9-19-2023 1:00 PM

Nonparametric Methods for Analysis and Sizing of Cluster Randomization Trials with Baseline Measurements

chengchun yu, *Western University*

Supervisor: Choi, Yun-hee, *The University of Western Ontario*
: Zou, Guangyong, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Epidemiology and Biostatistics

Recommended Citation

yu, chengchun, "Nonparametric Methods for Analysis and Sizing of Cluster Randomization Trials with Baseline Measurements" (2023). *Electronic Thesis and Dissertation Repository*. 9697.
https://ir.lib.uwo.ca/etd/9697

Abstract

Cluster randomization trials are popular in situations where the intervention needs to be implemented at the cluster level, or logistical and/or financial reasons require the choice of randomization at the cluster level, or minimization of contamination is needed. It is very common for cluster trials to take measurements before randomization and again at follow-up, resulting in a clustered pretest-posttest design. For continuous outcomes, the cluster-adjusted analysis of covariance approach can be used to adjust for accidental bias and improve efficiency. However, a direct application of this method is inappropriate if the measurements are not on an interval scale, yet such data are very common in practice.

In this thesis, we propose nonparametric methods for trials with a clustered pretest-posttest design, focusing on estimation of the treatment effect. We quantify treatment effects using the win probability, defined as the probability that a randomly selected subject in the treatment group has a more favourable outcome than one in the control group. The methods for data analysis and sample size planning for estimating win probability rely on subject-specific win fractions created from outcome measurements at baseline and follow-up. Specifically, the win fraction for a subject is given by the difference between the rank of the observation among all observations in the combined sample of two groups and its rank among observations in its own group divided by the sample size of the comparison group. The cluster-adjusted analysis of covariance is then applied to win fractions created from baseline and follow-up measurements. The proposed methods, which may be considered as an extension of Zou (2021) for follow-up measurements, are applicable to studies with binary, ordinal, count, and continuous outcomes without making parametric assumptions.

Simulation results demonstrated that the methods for constructing confidence intervals for the win probability performed well in terms of coverage and average interval width, even when the number of clusters is small as 5 clusters per arm. The methods for sample size estimation also performed well in terms of the probability of achieving a pre-specified precision.

The methods are illustrated using data from two published cluster randomization trials with

SAS code provided.

Keywords: Cluster randomization trials, nonparametric, receiver operating characteristic curve, rank sum test, Wilcoxon-Mann-Whitney statistic, interval estimation, sample size.

Summary for lay audience

Cluster randomization trials are popular in situations where the intervention needs to be implemented at the cluster level, or logistical and/or financial reasons require the choice of randomization at the cluster level, or minimization of contamination is needed. It is very common for cluster trials to take measurements before randomization and again at follow-up, resulting in a clustered pretest-posttest design. For continuous outcomes, the cluster-adjusted analysis of covariance approach can be used to adjust for accidental bias and improve efficiency. However, a direct application of this method is inappropriate if the measurements are not on an interval scale, yet such data are very common in practice.

In this thesis, we propose nonparametric methods for trials with a clustered pretest-posttest design, focusing on estimation of the treatment effect. We quantify treatment effects using the win probability, defined as the probability that a randomly selected subject in the treatment group has a more favourable outcome than one in the control group. The methods for data analysis and sample size planning for estimating win probability rely on subject-specific win fractions created from outcome measurements at baseline and follow-up. Specifically, the win fraction for a subject is given by the difference between the rank of the observation among all observations in the combined sample of two groups and its rank among observations in its own group divided by the sample size of the comparison group. The cluster-adjusted analysis of covariance is then applied to win fractions created from baseline and follow-up measurements. The proposed methods, which may be considered as an extension of Zou (2021) for follow-up measurements, are applicable to studies with binary, ordinal, count, and continuous outcomes without making parametric assumptions.

Simulation results demonstrated that the methods for constructing confidence intervals for the win probability performed well in terms of coverage and average interval width, even when the number of clusters is small as 5 clusters per arm. The methods for sample size estimation also performed well in terms of the probability of achieving a pre-specified precision.

Acknowledgment

I am grateful to Dr. Choi and Dr. Zou for guiding and supervising me in the completion of this thesis and my degree. I would not have completed this thesis without their support and encouragement. They have taught me the philosophy and passion of statistics in addition to valuable research skills. I believe these skills will be a valuable asset in my career. I am also grateful for the financial support from Dr. Montero-Odasso. Working with his Dementia study group was exciting and inspiring.

I cannot express more of my appreciation toward my parents for supporting, both financially and mentally. I especially thank my mother for her encouragement during my downs and my father for his wisdom in facing challenges. I also thank Chloe for accompanying me during this marathon to complete my thesis. Finally, I would like to thank the friends from the church who kept encouraging me throughout my study.

Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA	Analysis of variance
ANCOVA	Analysis of covariance
AUC	Area under the receiver operating characteristic curve
AUDIT	Alcohol use disorders identification test
CDF	Cumulative distribution function
CONSORT	Consolidated standards of reporting trials
CV	Coefficient of variation
ECI	Early combined immunosuppression
GEE	Generalized estimating equation
HBI	Harvey-Bradshaw index
ICC	Intraclass correlation coefficient
MW	Mann-Whitney
NNT	Number needed to treat
PIM	Probabilistic index model
REACT	Randomised evaluation of an algorithm for Crohn's treatment
TVSFP	Television, school and family cessation project
WinP	Win probability
WLS	Weighted least square

Chapter 1

Introduction

Randomized controlled trials are important tools to assess the magnitude of treatment effects in evidence-based medicine, where individual subjects are allocated into different intervention arms, and the outcomes between study arms are compared. Conventional trials randomize individuals into different study arms; however, randomizing clusters (groups) of individuals into different intervention arms could be the more feasible design due to practical reasons, such as avoiding logistic convenience or intervention contamination (Donner *et al.*, 1981). Trials with such a design are referred to as cluster randomization trials and are often chosen to evaluate the effectiveness of educational programs or policy change because the implementation of the intervention occurs at the cluster level.

As in individually randomized trials, it is common for cluster randomization trials to measure the outcome variable at baseline and include it in the analysis to increase the efficiency and control for accidental imbalance at baseline. Analysis of covariance (ANCOVA) with adjustment for clustering is generally recommended to analyze data from cluster trials with such a design (Hooper *et al.*, 2018). However, this approach could be invalid or inappropriate when mean comparisons of the outcome have no clear interpretation, which is common when the outcome is measured on an ordinal scale, such as the Likert scale or the modified Rankin scale. We consider an effect measure that requires no unit of the outcomes for interpretation, defined as the probability of one randomly chosen participant in the treatment arm having a better outcome over a randomly selected participant in the control arm (Zou, 2021). Although Zou (2021) refers to this probability as the Mann-Whitney (MW) probability, reflecting the underlying parameter in the MW test, we prefer the term win probability (WinP) and focus on its estimation for cluster randomization trials with baseline measurements in this thesis. We also develop corresponding formulas for sample size estimation. The development of our methods is motivated by Zou (2021) based on ranks and extended to include baseline measurements to increase efficiency and reduce bias from accidental baseline imbalance.

We will briefly introduce randomized controlled trials, followed by a brief review of analyzing trials that involve ordinal scale outcomes and their limitations in current literature. We will also review sample size estimation and discuss the statistical challenges. Finally, we will finish this chapter with the objectives and organization of this thesis.

1.1 Randomized controlled trials

A randomized controlled trial allocates participants randomly into the control group or the intervention group, such that the group assignment for a participant is purely due to chance and free of confounding from other factors. Randomization enables the statistical testing of no treatment effect by creating comparable groups that only differ by interventions; hence, we can quantify how likely the differences between study groups are merely due to chance. Statistical theories based on identical distribution and independent observations are often applied to analyze data from such trials to estimate an effect measure that answers clinically important questions.

Three fundamental issues in designing a trial include: specifying eligibility criteria for enrolling subjects, choosing meaningful interventions and defining outcome measurements that are reliable and responsive (Pocock, 2013). The treatment effect is then quantified by comparing the outcomes between studying arms according to the scientific questions. Common effect measures to quantify treatment effects are mean difference for continuous outcomes, risk difference/ratio for binary outcomes and hazard ratio for survival outcomes.

Although the focus of randomized controlled trials is to compare the outcome after the participants received the intervention for some defined time periods, it is common to utilize the information of the same outcome variable measured before randomization. The benefit of including the baseline measurement in an analysis of covariance is that it improves efficiency and adjusts for accidental bias caused by baseline imbalance (Vickers and Altman, 2001). However, such an approach could yield mean comparisons which are challenging to interpret when the outcome is measured in an ordinal scale, or the outcomes do not have a clear and interpretable unit.

1.2 Measurement scale and effect measure

To define a meaningful effect measure, it is important to consider the measurement scale of the outcomes. The measurement scale can be categorized into four types: nominal, ordinal, interval and ratio, where the latter measurement scales contain more information (Stevens, 1946). Nominal outcomes are mutually exclusive categories that cannot be compared numerically, such as region, gender or blood type. Ordinal outcomes allow the determination of greater, equal or lesser for any two outcomes, but the equality of intervals does not hold. For example, in the stage of cancer, the later stage implies severer cancer, but a progression from stage III to stage IV is not the same as a progression from stage I to stage II (intervals are not equal). The well-known Likert scale that each item goes from 1=‘strongly disagree’ to 5=‘strongly agree’ also falls into this category (Likert, 1932). Interval or ratio scales are often dimensional phenomena that can be directly measured, such as body temperature in Celsius and blood pressure in millimetres of mercury (mmHg). Differences in the numeric values in the interval and ratio scale are meaningful and can be compared; however, we cannot compare the differences in the numeric values for ordinal outcomes. Hence, statistical methods for ordinal outcomes can be applied to interval and ratio outcomes, but not vice versa.

The development of methods in this thesis will be focused on ordinal and the methods are also applicable to interval and ratio scales, without relying on the normality assumption. Ordinal outcomes are commonly used in medical research when the phenomenon cannot be directly measured, such as the intensity of pain, depression and lifestyle changes which often involve self-reported questionnaires. With more and more rating scales being developed in different specialties, ordinal scales are often used in medical studies (Svensson, 2000). For example, the visual analogue scale measures the participants' pain intensity on a 10-point scale (continuous), with 0 being no pain and 10 being the worst pain (Huskisson, 1974). Another example is the modified Rankin Scale widely used in stroke trials (Broderick *et al.*, 2017). It is a 7-point scale representing 'no symptom at all, 'no significant disability despite symptoms', 'slight disability', 'moderate disability', 'moderately severe disability', 'severe disability' and 'death'.

Ordinal outcomes do not have meaningful units but can be comparable, where greater, lesser or equal can be determined between two outcomes. The numbers assigned to the ordinal outcomes are arbitrary, leading to different mean differences when different numbers are assigned. Stevens (1946) suggests that outcomes should be analyzed with statistics invariant under the measurement scale, in other words, the analysis method should yield the same result regardless of the numbers assigned to the ordinal outcomes. A natural way to analyze ordinal outcomes is to use only the rank information of the outcomes, which guarantees invariance after rescaling.

In practice, ordinal outcomes are often tested by a two-sample *t*-test or Mann-Whitney (MW) test representing parametric and nonparametric approaches, respectively (Forrest and Andersen, 1986). The *t*-test tests whether the two intervention arms have the same mean, whereas the MW tests whether the outcomes from the two intervention arms come from the same distribution. Parametric methods such as a *t*-test could increase the chance of erroneous conclusions when applied to the data on ordinal scale (Jamieson, 2004). Additionally, the assignment of different numeric to ordinal outcomes can result in different magnitudes of mean differences, complicating the interpretation of the treatment effect.

Another problem using parametric methods on ordinal outcomes is that averaging ordinal outcomes is not sensible when the interval between each unit could have a different meaning (Kuzon *et al.*, 1996). For example, averaging 0=‘no symptom at all’ and 6=‘death’ in the modified Rankin Scale results in 3=‘moderate disability’. It is hard to argue that going from no symptoms to moderate disability is the same as going from moderate disability to death. Another difficulty arises in interpreting when a non-integer value appears in mean difference for ordinal outcomes. For example, it is hard to explain what a 0.7-point improvement means for a patient. For these two reasons nonparametric methods are more favourable for ordinal outcomes.

The consolidated standards of reporting trials (CONSORT) have encouraged researchers to not only report hypothesis testing but also report the effect size with a confidence interval (Moher *et al.*, 2010). For ordinal outcomes, the probability of a treated outcome is more favourable than a randomly untreated outcome can be more easily understood by most people compared to the mean difference (Moses *et al.*, 1984). This probability has various names in the medical research literature, such as the area under the receiver operating characteristic curve (AUC) literature (Bamber, 1975), relative effect (Brunner and Munzel, 2000), concordance statistic (Harrell *et al.*, 1996; Zou *et al.*, 2022), the probability of a better outcome (Colditz *et al.*, 1988) and the win probability (Hayter, 2012; Zou *et al.*, 2022, 2023). We will refer to this effect measure as the win probability (WinP) throughout this thesis because it better conveys the core idea of this measure. The WinP came from the estimand of the MW test and was recognized long ago to be a suitable effect size measure for ordinal outcomes (Moses *et al.*, 1984). The journal *Statistics in Medicine* in 2006 devoted a special issue to confidence intervals for the WinP with independent outcomes (D’Agostino *et al.*, 2006).

1.3 Cluster randomization trials

A cluster randomization trial, also referred to as a group randomization trial, randomizes groups of individuals into different intervention arms, where individuals in the assigned groups receive the same intervention (Donner *et al.*, 1981). Clusters can be formed from various types of intact units, such as communities, patients registered under the same general practitioners, or groups formed temporarily (Donner and Klar, 2000).

Randomizing clusters is done for practical reasons such as increasing intervention compliance, reducing the risk of contamination, and avoiding logistic inconveniences of the trial (Donner and Klar, 2000). Such design has become the standard for evaluating educational and healthcare programs. It is also popular for vaccine efficacy trials to capture population-level effects (Hayes and Moulton, 2017).

A cluster randomization trial is less efficient than an individual randomization trial with the same number of individuals because the outcomes within the same cluster are more likely to be similar to each other than the outcomes from different clusters (Donner and Klar, 2000). Such correlation could be due to the similarities of characteristics within clusters, such as socioeconomic status, disease severity or other demographics because individuals within the same cluster could interact with each other. For example, if households were recruited in a trial, the members would likely influence each others' outcomes. Due to the correlated nature of individuals within clusters, statistical methods for cluster randomization trials must account for within-cluster correlation (Donner *et al.*, 1981).

Designing and analyzing cluster randomization trials have largely been based on parametric methods for continuous or binary outcomes (Donner and Klar, 2000; Hayes and Moulton, 2017). However, analyzing ordinal outcomes in cluster randomization trials based on parametric methods that aggregate the outcomes as cluster-specific means and compare them could be questionable because comparing cluster-specific means is meaningless with ordinal outcomes.

Zou (2021) proposed to use nonparametric methods to quantify the treatment effect in cluster randomization trials using WinP, where WinP was referred to as the MW probability. The method can be summarized in two steps. First, outcomes are transformed into the win fractions, obtained by subtracting the rank of such outcome in its own arm from its rank among the whole sample and divided by the sample size of the comparison arm. Second, the WinP and its variance are estimated with mean win fractions and variance of win fractions, respectively. This method is compatible with any type of outcome, including continuous, ordered category, binary and count outcomes, as they can all be transformed into win fractions from rankings.

1.4 Baseline adjustment for cluster randomized trials

It is common for participants to be assessed before and after receiving the intervention in cluster randomization trials, resulting in a pretest-posttest design. There are three approaches for analyzing the outcomes of such trials. The first focuses on analyzing the posttest outcome, ignoring the pretest outcome. The second approach analyzes the difference between pretest and posttest outcomes, resulting in a change from baseline analysis. The third approach analyzes posttest outcomes, treating pretest outcomes as a covariate, commonly known as the analysis of covariance (ANCOVA). Although all three approaches provide unbiased estimates of the treatment effect in randomized controlled trials, the ANCOVA approach provides the highest power because some of the posttest variances are explained by the pretest, hence reducing the residual variance (Van Breukelen, 2006). Another appealing property of ANCOVA in a randomized study is that it provides a consistent estimate of the treatment effect even if the effect between the pretest and postest is non-linear or the interaction term is misspecified (Yang and Tsiatis, 2001), and the variance estimate is also robust to misspecification (Wang *et al.*, 2019) if randomization ratio is 1:1, or if the sandwich estimator is used for variance estimation (Bartlett, 2020).

Accounting for baseline is more appealing and important for cluster randomization trials than individual randomization trials for a few reasons. Accidental imbalance is more likely to occur in cluster randomization trials when a small number of clusters are randomized. Additionally, cluster randomization trials require more participants to maintain the same power as individual randomization trials. Thus, increasing power through baseline adjustment could reduce the cost of the trial.

For continous outcomes, the ANCOVA approach can be implemented with the mixed model approach (Laird and Ware, 1982) or the generalized estimating equation (GEE) method (Liang and Zeger, 1986) to account for the correlated outcomes in cluster randomization trials. The mixed model includes a fixed effect from the treatment and a random effect from clustering to account for correlation within clusters.

The GEE method treats the clustering effect as a nuisance parameter and focuses on estimating the marginal effect of the treatment on the response. The method requires specifying a covariance structure of the marginal means for estimation. However, the treatment effect can be consistently estimated even if the covariance structure is misspecified. The treatment effect estimated from the GEE is the population-averaged effect regardless of which cluster the individuals belong to. On the other hand, the treatment effect from the mixed model is the conditional effect of the individuals belonging to the same cluster. When the treatment effect is expressed as the mean difference on the raw scale, the population-averaged effect and the conditioned effect are the same (Hubbard *et al.*, 2010). However, for binary and ordered category outcomes, this property does not hold when the effect measure requires a non-linear link function between the outcome and covariates.

There is currently little research on nonparametric methods for baseline adjustment in cluster randomization trials. Based on the methods by Zou (2021), we propose two baseline adjustment approaches for the WinP. The first one extends the weighted least square method proposed by Koch *et al.* (1998) to cluster randomization trials with (co)variance estimators incorporating the clustering effect (Zou, 2021). The WinP is adjusted by the magnitude of the baseline imbalance and the strength of the correlation between baseline and follow-up in the weighted least square method. It is similar to ANCOVA approach but uses a weighted least square estimator instead, resulting in a different correlation estimate between baseline and follow-up. The second one extends the mixed model ANOVA for estimating WinP in Zou (2021) by including the baseline win fractions as a covariate. Although it is known that the weighted least square method is asymptotically equivalent to ANCOVA for independent outcomes (Lesaffre and Senn, 2003), such a result does not hold for correlated outcomes, which will be discussed in Chapter 3.

1.5 Sample size for Mann-Whitney test

There are a large number of sample size formulas for the MW test. The most simple formula involves only the variance of the MW test under the null hypothesis and assumes the data has no ties (Noether, 1987). The variance under the null is a function of the sizes of both arms. Therefore, the formula by Noether (1987) is convenient to use but at the cost of imprecision when the true effect is far from the null because the variance under the alternative depends on the distribution function of the outcomes and could be far from the variance under the null. Additionally, omitting ties can result in increasing variance for the treatment effect hence increasing the required sample size. The remedy for this problem is to incorporate the variance of the MW test under both the null and alternative hypotheses while accounting for ties. Most work in the literature assumes the outcome follows certain distributions to derive the variance of the MW test under the alternative hypothesis and derive sample size formulas based on it (Rahardja *et al.*, 2009; Happ *et al.*, 2019).

