation distribution, which is in concordance with the results from Gehan [10]. Slightly larger sample sizes are required for the null bootstrap distribution, which has been shown to be normal for U-statistics [46], such as the net treatment benefit. Further research is required to improve the small sample behavior of the exact bootstrap test, which may include an Edgeworth expansion [29], alternative transformations [47], small sample corrections [48], wild bootstrap [49] or fractional-random-weighted bootstrap methods [50].

Even though the mean and variance for the null permutation distribution for absolute GPC statistics, such as the Finkelstein–Schoenfeld statistic and net treatment benefit, has been established in the literature [6,10,11,20], using the graph theory notation allows easy extension to the relative GPC statistic win ratio, allows extension to the bootstrap distribution with sampling within the treatment group and sampling over the entire population and allows the generalization to score entries in  $\mathbb{R}$ , which result from non-prioritized GPC and censoring correcting algorithms. Because of the special structure of GPC, we are able to determine the expected value and variance from the bootstrap distribution without actually using the randomization. This is not a unique feature, since Efron has shown that in some other statistics, the exact values for the bootstrap distribution can be computed [51] (Section 10.3). Our proposed approach means an improvement in both accuracy and speed for the win ratio bootstrap test [40–45].

In the case of a single outcome variable without censoring, the Mann–Whitney test [9], the Finkelstein–Schoenfeld formula (2) and the exact formula will all agree, because all are based on counting arguments using the exact permutation distribution of the trial arms. Similarly, in the case of a single outcome variable with censored observations, the Gehan–Wilcoxon test [10], Finkelstein–Schoenfeld formula (2) and the exact formula will agree. The exact computation of the bootstrap mean and variance is as easy as that for the permutation mean and variance used for the Mann–Whitney, Wilcoxon, or Gehan–Wilcoxon analyses. Accordingly, the bootstrap evaluation could be considered a practical alternative to the original permutation test-based analyses. It is important to note that for the GPC permutation and bootstrap test, the slightly different null hypotheses (see introduction) are subject to different assumptions. For example, when the observations in each treatment arm come from distributions with equal location but highly different variability, then we are under the null hypothesis for the bootstrap test but not for the permutation test [52]. Although, both tests are only consistent against location shift alternatives [21,22].

In addition to randomization hypothesis tests for the GPC statistics, asymptotic tests have been proposed as well [14,26,27]. In a direct comparison of the asymptotic tests and the exact methods for the net benefit and win ratio statistics under location shift models, with respect to type I error control, small sample bias and 95% confidence interval coverage, the exact methods are more accurate and at least as fast [30]. Especially in sample sizes below 100–200 subjects, the exact permutation test clearly outperforms the asymptotic based tests.

Recently, a hypothesis test for the net treatment benefit has been suggested based on U-statistic decomposition [52]. Interestingly, the variance estimator obtained from the second-order Hoeffding decomposition of the GPC statistics exactly equals the variance of the bootstrap distribution (Appendix B).

Importantly, the application of the variance of the permutation and bootstrap distribution assumes that the score in a pair of subjects remains constant over all possible permutation and bootstrap samples. For example, in the GPC scoring algorithms that use Kaplan–Meier estimators within treatment arms to correct for censored observations [15,17], the score in a pair differs between permutation and bootstrap samples. It is therefore not recommended to use the exact permutation and bootstrap inference for these scoring algorithms. Similarly, in the inverse probability censoring weighting algorithms [18,53] and the algorithm using the joined Kaplan–Meier estimators over both treatment arms [16], the

score in a pair remains constant in all permutation samples but not in all bootstrap samples. Only the exact permutation inference is thus recommended for these scoring algorithms.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/math11061502/s1>.

**Author Contributions:** Conceptualization, W.N.A. and J.V.; methodology, W.N.A.; software, W.N.A. and J.V.; validation, W.N.A. and J.V.; formal analysis and simulations main paper, J.V.; simulations supplementary material, W.N.A.; proofs in appendix, W.N.A. and J.V.; writing—original draft preparation, W.N.A. and J.V.; writing—review and editing, W.N.A. and J.V.; visualization, J.V.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Data and code supporting the reported results can be found in the Supplementary Materials.

