Perioperative Medicine

REVIEW

Open Access

Efficient statistical analysis of trial designs: win ratio and related approaches for composite outcomes

Wilson Fandino[1], Matthew Dodd[2], Gudrun Kunst[3*] and Tim Clayton[2*]

Abstract

In randomized controlled clinical trials, composite outcomes are often used to study treatment effects. This approach is popular because it increases the number of observed events, enhancing statistical power while reducing the required patient sample size. However, composite outcomes do not provide insight into the effect of individual endpoints. This becomes particularly relevant when mortality is combined with less critical but clinically relevant endpoints or when the clinical importance of individual endpoints varies significantly. As a result, interpreting composite outcomes can be challenging.

This narrative review introduces the win ratio (WR), a method for prioritizing individual endpoints within a composite outcome. The WR offers an alternative to composite outcomes by considering the clinical importance of each component and prioritizing the most critical endpoint, such as death, over less significant events.

Despite the popularity of the WR among cardiovascular trialists, this approach has not been extensively used in other areas of clinical research. We contend, that perioperative and periprocedural researchers could consider the WR and related approaches when the outcomes of interest are not of similar clinical importance. To this end, understanding the benefits and limitations of the WR will be essential to exploit its benefits, while avoiding potential misuses of the technique.

One critical step in the design of clinical trials is the computation of sample size required to address a specific research question. In general, this calculation relies on pre-existing subject-matter knowledge of the levels of outcome in the control arm, the minimal effect size deemed clinically relevant, and the expected variability. In addition, researchers need to specify the alpha level and statistical power in accordance with the desired risk of type I and type II errors, respectively.

Accordingly, the ability of clinical trials to provide meaningful results can be threatened when the number of participants needed to achieve the desired statistical power is insufficient. For example, suppose that researchers are interested in investigating the 1-year mortality of patients with symptomatic peripheral arterial disease undergoing lower limb revascularization, when comparing rivaroxaban plus aspirin (treatment group) vs. placebo plus aspirin (control group). Assuming an alpha critical level of 0.05 and statistical power of 80%, approximately 3000 participants would be needed to demonstrate a clinically relevant decrease in four percentage

*Correspondence:
Gudrun Kunst
gudrun.kunst@kcl.ac.uk
Tim Clayton
Tim.Clayton@lshtm.ac.uk
[1] Anaesthetics Department, Guy's and St Thomas' Hospital NHS Foundation Trust, London, UK
[2] Clinical Trials Unit, London School of Hygiene and Tropical Medicine, London, UK
[3] Department of Anaesthetics, King's College Hospital NHS Foundation Trust and School of Cardiovascular and Metabolic Medicine & Sciences, King's College London, London, UK

points for 1-year mortality (from 20.0 to 16.0%), when comparing the treatment group with the control group. Accounting for losses to follow-up, the number required would be even higher. While this sample size may be accomplished in adequately funded, multicentre clinical trials, in many scenarios these numbers are unrealistic. Had the study been conducted with, say, 2400 participants, the estimated statistical power would have been reduced from 80 to 70%. The problem of lack of statistical power becomes more pronounced when the frequency of the event of interest is relatively rare.

One statistically efficient alternative to overcome the problem of conducting clinical trials to evaluate treatment effects when outcomes are relatively rare is to use composite outcomes. They combine different endpoints related to the primary objective of the study, thus optimizing the statistical power by increasing the number of events observed (Freemantle et al. 2003). However, the interpretation of composite outcomes can be problematic, particularly when the clinical relevance of individual endpoints is substantially different. When these endpoints need to be clinically prioritized, the win ratio (WR) approach has been proposed as an alternative to analyze composite outcomes while accounting for the clinical importance of each individual component (Pocock et al. 2012; Baracaldo-Santamaria et al. 2023).

In this narrative review, we discuss main opportunities and challenges of the WR approach. In addition, we introduce alternative forms of the WR, which have been described in the literature to analyze composite outcomes.

Composite outcomes

Composite outcomes have been extensively used, particularly in small randomized clinical trials or where the number of events is low (Freemantle et al. 2003). They combine two or more endpoints into a single measure, thereby allowing researchers to improve statistical power, while avoiding the problem of multiple testing when evaluating individual components of composite outcomes separately (Multiple endpoints in clinical trials: guidance for industry 2022). From this perspective, the statistical analysis does not require adjustment for type I error (Freemantle et al. 2003). In addition, the combination of endpoints can substantially reduce the required sample size, thus improving the statistical efficiency by increasing the event occurrence (Cordoba et al. 2010; Redfors et al. 2020). This is particularly true when the treatment of interest has a consistent impact across the individual components of the composite outcome (Baracaldo-Santamaria et al. 2023).

When combining outcomes of similar clinical relevance and related to the primary objective of interest, a conventional analysis of composite endpoints is usually deemed appropriate [1]. For example, in comparing epinephrine with phenylephrine infusion for the prevention of hypotension in patients undergoing spinal anesthesia for cesarean delivery, outcomes such as the occurrence of hypotension, hypertension, bradycardia, and/or tachycardia are comparable, and therefore, the combination of these endpoints seems a sensible choice (Hassabelnaby et al. 2024).