Another sample size estimation approach uses proportional odds assumption for the ordinal data (Whitehead, 1993). The proportional odds assumption implies that if we collapse the $2\times K$ frequency table of the outcomes (assuming K categories for the outcomes) into a 2×2 table by selecting a category as the cutoff, the odds ratios are the same regardless of which category is selected. The proportional odds assumption is usually hard to check beforehand, and it could deteriorate the precision of the sample size formula when such an assumption is violated.

The CONSORT, the standard guideline for reporting clinical trials, emphasizes the limit of reporting a single p-value and encourages reporting confidence intervals. The sample size formulas covered by Rahardja *et al.* (2009) and Happ *et al.* (2019) are based on hypothesis testing, which focuses on detecting the treatment effect away from the null hypothesis. Those sample size formulas could be less useful for trials focusing on estimating the treatment effect as the null hypothesis is not of interest.

For individually randomized trials, Zou *et al.* (2022, 2023) developed sample size formulas focused on the estimation of WinP. They use win fraction from pilot data to obtain the variance components for sample size estimation. One can also generate hypothetical pilot data based on expert knowledge of the treatment effect on the outcome distribution, the win fractions can then be derived and their variance can be used for sample size estimation. Since their method uses ranks to calculate the win fractions, it can be easily applicable to continuous, binary and ordered category outcomes, as long as they can be ranked. To our knowledge, there is no sample size formula for estimating WinP in cluster randomization trials.

1.6 Objectives and organization of the thesis

This thesis aims to provide statistical methods for analyzing and determining sample size for cluster randomization trials with baseline measurements. The specific objectives are:

1. Propose the WinP estimator adjusted for baseline measurements in cluster randomization trials using the weighted least square approach and mixed model approach.

2. Derive the variance estimators for the adjusted WinP estimates and their asymptotic properties. Compare the efficiency of the weighted least square approach to the mixed model approach.

3. Derive sample size formulas focusing on confidence interval estimation for cluster randomization trials with baseline adjustments. Discuss important design considerations, such as the properties of the correlation coefficient associated with WinP.

4. Evaluate the finite sample performance of our proposed methods to estimate the WinP in cluster randomization trials using simulation studies focusing on the coverage rate and confidence interval width.

5. Evaluate the performance of sample size formulas using simulation studies focusing on the assurance probability, that is the probability of the lower limit of WinP exceeds a prespecified precision.

The thesis is organized as follows. We review the literature on WinP in Chapter 2, followed by proposed methods for confidence interval estimation of WinP in Chapter 3 and sample size estimation in Chapter 4. The performance of the proposed methods is evaluated through simulation in Chapter 5. Chapter 6 illustrates the methods using data from two cluster randomization trials. Finally, we summarize the thesis, discuss our findings, and suggest future research directions in Chapter 7.

Chapter 2

Win probability in randomized controlled trials

The purpose of this chapter is to review relevant or essential methods for the win probability (WinP) in randomized controlled trials. We will provide a formal definition of the WinP and review its relation to other effect measures in clinical trials. We will first review methods of WinP for an individual randomization trial in Section 2.1, which includes baseline adjustment methods and sample size estimation, followed by methods of WinP for cluster randomization trials in Section 2.2.

2.1 Definition of the win probability

Consider a two-arm individual randomization controlled trial. Suppose n_i subjects are allocated to arm i, with $i = 1$ for control and $i = 2$ for treatment. The size of the trial is therefore $N = n_1 + n_2$. Let Y_{ij} denote the outcome of the jth subject, $j = 1, 2, \cdots, n_i$, in the ith arm. Denote the left-continuous distribution function as $F^-(x) = \mathrm{P}(X < x)$ and right-continuous as $F^+(x) = \mathrm{P}(X \leq x)$, we define the distribution function as the average of right- and left-continuous distribution functions

$$F(x) = 0.5[F^{-}(x) + F^{+}(x)] = P(X < x) + 0.5P(X = x) \tag{2.1}$$

to handle data with ties (Akritas *et al.*, 1997). Denote the outcomes from arm i follows Y_i, the win probability (WinP) is defined as,

$$\text{WinP} = \int F_1(x)dF_2(x) = \Pr(Y_2 > Y_1) + 0.5\Pr(Y_2 = Y_1)\,, \tag{2.2}$$

which is the probability that a randomly chosen outcome in the treatment arm wins (or is better than) a randomly chosen outcome in the control arm plus half the probability of a tie. This definition includes ties making it more useful for ordinal outcomes where ties occur naturally. The assigned weight of 0.5 for tied outcomes is a result of randomly breaking the ties without favouring any arm (Putter, 1955). When there is no treatment effect, the distributions of both intervention arms are the same, hence WinP=0.5. A higher WinP indicates a stronger positive treatment effect, with a value of one implying that all participants in the treatment arm have a better outcome than the participants in the control arm.

There are various names for WinP in the literature due to its usefulness in comparing two groups in medical studies. Zou *et al.* (2023) identified at least 12 terms for WinP including the area under the receiver operating characteristic curve in diagnostic literature (Bamber, 1975), the concordance statistic in evaluating the performance of prediction models (Harrell *et al.*, 1996; Zou *et al.*, 2022), the relative treatment effect (Brunner and Munzel, 2000) and probabilistic index (Thas *et al.*, 2012). Since the WinP is an underlying measure of the Mann-Whitney (MW) test, it is also referred to as the MW probability (Newcombe, 2006; Zou, 2021). We prefer the term WinP because it does not require a statistics background to understand winning and it better conveys the benefit of receiving the treatment.

The WinP is related to many other effect measures. For binary outcomes, the WinP is related to risk difference Δ by

$$\text{WinP} = (\Delta + 1)/2\,.$$

For ordered categories with p_i as the row vector of proportions for each category for arm i,

$$\text{WinP} = p_1 \Omega p_2'\,,$$

where p_2' denotes the transpose of p_2 and Ω is an upper triangle matrix of ones but half on the diagonal (Zou, 2021). For continuous outcomes from normal distributions $Y_i \sim N(\mu_i, \sigma_i^2)$, WinP can also be written as

$$\begin{aligned}
\text{WinP} &= \Pr(Y_1 < Y_2) = \Pr(Y_1 - Y_2 < 0) \\
&= \Pr\left[\frac{Y_1 - Y_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}} < \frac{-(\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right] \\
&= \Pr\left[Z < \frac{\mu_2 - \mu_1)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right] \\
&= \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right),
\end{aligned}$$

where Φ is the standard normal cumulative distribution function. The special case of equal variance ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) yields WinP $= \Phi(\text{ES}/\sqrt{2})$ where ES $= (\mu_2 - \mu_1)/\sigma$ is the Cohen's effect size (Cohen, 1988). As a reference benchmark for the Cohen's effect size, values 0.2, 0.5 and 0.8 are considered as small, medium and large effect size, which corresponds to 0.56, 0.64 and 0.71, respectively, for WinP.

Another useful treatment effect measure is the number needed to treat (NNT), which is the expected number of patients needed to be treated until there is one more patient with a better outcome. It is the reciprocal of risk difference for binary outcomes (Laupacis *et al.*, 1988).

Therefore, NNT can also be obtained from WinP for binary outcomes by

$$\text{NNT} = \frac{1}{2\text{WinP} - 1}.$$

For continuous and ordered categories outcomes, Zimmermann and Rahlfs (2014) showed that

$$\text{NNT} = \frac{1}{\text{WinP} - 0.5},$$

by calculating areas of risk differences on the percentile-percentile plot comparing the distributions between treatment and control arms. The NNT is interpreted as the number needed to treat to see an improvement of one category on average for ordered category outcomes. The linear relationship between WinP and other treatment effect measures suggests that WinP can be easily transformed for other effect size measures.

2.1.1 Estimators of the win probability and variance

Most of the literature relates WinP estimation to the U-statistic theory where WinP can be estimated by

$$\widehat{\text{WinP}} = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} H_{ij}}{n_1 n_2}. \tag{2.3}$$

where $H_{ij} = H(Y_{1i}, Y_{2j}) = I(Y_{1i} < Y_{2j}) + 0.5I(Y_{1i} = Y_{2j})$ is the Heaviside function yielding values of 1, 0, 0.5 for a win if $Y_{2j} > Y_{1i}$, a loss if $Y_{2j} < Y_{1i}$ and ties if $Y_{2j} = Y_{1i}$, respectively. The Heaviside function is referred to as the kernel in U-statistic theory (DeLong *et al.*, 1988), where asymptotic properties are derived with the kernel as the basic unit (Lee, 1990). The values of the kernel are correlated to each other by definition since they share the same indices. The variance formula is given by (Bamber, 1975)

$$\text{Var}(\widehat{\text{WinP}}) = \left[\text{WinP}(1 - \text{WinP}) + (n_1 - 1)Q_1 + (n_2 - 1)Q_2\right]/n_1 n_2,$$

where $Q_1 = \mathrm{Cor}(H_{ij}, H_{kj})$ and $Q_2 = \mathrm{Cor}(H_{ij}, H_{ik})$ are the correlations between two Heaviside functions comparing one outcome from one arm to two outcomes from the other arm. We can estimate Q_i empirically by

$$\begin{aligned}\widehat{Q}_1 &= \frac{\sum_{i=1}^{n_1}\sum_{k\neq i}^{n_1}\sum_{j=1}^{n_2} H_{ij}H_{kj}}{n_1 n_2(n_1-1)} - \widehat{\mathrm{WinP}}^2\\ \widehat{Q}_2 &= \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k\neq j}^{n_2} H_{ij}H_{ik}}{n_1 n_2(n_2-1)} - \widehat{\mathrm{WinP}}^2,\end{aligned}$$

where the calculation can be demanding because of the triple summation over the Heaviside functions.

A common approach to reducing the calculation complexity is to assume parametric assumptions of the outcomes, where Q_1 and Q_2 can be expressed as a function of WinP. Hanley and McNeil (1982) used the negative exponential distribution for Y_1 and Y_2 to yield $Q_1 = 2\mathrm{WinP}^2/(1+\mathrm{WinP}) - \mathrm{WinP}^2$ and $Q_2 = \mathrm{WinP}/(2-\mathrm{WinP}) - \mathrm{WinP}^2$, which provide the most conservative variance estimate among the exponential family. Newcombe (2006) used the beta distribution for $Y_1 \sim \mathrm{Beta}(1, \gamma+1)$ and $Y_2 \sim \mathrm{Beta}(\gamma, 1)$, where Γ is the gamma function and γ is solved in $\widehat{\mathrm{WinP}} = 1 - \Gamma^2(\gamma+1)/\Gamma(2\gamma+1)$, where $\widehat{\mathrm{WinP}}$ is estimated by equation (2.3). The variance of $\widehat{\mathrm{WinP}}$ in this model is

$$\begin{aligned}\mathrm{Var}(\widehat{\mathrm{WinP}}) = &\frac{(n_1+n_2-2)[1-2\Gamma^2(\gamma+1)/\Gamma(2\gamma+1)+\Gamma(2\gamma+1)\Gamma(\gamma+1)/\Gamma(3\gamma+1)]}{n_1 n_2}\\ &+\frac{\mathrm{WinP}(1-\mathrm{WinP})}{n_1 n_2}.\end{aligned}$$

These variance formulas can be used to construct Wald-type confidence interval by substituting WinP with $\widehat{\mathrm{WinP}}$ in the variance formulas for $\widehat{\mathrm{WinP}} \mp z_{\alpha/2}\sqrt{\mathrm{Var}(\widehat{\mathrm{WinP}})}$, or Wilson score confidence interval by solving $|\widehat{\mathrm{WinP}} - \mathrm{WinP}|/\sqrt{\mathrm{Var}(\widehat{\mathrm{WinP}})} \leq z_{\alpha/2}$ for WinP, where z_x is the upper quantile of a standard normal distribution.

Newcombe (2006) conducted simulation studies to compare Wald-type interval to Wilson score interval with the negative exponential variance estimator. His simulation study showed

Table 2.1: Relation between the win fractions and the Heaviside function for estimating WinP.

Control Treatment	Y_{11}	Y_{12}	$\cdots$	Y_{1n_1}	Win fraction
Y_{21}	H_{11}	H_{12}	$\cdots$	H_{1n_1}	$w_{21} = \overline{H}_{1.}$
Y_{22}	H_{21}	H_{22}	$\cdots$	H_{2n_1}	$w_{22} = \overline{H}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Y_{2n_2}	H_{n_21}	H_{n_22}	$\cdots$	$H_{n_2n_1}$	$w_{2n_2} = \overline{H}_{n_2.}$
Win fraction	$w_{11} = 1 - \overline{H}_{.1}$	$w_{12} = 1 - \overline{H}_{.2}$	$\cdots$	$w_{1n_1} = 1 - \overline{H}_{.n_1}$	$\widehat{\text{WinP}} = \overline{H}_{..}$

Note: $H_{ij} = I(Y_{1i} < Y_{2j}) + 0.5I(Y_{1i} = Y_{2j})$ and $\widehat{\text{WinP}} = \overline{H}_{..} = \overline{w}_{2.} = 1 - \overline{w}_{1.}$.

that Wald-type interval yields under-coverage and wider intervals, whereas Wilson score interval with the negative exponential variance estimator and the average size of two arms by $n_1^* = n_2^* = (n_1 + n_2)/2 - 1$ (use n_1^* and n_2^* instead of n_1 and n_2 in variance formula) maintains the coverage rate for most combinations of WinP and the distribution of the outcomes. However, the nuisance parameters Q_1 and Q_2 could be poorly estimated when the parametric assumptions are violated (Perme and Manevski, 2019). It is unreliable to estimate the variance without empirically estimating Q_1 and Q_2 as the correlation between the kernels H_{ij} could vary by the choice of outcome distributions.

To reduce the calculation complexity without parametric assumption, one can first transform the outcome to 'win fraction' and then estimate the variance of $\widehat{\text{WinP}}$ with win fractions. The win fraction, denoted by w_{ij}, of an outcome is the proportion of 'wins' it achieves compared to all the outcomes in the other arm. The win fraction can be calculated from the Heaviside function, and the WinP can be estimated by averaging the win fractions. We have demonstrated the calculation of the win fractions and the WinP in Table 2.1. Note that the win fraction is related to the distribution functions that the win fraction of subject j in the treatment arm is $w_{2j} = P(Y_1 < Y_{2j}) + 0.5P(Y_1 = Y_{2j}) = \widehat{F}_1(Y_{2j})$ and can be viewed as the probability for subject j in the treatment arm winning a randomly selected subject in the control arm (Zou, 2021). Similarly, $w_{1j} = \widehat{F}_2(Y_{1j})$, as the probability for subject j in the control arm winning a randomly selected subject in the treatment arm. The calculation of the win fraction can then be simplified using ranks by utilizing the relationship between the win fraction and the empirical distribution function. The rank of Y_{ij} in arm i, denoted by r_{ij}, is related to the empirical distribution function F_i by

$$\begin{aligned} r_{ij} &= \sum_{k=1}^{n_i} I(Y_{ik} < Y_{ij}) + 0.5 \sum_{k=1}^{n_i} I(Y_{ik} = Y_{ij}) + 0.5 \\ &= n_i \frac{\sum_{k=1}^{n_i} \left[I(Y_{ik} < Y_{ij}) + 0.5 I(Y_{ik} = Y_{ij}) \right]}{n_i} + 0.5 \\ &= n_i \widehat{F}_i(Y_{ij}) + 0.5\,. \end{aligned}$$

The first equality follows by the counting process occurred in ranking Y_{ij} in arm i. We first count how many outcomes are less than our given outcome as $C_1 = \sum_{k=1}^{n_i} I(Y_{ik} < Y_{ij})$, and count the number of tied outcomes as $C_2 = \sum_{k=1}^{n_i} I(Y_{ik} = Y_{ij})$. The rank of Y_{ij} is then $C_1 + C_2/2 + 0.5$. The second equality then follows naturally by the empirical estimator of the distribution function

$$\widehat{F}_i(Y_{ij}) = \frac{\sum_{k=1}^{n_i} \left[I(Y_{ik} < Y_{ij}) + 0.5 I(Y_{il} = Y_{ij}) \right]}{n_i}\,.$$

Similarly, the rank of Y_{ij} in the whole sample is related to the combined distribution by $R_{ij} = N\widehat{F}(Y_{ij}) + \frac{1}{2}$, where $\widehat{F}(x)$ is the empirically estimated distribution function from combining both arms, $N\widehat{F}(x) = n_1\widehat{F}_1(x) + n_2\widehat{F}_2(x)$. We can estimate the win fraction w_{ij} from R_{ij} and r_{ij} by

$$\begin{aligned} w_{2j} &= \widehat{F}_1(Y_{2j}) = \frac{N\widehat{F}(Y_{2j}) - n_2\widehat{F}_2(Y_{2j})}{n_1} = \frac{R_{2j} - r_{2j}}{n_1} \\ w_{1j} &= \widehat{F}_2(Y_{1j}) = \frac{N\widehat{F}(Y_{1j}) - n_1\widehat{F}_1(Y_{1j})}{n_2} = \frac{R_{1j} - r_{1j}}{n_2}\,. \end{aligned}$$

The quantities w_{2j} and $1 - w_{1j}$ are also referred to as the placement value(Hanley and Hajian-Tilaki, 1997), structural components of U-statistics and ‘ridits’ (relative to an identified distribution) as pointed out by Zou *et al.* (2023).