**Acknowledgments:** We would like to acknowledge Johann Bauer for granting permission to use the EB trial data and Brice Ozenne for fruitful discussion on the scoring algorithms and software development.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

Cov	covariance
FS	Finkelstein–Schoenfeld
GPC	generalized pairwise comparisons
NTB	net treatment benefit
Var	variance
WO	win odds

## Appendix A. Expectation and Variance of Finkelstein–Schoenfeld Statistic

We find the expectation and variance of the *FS* statistic by treating  $D_1, \dots, D_N$  as random variables and the scores  $U_1, \dots, U_N$  as fixed quantities. Then

$$\mathbb{E}(D_i) = m/N. \quad (\text{A1})$$

For the variance, we have

$$\begin{aligned} \text{Var}(D_i) &= \mathbb{E}(D_i^2) - \mathbb{E}(D_i)^2 \\ &= m/N - (m/N)^2 \\ &= \frac{mn}{N^2} \end{aligned} \quad (\text{A2})$$

and for  $i \neq j$

$$\begin{aligned} \text{Cov}(D_i, D_j) &= \mathbb{E}(D_i D_j) - \mathbb{E}(D_i)\mathbb{E}(D_j) \\ &= \frac{m(m-1)}{N(N-1)} - \frac{m^2}{N^2} \\ &= -\frac{mn}{N^2(N-1)} \end{aligned} \quad (\text{A3})$$



For the test statistic  $FS$ , use (A1) to obtain

$$\begin{aligned}\mathbb{E}(FS) &= \mathbb{E}\left(\sum_{i=1}^N D_i U_i\right) \\ &= \sum_{i=1}^N \mathbb{E}(D_i) U_i \\ &= \frac{m}{N} \sum_{i=1}^N U_i \\ &= 0,\end{aligned}$$

and using (A2) and (A3)

$$\begin{aligned}\text{Var}(FS) &= \text{Var}\left(\sum_{i=1}^N D_i U_i\right) \\ &= \sum_{i=1}^N \sum_{j=1}^N U_i U_j \text{Cov}(D_i, D_j) \\ &= \sum_{i=1}^N \text{Var}(D_i) U_i^2 + \sum_{i=1}^N \sum_{j \neq i}^N \text{Cov}(D_i, D_j) U_i U_j \\ &= \frac{mn}{N^2} \sum_{i=1}^N U_i^2 + \frac{mn}{N^2(N-1)} \sum_{i=1}^N U_i^2 \\ &= \frac{mn}{N(N-1)} \sum_{i=1}^N U_i^2.\end{aligned}$$

## Appendix B. Equality of the Variance of the Bootstrap Distribution and the U-Statistic Second-Order Hoeffding Decomposition Estimator

In the following, we will show that the variance estimator from the second-order Hoeffding decomposition in Ozenne et al. [52] equals the variance of the bootstrap distribution. We will restrict ourselves to a  $U$ -matrix with entries in  $\{-1, 0, 1\}$ , following Ozenne et al. [52].

Using the second-order Hoeffding decomposition, the variance of  $W_T$ , which is the variance estimator  $(nm^2)\sigma_{U^+, U^+}$  in the notation of Ozenne et al. [52], equals:

$$\text{Var}(W_T) = nm \left[ \frac{n-1}{m} \sum_{i=1}^m (\hat{h}_1^+(i))^2 + \frac{m-1}{n} \sum_{j=1}^n (\hat{h}_1^+(j))^2 + U^+(1 - U^+) \right]$$

Following Ozenne et al. [52],  $\hat{h}_1^+(i) = \frac{\#\mathbb{T}_{D_U=1}}{n} - \frac{W_T}{nm}$ ,  $\hat{h}_1^+(j) = \frac{\#\mathbb{T}_{D_U=0}}{m} - \frac{W_T}{nm}$  and  $U^+ = \frac{W_T}{nm}$ . Thus:

$$\begin{aligned}
\text{Var}(W_T) &= nm \left[ \frac{n-1}{m} \sum_{i=1}^m \left( \frac{\#\mathbb{T}_{D_v=1}}{n} - \frac{W_T}{nm} \right)^2 \right. \\
&\quad \left. + \frac{m-1}{n} \sum_{j=1}^n \left( \frac{\#\mathbb{T}_{D_v=0}}{m} - \frac{W_T}{nm} \right)^2 + \frac{W_T}{nm} - \left( \frac{W_T}{nm} \right)^2 \right] \\
&= nm \left[ \frac{n-1}{m} \left( \frac{\sum_{D_v=1} \#\mathbb{T}_v^2}{n^2} - \frac{2}{n^2 m} W_T \sum_{D_v=1} \#\mathbb{T}_v + m \left( \frac{W_T}{nm} \right)^2 \right) \right. \\
&\quad \left. + \frac{m-1}{n} \left( \frac{\sum_{D_v=0} \#\mathbb{T}_v^2}{m^2} - \frac{2}{nm^2} W_T \sum_{D_v=0} \#\mathbb{T}_v + n \left( \frac{W_T}{nm} \right)^2 \right) + \frac{W_T}{nm} - \left( \frac{W_T}{nm} \right)^2 \right] \\
&= \frac{n-1}{n} \sum_{D_v=1} \#\mathbb{T}_v^2 - \frac{2(n-1)}{nm} W_T \sum_{D_v=1} \#\mathbb{T}_v + \frac{(n-1)}{nm} W_T^2 \\
&\quad + \frac{m-1}{m} \sum_{D_v=0} \#\mathbb{T}_v^2 - \frac{2(m-1)}{nm} W_T \sum_{D_v=0} \#\mathbb{T}_v + \frac{(m-1)}{nm} W_T^2 + W_T - \frac{1}{nm} W_T^2
\end{aligned}$$

Since due to symmetry  $\sum_{D_v=1} \#\mathbb{T}_v = \sum_{D_v=0} \#\mathbb{T}_v = W_T$ :

$$\begin{aligned}
\text{Var}(W_T) &= W_T + \frac{n-1}{n} \sum_{D_v=1} \#\mathbb{T}_v^2 + \frac{m-1}{m} \sum_{D_v=0} \#\mathbb{T}_v^2 - \frac{2(n-1)}{nm} W_T^2 + \frac{(n-1)}{nm} W_T^2 \\
&\quad - \frac{2(m-1)}{nm} W_T^2 + \frac{(m-1)}{nm} W_T^2 - \frac{1}{nm} W_T^2 \\
&= W_T + \frac{n-1}{n} \sum_{D_v=1} \#\mathbb{T}_v^2 + \frac{m-1}{m} \sum_{D_v=0} \#\mathbb{T}_v^2 - \frac{m+n-1}{nm} W_T^2
\end{aligned}$$

Since  $\sum_{D_v=1} \#\mathbb{T}_v^s = \sum_{D_v=1} \#\mathbb{T}_v = W_T$  when entries in the  $U$ -matrix are restricted to  $\{-1, 0, 1\}$ , it follows that  $\text{Var}(W_T) = \text{Var}_B(W_T)$  (see Equation (9)).

Similarly, it can be shown that the variance of  $W_C$ , which is the variance estimator  $(nm^2)\sigma_{U^-, U^-}$  in the notation of Ozenne et al. [52] equals  $\text{Var}_B(W_C)$ .

Finally, the  $\text{Cov}(W_T W_C)$  or  $(nm^2)\sigma_{U^+, U^-}$  in the notation of Ozenne et al. [52] is:

$$\text{Cov}(W_T W_C) = nm \left[ (n-1) \sum_{i=1}^m \hat{h}_1^+(i) \hat{h}_1^-(i) + \frac{m-1}{n} \sum_{j=1}^n \hat{h}_1^+(j) \hat{h}_1^-(j) - U^+ U^- \right]$$