On the other hand, the combination of safety and efficacy outcomes is generally not recommended, because the overall effect of composite endpoints with such heterogeneous constituents can be hard to interpret (Pocock et al. 2015). Furthermore, the conventional analysis of composite outcomes may not be appropriate when the clinical relevance of the outcomes involved are substantially different (Baracaldo-Santamaria et al. 2023). This is because important information as to whether nonfatal outcomes occur more than once, or are followed by a fatal event, is disregarded with traditional time-to-event analysis techniques (Pocock et al. 2012; Cordoba et al. 2010). More importantly, alternative approaches are needed when fatal outcomes are included in the analysis of composite outcomes, in order to prioritize their individual contribution to the overall effect.

Despite these potential limitations, the literature is replete with examples of conventional analysis of composite outcomes involving mortality (Cardoza et al. 2024; Perkovic et al. 2019). Consider, for example, a randomized clinical trial comparing renal outcomes in patients with diabetic nephropathy receiving canagliflozin—a sodium-glucose co-transporter 2 (SGLT-2) inhibitor—or placebo. In this study, the primary outcome was a composite of end-stage kidney disease, increasing serum creatinine levels by twice as much, or death from renal or cardiovascular causes. This study demonstrated a beneficial treatment effect with the intervention (Perkovic et al. 2019). Notably, the occurrence of outcomes such as mortality should be prioritised over endpoints related to the kidney function decline.

While conventional analyses effectively optimize the statistical power and are appropriate when the individual outcomes are of similar clinical importance, alternative approaches are warranted when outcomes such as mortality are incorporated. Table 1 outlines some of the pros and cons of using conventional analysis for composite outcomes in clinical trials (Pocock et al. 2015). In the next section, we explore the issues encountered when mortality endpoints are included in composite outcomes.

Evaluating mortality outcomes

The inclusion of composite outcomes may raise concerns when the measures of interest are not of similar

Table 1 Pros and cons of using conventional analysis for composite outcomes in clinical trials (Pocock et al. 2015)

Pros	Cons
• Improvement of statistical power by increasing the number of events • Statistically efficient technique when outcomes involved are relatively infrequent • Statistical analysis does not require adjustment for multiple testing • Indicated when outcomes included have similar clinical importance	• Attention is focused on the occurrence of the first outcome, which is often the least important measure • Misleading interpretation when the clinical importance of outcomes involved is substantially different • Not indicated when outcomes need to be clinically prioritized • May be less appropriate for the analysis of outcomes involving fatal endpoints

clinical importance (Cordoba et al. 2010). In this regard, composite outcomes often combine fatal events (usually occurring later in follow-up, and analyzed with time-to-event techniques), with non-fatal events (typically involving recurring measures that require longitudinal analysis), (Cordoba et al. 2010).Thus, the inclusion of mortality endpoints can be problematic with usually less clinically important non-fatal events occurring prior to the fatal event. In addition, when the effect on mortality predominates over other components, the observed mortality effect can be diluted, thereby misleading the interpretation of the composite outcome.Furthermore, when treatment has a true effect on other components of the composite outcome, but the study is underpowered to demonstrate any effect on mortality, any overall positive results might be misinterpreted as similar improvement on all individual endpoints. That is, composite outcomes do not quantify the effect of individual endpoints. This issue has led the US Food and Drug Administration (FDA) and other authorities to recommend that individual constituents of composite outcomes should be analyzed and reported separately as secondary outcomes (Multiple endpoints in clinical trials: guidance for industry 2022; Butcher et al. 2022). Accordingly, when reporting the effect of a novel treatment with a composite outcome that combines mortality with less clinically important events, caution must be exercised when interpretating the overall effect.

Returning to the example provided in the introduction, consider a clinical trial involving 6564 patients with peripheral artery disease undergoing lower limb revascularization. Patients were randomized to receive 2.5 mg of rivaroxaban plus aspirin, or placebo plus aspirin, to investigate the effect of rivaroxaban on the incidence of ischemic limb and cardiovascular events, and bleeding (as the principle saftey outcome), (Bonaca et al. 2020). A composite outcome of ischemic risk,including acute limb ischemia, major limb amputation, myocardial infarction, ischemic stroke, and cardiovascular death,was defined as the primary efficacy outcome. The results indicated that rivaroxaban effectively decreased ischemic risk, at the cost of an increasing bleeding risk in these patients. However, when interpreting the overall effect of the composite outcome, the beneficial effects of rivaroxaban were mainly driven by reducing the incidence of acute limb ischemia, major amputation for vascular causes, myocardial infarction, and ischemic stroke, but less so by cardiovascular death (Bonaca et al. 2020).

A perceived complexity in interpreting composite outcome results has led some authors to suggest avoiding the use of composite outcomes, in particular when mortality is combined with other clinically relevant, but less important endpoints (Cordoba et al. 2010; Butcher et al. 2022).

Developments beyond composite outcome as the primary endpoint

Motivated by the need for combining outcomes such as mortality with outcomes involving repeated measures, Finkelstein and Schoenfeld proposed in 1999 a novel approach to handle composite outcomes. The methodology suggested by these authors is based on the non-parametric Wilcoxon–Mann–Whitney test to compare the sum of ranks of two continuous outcomes (Finkelstein and Schoenfeld 1999).