To derive the variance of $\widehat{\text{WinP}}$, the $\widehat{\text{WinP}}$ is decomposed by Zou (2021) as

$$\begin{aligned}
\widehat{\text{WinP}} &= \int \widehat{F}_1(x)d\widehat{F}_2(x) \\
&= \int \widehat{F}_1(x) - F_1(x) + F_1(x)d\left(\widehat{F}_2(x) - F_2(x) + F_2(x)\right) \\
&= \int \widehat{F}_1(x) - F_1(x) + F_1(x)d\left(\widehat{F}_2(x) - F_2(x)\right) + \int \widehat{F}_1(x) - F_1(x) + F_1(x)dF_2(x) \\
&= \int \widehat{F}_1(x) - F_1(x)d\left(\widehat{F}_2(x) - F_2(x)\right) + \int F_1(x)d\widehat{F}_2(x) - \int F_1(x)dF_2(x) \\
&\quad + \int \widehat{F}_1(x)dF_2(x) \\
&= \int F_1(x)d\widehat{F}_2(x) + \int \widehat{F}_1(x)dF_2(x) - \text{WinP} + \int \widehat{F}_1(x) - F_1(x)d\left(\widehat{F}_2(x) - F_2(x)\right) \\
&= \int F_1(x)d\widehat{F}_2(x) + 1 - \int F_2(x)d\widehat{F}_1(x) - \text{WinP} + \int \widehat{F}_1(x) - F_1(x)d\left(\widehat{F}_2(x) - F_2(x)\right) \\
&\approx 1 - \text{WinP} + \int F_1(x)d\widehat{F}_2(x) - \int F_2(x)d\widehat{F}_1(x)\,, \qquad (2.4)
\end{aligned}$$

because $\int \widehat{F}_1(x)dF_2(x) = 1 - \int F_2(x)d\widehat{F}_1(x)$ by integration by parts and $\int \widehat{F}_1(x) - F_1(x)d\left(\widehat{F}_2(x) - F_2(x)\right)$ converges to zero in probability since empirical distributions are consistent estimators. The equation (2.4) implies the variance of $\widehat{\text{WinP}}$ is obtained by

$$\begin{aligned}
\text{Var}(\widehat{\text{WinP}}) &= \text{Var}\left[1 - \text{WinP} + \int F_1(x)d\widehat{F}_2(x) + \int F_2(x)d\widehat{F}_1(x)\right] \\
&= \text{Var}\left[\int F_1(x)d\widehat{F}_2(x)\right] + \text{Var}\left[\int F_2(x)d\widehat{F}_1(x)\right] \\
&= \text{Var}\left[\frac{\sum_{j=1}^{n_1} \widehat{F}_2(Y_{1j})}{n_1}\right] + \text{Var}\left[\frac{\sum_{j=1}^{n_2} \widehat{F}_1(Y_{2j})}{n_2}\right] \\
&= \text{Var}(\overline{w}_{1.}) + \text{Var}(\overline{w}_{2.})\,, \qquad (2.5)
\end{aligned}$$

where $\text{Var}(\overline{w}_{i.})$ can be estimated consistently by the sample variance of win fractions. The win fractions are asymptotically independent as long as $n_1 \to \infty$ and $n_2 \to \infty$ and w_{1j} and w_{2j} converge to $F_2(Y_{1j})$ and $F_1(Y_{2j})$, respectively. A rigorous proof of the asymptotic independence is provided by Sen (1967). Zou *et al.* (2023) also showed that the covariance between two win fractions within the same arm is bounded by the inverse sample size of the other arm.

2.1.2 Baseline adjustment for win probability

In randomized controlled trials, the outcomes are often measured at baseline and the follow-up after interventions. Adjustment of the baseline outcome is recommended in the CONSORT statement for clinical trials (Moher *et al.*, 2010). The main purpose of baseline adjustment in randomized controlled trials is to improve efficiency by variance reduction. The stronger the correlation between baseline and follow-up, the more efficiency is gained. In addition, baseline adjustment reduces accidental bias caused by the imbalance of baseline measurements that occurred by chance.

Common strategies of baseline adjustment include analysis of covariance (ANCOVA), which treats the baseline outcome a covariate, t-test on change from baseline to follow-up, and repeated-measure analysis of variance (ANOVA) with constraints on the baseline. The ANCOVA is generally considered as the most powerful method for randomized studies in the literature (Vickers and Altman, 2001). However, direct application of ANCOVA to ordinal outcomes could yield mean comparisons that is difficult to interpret.

A nonparametric way of adjusting the baseline for the WinP is the weighted least square method by Koch *et al.* (1998), which is essentially a regression of the arm-specific mean. Denote the baseline assessment by X_{ij} and the baseline imbalance as $\hat{\delta} = \overline{X}_{2.} - \overline{X}_{1.}$. The weighted least square method regresses $Y = (\widehat{\text{WinP}}, \hat{\delta})'$ to $X = (1, 0)'$ by the following model

$$Y = \beta X\,, \tag{2.6}$$

where β is the adjusted WinP after constraining $\delta = 0$. Since $\widehat{\text{WinP}}$ and $\hat{\delta}$ are correlated and have different variances, it is reasonable to weight them based on the inverse of their covariance matrix given by

$$
\begin{aligned}
\Sigma^{-1} &= \begin{pmatrix} \widehat{\text{Var}}(\widehat{\text{WinP}}) & \widehat{\text{Cov}}(\widehat{\text{WinP}}, \hat{\delta}) \\ \widehat{\text{Cov}}(\widehat{\text{WinP}}, \hat{\delta}) & \widehat{\text{Var}}(\hat{\delta}) \end{pmatrix}^{-1} \\
&= \frac{1}{\det(\Sigma)} \begin{pmatrix} \widehat{\text{Var}}(\hat{\delta}) & -\widehat{\text{Cov}}(\widehat{\text{WinP}}, \hat{\delta}) \\ -\widehat{\text{Cov}}(\widehat{\text{WinP}}, \hat{\delta}) & \widehat{\text{Var}}(\widehat{\text{WinP}}) \end{pmatrix},
\end{aligned}
$$

where $\det(\Sigma) = \widehat{\text{Var}}(\widehat{\text{WinP}})\widehat{\text{Var}}(\hat{\delta}) - \widehat{\text{Cov}}^{2}(\widehat{\text{WinP}}, \hat{\delta})$ is the determinant of the covariance matrix of Y. The weighted least square estimator of adjusted WinP, denoted as $\widehat{\text{WinP}}^{*}$, is hence obtained as

$$
\begin{aligned}
\widehat{\text{WinP}}^{*} &= \hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y \\
&= \widehat{\text{WinP}} - \frac{\widehat{\text{Cov}}(\widehat{\text{WinP}}, \hat{\delta})}{\widehat{\text{Var}}(\hat{\delta})}\hat{\delta} \qquad (2.7)
\end{aligned}
$$

and the variance for this estimator can be estimated with

$$
\begin{aligned}
\widehat{\text{Var}}(\widehat{\text{WinP}}^{*}) &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\widehat{\text{Var}}(Y)(\Sigma')^{-1}X(X'\Sigma^{-1}X)^{-1} \\
&= \text{Var}(\widehat{\text{WinP}}) - \frac{\text{Cov}^{2}(\widehat{\text{WinP}}, \hat{\delta})}{\text{Var}(\hat{\delta})}, \qquad (2.8)
\end{aligned}
$$

where the covariance between $\widehat{\text{WinP}}$ and $\hat{\delta}$ can be estimated by decomposing it into the summation of sample covariances between win fractions and baseline measurements for both arms; detailed proof is provided in Chapter 3.

Alternatively, the weighted least square method for adjusting WinP can be explained by the conditional distribution of $\widehat{\text{WinP}}$. Consider $\widehat{\text{WinP}}$ and $\hat{\delta}$ follow a bivariate normal with mean $u = (\text{WinP}, 0)'$ and the covariance matrix Σ. The conditonal distribution of $\widehat{\text{WinP}}$ given $\hat{\delta}$

follows a normal distribution with a mean of

$$u_c = \text{WinP} + \frac{\text{Cov}(\widehat{\text{WinP}}, \hat{\delta})}{\text{Var}(\hat{\delta})}\hat{\delta},$$

and variance of $\text{Var}(\widehat{\text{WinP}}) - \text{Cov}^2(\widehat{\text{WinP}}, \hat{\delta})/\text{Var}(\hat{\delta})$. This implies the unadjusted estimator of WinP is biased by the difference between marginal and conditional mean for WinP. An unbiased adjusted estimator for WinP can then be obtained by

$$\widehat{\text{WinP}}^* = \widehat{\text{WinP}} - \frac{\widehat{\text{Cov}}(\widehat{\text{WinP}}, \hat{\delta})}{\widehat{\text{Var}}(\hat{\delta})}\hat{\delta}$$

and the variance of this estimator can be estimated by

$$\widehat{\text{Var}}(\widehat{\text{WinP}}^*) = \widehat{\text{Var}}(\widehat{\text{WinP}}) - \frac{\left[\widehat{\text{Cov}}(\widehat{\text{WinP}}, \hat{\delta})\right]^2}{\text{Var}(\hat{\delta})}.$$

These are equivalent to weighted least square estimators shown in equations (2.7) and (2.8), respectively.

Another way to quantify the baseline imbalance is to use WinP of baseline instead of the mean difference (Schacht *et al.*, 2008). In such a case, the term $\hat{\delta}$ in (2.7) is substituted by $\widehat{\text{WinP}}_X - 0.5$ where WinP_X is the win probability of the baseline measurement.

Recently, Zou *et al.* (2023) proposed regressing the win fractions of follow-up outcomes, w_{ij}, by the treatment indicator and the win fractions for baseline measurement using the following linear model

$$w_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 w_{ij}^x + \epsilon_{ij},$$

where t_{ij} is the treatment indicator (1 = treatment, 0 = control), w_{ij}^x is the win fraction for baseline assessment and ϵ_{ij} is the error term. Since $\overline{w}_{2.} + \overline{w}_{1.} = 1$ and $\overline{w}_{2.} = \widehat{\text{WinP}}$, the mean difference in win fractions is related to WinP by $\overline{w}_{2.} - \overline{w}_{1.} = 2\widehat{\text{WinP}} - 1$. It then follows that the adjusted WinP is estimated by $\widehat{\text{WinP}}^* = \hat{\beta}_1/2 + 0.5$ because $\hat{\beta}_1$ is the adjusted mean difference of win fractions between the treatment and control arms for follow-up.

Another regression framework for adjusting WinP is the probabilistic index model (PIM) by Thas *et al.* (2012). This model uses the pairwise comparison H_{ij} as the dependent variable and the treatment indicator and baseline assessment as independent variables in a regression model. The PIM approach is much more computationally intensive because they use $n_1 n_0$ pairwise comparisons to regress, whereas the approach proposed by Zou *et al.* (2023) only involves $n_1 + n_0$ win fractions to regress. In addition, the PIM requires a logistic link function to connect WinP and the covariates, which can cause noncollapsibility problem yielding different interpretation between adjusted and unadjusted effects (Robinson and Jewell, 1991). Another drawback is there is no closed-form formula for the variance; bootstrap could be used to estimate the variance but it is unclear how to extend it to cluster randomization trials.

2.1.3 Confidence interval estimation

Once the adjusted WinP and its variance are estimated, several strategies are available to construct confidence interval for WinP. DeLong *et al.* (1988) proposed a large sample 2-sided confidence interval given by

$$(L_1, U_1) = \widehat{\text{WinP}}^* \mp z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\widehat{\text{WinP}}^*)},$$

where z_x is the upper quantile of the standard normal distribution. A small-sample correction can be made by using the t-statistic instead of z-statistic, where the degrees of freedom follows by Satterthwaite's approximation (Brunner and Munzel, 2000),

$$\text{df} = \frac{\left[\widehat{\text{Var}}(\overline{w}_{1.}) + \widehat{\text{Var}}(\overline{w}_{2.})\right]^2}{\widehat{\text{Var}}^2(\overline{w}_{1.})/(n_1 - 1) + \widehat{\text{Var}}^2(\overline{w}_{2.})/(n_2 - 1)}.$$

Another approach is to use df $= N - 2$ because the variances of two arm-specific means are estimated; hence two degrees of freedom are lost.

The range restriction of WinP from 0.5 to 1 suggests that the sampling distribution could be asymmetric. Therefore, symmetric confidence intervals of WinP could suffer from under coverage and imbalanced tail errors (Newcombe, 1998), leading to disagreement between hypothesis testing and confidence intervals of the treatment effect. The logit-transformed intervals that adjust the location can be employed as they yield higher coverage and narrower intervals (Newcombe, 1998), where the empirical coverage is close to nominal coverage, and the tail errors are more balanced. The logit-transformed confidence interval (L_2, U_2) is given by

$$L_2 = \frac{\exp(l_2)}{1+\exp(l_2)}, \qquad U_2 = \frac{\exp(u_2)}{1+\exp(u_2)},$$

where

$$l_2, u_2 = \ln\frac{\widehat{\text{WinP}}^*}{1-\widehat{\text{WinP}}^*} \mp t_{alpha/2,\text{df}}\frac{\sqrt{\widehat{\text{Var}}(\widehat{\text{WinP}}^*)}}{\widehat{\text{WinP}}^*(1-\widehat{\text{WinP}}^*)}.$$

Newcombe (2001) also proposed the arsinh-transformed confidence interval on the logit scale for binomial proportions, which is essentially the Wilson interval on the logit scale. The arsinh transformed confidence interval is given by

$$L_3 = \frac{\exp(l_3)}{1+\exp(l_3)}, \qquad U_3 = \frac{\exp(u_3)}{1+\exp(u_3)},$$

where

$$l_3, u_3 = \ln\frac{\widehat{\text{WinP}}^*}{1-\widehat{\text{WinP}}^*} \mp 2\text{arsinh}\left[t_{\alpha/2,\text{df}}\frac{\sqrt{\widehat{\text{Var}}(\widehat{\text{WinP}}^*)}}{\widehat{\text{WinP}}^*(1-\widehat{\text{WinP}}^*)}\right],$$

and $\text{arsinh}(x) = \ln(x + \sqrt{x^2+1})$ is the arsinh transformation. Newcombe (2001) showed that the asrinh-transformed interval (L_3, U_3) is always contained in the logit-transformed interval

(L_2, U_2), implying the former one has higher efficiency. The degrees of freedom for confidence interval construction of adjusted WinP follow from the Satterthwaite approximation by Schacht *et al.* (2008)

$$\mathrm{df}_s = \frac{(\sum_{i=1}^{2} \tau_i^2)^2}{\sum_{i=1}^{2} \tau_i^2/(n_i - 1)},$$

where

$$\tau_i^2 = \frac{1}{n_i}\left\{\widehat{\mathrm{Var}}(w_{ij}) + \left[\frac{\widehat{\mathrm{Cov}}(\widehat{\mathrm{WinP}}, \hat{\delta})}{\widehat{\mathrm{Var}}(\hat{\delta})}\right]^2 \widehat{\mathrm{Var}}(X_{ij}) - 2\frac{\widehat{\mathrm{Cov}}(\widehat{\mathrm{WinP}}, \hat{\delta})}{\widehat{\mathrm{Var}}(\hat{\delta})}\widehat{\mathrm{Cov}}(w_{ij}, X_{ij})\right\}.$$

In case where $\mathrm{df}_s < 1$, we use one as degrees of freedom instead.

2.2 Sample size for the Mann-Whitney test

Most sample size formulas related to WinP were developed focused on the Mann-Whitney (MW) test, which tests whether two groups are from the same distribution against one is stochastic superior to the other. A distribution is stochastic superior to the other if its cumulative distribution function (CDF) is greater than the other CDF at any point, or graphically one CDF completely falls under the other.

Denote the size of the trial as N and the fraction of participants in the treatment arm as f. The treatment and control arm sizes are therefore Nf and $N(1 - f)$, respectively. Although the sample size formula by Noether (1987) is derived from the z-test of the MW test statistic U, it is the same as sizing of WinP because $\mathrm{WinP} = U/(n_1n_2)$ and n_1 and n_2 are factored out from U to derive the sample size. Hence, the sample size for the MW test at α level with $1 - \beta$ power can be derived from a z-test of WinP = 0.5, which can be written as

$$(z_{\alpha/2}\sigma_1 + z_\beta\sigma_2)^2 = (\mathrm{WinP} - 0.5)^2, \tag{2.9}$$

where z_x is the upper quantile of a standard normal distribution, σ_1^2 and σ_2^2 are the variances of $\widehat{\mathrm{WinP}}$ under the null and alternative hypothesis, respectively. The next section presents the estimation of σ_i^2 and sample size requirements based on different parametric assumptions.

2.2.1 Continuous outcomes

The variance of $\widehat{\text{WinP}}$ under the null for continuous outcome was derived by Mann and Whitney (1947) as $\tilde{\sigma}_1^2 = (N+1)/[12N^2 f(1-f)]$. One can also arrive with a similar variance formula using win fractions under the null hypothesis that the outcomes from both groups follow the same distribution. Under the null hypothesis, the win fractions follow a continuous uniform distribution $U(0,1)$ and the variance of win fractions w_{ij} are $1/12$ for both group, $i = 1, 2$. The variance of $\widehat{\text{WinP}}$ under the null hypothesis can be derived as the following according to equation (2.5)

$$\begin{aligned}\text{Var}(\widehat{\text{WinP}}|\text{H}_0) &= \text{Var}(\overline{w}_{1.}) + \text{Var}(\overline{w}_{2.}) \\ &= \frac{1}{12Nf} + \frac{1}{12N(1-f)} \\ &= \frac{1}{12Nf(1-f)}\,.\end{aligned}$$

Assuming the variance of $\widehat{\text{WinP}}$ under the alternative does not differ much from the variance under the null, σ_1 and σ_2 are both substituted by $\tilde{\sigma}_1$ into equation (2.9) and solving for N yields the sample size formula by Noether (1987):

$$N_1 = \frac{(z_{\alpha/2} + z_\beta)^2}{12f(1-f)(\text{WinP} - 0.5)^2}\,. \tag{2.10}$$

This formula emphasizes maintaining the significance level, so it is only suitable when WinP is close to 0.5. When WinP is far from 0.5, the approximation of σ_2 by σ_1 could be inaccurate, resulting in over/underestimates of sample size (Shieh *et al.*, 2006). Therefore, it is reasonable to shift the focus to the distribution of $\widehat{\text{WinP}}$ under the alternative when the trials focus on effect estimation, as most trials usually will not have WinP close to 0.5. The variance under the alternative was derived by Bamber (1975)

$$\tilde{\sigma}_2^2 = [\text{WinP}(1-\text{WinP}) + (n_1-1)Q_1 + (n_2-1)Q_2]/(n_1n_2),$$

where $Q_1 = P(Y_{11} < Y_{21}, Y_{11} < Y_{22}) - \text{WinP}^2$ and $Q_2 = P(Y_{11} < Y_{21}, Y_{12} < Y_{21}) - \text{WinP}^2$, respectively. This formula requires estimation of Q_1 and Q_2 from pilot studies. The probability $P(Y_{11} < Y_{21}, Y_{11} < Y_{22})$ is estimated empirically by $n_1^{-1}n_2^{-1}(n_2-1)^{-1}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k\neq j}^{n_2} H_{ij}H_{ik}$, and $P(Y_{11} < Y_{21}, Y_{12} < Y_{21})$ is estimated similarly by $n_1^{-1}n_2^{-1}(n_1-1)^{-1}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k\neq i}^{n_1} H_{ij}H_{kj}$. Solving the equation (2.9) for N with $\tilde{\sigma}_2^2$ yields a sample size formula similar to Wang *et al.* (2003):

$$N_2 = \frac{\left[(z_{\alpha/2} + z_\beta)\sqrt{12fQ_1 + 12(1-f)Q_2}\right]^2}{12f(1-f)(\text{WinP}-0.5)^2}. \tag{2.11}$$

The major difference between N_2 and N_1 is that N_2 incorporates the variance under the alternative, which improves in maintaining power (Shieh *et al.*, 2006), making it more appropriate when WinP is not close to 0.5.