Following Ozenne et al. [52],  $\hat{h}_1^+(i) = \frac{\#\mathbb{T}_{D_v=1}}{n} - \frac{W_T}{nm}$ ,  $\hat{h}_1^-(j) = \frac{\#\mathbb{T}_{D_v=0}}{m} - \frac{W_T}{nm}$ ,  $U^+ = \frac{W_T}{nm}$ ,  $\hat{h}_1^-(i) = \frac{\#\mathbb{C}_{D_v=1}}{n} - \frac{W_C}{nm}$ ,  $\hat{h}_1^-(j) = \frac{\#\mathbb{C}_{D_v=0}}{m} - \frac{W_C}{nm}$ ,  $U^- = \frac{W_C}{nm}$ . Thus:

$$\begin{aligned}
\text{Cov}(W_T W_C) &= nm \left[ \frac{n-1}{m} \sum_{i=1}^m \left( \frac{\#\mathbb{T}_{D_v=1}}{n} - \frac{W_T}{nm} \right) \left( \frac{\#\mathbb{C}_{D_v=1}}{n} - \frac{W_C}{nm} \right) \right. \\
&\quad \left. + \frac{m-1}{n} \sum_{j=1}^n \left( \frac{\#\mathbb{T}_{D_v=0}}{m} - \frac{W_T}{nm} \right) \left( \frac{\#\mathbb{C}_{D_v=0}}{m} - \frac{W_C}{nm} \right) - \frac{W_T W_C}{(nm)^2} \right] \\
&= nm \left[ \frac{n-1}{m} \left( \frac{\sum_{D_v=1} \#\mathbb{T}_v \#\mathbb{C}_v}{n^2} - \frac{\sum_{D_v=1} \#\mathbb{T}_v W_C}{n^2 m} - \frac{\sum_{D_v=1} \#\mathbb{C}_v W_T}{n^2 m} + m \frac{W_T W_C}{(nm)^2} \right) \right. \\
&\quad \left. + \frac{m-1}{n} \left( \frac{\sum_{D_v=0} \#\mathbb{T}_v \#\mathbb{C}_v}{m^2} - \frac{\sum_{D_v=0} \#\mathbb{T}_v W_C}{nm^2} - \frac{\sum_{D_v=0} \#\mathbb{C}_v W_T}{nm^2} + n \frac{W_T W_C}{(nm)^2} \right) - \frac{W_T W_C}{(nm)^2} \right]
\end{aligned}$$

Since due to symmetry  $\sum_{D_v=1} \#\mathbb{T}_v = \sum_{D_v=0} \#\mathbb{T}_v = W_T$  and  $\sum_{D_v=1} \#\mathbb{C}_v = \sum_{D_v=0} \#\mathbb{C}_v = W_C$ :

$$\begin{aligned}
\text{Cov}(W_T W_C) &= nm \left[ \frac{n-1}{m} \left( \frac{\sum_{D_v=1} \#T_v \#C_v}{n^2} - 2 \frac{W_T W_C}{n^2 m} + m \frac{W_T W_C}{(nm)^2} \right) \right. \\
&\quad \left. + \frac{m-1}{n} \left( \frac{\sum_{D_v=0} \#T_v \#C_v}{m^2} - 2 \frac{W_T W_C}{nm^2} + n \frac{W_T W_C}{(nm)^2} \right) - \frac{W_T W_C}{(nm)^2} \right] \\
&= \frac{n-1}{n} \sum_{D_v=1} \#T_v \#C_v + \frac{m-1}{m} \sum_{D_v=0} \#T_v \#C_v - \frac{(n-1) + (m-1) + 1}{nm} W_T W_C \\
&= \frac{n-1}{n} \sum_{D_v=1} \#T_v \#C_v + \frac{m-1}{m} \sum_{D_v=0} \#T_v \#C_v - \frac{(m+n-1)}{nm} W_T W_C
\end{aligned}$$

From which it follows that  $\text{Cov}(W_T W_C) = \text{Cov}_B(W_T W_C)$  (see Equation (10)).