With the Finkelstein-Schoenfeld (FS) test, mortality is first analyzed by making pairwise comparisons between each participant allocated to the treatment group and each participant allocated to the control group, and assigning a score of 1 if the participant belonging to the treatment group survived and the participant assigned to the control group died, or assigning a score of – 1 in the opposed situation; in a second stage, in the event that both participants died, the time-to-death is analyzed, and scores are assigned according to which of the pair died first within a common follow-up period. When there is not enough information available as to who died first, scores may be assigned with respect to the outcome involving repeated measures or recurrent events (e.g., stroke) within a common follow-up period for that pair. The FS score is then computed by summing the obtained scores for the treated group and converted to a p-value. The details of this test are beyond the scope of this review

and are provided by Finkelstein and Schoenfeld (Finkelstein and Schoenfeld 1999).

The win ratio approach

In evaluating the effectivity of composite outcomes with win ratio methodology, matched and unmatched analyses have been described. With the matched approach, each individual in the treatment group is paired with a single individual in the control group according to their underlying risk of two or more individual outcomes (i.e., a composite outcome). Similar to any matching technique, this methodology generally increases the statistical power of the test by making comparisons between participants with similar risks (Pocock et al. 2012). However, one drawback of this method is that not all individuals can be matched, and in consequence, a variable number of observations needs to be removed from the analysis. Hence, the matched win ratio approach has not gained wide acceptance among researchers (Redfors et al. 2020), and its use is generally not recommended for clinical interventional trials, although it may be useful in observational studies (Multiple endpoints in clinical trials: guidance for industry 2022).

In contrast, with the unmatched approach every individual in the treated group is compared with every individual in the control group for the hierarchical composite endpoint, thus including every participant in the analysis. The methodology proposed by Finkelstein and Schoenfeld was later used by Pocock to further develop the unmatched approach of a new approach, which was introduced in 2012 as win ratio (Pocock et al. 2012). For the purposes of the present review, we refer to the unmatched win ratio as WR. The WR can be interpreted as an extension of Wilcoxon–Mann–Whitney test for single continuous outcomes to a more generalized test that accommodates different types of outcomes with missing data, and provides a measure effect with its confidence interval (Pocock et al. 2012, Redfors et al. 2020, Pocock et al. 2023). The key advantage of WR approach is that priority is given to the most important endpoint such as, e.g., death, instead of the first event to happen (Pocock et al. 2012).

With the aim of understanding the rationale behind the WR, we provide an example from the literature. In a randomized multicenter clinical trial, Tavares et al. conducted a trial to evaluate the efficacy of dapaglifozin, a SGLT-2 inhibitor, to improve a composite outcome involving (1) mortality, (2) use of continuous renal replacement therapy (CRRT), and (3) length of stay (LOS) in critical care patients. With this aim, a total of 507 patients (admitted with acute organ dysfunction to 22 different critical care units in Brazil) were randomized to receive 10 mg of dapaglifozin along with the standard of care ($n=248$), or standard of care alone ($n=259$), (Tavares et al. 2024). In analyzing the primary composite outcome with the WR, Tavares et al. reported a total of 27,143 wins and 26,929 losses from a total of 64,232 (248×259) pairwise comparisons, yielding a WR statistic of 27,143/26,929 = 1.01 (95% CI 0.90 to 1.13, p-value = 0.89). Therefore, the authors concluded that the addition of dapaglifozin to the standard of care of critical care patients with acute organ dysfunction did not improve the proposed clinical outcomes, after accounting for their clinical relevance (Tavares et al. 2024).

A worked example

In this section, we develop a worked example adapted from the clinical trial conducted by Tavares and colleagues (Tavares et al. 2024), with the aim of illustrating the methodology used by unmatched WR to prioritize outcomes in the setting of trial designs with composite outcomes.

For simplicity, assume that in the clinical trial above described only 8 patients were assigned to receive dapagliflozin (group = 1), whereas 12 patients did not receive this treatment (group = 0). We compared 8 against 12 patients to emphasize that the number of participants allocated to the treatment and control groups does not need to be identical to conduct a WR.

Figure 1 and the supplementary information illustrate the methodology used in this example to compare the outcomes mortality, CRRT, and LOS among 20 patients. A total of 96 pairwise comparisons are made between each patient allocated to the treatment group (dapaglifozin plus standard of care) and all patients allocated to the control group (standard of care alone). A total of 45 wins, 49 losses, and 2 ties were obtained with respect to mortality, CRRT, and LOS. The details for the statistical analysis of composite outcomes with WR are provided in Figs. 1, 2, 3, and 4 of the Supplementary information.

In the example illustrated in Fig. 1, the hierarchical levels for each component of the composite outcome were pre-defined in accordance with the clinical importance of the outcome. Thus, mortality was prioritized over CRRT, and these two outcomes were deemed clinically more important than LOS. This step is critical for the evaluation of composite outcomes, because the results will be driven by the hierarchical arrangement of these variables (Tavares et al. 2024).