When the treatment effect causes a location-shift of the distribution of outcomes, such as a reduction in blood pressure, it is common to test the group means using a two-sample t-test due to its high power in most settings (Fay and Proschan, 2010). However, mean comparison in the t-test requires the outcome to possess interval properties, and the t-test could be less powerful than the MW test when the treatment also affects the shape of the distributions, such as skewness or spread. Some authors considered the location-shift alternative hypothesis of the MW test for sample size planning. Chakraborti *et al.* (2006) proposed using pilot data to construct the empirical distribution functions of both intervention arms and then estimate WinP from the distribution functions. Sample size can then be estimated using either formula (2.10) or (2.11). Rosner and Glynn (2011) showed that under the normal distribution location-shift model, the two quantities in σ_2^2 can be written as $Q_1 = Q_2 = \Phi_2(\Phi^{-1}(\text{WinP}), \Phi^{-1}(\text{WinP}), 0.5) - \text{WinP}^2$, where $\Phi_2(a, b, 0.5) = \text{P}(\text{Z}_1 < a, \text{Z}_2 < b)$ is the cumulative standard bivariate normal distribution function of Z_1 and Z_2 with their correlation coefficient being 0.5 and Φ^{-1}(WinP) is the standardized mean difference under normality assumption. Rosner and Glynn's variance estimator depends solely on WinP, which could be helpful in the planning stage when Q_1 and Q_2 cannot be reliably estimated and the sizes of both groups differ substantially. A disadvantage of this method is that they are only developed for the location-shift model; when treatment also affects the shape of the distribution, the validity of this method is unclear. Additionally, the location-shift assumption may not always be appropriate. For example, suppose a trial aimed at reducing alcohol consumption and an investigator expects the new intervention to cut alcohol consumption by half. In that case, a change in the quantile of the distribution by 50%, i.e $F_2(x/2) = F_1(x)$, fits better than a location-shift model.

2.2.2 Ordered category outcomes

We have described several sample size formulas for outcomes without ties, which are incompatible with ordered category outcomes that have ties naturally. Suppose there are K categories where category c, $c = 1, \cdots, K$, has proportion p_c in the treatment arm and q_c in the control arm. The variance of $\widehat{\text{WinP}}$ under the null can then be estimated by (Emerson and Moses, 1985)

$$\tilde{\sigma}_1^2 = \frac{N+1}{12N^2 f(1-f)} - \frac{1}{12N^3(N-1)f(1-f)} \sum_{c=1}^{K} T_c\,,$$

where $T_c = N^3[fp_c+(1-f)q_c]^3 - N[fp_c+(1-f)q_c]$ is the variance reduction from conditioning on the total count of each tie, hence $\tilde{\sigma}_1^2$ is a conditional estimator. Zhao *et al.* (2008) used a simpler formula to approximate $\tilde{\sigma}_1^2$ by dropping the second term in T_c yielding

$$\tilde{\sigma}_1^2 = \frac{1 - \sum_{c=1}^{K} fp_c + (1-f)q_c)^3}{12Nf(1-f)}$$

Using the simpler variance for both null and alternative hypothesis in (2.9) and solving for N yields:

$$N_3 = \frac{(z_{\alpha/2} + z_\beta)^2 \left[1 - \sum_{c=1}^{K} \{f p_c + (1-f) q_c\}^3\right]}{12 f(1-f)(\text{WinP} - 0.5)^2}. \tag{2.12}$$

Wellek (2017) derived the variance under alternative variance by separately estimating $\Pr(Y_1 < Y_2)$ and $\Pr(Y_1 = Y_2)$ and considering the correlation between $\Pr(Y_1 < Y_2)$ and $\Pr(Y_1 = Y_2)$ in addition to the correlation of the pairwise comparison of outcomes. Although this method provides an unbiased formula, it brings in the complicated correlated nature of U-statistics. An advantage of his formula is that the variance is derived without conditioning on the ties; hence the variance is usually smaller, leading to a smaller sample size.

Lachin (2011) proposed to plan sample size for categorical outcomes based on the Cochran – Mantel – Haenszel (CMH) mean score test (Cochran, 1954; Mantel, 1963; Mantel and Haenszel, 1959). When the rank of each categories in the whole sample is used as the scores, the test statistic is a function of the mean rank difference and the variance of the test statistic is derived by assuming the proportions of each categories follow a multinomial distribution and the scores for each category are assumed fixed with given sample size. Under the alternative hypothesis, the test statistic is a squared normal distribution with unit variance and a mean of

$$N\tau^2 = \frac{\left[\sum_{c=1}^{K} v_c (p_c - q_c)\right]^2}{[f^{-1} + (1-f)^{-1}]},$$

where $v_c = \sum_{i=1}^{c-1} h_i + 0.5 h_c$ is the cumulative probability of category c of combining both arms with $h_c = f p_c + (1-f) q_c$ and $V = v \Sigma v'$ with $v = (v_1, \cdots, v_K)$ and Σ is the (co)variance matrix of v from a multinomial distribution. The statistic $N\tau^2$ follows a noncentral chi-squared distribution with one degrees of freedom, hence sample size can be estimated by

$$N_4 = \frac{\Phi^2(1, \alpha, \beta)}{\tau^2}, \tag{2.13}$$

where $\Phi^2(1, \alpha, \beta)$ is the noncentrality parameter (ncp) for the noncentral chi-squared distribution with one degrees of freedom to have $1 - \beta$ power in a α level test. In other words, $\Pr(\chi^2_{1,\text{ncp}} < q_\alpha) = \beta$ where $\chi^2_{1,\text{ncp}}$ is a noncentral chi-squared distribution with one degrees of freedom, $\Phi^2(1, \alpha, \beta)$ as the noncentrality parameter and q_α is the critical value for a chi-squared distribution with α level.

Happ *et al.* (2019) proposed to use placement values to estimate the variance of $\widehat{\text{WinP}}$ under the alternative as the summation of the arm-specific sample variances of placement values. Note that placement values are the same as the win fractions in the treatment arm but equal to one minus the win fractions in the control arm. The variance under the null is obtained by a pooled sample variance of placement values, with the mean of placement values being forced to 0.5 under the null hypothesis. They then used these two variances with equation (2.9) for sample size estimation. Their method also focused on hypothesis testing but is more accurate when the WinP is not close to the null.

Recently, Zou *et al.* (2023) proposed to use win fractions to extend sample size formulas for continuous outcomes to be compatible with any outcomes that can be ranked. The formula is derived based on the result that the variance for the estimator of WinP can be expressed by the mean difference of win fractions. Hence, planning sample size for WinP is essentially planning sample size for the mean difference of win fractions. Since win fractions can be obtained as long as the outcomes can be ranked, the method by Zou *et al.* (2023) can be used for continuous, binary and ordered category outcomes. Additionally, the formula by Zou *et al.* (2023) was derived with the aim of confidence interval estimation of WinP; therefore, the assurance probability of the lower limit above a certain value was used in the sample size formula, where the statistical power of testing WinP $= 0.5$ is a special case by specifying the lower limit of WinP as 0.5 in the sample size formula.

2.3 Estimation of the win probability for cluster randomization trials

The usefulness of WinP in cluster randomization trials as an effect measure was only noticed recently by Zou (2021). We cannot apply the methods reviewed in previous section for individual randomization trials to data from a cluster randomization trial as the variance of $\widehat{\text{WinP}}$ could be underestimated due to the correlation between individuals within the same cluster. In this section, we will review related methods for WinP in cluster randomization trials to identify the gaps in the literature.

2.3.1 Estimating the win probability and its variance

Zou (2021) proposed methods estimating WinP and its variance for cluster randomization trials. We summarize the methods here. Denote k_i as the number of clusters in each arm ($i = 1$ for control, $i = 2$ for treatment), and m_{ij} as the size of cluster j in arm i, $j = 1, \cdots, k_i$. The total number of clusters is $k = k_1 + k_2$, the size of arm i is $M_i = \sum_{j=1}^{k_i} m_{ij}$, and the size of the trial is $N = M_1 + M_2$. The outcome of the lth subject in the jth cluster in arm i is denoted by Y_{ijl}. The win fraction of Y_{ijl} is obtained similarly as the case for the independent outcomes ignoring clusters. In short, one ranks the outcomes Y_{ijl} with the whole sample and its intervention arm, which are denoted as R_{ijl} and r_{ijl}, respectively. The win fraction for Y_{ijl} is then calculated by $w_{ijl} = (R_{ijl} - r_{ijl})/(N - M_i)$. The point estimator of WinP is obtained by the mean of win fractions in the treatment arm

$$\widehat{\text{WinP}} = \frac{\sum_{j=1}^{k_2} \sum_{l=1}^{m_{1j}} w_{2jl}}{M_2} = \frac{\sum_{j=1}^{k_2} w_{2j.}}{M_2} = \frac{\sum_{j=1}^{k_2} m_{2j}\overline{w}_{2j.}}{\sum_{j=1}^{k_2} m_{2j}}, \tag{2.14}$$

where $w_{2j.} = \sum_{l=1}^{m_{2j}} w_{2jl}$ and $\overline{w}_{2j.} = \sum_{l=1}^{m_{2j}} w_{2jl}/m_{2j}$.

The expression (2.14) implies that $\widehat{\text{WinP}}$ is also a ratio estimator or a weighted mean estimator. Following the decomposition of $\widehat{\text{WinP}}$ for independent outcomes as shown in equation

(2.4), the variance of $\widehat{\text{WinP}}$ can be obtained as

$$\begin{aligned}\text{Var}(\widehat{\text{WinP}}) &= \text{Var}\left(\frac{\sum_{j=1}^{k_1}\sum_{l=1}^{m_{1j}} w_{1jl}}{M_1}\right) + \text{Var}\left(\frac{\sum_{j=1}^{k_2}\sum_{l=1}^{m_{2j}} w_{2jl}}{M_2}\right) \\ &= \text{Var}(\overline{w}_{1..}) + \text{Var}(\overline{w}_{2..})\,,\end{aligned}$$

where $\overline{w}_{i..} = \sum_{j=1}^{k_i}\sum_{l=1}^{m_{ij}} w_{ijl} / \sum_{j=1}^{k_i} m_{ij}$ is the mean win fractions of arm i.

To account for the clustering effect, Zou (2021) proposed three estimators of $\text{Var}(\overline{w}_{i..})$. The first variance estimator is from the sampling survey literature, recognizing $\widehat{\text{WinP}}$ as a ratio estimator,

$$\widehat{\text{Var}}\,(\overline{w}_{i..}) = \frac{k_i}{(k_i - 1)\,M_i^2}\sum_{j=1}^{k_i}\left(w_{ij.} - m_{ij}\overline{w}_{i..}\right)^2 .$$

This estimator is identical to the estimator from the area under the receiver operating characteristic curve literature (Obuchowski, 1997), which was derived using the U-statistic theory.

The second variance estimator for $\widehat{\text{WinP}}$ is based on a weighted mean estimator,

$$\widehat{\text{Var}}\,(\overline{w}_{i..}) = \frac{1}{(k_i - 1)\,M_i}\sum_{j=1}^{k_i} m_{ij}\left(\overline{w}_{ij.} - \overline{w}_{i..}\right)^2 ,$$

where each individual in the trial is equally weighted instead of clusters being equally weighted. Weighting each cluster equally is less common for cluster randomization trials because it ignores the fact that including an additional large cluster reduces the variance more than including an additional small cluster.

Lastly, Zou (2021) proposed to analyze the win fractions with a mixed model that yields the variance estimator,

$$\widehat{\text{Var}}(\overline{w}_{i..}) = \widehat{\sigma}^2/M_i + \widehat{\sigma_c}^2/k_i\,,$$

where $\widehat{\sigma}^2$ and $\widehat{\sigma_c}^2$ are respectively the within and between cluster variance components estimated from the mixed model. This approach can also yield an estimate of the intraclass correlation coefficient (ICC) from the mixed model, which may be useful for planning future trials. We can use a t-statistic with $k - 2$ degrees of freedom to construct confidence intervals because two variances are estimated from cluster means. Simulation studies by Zou (2021) showed that the three variance estimators performed equally well with the logit-transformed and arcsine-transformed intervals compared with Wald-type intervals.

Other related works of WinP for cluster randomization trials mostly focus on the MW test. Rosner and Grove (1999) derived the variance of the test statistic under the null while accounting for clustering. The method is computationally extensive as the correlation induced by pairwise comparison is more complicated for clustered samples; the pairwise correlation is likely different for the outcomes sampled from different clusters. Dutta and Datta (2016) proposed to estimate the variance through a jackknife approach by deleting one cluster at a time for cluster means of raw scale. However, their method is inappropriate for ordinal outcomes because they require averaging the raw outcomes for each cluster.

2.3.2 Sample size based on Mann-Whitney test

Rosner and Glynn (2011) proposed to estimate the sample size for the MW test based on a location-shift model with normality assumption for cluster randomization trials. Their method extends the method for individually randomized controlled trials (Rosner and Glynn, 2009) to cluster randomization trials. Their method only requires the specification of WinP, a constant cluster size, and the intraclass correlation. They only evaluated the performance with small constant cluster sizes of 2 for ophthalmological studies. It is less known how it performs in other cluster randomization trial settings.

For sample size planning for the WinP, Obuchowski (1997) suggested first calculating the sample size as if WinP is estimated from an individually randomized controlled trial, then increasing it by a factor to account for the variance increased due to cluster randomization, which is a common approach to (Donner *et al.*, 1981).

Parametric assumptions are usually hard to justify, especially for ordinal outcomes. Generally it is advocated to analyze the data with the same methods used for sample size planning. Therefore, it is desirable to have a sample size formula that does not require parametric assumptions.

2.4 Summary

The WinP is a natural way to quantify the treatment effect for ordinal outcomes. It can be estimated with the win fractions of the outcomes, where the win fraction of an outcome is the proportion of wins that outcome achieved by comparing to all the outcomes in the other arm. The estimator for WinP is the mean of win fractions for the treatment arm, and the variance of this estimator is estimated by the sum of sample variances of mean win fractions from both arms. We can treat win fractions as independent observations for individually randomized trials but have to adjust for their correlations to each other within a cluster for cluster randomization trials. The essence of the win fraction approach is to transform outcomes into win fractions and then analyze them as continuous outcomes.

The advantage of the win fraction approach is that only the rank of the outcome is used, being consistent with the ordinal property. Additionally, the win fraction approach unifies the calculation for continuous, binary and ordered category outcomes regardless of the presence of ties, thus makes planning sample size with trials focusing on estimating the WinP more straightforward, as one does not need to worry about the ties and the type of data. This is particularly useful for sample size estimation, as there is usually limited information available in the planning stage of a trial.

Adjusting for baseline assessments could increase the efficiency of estimating the treatment effect, reducing the required sample size to maintain the same power. We reviewed the weighted least square approach and the regression approach to adjust WinP for baseline in individually randomized trials, where the regression approach is steadily available in most statistical software. However, methods to estimate baseline adjusted WinP have not been available for cluster randomization trials. We will propose methods for cluster randomization trials by building on Zou (2022c)'s work in Chapter 3. Sample size formulas focused on confidence interval precision will be proposed in Chapter 4.

Chapter 3

Estimation of treatment effect in cluster randomization trials with baseline measurements

In the previous chapter, we reviewed the literature on estimating the WinP as the treatment effect in randomized controlled trials. Methods for baseline adjustment are only available for individually randomized trials and lacking for cluster randomization trials.

This chapter develops methods for baseline adjustment of the WinP estimation in cluster randomization trials with a focus on interval estimation. We will extend the weighted least square method for (co)variance estimators to account for the correlated outcomes in a cluster randomization trial. Large sample properties of those (co)variance estimators will be derived. We will also propose a mixed model method for baseline adjustment as an analysis of covariance model that accounts for clustering.

3.1 Notations

We focus on two-arm cluster randomization trials and define notations accordingly. Suppose there are k_i clusters randomized to arm i ($i = 1$ for control, $i = 2$ for treatment), where each cluster consists of m_{ij} members ($j = 1, ..., k_i$). The size of arm i is $N_i = \sum_{j=1}^{k_i} m_{ij}$, and the size of the trial is $N = M_1 + M_2$. The outcomes measured at follow-up are denoted as:

$$Y_{ijl} = \text{The outcomes for the } l\text{th participant in the } j\text{th cluster in the } i\text{th arm}\,,$$

where $Y_{ijl} \sim Y_i$. We denote the distribution function for Y_i as F_i, which is defined as the average of the left- and right-continuous distribution function

$$F_i(x) = 0.5[F_i^-(x) + F_i^+(x)]\,,$$

where $F_i^-(x) = \Pr(Y_i < x)$ denotes the left-continuous distribution function and $F_i^+(x) = \Pr(Y_i \leq x)$ denotes the right-continuous distribution function. The WinP is the probability that a randomly chosen participant in the treatment arm has a better outcome compared to a randomly chosen participant in the control arm,

$$\text{WinP} = \int F_1(x)dF_2(x) = \Pr(Y_2 > Y_1) + 0.5\Pr(Y_2 = Y_1)\,, \tag{3.1}$$

where ties are counted as both participants have an equal chance of winning each other. We denote the baseline measurement by

$$X_{ijl} = \text{The baseline measurement for the } l\text{th participant in the } j\text{th cluster in the } i\text{th arm}\,,$$

where $X_{ijl} \sim X_i$ and the distribution function for X_i is G_i. Although the t-test is often used for testing the baseline imbalance in the literature, it ignores the ordinal nature of the baseline measurement by calculating mean and standard errors on the original scale. We can better quantify the baseline imbalance using the WinP for baseline measurements

$$\text{WinP}_X = \int G_1(x)dG_2(x) = P(X_2 > X_1) + 0.5P(X_2 = X_1)\,,$$

where randomization implies that $G_2 = G_1$; hence, $\text{WinP}_X = 0.5$.

We briefly review how WinP is estimated, where details are presented in Chapter 2. We first transform the outcomes into win fractions. For each outcome Y_{ijl}, the win fraction for it is obtained from the rank of Y_{ijl} in its arm, denoted as r_{ijl}, and the rank of Y_{ijl} in the whole sample, denoted as R_{ijl}. The win fractions for outcomes in the treatment and control arm are obtained respectively,

$$w^Y_{2jl} = \widehat{F}_1(Y_{2jl}) = \frac{R^Y_{2jl} - r^Y_{2jl}}{N - M_2} = \text{Win fraction of } Y_{2jl}$$
$$w^Y_{1jl} = \widehat{F}_2(Y_{1jl}) = \frac{R^Y_{1jl} - r^Y_{1jl}}{N - M_1} = \text{Win fraction of } Y_{1jl}\,.$$

Similarly, the win fractions for baseline measurements are obtained by

$$w^X_{2jl} = \widehat{G}_1(X_{2jl}) = \frac{R^X_{2jl} - r^X_{2jl}}{N - M_2} = \text{Win fraction of } X_{2jl}$$
$$w^X_{1jl} = \widehat{G}_2(X_{1jl}) = \frac{R^X_{1jl} - r^X_{1jl}}{N - M_1} = \text{Win fraction of } X_{1jl}\,,$$

where R^X_{ijl} and r^X_{ijl} are the ranks of X_{ijl} in the whole sample and its intervention arm, respectively. The point estimator of WinP is the mean of win fractions in the treatment arm

$$\widehat{\text{WinP}} = \overline{w}^Y_{2..} = \sum_{j=1}^{k_2} \sum_{l=1}^{n_{2j}} w^Y_{2jl}/M_2\,,$$

and the baseline imbalance is the mean of win fractions for baseline measurements

$$\widehat{\text{WinP}}_X = \overline{w}^X_{2..} = \sum_{j=1}^{k_2} \sum_{l=1}^{m_{2j}} w^X_{2jl}/M_2\,.$$

The variance of the estimators is estimated by the sum of the sample variances of win fractions for each arm,

$$\begin{aligned}
\widehat{\text{Var}}(\widehat{\text{WinP}}) &= \frac{\sum_{j=1}^{k_1}\sum_{l=1}^{m_{1j}}(w_{1jl}^Y - 1 + \widehat{\text{WinP}})^2}{M_1} + \frac{\sum_{j=1}^{k_2}\sum_{l=1}^{m_{2j}}(w_{2jl}^Y - \widehat{\text{WinP}})^2}{M_2} \\
\widehat{\text{Var}}(\widehat{\text{WinP}}_X) &= \frac{\sum_{j=1}^{k_1}\sum_{l=1}^{m_{1j}}(w_{1jl}^X - 1 + \widehat{\text{WinP}}_X)^2}{M_1} + \frac{\sum_{j=1}^{k_2}\sum_{l=1}^{m_{2j}}(w_{2jl}^X - \widehat{\text{WinP}}_X)^2}{M_2}\,.
\end{aligned} \tag{3.2}$$

We will denote the WinP estimate adjusted by baseline measurements as $\widehat{\text{WinP}}^*$ to avoid confusion with $\widehat{\text{WinP}}$ for follow-up outcomes in the rest of the chapter.