## References

- Freemantle, N.; Calvert, M.; Wood, J.; Eastaugh, J.; Griffin, C. Composite outcomes in randomized trials: Greater precision but with greater uncertainty? *JAMA* **2003**, *289*, 2554–2559. [\[CrossRef\]](#)
- Ferreira-González, I.; Busse, J.W.; Heels-Ansdell, D.; Montori, V.M.; Akl, E.A.; Bryant, D.; Alonso, J.; Jaeschke, R.; Schünemann, H.J.; Permyer-Miralda, G.; et al. Problems with Use of Composite End Points in Cardiovascular Trials: Systematic Review of Randomised Controlled Trials. *BMJ Br. Med. J.* **2007**, *334*, 786–788. [\[CrossRef\]](#)
- Lim, E.; Brown, A.; Helmy, A.; Mussa, S.; Altman, D. Composite Outcomes in Cardiovascular Research: A Survey of Randomized Trials. *Ann. Intern. Med.* **2008**, *149*, 612–617. [\[CrossRef\]](#) [\[PubMed\]](#)
- Armstrong, P.; Westerhout, C. Composite End Points in Clinical Research. A Time for Reappraisal. *Circulation* **2017**, *135*, 2299–2307. [\[CrossRef\]](#)
- Cox, D. Regression models and life-tables. *J. R. Stat. Soc. Ser. B* **1972**, *34*, 187–220. [\[CrossRef\]](#)
- Finkelstein, D.; Schoenfeld, D. Combining Mortality and Longitudinal Measures in Clinical Trials. *Stat. Med.* **1999**, *18*, 1341–1354. [\[CrossRef\]](#)
- Buyse, M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat. Med.* **2010**, *29*, 3245–3257. [\[CrossRef\]](#)
- Pocock, S.; Ariti, C.; Collier, T.; 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. [\[CrossRef\]](#)
- Mann, H.; Whitney, D. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math Stat.* **1947**, *18*, 50–60. [\[CrossRef\]](#)
- Gehan, E. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. *Biometrika* **1965**, *52*, 203–223. [\[CrossRef\]](#) [\[PubMed\]](#)
- Gilbert, J.P. Random Censorship. Ph.D. Thesis, University of Chicago, Chicago, IL, USA, 1962.
- Dong, G.; Huang, B.; Wang, D.; Verbeeck, J.; Wang, J.; Hoaglin, D. Adjusting win statistics for dependent censoring. *Pharm. Stat.* **2021**, *20*, 440–450. [\[CrossRef\]](#) [\[PubMed\]](#)
- Verbeeck, J.; Spitzer, E.; de Vries, T.; van Es, G.A.; Anderson, W.; Van Mieghem, N.; Leon, M.; Molenberghs, G.; Tijssen, J. Generalized Pairwise Comparison Methods to Analyze (non)-Hierarchical Composite Endpoints. *Stat. Med.* **2019**, *38*, 5641–5646. [\[CrossRef\]](#)
- Ramchandani, R.; Schoenfeld, D.; Finkelstein, D. Global rank tests for multiple, possibly censored, outcomes. *Biometrics* **2016**, *72*, 926–935. [\[CrossRef\]](#) [\[PubMed\]](#)
- Efron, B. The two sample problem with censored data. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June–18 July 1965 and 27 December 1965–7 January 1966; Volume 4, pp. 831–853.
- Latta, R. Generalized Wilcoxon statistics for the two sample problem with censored data. *Biometrika* **1977**, *63*, 633–635. [\[CrossRef\]](#)
- Péron, J.; Buyse, M.; Ozenne, B.; Roche, L.; Roy, P. An extension of generalized pairwise comparisons of prioritized outcomes in the presence of censoring. *Stat. Methods Med Res.* **2016**, *27*, 1230–1239. [\[CrossRef\]](#)
- Deltuvaite-Thomas, V.; Verbeeck, J.; Burzykowski, T.; Buyse, M.; Tournigand, C.; Molenberghs, G.; Thas, O. Generalized pairwise comparisons for censored data: An overview. *Biom. J.* **2022**, *65*, 2100354. [\[CrossRef\]](#)
- Dong, G.; Hoaglin, D.; Qiu, J. The win ratio: On interpretation and handling of ties. *Biopharm. Stat.* **2020**, *12*, 99–106. [\[CrossRef\]](#)
- Mantel, N. Ranking Procedures for Arbitrarily Restricted Observation. *Biometrics* **1967**, *23*, 65–78. [\[CrossRef\]](#)
- Thas, O. *Comparing Distributions*; Springer Science+Business Media: New York, NY, USA, 2010.
- Brunner, E.; Bathke, A.C.; Konietzschke, F. *Rank and PseudoRank Procedures for Independent Observations in Factorial Designs*; Springer Nature: Cham, Switzerland, 2019.
- Lee, A. *U-Statistics: Theory and Practice*; Chapman & Hall/CRC: New York, NY, USA, 1990.