The win difference (WD) is used as a measure to compare the treatment with standard of care in individual outcomes, and it represents the absolute difference between the number of"wins"in the treatment group and the control group. A positive WD indicates benefit with the treatment, i.e., a “win”; a negative WD indicates benefit with standard care, i.e., a “loss”; and a WD of 0

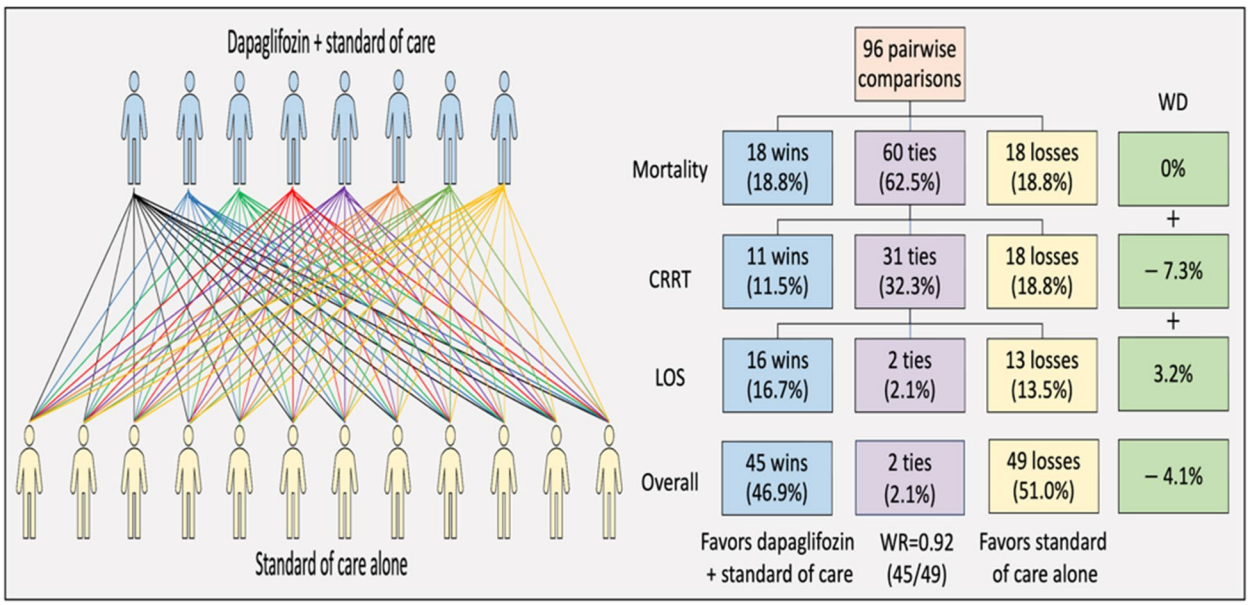

Fig. 1 Unmatched win ratio analysis (WR) and win difference (WD) for 20 randomized participants allocated to the treatment group (dapaglifozin plus standard of care) or control group (standard of care alone), with respect to a hierarchical composite outcome encompassing (1) mortality, (2) continuous renal replacement therapy (CRRT), and (3) length of stay in critical care (LOS). Adapted from Pocock et al. 2023

indicates no benefit with either the treatment or control, thus representing a "tie".

The comparison between 8 participants assigned to the treatment group against 12 participants assigned to the control group produced a total of 96 pairwise comparisons, as depicted in Fig. 1. For every discordant comparison (i.e., when the outcome differed for the participant assigned to treatment group, as compared to the one assigned to the control group), a win is assigned to the treatment group if the treated individual showed a better outcome. Otherwise, it will be counted as a loss. For example, when comparing the CRRT status between two participants (one allocated to the treatment group, and one allocated to the control group), there are four possible scenarios: only the treated patient required CRRT, only the control patient required CRRT, both patients required CRRT, or none of them required CRRT. For the first two cases, the pairwise comparisons would be regarded as loss and win (i.e., discordant pairwise comparisons), respectively; the last two cases are regarded as concordant pairwise comparisons. These tied results are then carried forward for the evaluation of the next outcome, in this case LOS, and the procedure continues until all outcomes are exhausted. At this stage, the residual concordant pairwise comparisons are regarded as ties.

According to this example, when accounting for mortality outcome, there were 18 wins and 18 losses; when combining mortality and CRRT, there were 11 wins and 18 losses; and after accounting for mortality, CRRT and LOS, there were 16 wins and 13 losses. In consequence, the total number of wins and losses for a composite outcome involving mortality, CRRT and LOS was 45 and 49, respectively, whereas there were 2 final ties after all pairwise comparisons were analyzed.

It follows that the WR statistic, computed as the number of wins divided by the number of losses, is 0.92 (45/49). The interpretation is that, if the treated and control patient differ in the outcome (i.e. a discordant pair), the odds for the treatment group to do better than the control group are 0.92. Equivalently, the odds for the control group to do better than the treatment group are 1.09 (1/0.92). Stated in a different way, the probability that the participant on dapagliflozin wins is $92/(1+0.92)=0.48$ (Pocock et al. 2012).

The WR for this small hypothetical example suggests that dapaglifozin might not be of benefit in patients admitted with acute organ dysfunction. As expected, given the small sample size ($n=20$), the level of evidence for an improvement of this composite outcome in patients prescribed dapaglifozin was poor (95% CI 0.31 to 2.71, p-value 0.88).