3.2 The weighted least square approach

The weighted least square approach of adjusting for baseline assessment is a regression model with arm-specific means as the observations. The model assumes baseline imbalance occurs only by chance in a randomized study hence constrains WinP_X to 0.5. Denote $Y = (\widehat{\text{WinP}}, \widehat{\text{WinP}}_X)'$ and $X = (1, 0)'$, the adjustment model can be written as $Y = \text{WinP}^* X$, where WinP^* denotes the baseline adjusted WinP. The weighted least square estimator for WinP^* is

$$\begin{aligned}
\widehat{\text{WinP}}^* &= \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}Y \\
&= \widehat{\text{WinP}} - \frac{\widehat{\text{Cov}}(\widehat{\text{WinP}}, \widehat{\text{WinP}}_X)}{\widehat{\text{Var}}(\widehat{\text{WinP}}_X)}\left(\widehat{\text{WinP}}_X - 0.5\right),
\end{aligned} \tag{3.3}$$

where Σ denotes the covariance matrix of $\widehat{\text{WinP}}$ and $\widehat{\text{WinP}}_X$. The estimator (3.3) indicates the adjustment is determined by the strength of correlation between baseline and follow-up and the magnitude of baseline imbalance by $\widehat{\text{WinP}}_X - 0.5$.

The variance of $\widehat{\text{WinP}}^*$ follows from the weighted least square method,

$$\begin{aligned}
\text{Var}(\widehat{\text{WinP}}^*) &= (X'WX)^{-1}X'W\text{Var}(Y)W'X(X'WX)^{-1} \\
&= \text{Var}(\widehat{\text{WinP}}) - \frac{\text{Cov}^2(\widehat{\text{WinP}}, \widehat{\text{WinP}}_X)}{\text{Var}(\widehat{\text{WinP}}_X)},
\end{aligned} \tag{3.4}$$

where the variances are estimated by using sample variances shown in equation (3.2), and the covariance of $\widehat{\text{WinP}}$ and $\widehat{\text{WinP}}_X$ is estimated using sample covariances as shown in the following section.

3.2.1 Covariance between $\widehat{\text{WinP}}$ and $\widehat{\text{WinP}}_X$

The covariance of the estimators $\widehat{\text{WinP}}$ and $\widehat{\text{WinP}}_X$ is derived using the decomposition of $\widehat{\text{WinP}}$ into win fractions as shown by Zou (2021) as $\widehat{\text{WinP}} = 1 - \text{WinP} + \int F_1(x)d\widehat{F}_2(x) + \int F_2(x)d\widehat{F}_1(x)$ and similarly, $\widehat{\text{WinP}}_X = 0.5 + \int G_1(x)d\widehat{G}_2(x) + \int G_2(x)d\widehat{G}_1(x)$. The asymptotic covariance between $\widehat{\text{WinP}}$ and $\widehat{\text{WinP}}_X$ is hence:

$$
\begin{aligned}
&\text{Cov}(\widehat{\text{WinP}}, \widehat{\text{WinP}}_X) \\
=&\text{Cov}\left[\int F_1(x)d\widehat{F}_2(x) + \int F_2(x)d\widehat{F}_1(x), \int G_1(x)d\widehat{G}_2(x) + \int G_2(x)d\widehat{G}_1(x)\right] \\
=&\text{Cov}\left[\int F_1(x)d\widehat{F}_2(x), \int G_1(x)d\widehat{G}_2(x)\right] + \underbrace{\text{Cov}\left[\int F_1(x)d\widehat{F}_2(x), \int G_2(x)d\widehat{G}_1(x)\right]}_{=0} \\
+&\text{Cov}\left[\int F_2(x)d\widehat{F}_1(x), \int G_2(x)d\widehat{G}_1(x)\right] + \underbrace{\text{Cov}\left[\int F_2(x)d\widehat{F}_1(x), \int G_1(x)d\widehat{G}_2(x)\right]}_{=0} \\
=&\text{Cov}\left[\int F_1(x)d\widehat{F}_2(x), \int G_1(x)d\widehat{G}_2(x)\right] + \text{Cov}\left[\int F_2(x)d\widehat{F}_1(x), \int G_2(x)d\widehat{G}_1(x)\right] \\
=&\text{Cov}\left[\widehat{F}_2(Y_{1jl}), \widehat{G}_2(X_{1jl})\right] + \text{Cov}\left[\widehat{F}_1(Y_{2jl}), \widehat{G}_1(X_{2jl})\right] \\
=&\text{Cov}\left[\frac{\sum_{j=1}^{k_1}\sum_{l=1}^{m_{1j}} w_{1jl}^Y}{M_1}, \frac{\sum_{j=1}^{k_1}\sum_{l=1}^{m_{1j}} w_{1jl}^X}{M_1}\right] + \text{Cov}\left[\frac{\sum_{j=1}^{k_2}\sum_{l=1}^{m_{2j}} w_{2jl}^Y}{M_2}, \frac{\sum_{j=1}^{k_2}\sum_{l=1}^{m_{2j}} w_{2jl}^X}{M_2}\right],
\end{aligned}
$$

where $\text{Cov}\left[\int F_1(x)d\widehat{F}_2(x), \int G_2(x)d\widehat{G}_1(x)\right] = \text{Cov}\left[\int F_2(x)d\widehat{F}_1(x), \int G_1(x)d\widehat{G}_2(x)\right] = 0$ because the outcomes in different arms are independent. The derivation implies that the covariance can be estimated by the sample covariances between arm-specific mean win fractions at the follow-up and baseline:

$$
\widehat{\text{Cov}}(\widehat{\text{WinP}}, \widehat{\text{WinP}}_X) = \widehat{\text{Cov}}(\overline{w}_{1..}^Y, \overline{w}_{1..}^X) + \widehat{\text{Cov}}(\overline{w}_{2..}^Y, \overline{w}_{2..}^X)\,. \tag{3.5}
$$

Although win fractions estimated from a finite sample are weakly correlated, they are asymptotically independent because $w^Y_{2jl} = \widehat{F}_1(Y_{2jl}) \rightarrow F_1(Y_{2jl})$ as $M_1 \rightarrow \infty$ (Akritas, 1990), where a rigorous proof can be found in Sen (1967). We now proceed to propose estimators of $\widehat{\text{Cov}}(\overline{w}^Y_{i..}, \overline{w}^X_{i..})$ that accounts for clusters.

3.2.2 Weighted covariance estimator

Denote the variance of win fractions at follow-up in arm i by $\text{Var}(w^Y_{ijl}) = \sigma^2_i$ and the intraclass correlation coefficient (ICC) for arm i by ρ_i. Standard theory from cluster sampling suggests that the variance of cluster-specific mean of win fractions can be written as (Cochran, 1976),

$$\text{Var}(\overline{w}^Y_{ij.}) = \frac{\sigma^2_i}{m_{ij}} \left[1 + (m_{ij} - 1)\rho_i\right] .$$

The most efficient way of combining the k_i cluster-specific means is by weighting them inversely proportional to their variance (Casella and Berger, 2001, p.303), suggesting the ICC weight

$$\omega_{ij} = \frac{m_{ij}/\left[1 + (m_{ij} - 1)\rho_i\right]}{\sum_{j=1}^{k_i} m_{ij}/\left[1 + (m_{ij} - 1)\rho_i\right]} .$$

The ICC weighted covariance estimator is then

$$\widehat{\text{Cov}}^{\text{icc}}(\overline{w}^Y_{i..}, \overline{w}^X_{i..}) = \frac{1}{k_i - 1} \sum_{j=1}^{k_i} \omega_{ij}(\overline{w}^Y_{ij.} - \overline{w}^Y_{i..})(\overline{w}^X_{ij.} - \overline{w}^X_{i..}) . \tag{3.6}$$

The efficiency of the estimator depends on how accurate the weights are (Donner and Klar, 2000, p.82). However, the weights are rarely known in advance since it involves the ICC of win fractions. We can estimate the ICC using original scale, assuming the ICC is the same for both arms.

Another method to estimate the ICC of win fractions is from an analysis of variance model to obtain the variance among and within clusters for each arm (Donner and Klar, 2000, p.9). Denote the mean square error among and within clusters by MSC_i and MSW_i from arm i, respectively , the ICC is given by

$$\widehat{\rho_i} = \frac{\mathrm{MSC}_i - \mathrm{MSW}_i}{\mathrm{MSC}_i + (m_i - 1)\mathrm{MSW}_i},$$

where

$$m_i = \frac{\left(M_i - \sum_{j=1}^{k_i} m_{ij}/M_i\right)}{k_i - 1}.$$

One can also use the cluster size as weights, yielding $\omega_{ij} = m_{ij}/M_i$ and the cluster-size weighted covariance estimator (Bland and Altman, 1995*b*)

$$\widehat{\mathrm{Cov}}^{\mathrm{size}}(\overline{w}_{i..}^{Y}, \overline{w}_{i..}^{X}) = \frac{1}{k_i - 1} \sum_{j=1}^{k_i} \frac{m_{ij}}{M_i} (\overline{w}_{ij.}^{Y} - \overline{w}_{i..}^{Y})(\overline{w}_{ij.}^{X} - \overline{w}_{i..}^{X}). \tag{3.7}$$

This weighting is more intuitive as every participants are equally weighted and does not require knowledge of ICC. However, it could be less efficient than the ICC weighted estimator when there is substantial variation in the cluster size (Kerry and Bland, 2001).

We will not consider assigning equal weights to cluster because it ignores larger clusters contributing more information. Additionally, when the cluster size does not vary, cluster size weight is the same to equal weights.

3.2.3 Asymptotic properties of the weighted covariance estimator

We now show the asymptotic properties of the weighted covariance estimator in equation (3.6). The summation term in the covariance estimator for the treatment arm ($i = 2$) can be decomposed as:

$$\sum_{j=1}^{k_2} \omega_{2j} \left(\overline{w}_{2j.}^{Y} - \overline{w}_{2..}^{Y}\right)\left(\overline{w}_{2j.}^{X} - \overline{w}_{2..}^{X}\right) = \sum_{j=1}^{k_2} \omega_{2j}(\overline{w}_{2j.}^{Y} - \text{WinP} + \text{WinP} - \overline{w}_{2..}^{Y})(\overline{w}_{2j.}^{X} - 0.5 + 0.5 - \overline{w}_{2..}^{X})$$

$$= \underbrace{\sum_{j=1}^{k_2} \omega_{2j}(\overline{w}_{2j.}^{Y} - \text{WinP})(\overline{w}_{2j.}^{X} - 0.5)}_{A} - \underbrace{\sum_{j=1}^{k_2} \omega_{2j}(\overline{w}_{2j.}^{Y} - \text{WinP})(\overline{w}_{2..}^{X} - 0.5)}_{B} - \underbrace{\sum_{j=1}^{k_2} \omega_{2j}(\overline{w}_{2..}^{Y} - \text{WinP})(\overline{w}_{2j.}^{X} - 0.5)}_{C}$$

$$+ \underbrace{\sum_{j=1}^{k_2}(\overline{w}_{2..}^{Y} - \text{WinP})(\overline{w}_{2..}^{X} - 0.5)}_{D},$$

where D converges to 0 in probability since $\overline{w}_{2..}^{Y} \rightarrow \text{WinP}$ and $\overline{w}_{2..}^{X} \rightarrow 0.5$ in probability. Hence the limit distribution of the weighted covariance estimator determined by the limit distribution of A, B and C. Denote the correlation of cluster-specific mean win fractions between the baseline and follow-up for arm i as

$$r_i = E[(\overline{w}_{ij.}^{Y} - \text{WinP})(\overline{w}_{ij.}^{X} - 0.5)]\,.$$

We refer r_i as a temporal correlation because it matches clusters for two different time points of measurements. Assuming the weights of each cluster are independent to the cluster-specific means of win fractions, we have

$$\begin{aligned} E[A] &= E\left[\sum_{j=1}^{k_2} \omega_{2j}(\overline{w}_{2j.}^{Y} - \text{WinP})(\overline{w}_{2j.}^{X} - 0.5)\right] \\ &= \sum_{j=1}^{k_2} \omega_{2j} E\left[(\overline{w}_{2j.}^{Y} - \text{WinP})(\overline{w}_{2j.}^{X} - 0.5)\right] \\ &= \sum_{j=1}^{k_2} \omega_{2j} r_2 \\ &= r_2\,. \end{aligned}$$

By observing that $\overline{w}^X_{2..} \approx \sum_{l=1}^{k_2} \omega_{2l}\overline{w}^X_{2l.}$ for the ICC weight, we have

$$\begin{aligned}
E[B] &= E\left[\sum_{j=1}^{k_2} \omega_{2j}(\overline{w}^Y_{2j.} - \text{WinP})(\overline{w}^X_{2..} - 0.5)\right] \\
&\approx E\left[\sum_{j=1}^{k_2} \omega_{2j}(\overline{w}^Y_{2j.} - \text{WinP})\sum_{l=1}^{k_2} \omega_{2l}(\overline{w}^X_{2l.} - 0.5)\right] \\
&= E\left[\sum_{j=1}^{k_2} \omega_{2j}^2(\overline{w}^Y_{2j.} - \text{WinP})(\overline{w}^X_{2j.} - 0.5)\right] + \underbrace{E\left[\sum_{j\neq l} \omega_{2j}\omega_{2l}(\overline{w}^Y_{2j.} - \text{WinP})(\overline{w}^X_{2l.} - 0.5)\right]}_{\text{=0 because clusters are independent}} \\
&= \sum_{j=1}^{k_2} \omega_{2j}^2 E\left[(\overline{w}^Y_{2j.} - \text{WinP})(\overline{w}^X_{2j.} - 0.5)\right] \\
&= \sum_{j=1}^{k_2} \omega_{2j}^2 r_2\,.
\end{aligned}$$

Note that $\omega_{ij} = m_{ij}/M_i$, $\sum_{l=1}^{k_2} \omega_{2l}\overline{w}^X_{2l.} = \overline{w}^X_{2..}$ when cluster size weight is used. Similarly, $E[C] \approx \sum_{j=1}^{k_2} \omega_{2j}^2 r_2$. Therefore the expectation of the weighted covariance estimator is approximated to $(1 - 2\sum_{j=1}^{k_2} \omega_{2j}^2)r_2$. Furthermore, when the cluster sizes are large, $m_{ij}/\left[1 + (m_{ij} - 1)\rho_i\right] \approx 1/\rho_i$ (Donner and Klar, 2000, p.88); hence, $\omega_{ij} \to 1/k_i$ the covariance term for the treatment arm converges to r_2 as $k_2 \to \infty$. A similar inference for the control arm can be made, yielding the weighted covariance estimator for the control arm converges to r_1.

3.2.4 Ratio covariance estimator

Another strategy to account for clustering is to consider the arm-specific mean win fractions as a ratio of the sum of win fractions over the sum of the size of clusters,

$$\widehat{\text{WinP}} = \overline{w}^Y_{2..} = \frac{\sum_{j=1}^{k_2}\sum_{l=1}^{m_{ij}} w^Y_{2jl}}{\sum_{j=1}^{k_2} m_{2j}}\,.$$

The covariance between WinP and WinP_X is

$$\text{Cov}(\widehat{\text{WinP}}, \widehat{\text{WinP}}_X) = \text{Cov}(\overline{w}^Y_{1..}, \overline{w}^X_{1..}) + \text{Cov}(\overline{w}^Y_{2..}, \overline{w}^X_{2..})\,,$$

The covariance between $\overline{w}_{2..}^{Y}$ and $\overline{w}_{2..}^{X}$ can be expressed with ratios by

$$\begin{aligned}
\mathrm{Cov}(\overline{w}_{2..}^{Y}, \overline{w}_{2..}^{X}) &= \mathrm{Cov}\left(\frac{\sum_{j=1}^{k_2} \sum_{l=1}^{m_{2j}} w_{2jl}^{Y}}{\sum_{j=1}^{k_2} m_{2j}}, \frac{\sum_{j=1}^{k_2} \sum_{l=1}^{m_{2j}} w_{2jl}^{X}}{\sum_{j=1}^{k_2} m_{2j}} \right) \\
&= \mathrm{Cov}\left(\frac{\sum_{j=1}^{k_2} \sum_{l=1}^{m_{ij}} w_{2jl}^{Y}/k_2}{\sum_{j=1}^{k_2} m_{2j}/k_2}, \frac{\sum_{j=1}^{k_2} \sum_{l=1}^{m_{ij}} w_{2jl}^{X}/k_2}{\sum_{j=1}^{k_2} m_{2j}/k_2} \right) \\
&= \mathrm{E}\left\{ \left[\frac{\sum_{j=1}^{k_2} \sum_{l=1}^{m_{ij}} w_{2jl}^{Y}/k_2}{\sum_{j=1}^{k_2} m_{2j}/k_2} - \mathrm{E}(\overline{w}_{2..}^{Y}) \right] \left[\frac{\sum_{j=1}^{k_2} \sum_{l=1}^{m_{ij}} w_{2jl}^{X}/k_2}{\sum_{j=1}^{k_2} m_{2j}/k_2} - \mathrm{E}(\overline{w}_{2..}^{X}) \right] \right\}
\end{aligned}$$

When k_2 is large, $\sum_{j=1}^{k_2} m_{2j}/k_2$ would be close to the mean cluster size of the population; hence we can treat $\sum_{j=1}^{k_2} m_{2j}/k_2 = M_2/k_2$ as a constant. The covariance can be approximated by

$$\mathrm{Cov}(\overline{w}_{2..}^{Y}, \overline{w}_{2..}^{X}) \approx \frac{k_2^2}{M_2^2} E\left[\left(\frac{\sum_{j=1}^{k_2} \sum_{l=1}^{m_{2j}} w_{2jl}^{Y}}{k_2} - \frac{\mathrm{WinP} \sum_{j=1}^{k_2} m_{2j}}{k_2} \right) \left(\frac{\sum_{j=1}^{k_2} \sum_{l=1}^{m_{2j}} w_{2jl}^{X}}{k_2} - \frac{\mathrm{WinP}_X \sum_{j=1}^{k_2} m_{2j}}{k_2} \right) \right], \tag{3.8}$$

since $\mathrm{E}[\overline{w}_{2..}^{Y}] = \mathrm{WinP}$ and $\mathrm{E}[\overline{w}_{2..}^{X}] = \mathrm{WinP}_X$. Considering the cluster-specific summations as the sampling unit from the population, the expectation term can be estimated with sample covariance, implying equation (3.8) can be estimated by

$$\widehat{\mathrm{Cov}}(\overline{w}_{2..}^{Y}, \overline{w}_{2..}^{X}) = \frac{k_2}{(k_2 - 1)M_2^2} \sum_{j=1}^{k_2} \left(\sum_{l=1}^{m_{2j}} w_{2jl}^{Y} - m_{2j}\widehat{\mathrm{WinP}} \right) \left(\sum_{l=1}^{m_{2j}} w_{2jl}^{X} - m_{2j}\widehat{\mathrm{WinP}}_X \right).$$

Similarly, we can estimate $\mathrm{Cov}(\overline{w}_{1..}^{Y}, \overline{w}_{1..}^{X})$ by

$$\widehat{\mathrm{Cov}}(\overline{w}_{1..}^{Y}, \overline{w}_{1..}^{X}) = \frac{k_1}{(k_1 - 1)M_1^2} \sum_{j=1}^{k_1} \left(\sum_{l=1}^{m_{1j}} w_{1jl}^{Y} - m_{1j}\widehat{\mathrm{WinP}} \right) \left(\sum_{l=1}^{m_{1j}} w_{1jl}^{X} - m_{1j}\widehat{\mathrm{WinP}}_X \right).$$

Thus, the ratio covariance estimator between $\widehat{\mathrm{WinP}}$ and $\widehat{\mathrm{WinP}}_X$ can be estimated by

$$\widehat{\mathrm{Cov}}^{r}(\widehat{\mathrm{WinP}}, \widehat{\mathrm{WinP}}_X) = \sum_{i=1}^{2} \frac{k_i}{(k_i - 1)M_i^2} \sum_{j=1}^{k_i} \left(\sum_{l=1}^{m_{ij}} w_{ijl}^{Y} - m_{ij}\overline{w}_{i..}^{Y} \right) \left(\sum_{l=1}^{m_{ij}} w_{ijl}^{X} - m_{ij}\overline{w}_{i..}^{X} \right). \tag{3.9}$$

The estimator (3.9) is the same as the covariance estimator for comparing two ratios in the survey literature (Cochran, 1976, p.181); hence we refer to it as the ratio covariance estimator.