24. Verbeeck, J.; Deltuvaite-Thomas, V.; Berckmoes, B.; Burzykowski, T.; Aerts, M.; Thas, O.; Buyse, M.; Molenberghs, G. Unbiasedness and efficiency of non-parametric and UMVUE estimators of the probabilistic index and related statistics. *Stat. Methods Med. Res.* **2021**, *30*, 747–768. [[CrossRef](#)]
25. Brunner, E.; Vandemeulebroecke, M.; Mütze, T. Win odds: An adaptation of the win ratio to include ties. *Stat. Med.* **2021**, *40*, 3367–3384. [[CrossRef](#)]
26. 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*, 430–437. [[CrossRef](#)]
27. Bebu, I.; Lachin, J. Large sample inference for a win ratio analysis of a composite outcome based on prioritized components. *Biostatistics* **2016**, *17*, 178–187. [[CrossRef](#)]
28. Harris, J.M.; Hirst, J.L.; Mossinghoff, M.J. *Combinatorics and Graph Theory*, 3rd ed.; Springer: New York, NY, USA, 2014.
29. Hall, P. *The Bootstrap and Edgeworth Expansion*; Springer: New York, NY, USA, 1992.
30. Verbeeck, J.; Ozenne, B.; Anderson, W. Evaluation of inferential methods for the net benefit and win ratio statistics. *J. Biopharm. Stat.* **2020**, *30*, 765–782. [[CrossRef](#)] [[PubMed](#)]
31. Coulombe, P.; Lee, C. Defining keratin protein function in skin epithelia: Epidermolysis bullosa simplex and its aftermath. *J. Invest. Dermatol.* **2012**, *132*, 763–775. [[CrossRef](#)] [[PubMed](#)]
32. Wally, V.; Hovnanian, A.; Ly, J.; Buckova, H.; Brunner, V.; Lettner, T.; Ablinger, M.; Felder, T.; Hofbauer, P.; Wolkersdorfer, M.; et al. Diacerein orphan drug development for epidermolysis bullosa simplex: A phase 2/3 randomized, placebo-controlled, double-blind clinical trial. *J. Am. Acad. Dermatol.* **2018**, *78*, 892–901. [[CrossRef](#)]
33. Barnard, G. Significance Tests for  $2 \times 2$  Tables. *Biometrika* **1947**, *34*, 123–138. [[PubMed](#)]
34. Matsouaka, R. Robust statistical inference for matched win statistics. *Stat. Methods Med. Res.* **2022**, *31*, 1423–1438. [[CrossRef](#)]
35. Konietzschke, F.; Pauly, M. A studentized permutation test for the nonparametric Behrens-Fisher problem in paired data. *Electron. J. Stat.* **2012**, *6*, 1358–1372. [[CrossRef](#)]
36. Maurer, M.; Schwartz, J.; Gundapaneni, B.; Elliott, P.; Merlini, G.; Waddington-Cruz, M.; Kristen, A.; Grogan, M.; Witteles, R.; Damy, T.; et al. Tafamidis treatment for patients with transthyretin amyloid cardiomyopathy. *N. Engl. J. Med.* **2018**, *379*, 1007–1016. [[CrossRef](#)]
37. Berry, N.; Mauri, L.; Feldman, T.; Komtebedde, J.; van Veldhuisen, D.; Solomon, S.; Massaro, J.; Shah, S. Transcatheter InterAtrial Shunt Device for the treatment of heart failure: Rationale and design of the pivotal randomized trial to REDUCE Elevated Left Atrial Pressure in Patients with Heart Failure II (REDUCE LAP-HF II). *Am. Heart J.* **2020**, *226*, 222–231. [[CrossRef](#)]
38. Lansky, A.; Makkar, R.; Nazif, T.; Messé, S.; Forrest, J.; Sharma, R.; Schofer, J.; Linke, A.; Brown, D.; Dhoble, A.; et al. A randomized evaluation of the TriGuard™ HDH cerebral embolic protection device to Reduce the Impact of Cerebral Embolic LESions after TransCatheter Aortic Valve ImplanTation: The REFLECT I trial. *Eur. Heart J.* **2021**, *42*, 2670–2679. [[CrossRef](#)]