The methodology above described can be employed for combining binary as well as continuous, categorical and ordinal outcome measures. Moreover, one of the key features of the WR approach for clinical trials with

Table 2 Six proposed steps for the analysis of composite outcomes using the unmatched WR approach (Pocock et al. 2012)

Step 1	**Prioritize endpoints taking part of the composite outcome, according to their clinical importance**
Step 2	Make pairwise comparisons between each participant assigned to treatment group, and every participant assigned to the control group, starting with the most prioritized outcome
Step 3	Define wins, losses and ties with respect to the most prioritized outcome
Step 4	For every tied pairwise comparison obtained from Step 3, define wins, losses, and ties with respect to the next most clinically important endpoint. Repeat this process until all outcomes are exhausted
Step 5	Repeat Step 4 until the evaluation of all outcomes is exhausted
Step 6	Compute the WR from the division between the number of ties and the number of losses obtained from the last available outcome. Report uncertainty of the obtained WR, in the form of confidence intervals and a *p*-value

WR win ratio

composite outcomes is its flexibility to combine other types of outcomes (for example, involving time-to-event data, longitudinal data, or self-reported events), thereby endowing the researcher with an armament of options that accommodate to the specific requirements of any composite outcome. Further, outcomes can be analyzed from the perspective of the event occurrence (yes or no), the number of events occurred, the time elapsed until the first event occurs (time-to-event analysis), or the severity of events (Redfors et al. 2020).

In recent years, a variety of statistical software packages have been made readily available on the internet for the application of this methodology. See, for example, the WWR and WINS packages developed for the R® programming language (Qiu et al. 2017; Introduction to the R package WINS 2024) the "winratiotest" command developed for Stata® statistical software (Gregson et al. 2023), and the implementation of WR in SAS® statistical software (Dong et al. 2016). Mao et al. have also described a methodology for the calculation of sample size in win ratio analysis (Mao et al. 2022), which has recently been implemented in R® statistical software (Sample size calculation for standard win ratio test 2024).

The stages involved in the analysis of composite outcomes with unmatched WR approach are summarized in Table 2 (Pocock et al. 2012).

Pros of WR

Conventional analyses of composite outcomes do not account for the clinical importance of individual components, and therefore the use of alternative methods is warranted and has been proposed. The WR offers an attractive solution to this problem, thereby providing clinicians with a useful metric, which is relatively easy to compute (Pocock et al. 2012).

The key advantage of the WR and other related tests is that outcomes are prioritized in accordance with their clinical impact on individuals. For example, mortality assessed in the WR approach with the highest priority provides more weight to this specific outcome, which contrasts with mortality as an individual outcome component of a conventional composite outcome. Here, a potentially low incidence of postoperative mortality does not add much weight in comparison to higher incidences of other individual components of composite outcome variables. In addition, with this methodology, the sample size required may be smaller to achieve the same statistical power when compared to conventional approaches. This feature has been demonstrated with simulation studies, although it would depend on specific aspects of the trial and the interventions (Redfors et al. 2020; Pocock et al. 2023).

Individual outcome variables of the WR approach may include binary as well as continuous, categorical, and ordinal outcome measures. This has the advantage of facilitating a combination of clinical and patient-centred outcomes, such as organ failure (e.g., acute kidney injury) plus quality of life (e.g., days-alive -and-at-home).

Furthermore, with the advent of computer programs recently developed for a variety of statistical softwares (Qiu et al. 2017; Gregson et al. 2023; Dong et al. 2016), the estimation of confidence intervals for the WR allowing for the lack of independency of pairwise comparisons can be readily obtained as well as sample size estimations.

WR analyses may lead to gains in power, particularly with high patient heterogeneity and low rates of drug discontinuation in pharmacological trials; however, this is not guaranteed (Claggett et al. 2018).

Cons of WR

It is worth noting that the methodology described in Fig. 1 and Table 2 to calculate the WR systematically excludes tied pairwise comparisons. When the number of ties obtained is large, this approach may be seen as problematic, because estimated treatment effects could be overestimated. However, confidence intervals are typically wide (Ajufo et al. 2023). Furthermore, the reported WR may not represent the whole study population (it

only involves a sub-population of patients for whom the corresponding pairwise comparison was labeled either as a win or a loss). On the other hand, additional outcome variables as part of the WR, e.g., inclusion of a quality-of-life measure, can then be helpful to clarify the wins as well as incorporating other important patient-centred outcomes and act as a "tiebreaker" (Ajufo et al. 2023).

An alternative metric, known as win odds (WO), has recently been proposed to address the problem of ignoring tied pairwise comparisons (Brunner et al. 2021). The WO is computed by adding one-half of the total number of ties to the numerator and denominator of the WR. In the example summarized in Fig. 1, the WO corresponds to $(45+1)/(49+1)$. Thus, this quantity remains virtually unchanged as compared to the unmatched WR (45/49), because there were only 2 ties. In fact, in the absence of ties, WO reduces to WR. However, in a study with a higher number of ties, the WO can be substantially different, thus adding complexity to the interpretation.