The ratio estimator can be considered as the most statistically efficient way of combining the k_i cluster means if two conditions are satisfied (Cochran, 1976, p.158): (i) each cluster has almost identical mean win fractions (ii) the variance of the mean win fractions is proportional to the cluster size. The first condition implies that there is little variation between clusters, and the within-cluster variation is much larger, leading to a small intraclass correlation coefficient. This means that the ratio estimator is more useful in trials with a small number of large clusters because a low intraclass correlation coefficient is more common in those trials. The second condition implies that the optimal way to combine the cluster-specific means is by weighting them with their size, i.e., the weighted estimator with cluster size as the weights will be optimal. In the case of constant cluster size, both the weighted estimator and the ratio estimator are the same.

We have proposed the ICC weighted covariance estimator in equation (3.6), the cluster size weighted covariance estimator in equation (3.7) and the ratio covariance estimator in equation (3.9) for $\widehat{\text{WinP}}$ and $\widehat{\text{WinP}}_X$ using only independent cluster-specific summary statistics of win fractions. The three covariance estimators are used with the weighted least square method in equations (3.3) and (3.4) to estimate the adjusted WinP and its variance. The variance of $\widehat{\text{WinP}}$ can be written as the variance for independent outcomes inflated by the design effect, $1 + (m-1)\rho$ (Donner and Klar, 2000)

$$\text{Var}(\widehat{\text{WinP}}) = \frac{[1 + (m-1)\rho]}{2km} \left[\text{Var}(w^Y_{1jl}) + \text{Var}(w^Y_{2jl})\right] ,$$

assuming constant cluster size ($m_{ij} = m$), balanced design ($k_i = k$) and homogeneous of ICC ($\rho_i = \rho$). The asymptotic variance of the adjusted WinP from the weighted least square approach can be derived from equation (3.4)

$$
\begin{aligned}
\mathrm{Var}(\widehat{\mathrm{WinP}}^{*}) &= \mathrm{Var}(\widehat{\mathrm{WinP}}) - \frac{\mathrm{Cov}^2(\widehat{\mathrm{WinP}}, \widehat{\mathrm{WinP}}_X)}{\mathrm{Var}(\widehat{\mathrm{WinP}}_X)} \\
&= \mathrm{Var}(\widehat{\mathrm{WinP}}) \left(1 - \underbrace{\frac{\mathrm{Cov}^2(\widehat{\mathrm{WinP}}, \widehat{\mathrm{WinP}}_X)}{\mathrm{Var}(\widehat{\mathrm{WinP}}_X)\mathrm{Var}(\widehat{\mathrm{WinP}})}}_{r_c^2} \right) \\
&= \frac{[1 + (m-1)\rho]}{2km} \left[\mathrm{Var}(w_{1jl}^{Y}) + \mathrm{Var}(w_{2jl}^{Y}) \right] (1 - r_c^2) , \qquad (3.10)
\end{aligned}
$$

where r_c^2 is the cluster-level correlation of baseline and follow-up win fractions from the weighted least square approach.

3.3 The mixed model approach

The mixed model approach can be regarded as the analysis of the covariance (ANCOVA) of the win fractions with mixed models, extending the analysis of variance of the win fractions for cluster trials proposed by Zou (2021). The mixed model or generalized estimating equation is generally recommended for ANCOVA for cluster trials with baseline measurements Hooper *et al.* (2018). Since the WinP can be estimated with $\widehat{\mathrm{WinP}} = (\overline{w}_{2\cdot\cdot}^{Y} - \overline{w}_{1\cdot\cdot}^{Y})/2 + 0.5$, we can use an analysis of covariance to estimate the adjusted mean difference of win fractions. The adjustment is meaningful because of the linear relation between the mean difference of win fractions and the WinP. Denote the intervention indicator for a participant in arm i by Z_i, ($Z_i = 1$ for treatment, $Z_i = 0$ for control), where we drop the cluster- and individual-level indices because it is only determined by i in a cluster randomization trial. The mixed model can be written as

$$
w_{ijl}^{Y} = \beta_0 + \beta_1 Z_i + \beta_2 w_{ijl}^{X} + \alpha_{ij} + \epsilon_{ijl} , \qquad (3.11)
$$

where $\alpha_{ij} \sim N(0, \tau^2)$ and $\epsilon_{ijl} \sim N(0, \sigma^2)$ are the random effects on the cluster- and individual-level, respectively, and they are assumed to be independent to each other. The treatment coefficient $\widehat{\beta_1}$ can be written as adjusted mean difference of win fractions between two arms

$$\widehat{\beta_1} = \overline{w}^{Y}_{2..} - \overline{w}^{Y}_{1..} - \widehat{\beta_2}\left(\overline{w}^{X}_{2..} - \overline{w}^{X}_{1..}\right) . \tag{3.12}$$

Using a simple linear transformation of $\hat{\beta}_1$, we can estimate the WinP adjusted for baseline imbalance:

$$\widehat{\text{WinP}}^{*} = \widehat{\beta_1}/2 + 0.5 = \widehat{\text{WinP}} - \widehat{\beta_2}(\widehat{\text{WinP}}_X - 0.5) , \tag{3.13}$$

where $\widehat{\beta_2}$ is the correlation between baseline and follow-up win fractions. Equation (3.13) also implies the mixed model constraints the baseline imbalance to 0.5 for randomized studies. Although $\widehat{\text{WinP}}^{*}$ is obtained by dividing $\hat{\beta}_1$ by two and plus 0.5, the variance of $\widehat{\text{WinP}}^{*}$ is estimated by $\widehat{\text{Var}}(\widehat{\beta_1})$, which is the sum of arm-specific mean win fraction variances.

In the case of a balanced design with constant cluster size, $m_{ij} = m$, $\widehat{\beta_2}$ can be seen as a weighted sum of the correlations between win fractions at the cluster- and individual-levels according to Klar and Darlington (2004),

$$\widehat{\beta_2} = \hat{p}\hat{r}_c + (1 - \hat{p})\hat{r}_m ,$$

where r_c is the temporal correlation of cluster-level means, and r_m is the temporal correlation of win fraction at individual-level. Note that in weighted least square approach, only r_c is estimated. The weight p is the proportion of between cluster variability over total variability similar, it can be estimated by

$$\widehat{p} = \frac{\hat{\sigma}^2/m \sum_{i=1}^{2} \sum_{j=1}^{k_i} m(\overline{w}^{X}_{ij.} - \overline{w}^{X}_{i..})^2}{\hat{\sigma^2}/m \sum_{i=1}^{2} \sum_{j=1}^{k_i} m(\overline{w}^{X}_{ij.} - \overline{w}^{X}_{i..})^2 + (\hat{\tau}^2 + \hat{\sigma}^2/m) \sum_{i=1}^{2} \sum_{j=1}^{k_i} \sum_{l=1}^{m} (w^{X}_{ijl} - \overline{w}^{X}_{ij.})^2} ,$$

where $\hat{\tau}^2$ and $\hat{\sigma}^2$ are the variances estimated at cluster level and individual level, respectively, with restricted maximum likelihood. The temporal correlation at the cluster-level is estimated by

$$\hat{r}_c = \frac{\sum_{i=1}^{2} \sum_{j=1}^{k_i} (\overline{w}_{ij.}^{X} - \overline{w}_{i..}^{X})(\overline{w}_{ij.}^{Y} - \overline{w}_{i..}^{Y})}{\sum_{i=1}^{2} \sum_{j=1}^{k_i} (\overline{w}_{ij.}^{X} - \overline{w}_{i..}^{X})^2}\,, \tag{3.14}$$

where the estimator weights every cluster equally. This is the same correlation used in the weighted least square approach, when cluster sizes are constant. The correlation at the individual-level r_m is estimated with

$$\hat{r}_m = \frac{\sum_{i=1}^{2} \sum_{j=1}^{k_i} \sum_{l=1}^{m} (w_{ijl}^{X} - \overline{w}_{ij.}^{X})(w_{ijl}^{Y} - \overline{w}_{ij.}^{Y})}{\sum_{i=1}^{2} \sum_{j=1}^{k_i} \sum_{l=1}^{m} (w_{ijl}^{X} - \overline{w}_{ij.}^{X})^2}\,. \tag{3.15}$$

Including individual-level correlation in estimating the correlation between baseline and follow-up makes the mixed model approach different from the weighted least square approach. We expect the mixed model approach to have higher efficiency when the correlation at the individual-level is higher than the correlation at the cluster-level for win fractions.

Assuming constant cluster size and homogeneous ICC ($\rho = \rho_i$), the asymptotic variance of $\widehat{\text{WinP}}^*$ from the mixed model approach is similar to the asymptotic variance in equation (3.10)

$$\begin{aligned} \text{Var}(\widehat{\beta}_1) &= \text{Var}(\widehat{\text{WinP}}^*) \\ &= \frac{[1 + (m-1)\rho]}{2km} \left[\text{Var}(w_{1jl}^{Y}) + \text{Var}(w_{2jl}^{Y})\right] (1 - r_a^2)\,, \end{aligned} \tag{3.16}$$

where r_a is a combination of cluster-level and individual-level correlations (Teerenstra *et al.*, 2012)

$$r_a = \frac{m\rho}{1 + (m-1)\rho} r_c + \frac{1-\rho}{1 + (m-1)\rho} r_m\,.$$

Comparing equation (3.10) to equation (3.16) shows that the two approaches differs only by the correlation parameter for efficiency. The mixed model approach utilizes the correlation at both cluster- and individual-levels, whereas the weighted least square approach only utilizes cluster-level correlation. However, it should be noted that both approach estimates the treatment effect (WinP) for individual-level inferences.

3.4 Confidence interval for the treatment effect

We have proposed the weighted least square and mixed model approaches to estimate the variance of the adjusted treatment effect. In this section we derive confidence intervals using those variance estimates.

With the adjusted treatment effect $\widehat{\text{WinP}}^*$ and its variance $\widehat{\text{Var}}(\widehat{\text{WinP}}^*)$ being estimated from the previous sections, a two-sided $(1-\alpha)\%$ confidence interval for the treatment effect is given by

$$(L_1, U_1) = \widehat{\text{WinP}}^* \mp t_{\alpha/2,\text{df}} \sqrt{\widehat{\text{Var}}(\widehat{\text{WinP}}^*)},$$

where $t_{\alpha/2,\text{df}}$ denotes the upper $\alpha/2$ quantile of a t distribution with df degrees of freedom. We refer this to the Wald confidence interval, as it comes from the t-test of $\text{WinP}^* = 0.5$. If the mixed model approach was used to estimate WinP^*, the degrees of freedom of the regression coefficient for the treatment indicator can be used to construct confidence interval. When weighted least square approach is used to estimate WinP^*, we use the Satterthwaite approximation of the degrees of freedom, which is commonly used in analyzing cluster randomization trials to account for the heterogeneity of variances across clusters (Leyrat *et al.*, 2018). The variance of $\widehat{\text{WinP}}^*$ from the weighted least square approach from equation (3.4) can be decomposed as

$$\text{Var}(\widehat{\text{WinP}}^*) = \text{Var}(\widehat{\text{WinP}}) - \frac{\text{Cov}^2(\widehat{\text{WinP}}, \widehat{\text{WinP}}_X)}{\text{Var}(\widehat{\text{WinP}}_X)}$$

$$
\begin{aligned}
&= \mathrm{Var}(\widehat{\mathrm{WinP}}) - r^2\mathrm{Var}(\widehat{\mathrm{WinP}})\\
&= \widehat{\mathrm{Var}}(\overline{w}^Y_{1..}) + \widehat{\mathrm{Var}}(\overline{w}^Y_{2..}) - r^2\left[\widehat{\mathrm{Var}}(\overline{w}^Y_{1..}) + \widehat{\mathrm{Var}}(\overline{w}^Y_{2..})\right]\\
&= \widehat{\mathrm{Var}}(\overline{w}^Y_{1..})(1-r^2) + \widehat{\mathrm{Var}}(\overline{w}^Y_{2..})(1-r^2)\,.
\end{aligned}
$$

Therefore, the degrees of freedom of $\widehat{\mathrm{Var}}(\widehat{\mathrm{WinP}}^*)$ can be approximated by

$$
\mathrm{df} = \frac{s_1^2 + s_2^2}{s_1^2/(k_1-1) + s_2^2/(k_2-1)}\,, \tag{3.17}
$$

where s_i^2 is the variance component of $\widehat{\mathrm{Var}}(\widehat{\mathrm{WinP}}^*)$ from arm i that is given by

$$
s_i^2 = \frac{1}{k_i}\left[\widehat{\mathrm{Var}}(\overline{w}^Y_{i..})(1-r^2)\right]\,.
$$

This formula is an extension of the formula for individually randomized trials proposed by Schacht *et al.* (2008), and we assume the covariance between win fractions of baseline measurement and outcome does not differ by study arm. Equation (3.17) shows that in a balanced design $k = k_1 = k_2$ with small r, df is close to $k - 1$.

The Wald interval has been known to produce under-coverage intervals for probability parameters, and it could result in intervals outside [0,1] (Newcombe, 1998). Additionally, the tail errors (non-coverage of right- or left-confidence limits) are asymmetric because Wald-type intervals assume the variance of the estimator is the same across the whole interval.

An improvement can be made by building the interval on the logit scale by

$$
l_2, u_2 = \ln\frac{\widehat{\mathrm{WinP}}^*}{1-\widehat{\mathrm{WinP}}^*} \mp t_{\alpha/2,\mathrm{df}}\frac{\sqrt{\widehat{\mathrm{Var}}(\widehat{\mathrm{WinP}}^*)}}{\widehat{\mathrm{WinP}}^*(1-\widehat{\mathrm{WinP}}^*)}\,, \tag{3.18}
$$

and then transform back to the probability scale by the inverse logit function:

$$
L_2 = \frac{\exp(l_2)}{1+\exp(l_2)}, \quad U_2 = \frac{\exp(u_2)}{1+\exp(u_2)}\,, \tag{3.19}
$$

which we will refer as the logit transformed interval. This transformation is common in analyzing proportions due to the wide use of logistic regression.

The Wilson confidence interval for proportions may also be considered due to its good performance on coverage and avoidance of boundary problems. If WinP is treated as a proportion, the $(1-\alpha)\%$ level Wilson confidence interval consists all WinP that satisfies $|\text{WinP} - \widehat{\text{WinP}}|/\sqrt{\text{Var}(\widehat{\text{WinP}})} < t_{\alpha/2}$, where $\text{Var}(\widehat{\text{WinP}}) = \text{WinP}(1-\text{WinP})/N$ and N is the sample size. However, the variance of $\widehat{\text{WinP}}$ cannot be estimated as a proportion because it depends on the outcome distributions of both intervention arm. We could use the closed-form expression of the Wilson confidence interval on logit WinP as

$$\text{logit}(\widehat{\text{WinP}}) \mp 2\text{arsinh}\left\{0.5t_{\alpha/2,\text{df}}\sqrt{\text{Var}[\text{logit}(\widehat{\text{WinP}}^*)]}\right\},$$

where the variance of logit $\widehat{\text{WinP}}^*$ is given by $\widehat{\text{Var}}(\widehat{\text{WinP}}^*)/[\widehat{\text{WinP}}^*(1-\widehat{\text{WinP}}^*)]^2$ and arsinh denotes the inverse hyperbolic sine function $\text{arsinh}(x) = \ln(x+\sqrt{x^2+1})$. The arsinh transformed interval is hence constructed from

$$l_3, u_3 = \ln\frac{\widehat{\text{WinP}}^*}{1-\widehat{\text{WinP}}^*} \mp 2\text{arsinh}\left[t_{\alpha/2,\text{df}}\frac{\sqrt{\widehat{\text{Var}}(\widehat{\text{WinP}}^*)}}{2\widehat{\text{WinP}}^*(1-\widehat{\text{WinP}}^*)}\right] \tag{3.20}$$

and apply the inverse logit function yielding,

$$L_3 = \frac{\exp(l_3)}{1+\exp(l_3)}, \qquad U_3 = \frac{\exp(u_3)}{1+\exp(u_3)}. \tag{3.21}$$

Newcombe (2001) pointed out that the arsinh-transformed interval for proportions is always contained in the logit-transformed interval, hence having higher efficiency for estimating proportions. The work of estimating WinP for cluster trials using only follow-up outcome by Zou (2021) also observed that arsinh-transformed confidence interval for WinP is on average narrower than logit-transformed intervals in the simulation study.

3.5 Summary and discussion

In this chapter, we developed the weighted least square and the mixed model approaches to estimate WinP while adjusting for baseline measurements. We developed cluster size weighted, ICC weighted and ratio covariance estimators to be used with the weighted least square approach, making four estimators for adjusted WinP, including the mixed model approach. We derived the asymptotic properties for the weighted covariance estimators and the ratio covariance estimator. We also developed Wald type, logit transformed and arsinh transformed confidence interval for the adjusted WinP, with the degrees of freedom approximation for intervals constructed with weighted least square estimators.

Our methods can be summarized into the following steps. Transform the outcomes and baseline measurements into win fractions using overall and group-specific ranks, and apply either the weighted least square approach or the mixed model approach to estimate adjusted WinP and its variance. Construct confidence intervals based on either Wald, logit transformed or arsinh transformed confidence intervals. The mixed model approach is available for most statistical software, as one only needs to obtain win fractions based on ranks and then regress the win fractions of outcome on the win fractions of baseline measurement and treatment indicator with the mixed model, specifying heterogeneity variance. On the other hand, the weighted least square approach would require custom programming to obtain the estimates.