39. Tamim, M.N.; Moses, J.; Sharma, R.; Dhoble, A.; Rovin, J.; Brown, D.; Horwitz, P.; Makkar, R.; Stoler, R.; Forrest, J.; et al. Randomized Evaluation of TriGuard 3 Cerebral Embolic Protection After Transcatheter Aortic Valve Replacement: REFLECT II. *JACC Cardiovasc. Interv.* **2021**, *14*, 515–527.
40. Milojevic, M.; Head, S.; Andrinopoulou, E.; Serruys, P.; Mohr, F.; Tijssen, J.; Kappetein, A. Hierarchical testing of composite endpoints: Applying the win ratio to percutaneous coronary intervention versus coronary artery bypass grafting in the SYNTAX trial. *EuroIntervention* **2017**, *13*, 106–114. [[CrossRef](#)] [[PubMed](#)]
41. Fergusson, N.; Ramsay, T.; Chassé, M.; English, S.; Knoll, G. The win ratio approach did not alter study conclusions and may mitigate concerns regarding unequal composite end points in kidney transplant trials. *J. Clin. Epidemiol.* **2018**, *98*, 9–15. [[CrossRef](#)]
42. Kotalik, A.; Eaton, A.; Lian, Q.; Serrano, C.; Connett, J.; Neaton, J. A win ratio approach to the re-analysis of Multiple Risk Factor Intervention Trial. *Clin. Trials* **2019**, *16*, 626–634. [[CrossRef](#)]
43. Hironori, H.; van Klaveren, D.; Takahashi, K.; Kogame, N.; Chichareon, P.; Modolo, R.; Tomaniak, M.; Ono, M.; Kawashima, H.; Wang, R.; et al. Comparative Methodological Assessment of the Randomized GLOBAL LEADERS Trial Using Total Ischemic and Bleeding Events. *Circ. Cardiovasc. Qual. Outcomes* **2020**, *13*, e006660.
44. Ferreira, J.; Jhund, P.; Duarte, K.; Claggett, B.; Solomon, S.; Pocock, S.; Petrie, M.; Zannad, F.; McMurray, J. Use of the Win Ratio in Cardiovascular Trials. *JACC Heart Fail* **2020**, *8*, 441–450. [[CrossRef](#)] [[PubMed](#)]
45. Kandzari, D.; Hickey, G.; Pocock, S.; Weber, M.; Böhm, M.; Cohen, S.; Fahy, M.; Lamberti, G.; Mahfoud, F. Prioritised endpoints for device-based hypertension trials: The win ratio methodology. *EuroIntervention* **2021**, *16*, e1496–e1502. [[CrossRef](#)]
46. Arcones, M.; Gine, E. On the bootstrap of U and V statistics. *Ann. Statist.* **1992**, *20*, 655–674. [[CrossRef](#)]
47. van der Vaart, A. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 1998.
48. Perme, M.; Manevski, D. Confidence intervals for the Mann–Whitney test. *Stat. Methods Med Res.* **2018**, *28*, 3755–3768. [[CrossRef](#)]
49. Wu, C. Jackknife, bootstrap and other resampling methods in regression analysis (with discussions). *Ann. Stat.* **1986**, *14*, 1261–1350.
50. Rubin, D. The Bayesian bootstrap. *Ann. Stat.* **1981**, *9*, 130–134. [[CrossRef](#)]
51. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*; SIAM: Philadelphia, PA, USA, 1982.

52. Ozenne, B.; Budtz-Jørgensen, E.; Péron, J. The asymptotic distribution of the Net Benefit estimator in presence of right-censoring. *Stat. Methods Med. Res.* **2021**, *30*, 2399–2412. [[CrossRef](#)] [[PubMed](#)]
53. Dong, G.; Mao, L.; Huang, B.; Gamalo-Siebers, M.; Wang, J.; Yu, G.; Hoaglin, D. The inverse-probability-of-censoring weighting (IPCW) adjusted win ratio statistic: An unbiased estimator in the presence of independent censoring. *J. Biopharm. Stat.* **2020**, *30*, 882–899. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.