The situation with a large number of ties can be pictured with the following hypothetical example using simulation: Consider a randomized clinical trial where 50 patients were allocated to the treatment group, and 52 were allocated to the control group. From the resulting 2,600 pairwise comparisons, there were 147 wins, 49 losses, and 2404 ties. The WR would be computed as $147/49=3.0$ (95% CI 0.41–21.9, p-value 0.279), (Introduction to the R package WINS 2024), but importantly, 2404 comparisons would be ignored. Applying the above adjustment with tied comparisons equally allocated to each arm, the resulting score would be $(1202+147)/(1202+49)=1.08$ (95% CI 0.93–1.25, p-value 0.324), (Introduction to the R package WINS 2024). Thus, some authors have recommended that WO should be reported in the presence of a high number of ties (Ajufo et al. 2023; Dong et al. 2023).

One disadvantage of WO is that the interpretation can be less intuitive, as compared to WR (Pocock et al. 2023). In addition, the analysis of composite outcomes with many ties would result in WO that favors the null hypothesis of no benefit of the proposed treatment. In consequence, in the context of non-inferiority trials in particular, the use of WO is generally not recommended (Ajufo et al. 2023).

From a clinical standpoint, the observed differences between pairwise comparisons that ultimately define winners and losers may not necessarily be of practical or clinical relevance. This limitation has led some authors to propose a winner is declared only if the pairwise difference is of a clinically relevant given size, e.g., based on a given difference in a quality-of-life scale, which is clinical meaningful, or the amount of troponin release in defining a myocardial infarction. However, this approach would be detrimental for the statistical power given the increased number of ties, and therefore, the use of margins (or clinically relevant given sizes of pairwise differences) has been discouraged by other authors (Redfors et al. 2020).

Another caveat of WR is that comparisons are often made between individuals that are not necessarily under the same risk of developing the outcome (Pocock et al. 2012; Ajufo et al. 2023). To overcome this problem, data can be stratified according to the variables influencing the risk of the composite outcome, and the stratified WR can be obtained by combining WRs across strata (Pocock et al. 2012). For example, in a randomized clinical trial, researchers evaluated the benefit of empagliflozin (another type of SGLT-2) as compared with placebo, in patients with heart failure (HF) after initial stabilization (Voors et al. 2022). The clinical benefit was defined by a composite outcome involving mortality, number of HF events, time to first HF event, and a self-reported outcome evaluating quality of life. The HF outcome was stratified into patients with acute de-novo HF and those with decompensated chronic HF. The reported WR for de novo and decompensated HF patients was 1.29 (95% CI 0.89–1.89) and 1.39 (95% CI 1.08–1.81), respectively, and the combined WR was 1.36 (95% CI 1.09–1.68), (Voors et al. 2022).

It should be noted that the WR has been conceived as a relative measure of wins and losses. An alternative approach as defined earlier is to report the WD in an additive scale, expressed in terms of percentage (i.e., % of wins minus % of losses) instead of the relative measure given by WR. In the example summarized in Fig. 1, the percentage of wins (46.9%) minus the percentage of losses (51.0%) yields a WD of -4.1%. One advantage of this approach is that it provides an absolute measure of treatment benefit, as opposed to the relative measure given by WR. Although the interpretation of WR and WD can be analogous to odds ratio and risk difference respectively, further calculations including number needed to treat (or harm) cannot be immediately inferred, because the WR only includes pairwise comparisons that were not disparate among groups (Ajufo et al. 2023). Similarly, in the setting of time-to-event analysis, WRs are comparable to hazard ratios, with the exception that tied comparisons are excluded from the analysis (Pocock et al. 2023; Ajufo et al. 2023).

Given the relative novelty of WR score, particularly outside the area of trials in cardiovascular disease, clinicians may find it challenging to translate the results into clinical practice (Pocock et al. 2012). For example, in the empagliflozin trial (Voors et al. 2022), how can a WR of 1.36 be interpreted? In this clinical trial, authors reported that the superiority of empagliflozin over placebo was

Table 3 Challenges encountered in the analysis and interpretation of win ratio approach, and proposed solutions to overcome these problems

Disadvantage	Potential solution
Less intuitive interpretation, as compared with other measurements of treatment effect (e.g., hazard, risk ratio, and risk difference)	Report WR along with number of wins, losses and ties for each individual outcome, as illustrated in Fig. 1
Tied pairwise comparisons are excluded from the analysis	Compute win odds instead when the number of tied comparisons is high
Differences observed not clinically relevant	Define clinically meaningful differences when computing wins and losses, in accordance with the clinical relevance of the endpoints
Comparison between individuals with different risk factors	Consider using stratification techniques to estimate the win ratio
Win ratio is a relative measure of treatment effect	Consider reporting win differences along with win ratios
Outcomes are prioritized, but weighting is equally assigned to each individual endpoint	Acknowledge this limitation and interpret results with caution. Alternatively, report weighted wins and losses, at the expense of increased complexity in the interpretation

WR win ratio

mainly driven by self-reported quality of life outcomes. The differences observed for mortality and HF events, although clinically relevant, did not substantially change the overall WR. However, this information is not conveyed in a single WR score. Therefore, we advocate the use of flow charts, as the one provided in Fig. 1, when analyzing composite outcomes with WR approach, to ensure transparency in the results for readers (Pocock et al. 2023).