A statistical advantage of the mixed model approach of estimating WinP is its convenience to compute. An additional advantage is no link functions are required to accommodate different types of outcomes. For example, binary outcomes usually require the logit link with the mixed model to estimate odds ratios, resulting in odds ratios that can only be interpreted conditional on the cluster, which is different to the population-averaged odds ratio (Robinson and Jewell, 1991).

The weighted least square approach has the advantage that it uses only the cluster-level summary statistics of win fractions, implying that the methods can be applied to cross-sectional designs, where the baseline and follow-up consist of different individuals. To be specific, one obtains the win fractions for baseline and follow-up, respectively, and then uses the weighted least square approach with the cluster-level mean of win fractions.

We examined the efficiency of the weighted least square approach and the mixed model approach by their asymptotic properties and showed that the weighted least square approach depends only on cluster-level correlation, but the mixed model approach depends on both cluster-level and individual-level correlation. Therefore, it is necessary to use examine the efficiency of both approaches with finite samples.

Our method can be extended to stratified trials, where clusters are randomized within strata to combine the estimates of the WinP in each stratum through weights. Donner and Klar (1993) suggested weights as the product of intervention arm sizes over the size of the strata, which is more efficient when the ICC does not differ much between strata. The variance is then obtained as the sum of the weighted variances from each stratum.

Chapter 4

Sample size estimation for cluster randomization trials

Sample size planning is an important step in designing a randomized controlled trial that can meet the study objectives without wasting excessive resources from recruiting an overpowered trial or missing important findings from recruiting an under-powered trial.

To plan for a cluster randomization trial, it is common to first calculate the sample size as if independent individuals are randomized and then increase it to account for cluster randomization (Donner *et al.*, 1981). The sample size formulas for individually randomized trials presented in Chapter 2 focused on hypothesis testing rather than confidence interval estimation, and mostly focused on the distribution of the test statistic under the null hypothesis. It is unclear how those methods would work for estimation purposes, where the distribution of an effect measure under the alternative is required.

One approach of sample size formula for effect estimation focuses on the width of the confidence interval (Beal, 1989). However, sample size formulas developed by specifying the width of the confidence interval are prone to underestimate the sample size because normal approximation implies the sample variance is symmetrically distributed having zero skewness, but chi-squared distributions are more appropriate (Kupper and Hafner, 1989). Another approach is to consider the confidence limits as random variables and plan the sample size to ensure the confidence interval excluding a certain value with a given probability (Greenland, 1988). This approach was used in the sample size formula derived by Zou (2012) for estimating the intraclass correlation coefficient (ICC), focusing on the probability that the lower confidence limit of ICC exceeds a certain value.

It is common for randomized trials to evaluate the effect of a certain intervention such that the intervention can be applied if it is proven to improve the outcome by at least a minimal clinically important difference. Following the same principle, we intent to develop a sample size formula for WinP to ensure the lower limit of the WinP exceeding a certain threshold with a reasonable chance. For example, using the minimal clinically important difference determined by a small Cohen's effect size, the sample size should allow the trial to have a reasonable chance for the lower limits of the Win probability to exceed 0.56, which is the WinP corresponding to the small Cohen's effect size for normal outcomes.

4.1 Sample size for confidence interval estimation

We will develop a sample size formula such that the lower limit of $\widehat{\text{WinP}}$ is greater than a prespecified WinP with assurance probability $1-\beta$. A benchmark of choosing the prespecified lower bound for WinP can be 0.5, 0.56, 0.64 and 0.71 for zero, small, medium and large effect sizes as a correspondence to Cohen's effect size for normal distribution outcomes. We will derive the sample size formula based on the logit-transformed confidence interval of WinP because they were observed to have a better coverage and more balanced tail errors for WinP (Zou, 2021) compared to a Wald type confidence interval.

The logit-transformed confidence interval as shown in equation (3.18) is

$$\text{logit}(\text{WinP}_L),\ \text{logit}(\text{WinP}_U) = \text{logit}(\widehat{\text{WinP}}) \mp z_{\alpha/2} \frac{\sqrt{\text{Var}(\widehat{\text{WinP}})}}{\widehat{\text{WinP}}(1-\widehat{\text{WinP}})},$$

where z_x is the upper x quantile of the standard normal distribution. The assurance probability $(1-\beta)$ of the lower limit exceeding the specified minimal WinP, denoted by WinP_l, can be formulated as

$$
\begin{aligned}
1-\beta &= \Pr\left(\widehat{\text{WinP}}_L \geq \text{WinP}_l\right) \\
&= \Pr\left[\text{logit}(\widehat{\text{WinP}}) - z_{\alpha/2}\frac{\sqrt{\text{Var}(\widehat{\text{WinP}})}}{\widehat{\text{WinP}}(1-\widehat{\text{WinP}})} \geq \text{logit}(\text{WinP}_l)\right]. \qquad (4.1)
\end{aligned}
$$

The asymptotic distribution of $\text{logit}(\widehat{\text{WinP}})$ can be obtained by applying the delta method yielding

$$
\text{logit}(\widehat{\text{WinP}}) \sim \text{N}\left(\text{logit}(\text{WinP}), \frac{\text{Var}(\widehat{\text{WinP}})}{\text{WinP}^2(1-\text{WinP})^2}\right).
$$

Equation (4.1) can be written as

$$
\begin{aligned}
1-\beta =&\Pr\left[\frac{\text{logit}(\widehat{\text{WinP}}) - \text{logit}(\text{WinP})}{\text{WinP}^{-1}(1-\text{WinP})^{-1}\sqrt{\text{Var}(\widehat{\text{WinP}})}} \geq \frac{\text{logit}(\text{WinP}_l) - \text{logit}(\text{WinP})}{\text{WinP}^{-1}(1-\text{WinP})^{-1}\sqrt{\text{Var}(\widehat{\text{WinP}})}} + z_{\alpha/2}\right] \\
&= \Pr\left[Z \geq \text{WinP}(1-\text{WinP})\frac{\text{logit}(\text{WinP}_l) - \text{logit}(\text{WinP})}{\sqrt{\text{Var}(\widehat{\text{WinP}})}} + z_{\alpha/2}\right] \qquad (4.2)
\end{aligned}
$$

because

$$
\frac{\text{WinP}(1-\text{WinP})\sqrt{\widehat{\text{Var}}(\widehat{\text{WinP}})}}{\widehat{\text{WinP}}(1-\widehat{\text{WinP}})\sqrt{\text{Var}(\widehat{\text{WinP}})}} \to 1
$$

in probability. Applying the upper-quantile function of the standard normal distribution to both sides yield

$$
z_{\alpha/2} + z_\beta = [\text{logit}(\text{WinP}) - \text{logit}(\text{WinP}_l)]\frac{\text{WinP}(1-\text{WinP})}{\sqrt{\text{Var}(\widehat{\text{WinP}})}}. \qquad (4.3)
$$

To obtain the sample size, we need to factor out the total sample size N in $\mathrm{Var}(\widehat{\mathrm{WinP}})$, which can be written as the sum of the arm-specific sample variance of mean win fractions. Denoting ϕ_i^2 as the variance of win fractions for arm i ($i = 1$ for control, $i = 2$ for treatment) and n_i as the size of arm i, we can write $\mathrm{Var}(\widehat{\mathrm{WinP}})$ as

$$\begin{aligned}\mathrm{Var}(\widehat{\mathrm{WinP}}) &= \frac{\phi_1^2}{n_1} + \frac{\phi_2^2}{n_2} \\ &= \frac{1}{N}(1 + 1/s)\left(s\phi_1^2 + \phi_2^2\right), \end{aligned} \tag{4.4}$$

where s is the ratio of subjects in the treatment arm over the control arm. A sample size formula from equations (4.3) and (4.4) yields

$$N = \left(1 + \frac{1}{s}\right)\left\{\frac{z_{\alpha/2} + z_\beta}{[\mathrm{logit}(\mathrm{WinP}) - \mathrm{logit}(\mathrm{WinP}_l)]}\right\}^2 \frac{s\phi_1^2 + \phi_2^2}{\mathrm{WinP}^2(1 - \mathrm{WinP})^2}, \tag{4.5}$$

which is the sample size formula for individually randomized trials proposed by Zou *et al.* (2023).

We now derive sample size for cluster randomized trials with fixed cluster size m by increasing the variance (4.4) by the design effect D,

$$\mathrm{D} = 1 + (m - 1)\rho,$$

where ρ is the intraclass correlation coefficient (ICC) of follow-up win fractions assuming it is the same for both arms. Therefore, $\mathrm{Var}(\widehat{\mathrm{WinP}})$ for a cluster randomized trial is

$$\mathrm{Var}(\widehat{\mathrm{WinP}}) = \frac{1}{N}(1 + 1/s)(s\phi_1^2 + \phi_2^2)\{1 + (m - 1)\rho\}.$$

Substituting this variance formula into equation (4.3) yields

$$z_{\alpha/2} + z_\beta = \sqrt{N}\,[\mathrm{logit}(\mathrm{WinP}_l) - \mathrm{logit}(\mathrm{WinP})]\,\frac{\mathrm{WinP}(1 - \mathrm{WinP})}{\sqrt{(1 + 1/s)(s\phi_1^2 + \phi_2^2)\{1 + (m - 1)\rho\}}}. \tag{4.6}$$

It follows that,

$$N = \left(1 + \frac{1}{s}\right)\left\{\frac{z_{\alpha/2} + z_{\beta}}{[\text{logit}(\text{WinP}) - \text{logit}(\text{WinP}_l)]}\right\}^2 \left(\frac{s\phi_1^2 + \phi_2^2}{\text{WinP}^2(1 - \text{WinP})^2}\right)[1 + (m-1)\rho] \ , \qquad (4.7)$$

where ϕ_i^2 can be substituted by the variance of win fractions from a pilot study.

Dividing the total sample size N by the cluster size m yields the required number of clusters k to be randomized, $k = N/m$. The use of Z-statistic rather than t-statistic in (4.7) could underestimate the sample size when the number of clusters is small, but it can be adjusted by increasing the number of clusters by one in each arm with 95% confidence interval and by two with 99% confidence interval for balanced trials (Snedecor and Cochran, 1989, p.104). When variable cluster sizes are anticipated, a slightly anti-conservative approach can be taken by replacing the cluster size m with the average cluster size $\overline{m}$ (Donner and Klar, 2000, p.57).

With given values of N, s, m, ρ, WinP and WinP$_l$ the assurance probability $(1-\beta)$ can be derived from equation (4.6),

$$1 - \beta = \Phi\left\{\frac{\text{WinP}(1-\text{WinP})[\text{logit}(\text{WinP}_l) - \text{logit}(\text{WinP})]}{\sqrt{(1 + 1/s)(s\phi_1^2 + \phi_2^2)\{1 + (m-1)\rho\}}}\sqrt{N} - z_{\alpha/2}\right\} \qquad (4.8)$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. Since equation (4.8) uses Z-score instead of t-score, it could underestimate the assurance probability, a more exact calculation can be obtained by using the t distribution based on the number of clusters k,

$$1 - \beta = \Phi_{t,k-2}\left\{\frac{\text{WinP}(1-\text{WinP})[\text{logit}(\text{WinP}_l) - \text{logit}(\text{WinP})]}{(\sqrt{1 + 1/s)[1 + (m-1)\rho](s\phi_1^2 + \phi_2^2})}\sqrt{N} - z_{\alpha/2}\right\} \qquad (4.9)$$

where $\Phi_{t,k-2}$ is the cumulative distribution function of t-distribution with $k-2$ degrees of freedom. Note that our sample size formula (4.7) can also be applied when the trial aims for hypothesis by specifying $\theta_l = 0.5$, since the exclusion of 0.5 of the confidence interval is the same as rejecting the null hypothesis. Therefore, the assurance probability corresponds to power when $\theta_l = 0.5$.

4.1.1 Estimation of win fractions for sample size planning

The distributions of the outcomes are required to obtain win fractions for sample size planning. We can obtain the distributions of the outcomes depending on which information is available. First, if pilot data is available from both intervention arms we can use them as the hypothetical distributions. Second, if pilot data is available only for the control arm, we can create the hypothetical distribution for the treatment arm using the knowledge of how much the treatment will change the distribution of the outcome. Finally, instead of pilot data, if an outcome distribution is only available for the control arm, we can still create the hypothetical distribution for the treatment arm with background knowledge. For example, we can utilize a distribution shift for continuous outcomes, or a one-category improvement is assumed for a categorical distribution. In this section, we will discuss how to apply our sample size formula (4.7) in those different situations.

4.1.1.1 Using pilot data for both arms

Before a cluster randomization trial starts, pilot data might be available from individually randomization trials. We can then compute the win fractions for each of the individuals as reviewed (Happ *et al.*, 2019) in Chapter 2. We first rank each individual in their own arm and in the combined sample of two arms. The win fraction for an individual is obtained as the percentage of wins comparing their outcome to all outcomes one at a time in the other arm. It is calculated by subtracting the rank within their arm from the rank in the whole sample divided by the size of the opposite arm of the individual. The variance component ϕ_i^2 in equation (4.6) is then obtained as the variance of win fractions for each arm by

$$\phi_i^2 = \frac{\sum_{j=1}^{n_i}(w_{ij} - \overline{w}_{i.})^2}{n_i}$$

where w_{ij} is the win fraction of the jth outcome in the ith arm. We used n_i instead of $n_i - 1$ in the denominator because we treat the data as theoretical.

As an illustration, consider pilot data available with 4 participants in the control arm with outcomes of 1, 5, 5 and 7 and 3 participants with outcomes of 4, 6 and 8 in the treatment arm. We rank each of the outcomes in the combined sample of two arms and subtract it by the rank of the outcomes in its own arm to obtain the number of wins such outcome has compared to the other arm. The win fraction for an outcome is then obtained from dividing the number of wins by the number of comparisons made to such outcome, which is the number of participants in the other arm. The WinP is obtained by the mean win fractions in the treatment arm. The variance components ϕ_i^2 for sample size planning is obtained by the variance of win fractions for arm i. The calculations of win fractions for each outcome are listed in Table 4.1.

Table 4.1: Calculation of win fractions for sample size planning from pilot data.

	Treatment ($n_1 = 3$)	Control ($n_2 = 4$)
Outcome	4, 6, 8	1, 5, 5, 7
Overall rank	2, 5, 7	1, 3.5, 3.5, 6
Arm-specific rank	1, 2, 3	1, 2.5, 2.5, 4
Wins (overall minus arm rank)	1, 3, 4	0, 1, 1, 2
Win fraction	1/4, 3/4, 4/4	0/3, 1/3, 1/3, 2/3
Mean win fraction	2/3	1/3
ϕ_i^2 = Var(win fraction)	$[(1/4 - 2/3)^2 + (3/4 - 2/3)^2 + (4/4 - 2/3)^2]/3 = 0.097$	$[(0/3 - 1/3)^2 + (1/3 - 1/3)^2 + (1/3 - 1/3)^2 + (2/3 - 1/3)^2]/4 = 0.222$
WinP	2/3=0.66	

Using the values in Table 4.1, we can determine the sample size for individually randomized trials using equation (4.7). For example, the sample size in a balanced trial ($s = 1$) with 90% assurance probability of 95% confidence interval excluding 0.5 can be obtained by

$$N = (1 + 1/1)\,(1 \times 0.097 + 0.222)\left\{\frac{z_{0.1} + z_{0.025}}{[\operatorname{logit}(0.66) - \operatorname{logit}(0.5)]\,0.66(1 - 0.66)}\right\}^2 = 302.6\,,$$

yielding 303 participants in total. If the randomization were performed at clusters with a size of 10 and ICC of 0.1 and the required sample size is then increased to $N(1 + (10 - 1)0.1) =$

574.9, yielding 575 participants in total. If we adjust for baseline measurement that have a Spearman correlation with the follow-up outcomes as 0.3 for such cluster trial, the sample size is decreased to $N(1 + 1(10 - 1)0.1)(1 - 0.3^2) = 523.2$, yielding 524 participants in total.

4.1.1.2 Using data available only for the control arm

The intervention for the control arm is usually current standard treatments that have been accessible for the target population for a while, implying the distribution of outcomes in the control arm can be known from previous studies. With such information and the knowledge of how the treatment could change the distribution of outcomes, we can create hypothetical outcomes for the treatment arm based on the control arm outcomes. Most sample size formulas for continuous outcomes in the literature assume a location shift of the outcomes under the alternative hypothesis. However, the location-shift assumption is not always plausible or meaningful in specifying the treatment effect. For example, consider the alcohol use disorders identification test (AUDIT) that evaluates drinking habits and alcohol dependence used in the SIPS trial reported by Kaner *et al.* (2013). The test score ranges from 0 to 40, where there is no clear interpretation of a one-point increase in the AUDIT score, except a greater score indicates more severe drinking. Furthermore, there are no generally accepted cut-offs for the AUDIT score (Reinert and Allen, 2007), which complicates sample size estimation, as different cut-offs could result in different sample sizes.

One way to avoid cut-offs in sample size planning is to assume the intervention results in a change in the distribution of AUDIT score. However, an absolute change could be inappropriate because a reduction from three points to zero points is a considerable improvement, but such a three-point reduction would not be relevant for someone who scored 40. Therefore, a location-shift model cannot meaningfully describe the treatment effect, but a percentage reduction of the score would be more relevant.

As an example, consider five individuals with the following AUDIT score: 4, 16, 20, 24 and 40. If clinicians expect a 25% reduction with a new education program, we can create hypothetical outcomes for the treatment arm by reducing the scores by 25% for the five individuals, which yield 3, 12, 15, 18 and 30, respectively. Since a lower AUDIT score is preferable, we rank the outcomes in reverse order (a higher score has a lower rank), resulting in WinP = 0.68 for this hypothetical data, and the variance components are $\phi_1^2 = \phi_2^2 = 0.0736$ for both arms. Suppose the clinicians wish to detect if the treatment effect is at least 0.64 (corresponding to medium Cohen's effect size for normal outcome) with a 90% chance under 5% level; the required sample size for an individually randomized trial can be obtained by

$$N = (1 + 1/1)\,(1 \times 0.0736 + 0.0736)\left\{\frac{z_{0.1} + z_{0.025}}{[\text{logit}(0.68) - \text{logit}(0.64)]\,0.68(1 - 0.68)}\right\}^2 = 2052.5\,,$$

yielding 2,053 individuals need to be randomized. If the trial randomizes clusters with a size of 100 and ICC of 0.01, the sample size is then increased to $N(1 + (100 - 1)0.01) = 4084.5$, yielding 4085 participants in total, or 42 clusters in total without missing data.