Lastly, clinicians should be aware that although the WR approach effectively prioritize outcomes, wins and losses are equally weighted across outcomes of differing clinical importance, thereby making this methodology in some instances inappropriate [22]. For example, if mortality has the same weight as hospitalization for HF, the WR will likely be driven by hospitalization events, because they can occur more frequently and earlier during the follow-up, despite the fact that when the outcome is associated to mortality more deaths are expected as the trial lasts longer.

Table 3 outlines the main disadvantages of WR and proposes potential solutions to address these issues, including the possibility of applying pre-defined weights to each of the components of the WR.

Conclusion

The analysis of composite endpoints often requires the use of alternative methods that account for the clinical importance of outcomes involved. While several methodologies have been proposed, the WR approach appears to be a sensible choice and has been increasingly used in cardiovascular trials (Pocock et al. 2024). This method is not free of drawbacks, including difficulties in translating results into a clinical setting and making comparisons between participants that are not under the same risk of developing the outcome. Understanding the nuances and benefits of this methodology, while recognizing its limitations, may help researchers choose the best analysis strategy for a particular combination of outcomes.

With this review, we hope to give some insights to guide perioperative and periprocedural trialists towards the usage of the WR. We also highlight the need for establishing a benchmark to consistently analyze composite outcomes and unify consensus as to what should be the most appropriate way to report them.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13741-025-00550-8.

Supplementary Material 1. Supplementary Figure S1. A worked example of the use of unmatched win ratio approach for composite outcomes in clinical trials. Suppose that researchers are interested in evaluating the efficacy of dapaglifozin in the critical care setting [Tavares, 2024]. This medication has been previously shown to be of benefit in patients diagnosed with type 2 diabetes, heart failure, and chronic kidney disease. To this end, researchers defined a composite outcome to evaluate mortality, use of continuous renal replacement therapyand length of stay in critical care. These outcomes have been arranged in descending order of importance. In this example, 8 patients have been assigned to group 1, and 12 patients have been assigned to group 0. The win ratio analysis of this dataset is shown in Figs. 2–4. id: subject identity. Adapted from Tavares, 2024. Supplementary Figure S2. Unmatched win ratio analysis of the outcome mortality, for the data shown in Fig. 1. With this methodology, each patient assigned to the treatment group is compared against patients assigned to the control group. All discordant pairwise comparisons are highlighted in green or red. Assuming that the proposed intervention is expected to improve mortality, any comparison wherein the subject did not die in the intervention group, and the counterpart died in the control group, is labeled as a win. Analogously, any comparison resulting in the death of the subject allocated to the treatment group and the survival of the compared subject assigned to the control group, is considered a loss. Thus, in this example, there were 18 wins and 18 losses. id: subject identity. Adapted from Tavares, 2024. Supplementary Figure S3. Unmatched win ratio analysis of a composite outcome defined by mortality and continuous renal replacement therapy, for the data shown in Fig. 1. In this example, only pairwise comparisons that were concordant in death, but discordant in terms of the use of CRRT, are highlighted in green or red. Given that all highlighted comparisons are concordant for death, and assuming that the proposed intervention is expected to avoid the use of CRRT, any comparison wherein the subject was not prescribed CRRT in the

intervention group and the counterpart did need this therapy in the control group is labeled as a win. Analogously, any comparison resulting in the use of CRRT for the subject allocated to the treatment group, and the avoidance of this therapy for the subject assigned to the control group, is considered a loss. Consequently, when accounting for mortality and CRRT, there were 11 wins and 18 losses. id: subject identity. Adapted from Tavares, 2024. Supplementary Figure S4. Unmatched win ratio analysis of a composite outcome defined by mortality, continuous renal replacement therapy, and length of stay in critical care, for the data shown in Fig. 1. Only pairwise comparisons that were concordant in death and CRRT, but discordant in terms of LOS, are highlighted in green or red. Since all comparisons highlighted in green or red are concordant for mortality and CRRT, assuming that the proposed intervention is expected to shorten LOS, any pairwise comparison wherein the subject assigned to the intervention group had shorter LOS, as compared with their counterpart allocated to control group, is labeled as a win. Analogously, any comparison resulting in a shorter LOS for the subject assigned to the control group is considered a loss. It follows that when accounting for mortality, CRRT and LOS, there were 16 wins and 13 losses. Of note, there were also 2 ties. id: subject identity. Adapted from Tavares, 2024.

Acknowledgements
Not applicable

Authors' contributions
Literature search: WF, GK, TC. Drafting of paper: WF, GK, TC, MD. Revising and approving of final paper: TC, MD, WF, GK.

Funding
Gudrun Kunst is supported by an NIHR Senior Clinical and Practitioner Research Award (NIHR306274).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 17 January 2025 Accepted: 9 June 2025
Published online: 05 July 2025

References

Ajufo E, Nayak A, Mehra MR. Fallacies of using the win ratio in cardiovascular trials: challenges and solutions. Basic to Translational Science. 2023;8(6):720–7.
Baracaldo-Santamaría D, Feliciano-Alfonso JE, Ramirez-Grueso R, Rojas-Rodríguez LC, Dominguez-Dominguez CA, Calderon-Ospina CA. Making sense of composite endpoints in clinical research. J Clin Med. 2023;12(13):4371.
Bonaca MP, Bauersachs RM, Anand SS, Debus ES, Nehler MR, Patel MR, at al. Rivaroxaban in peripheral artery disease after revascularization. NEJM. 2020;382(21):1994–2004.
Brunner E, Vandemeulebroecke M, Mütze T. Win odds: an adaptation of the win ratio to include ties. Stat Med. 2021;40(14):3367–84.
Butcher NJ, Monsour A, Mew EJ, Chan AW, Moher D, Mayo-Wilson E, et al. Guidelines for reporting outcomes in trial reports: the CONSORT-outcomes 2022 extension. JAMA. 2022;328(22):2252–64.
Cardoza K, Kang A, Smyth B, Yi TW, Pollock C, Agarwal R, et al. Geographic and racial variability in kidney, cardiovascular and safety outcomes with canagliflozin: a secondary analysis of the CREDENCE randomized trial. Diabetes Obes Metab. 2024;26(9):3530–40.
Claggett B, Pocock SJ, Wie LJ, Pfeffer MA, McMurray JJV, Solomon SD. Comparison of time-to first event and recurrent-event methods in randomized clinical trials. Circulation. 2018;138(6):570–7.
Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. BMJ. 2010;341: c3920.
Dong G, Li D, Ballerstedt S, Vandemeulebroecke M. A generalized analytic solution to the win ratio to analyze a composite endpoint considering the clinical importance order among components. Pharm Stat. 2016;15(5):430–7.
Dong G, Huang B, Verbeeck J, Cui Y, Song J, Gamalo-Siebers M. Win statistics (win ratio, win odds, and net benefit) can complement one another to show the strength of the treatment effect on time-to-event outcomes. Pharm Stat. 2023;22(1):20–33.
Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. Stat Med. 1999;18(11):1341–54.
Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA. 2003;289(19):2554–9.
Gregson J, Ferreira JP, Collier T. Winratiotest: A command for implementing the win ratio and stratified win ratio in Stata. Stand Genomic Sci. 2023;23(3):835–50.
Hassabelnaby YS, Hasanin AM, Shamardal M, Mostafa M, Zaki RM, Elsherbiny M, et al. Epinephrine vs. phenylephrine infusion for prophylaxis against maternal hypotension after spinal anesthesia for cesarean delivery: a randomized controlled trial. Journal of Anesthesia. 2024;38(4):500–7.
Introduction to the R package WINS, Cui Y, Huang B. 2024. https://cran.uni-muenster.de/web/packages/WINS/vignettes/vignette.pdf. Accessed 17 Nov 2024.
Mao L, Kim K, Miao X. Sample size formula for general win ratio analysis. Biometrics. 2022;78(3):1257–68.
Multiple endpoints in clinical trials: guidance for industry. In US Food and Drug Administration. 2022. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm. Accessed 17 Nov 2024.
Perkovic V, Jardine MJ, Neal B, et al. Canagliflozin and Renal Outcomes in Type 2 Diabetes and Nephropathy. N Engl J Med. 2019;380(24):2295–306.
Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. Eur Heart J. 2012;33(2):176–82.
Pocock SJ, Clayton TC, Stone GW. Design of major randomized trials: part 3 of a 4-part series on statistics for clinical trials. J Am Coll Cardiol. 2015;66(24):2757–66.
Pocock SJ, Ferreira JP, Collier TJ, Angermann CE, Biegus J, Collins SP, et al. The win ratio method in heart failure trials: lessons learnt from EMPULSE. Eur J Heart Fail. 2023;25(5):632–41.
Pocock SJ, Gregson J, Collier TJ, Ferreira JP, Stone GW. The win ratio in cardiology trials: lessons learnt, new developments, and wise future use. Eur Heart J. 2024;45(44):4684–99.
Qiu J, Luo X, Bai S, Tian H, & Mikailov M. WWR: an R package for analyzing prioritized outcomes. J Med Stat Inform, 2017;5(4):1-7.
Redfors B, Gregson J, Crowley A, McAndrew T, Ben-Yehuda O, Stone GW, at al. The win ratio approach for composite endpoints: practical guidance based on previous experience. European Heart Journal. 2020;41(46):4391–4399.
Sample size calculation for standard win ratio test, Mao L. 2024. https://cran.r-project.org/web/packages/WR/vignettes/WR_sample_size.html. Accessed 17 Nov 2024.

Tavares CA, Azevedo LC, Rea-Neto Á, Campos NS, Amendola CP, Kozesinski-Nakatani AC, et al. Dapagliflozin for critically ill patients with acute organ dysfunction: the DEFENDER randomized clinical trial. JAMA. 2024;332(5):401–11.

Voors AA, Angermann CE, Teerlink JR, Collins SP, Kosiborod M, Biegus J, et al. The SGLT2 inhibitor empagliflozin in patients hospitalized for acute heart failure: a multinational randomized trial. Nat Med. 2022;28(3):568–74.

Publisher's Note