4.1.1.3 Using a categorical distribution available only for the control arm with odds ratio

When the distribution for the control arm and the odds ratios are known, the distribution for the treatment arm can be obtained. The odds for outcome Y being better than category j is defined by $\text{Odds}_j = \Pr(Y > j)/\Pr(Y \leq j)$. The odds ratio for each category could be specified, or proportional odds could be assumed such that the odds ratio is the same regardless of the selected reference category j, which is common in practice (Agresti, 1999). Consider there are q categories in the outcome, and denote $p_1 = (p_{11}, p_{12}, p_{13}, \cdots, p_{1q})$ as the row vector of proportions for the control arm and $p_2 = (p_{21}, p_{22}, p_{23}, \cdots, p_{2q})$ as the row vector of proportions for the treatment arm. The proportion of each category (p_{2j}) for the treatment arm can be obtained from the proportion of the control arm and the odds ratio with

$$p_{2j} = \frac{1 - s_j}{s_j\text{OR} + 1 - s_j} - \sum_{i=1}^{j-1} p_{2i}, \quad j = 1, 2, \cdots, q - 1\,, \tag{4.10}$$

where $s_j = \sum_{i=j+1}^{q} p_{1i}$. The WinP can be obtained by

$$\text{WinP} = p_1 \Omega p_2' , \tag{4.11}$$

where Ω is an upper triangle matrix of ones and halves on the diagonal, and p_i' is the transpose of p_i. Applying delta method to equation (4.11) with respect to p_i' yields the variance formula (Zou *et al.*, 2023)

$$\begin{aligned} \phi_1^2 &= p_2 \Omega' \Sigma_1 \Omega p_2' \\ \phi_2^2 &= p_1 \Omega \Sigma_2 \Omega' p_1' , \end{aligned} \tag{4.12}$$

where $\Sigma_i = \text{diag}(p_i) - p_i' p_i$ is the covariance matrix of the multinomial distribution for p_i. Note that one can derive the same variance formula from win fractions as $p_2\Omega$ and $p_1\Omega$ are the win fraction vectors for control arm and treatment arm, respectively.

The proportional odds model systematically shifts a proportion of the outcomes into a better category, given the odds ratio is greater than one, but the proportion is not the same for all categories. If a clinician expects the treatment to shift a fixed proportion (δ) of subjects into a better category, the proportional odds model cannot be satisfactory. In such case, $p_{2,j} = \delta p_{2,j-1} + p_{2,j}(1 - \delta)$ can be used to construct the distribution for the treatment arm.

As an example, we consider the control arm distribution from the diabetes treatment data (Lachin, 2011) where the outcome is the level of albumin that is classified into normal, micro and macro by the severity. Assuming a common odds ratio of 3, we can use equation (4.10) to obtain the distribution for the treatment arm and use equations (4.12) to obtain the win fractions, listed in Table 4.2. In this data, a lower albumin level is preferred, and the WinP is 0.60 with $\phi_1^2 = 0.031$ and $\phi_2^2 = 0.058$. Suppose the minimum WinP is 0.56, corresponding to a small Cohen's effect size for normally distributed outcome. An individually randomized trial would need 1,830 participants to be randomized to have a 90% power at 5% level, and a total of $3,642$ participants are required if the trial randomizes clusters of size 100 and ICC = 0.01.

We have discussed different ways to obtain the hypothetical distribution when the treatment has different effects on the outcome distribution. Our method unifies the calculation of sample size regardless of the distributional assumptions, whereas it generally requires different sample size formulas in the literature.

Table 4.2: Listing of the categorical probability of albumin level and its win fractions in the parentheses assuming the common odds ratio is 3.

	Albumin level		
Group	Normal	Micro	Macro
Control	0.85 (0.325)	0.10 (0.705)	0.05 (0.93)
Treatment	0.65 (0.425)	0.21 (0.900)	0.14 (0.975)

4.2 Design considerations

4.2.1 Baseline adjustment

When baseline assessment is anticipated in the analysis of a trial, accounting for the correlation between the baseline and follow-up can reduce the required sample size of the trial while maintaining the same accuracy. We have shown that the efficiency gained from baseline adjustment depends on the correlation between baseline and follow-up win fractions of the cluster level in Chapter 3.

Denote the temporal correlation of cluster-specific mean win fractions between baseline ($\overline{w}^X_{ij.}$) and follow-up ($\overline{w}^Y_{ij.}$) as $r_c = \mathrm{Cor}(\overline{w}^X_{ij.}, \overline{w}^Y_{ij.})$ and the correlation of individual-specific win fraction as $r = \mathrm{Cor}(w^X_{ijl}, w^Y_{ijl})$. Using the weighted least square approach, baseline adjustment reduces the variance of $\widehat{\mathrm{WinP}}^*$ to

$$\mathrm{Var}(\widehat{\mathrm{WinP}}^*) = \frac{1}{N}(1 + 1/s)(s\widehat{\phi_1^2} + \widehat{\phi_2^2})\{1 + (m-1)\rho\}(1 - r_c^2)\,.$$

Hence the sample size required to maintain the same assurance probability would be

$$N^* = (1 - r_c^2)N ,$$

where N is the sample size without baseline adjustment from equation (4.7).

The efficiency gained from baseline adjustment in the mixed model approach depends on the correlation between baseline and follow-up win fraction on both the cluster- and individual-level (Teerenstra *et al.*, 2012). To be specific, the correlation from the mixed model weights cluster- and individual-level correlation by

$$r_a = \frac{m\rho}{1 + (m-1)\rho} r_c + \frac{1 - \rho}{1 + (m-1)\rho} r , \tag{4.13}$$

where ρ is the ICC of follow-up win fractions and m is the cluster size. The corresponding sample size with such an analysis strategy is $N^* = (1 - r_a^2)N$. When the temporal correlation for individuals and the cluster means are the same, the efficiency of the weighted least square approach is the same as the efficiency of the mixed model approach.

4.2.2 Temporal correlation of win fractions

Our sample size formula is derived with the temporal correlation and ICC of win fractions; however, these are rarely available in practice but the temporal correlation and ICC could be available on the original scale. There is no closed-form relationship between the correlation of win fractions and the correlation of the original scale in most cases. The relationship may be examined with simulations. Since win fractions are obtained from ranks instead of the original scale, it is possible to derive the equivalence of the Spearman correlation of the original scale (Pearson correlation formula applied to ranks) and Pearson correlation of win fractions for individually randomized trials.

To be specific, the Pearson correlation of the win fractions is equivalent to the Spearman correlation of the original scale under the null hypothesis (WinP = 0.5). Suppose there are

N subjects randomized with the treatment to control ratio s, with $n_1 = N/(1+s)$ subjects in the control arm and $n_2 = Ns/(1+s)$ subjects in the treatment arm, where both are assumed to have integer values. Denote the outcome for baseline as $X_{ij} \sim G$, $i = 1,2$, $j = 1,2,\cdots,n_i$ and for follow-up as $Y_{ij} \sim F_i$. We drop the subscript for G because randomization implies the distribution of baseline measurements is the same for both arms. The combined distribution of the follow-up is hence,

$$F(x) = f_1F_1(x) + f_2F_2(x)\,,$$

where $f_1 = 1/(1+s)$ is the fraction of control arm and $f_2 = s/(1+s)$ is the fraction of treatment arm. The derivations in this section would make more sense and intuitive by using f_i instead of s. Therefore, we will use f_i for the rest of this section. The rank of subject j within its arm at baseline is denoted by r^X_{ij} and rank of the same subject in the whole sample combining both arms by R^X_{ij}. Similarly, we denote the within-arm rank and total rank for follow-up as r^Y_{ij} and R^Y_{ij}, respectively. The Spearman correlation r_s is estimated by

$$\hat{r}_s = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(R^X_{ij} - \overline{R}^X_{..})(R^Y_{ij} - \overline{R}^Y_{..})}{\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(R^X_{ij} - \overline{R}^X_{..})^2}\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(R^Y_{ij} - \overline{R}^Y_{..})^2}}\,.$$

Since the rank is related to the empirical distribution function by $R^X_{ij} = 0.5 + N\hat{G}(X_{ij})$ and $R^Y_{ij} = 0.5 + N\hat{F}(Y_{ij})$, and the mean overall rank is only related to sample size $\overline{R}^X_{..} = \overline{R}^Y_{..} = (1+N)/2$, the Spearman correlation can be written in terms of the empirical distribution function,

$$\hat{r}_s = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\widehat{G}(Y_{ij}) - 0.5\right)\left(\widehat{F}(X_{ij}) - 0.5\right)}{\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\widehat{G}(Y_{ij}) - 0.5\right)^2}\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\widehat{F}(X_{ij}) - 0.5\right)^2}}\,.$$

Using the win fractions for X_{ij}, denoted as w^X_{ij}, and for Y_{ij}, denoted as w^Y_{ij}, the Pearson correlation of win fractions can be written as

$$
\begin{aligned}
\hat{r}_p &= \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(w_{ij}^{X}-\overline{w}_{..}^{X})(w_{ij}^{Y}-\overline{w}_{..}^{Y})}{\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(w_{ij}^{X}-\overline{w}_{..}^{X})^2}\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n}(w_{ij}^{Y}-\overline{w}_{..}^{Y})^2}}\\
&= \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\frac{R_{ij}^{X}-r_{ij}^{X}}{N-n_i}-\overline{w}_{..}^{X}\right)\left(\frac{R_{ij}^{Y}-r_{ij}^{Y}}{N-n_i}-\overline{w}_{..}^{Y}\right)}{\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\frac{R_{ij}^{X}-r_{ij}^{X}}{N-n_i}-\overline{w}_{..}^{X}\right)^2}\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\frac{R_{ij}^{Y}-r_{ij}^{Y}}{N-n_i}-\overline{w}_{..}^{Y}\right)^2}}\\
&= \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\frac{N\widehat{G}(X_{ij})-\widehat{G}_i(X_{ij})}{N-n_i}-\overline{w}_{..}^{X}\right)\left(\frac{N\widehat{F}(Y_{ij})-\widehat{F}_i(Y_{ij})}{N-n_i}-\overline{w}_{..}^{Y}\right)}{\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\frac{N\widehat{G}(X_{ij})-\widehat{G}_i(X_{ij})}{N-n_i}-\overline{w}_{..}^{X}\right)}\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\frac{N\widehat{F}(Y_{ij})-\widehat{F}_i(Y_{ij})}{N-n_i}-\overline{w}_{..}^{Y}\right)^2}}\\
&= \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\widehat{G}(X_{ij})-0.5\right)\left(\frac{1}{1-f_i}\widehat{F}(Y_{ij})-\frac{f_i}{1-f_i}\widehat{F}_i(Y_{ij})-\overline{w}_{..}^{Y}\right)}{\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\widehat{G}(X_{ij})-0.5\right)^2}\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\frac{1}{1-f_i}\widehat{F}(Y_{ij})-\frac{f_i}{1-f_i}\widehat{F}_i(Y_{ij})-\overline{w}_{..}^{Y}\right)^2}}.
\end{aligned}
$$

The Pearson correlation of win fractions is also closely related to the Pearson correlation based on ranks but has a term that depends on the WinP ($\overline{w}_{..}^{Y}$). When no treatment effect is present, i.e $F = F_1 = F_2$, the two correlations are the same $\widehat{r_p} = \widehat{r_s}$ since $\overline{w}_{..}^{Y} = 0.5$. However, they are not the same under WinP $\neq 0.5$ because the combined distribution F depends on the magnitude of WinP. We evaluated their relationship via a simulation study in Chapter 5.

4.3 Summary and discussion

In this chapter, we developed sample size formulas for cluster randomization trials focusing on the confidence interval estimation of WinP. The sample size estimation can be summarized into three steps. The first step is to obtain the variance of win fractions for each arm and estimate the sample size as if independent individuals are randomized. The second step is to apply formula for individually randomized trial Zou *et al.* (2023), and finally multiply the sample size for independent outcomes by the design effect of randomizing clusters (and baseline adjustments). Our work extends the sample size planning of WinP for individually randomized trials by Zou *et al.* (2023) to cluster randomization trials.

There are several sample size formulas for cluster randomization trials in the literature, as reviewed by Rutterford *et al.* (2015) and Gao *et al.* (2015). The formulas in the literature depend on the distribution of the outcome because different parametric assumptions are required to analyze those outcomes. However, our formula can be used for any type of outcome without any distributional assumptions. Additionally, our formula is derived for trials focusing on effect estimation, whereas other formulas in the literature are mainly derived for hypothesis testing.

Our sample size formula is derived based on the correlation and ICC of win fractions, which are rarely known. The correlation and ICC of the original scale can be used with our sample size formula for practicability, similar to the approach proposed by Zou *et al.* (2023) for sample size estimation of individually randomized trials. Such an approach performed well for individually randomized trials in simulation studies (Zou *et al.*, 2023), but its performance for cluster randomization trials still needs to be investigated.

Chapter 5

Simulation study

We have proposed estimators of win probability (WinP) and its variance with baseline adjustments in Chapter 3 and thee corresponding sample size formulas in Chapter 4. We have shown the (co)variance estimators based on the weighted least square estimators are consistent but slightly biased in finite samples. In this chapter, we conduct simulation studies to evaluate the performance of confidence intervals of WinP proposed in Chapter 3 and the performance of sample size formulas proposed in Chapter 4 in finite sample settings. Our simulation studies have the following specific objectives:

1. Evaluate the performance of the confidence intervals for WinP in terms of empirical coverage rate and the empirical width. Confidence intervals based on the four variance estimator of adjusted WinP are compared to the confidence interval of unadjusted WinP.

2. Evaluate the validity of our sample size formula based on the empirical assurance probability and empirical coverage rate of the confidence intervals constructed from data generated with the size from our sample size formula.

3. Examine the relationship between temporal correlation of the raw outcome and its temporal correlation of win fractions.

We organize the rest of this chapter as follows. Section 5.1 provides scenarios of real-world cluster randomization trials. Section 5.2 describes the data generation model. We present the results of our simulation study in three sections. Section 5.3 focuses on evaluating the methods proposed in Chapter 3, confidence intervals for WinP adjusting for baseline measurement. Section 5.4 evaluates the sample size formula proposed in Chapter 4. Section 5.5 focuses on assessing the relationship between the correlation of ranks of the raw scale (Spearman correlation) and the correlation of win fractions. We discuss and summarize our findings of the simulation study at the end of this chapter.

5.1 Simulation settings

The methods in Chapters 3 and 4 focus on two-arm cluster randomization trials with baseline measurements by assuming large number of clusters. The performance of our methods may be affected by several factors, including the total number of clusters in the trial, average cluster size, the intraclass correlation coefficient (ICC), the magnitude of treatment effect and the correlation between baseline measurement and follow-up outcome. We will generate outcomes from continuous, binary and ordered category distributions to assess the robustness under finite samples.

As suggested by reviews of cluster randomization trials (Simpson *et al.*, 1995; Varnell *et al.*, 2004), an average cluster size of 50 should suffice to represent a moderate cluster size. We consider an average cluster size of 25 as a smaller cluster size. We generate variable cluster size from binomial distribution with $n = 50, 100$ and $p = 0.5$, which results in an average cluster size of 25 and 50, respectively. The variability of cluster size is often measured by the coefficient of variation (CV) of cluster size, which is the standard deviation of cluster sizes over mean cluster size. The CV of cluster size generated from binomial distribution is hence $\text{CV} = \sqrt{(1-p)/(np)}$ and we have $\text{CV} = 0.14$ and 0.1 for cluster size from binomial $n = 50, 100$ and $p = 0.5$, respectively. For cluster size with higher variability, cluster size was generated from a discrete uniform distribution $U(l, u)$, following Zou *et al.* (2005). The mean and variance of the distribution $U(l, u)$ are $(l + u)/2$ and $(u - l)(u - l + 2)/12$ respectively, implying the coefficient of variation is

$$\mathrm{CV} = \sqrt{(u - l)(u - l + 2)/3}/(u + l)\,.$$

The closest integers of l, u are chosen as $l = 24$ and $u = 76$ to have a mean of 50 and CV of 0.3. A higher coefficient of variation such as 0.5 can result in clusters with fewer than 10 participants; hence, it is not considered.

The ICC is set to be 0.01, 0.05 and 0.1 because these are more common in practice for clusters as socially intact units (Eldridge *et al.*, 2004). Higher ICC values are usually found in studies with small cluster sizes but have more clusters to satisfy the conditions for large sample theories.

We consider both balanced and unbalanced cluster assignments to treatment arms in our simulation. For the balanced case, we consider 5 clusters in each arm to evaluate the performance of our methods under a small sample size. We also consider 15 clusters in each arm as this is more commonly used from a review of 152 trials (Eldridge *et al.*, 2004). We will also consider 10 clusters in one arm and 20 clusters in the other arm representing the unbalanced assignment of clusters. The performance under those cases will be compared to the results with 15 clusters per arm to determine the consequence of unbalanced designs.

We will use Cohen's effect size of 0.2, 0.5, and 0.8 as a small to large effect size, which is equivalent to WinP $= 0.56, 0.64$, and 0.71, respectively, for outcomes following a normal distribution. We will also use these three values of WinP for binary and ordered category outcomes in our simulation study. Although small to medium effect sizes are more common in practice, we will still consider the large effect size in evaluating the performance of interval estimation of WinP from Chapter 3. However, for sample size estimation, we will only consider WinP $= 0.56, 0.60$ because WinP ≥ 0.64 will yield sample sizes too small to conduct a cluster randomization trial.

The correlation between baseline and follow-up at individual level is assumed to be 0.3 and 0.5 as weak to medium correlation. Cluster randomization trials usually have more extended follow-up periods than individually randomized trials hence a strong correlation of 0.7 is much less likely. Spearman correlation is used because we focus on ordinal outcomes, where rank is the only available information to compare outcomes. The parameters used in the simulation study are summarized in Table 5.1.

Table 5.1: Parameters used to generate baseline measurement and outcomes for a two-arm parallel cluster randomization trial.

Parameter	values
WinP[a]	0.56, 0.64, 0.71
Total number of clusters (k)	10[b],30
Average cluster size (n)	25, 50
Cluster size CV[c]	0, 0.1, 0.14, 0.3
ICC[d](ρ)	0.01, 0.05, 0.1
Correlation[e](r)	0.3, 0.5
Randomization ratio[f]	0.5, 1, 2

[a]WinP = $\Pr(Y_1 < Y_2)+0.5\Pr(Y_1 = Y_2)$. [b] only considered for randomization ratio = 1. [c]CV: coefficient of variation. [d]ICC: intraclass correlation coefficient. [e] Spearman correlation coefficient between baseline and follow-up. [f] the ratio of number of subjects in treatment arm over the control arm.

We now describe the covariance matrix used to simulate baseline measurement and follow-up outcomes within a cluster. Denote ρ as the ICC, which is assumed to be the same for both follow-up and baseline assessments, and r as the individual temporal correlation, which is the correlation of baseline and follow-up within the same individual. The correlation of the baseline measurements and follow-up outcomes within a cluster consisting of n subjects can be expressed by the Kronecker product $\mathbf{\Sigma} = \mathbf{R} \otimes \boldsymbol{P}$, where

$$\mathbf{R} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix},$$

and ***P*** a diagonal n × n matrix with one on the diagonals and ρ on the off-diagonal,

$$\mathbf{P} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}.$$

Hence, $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & \cdots & \rho & r & r\rho & \cdots & r\rho \\ \rho & 1 & \cdots & \rho & r\rho & r & \cdots & r\rho \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & r\rho \\ \rho & \rho & \cdots & 1 & r\rho & r\rho & \cdots & r \\ r & r\rho & \cdots & r\rho & 1 & \rho & \cdots & \rho \\ r\rho & r & \cdots & r\rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & r\rho & \vdots & \vdots & \ddots & \vdots \\ r\rho & r\rho & \cdots & r & \rho & \rho & \cdots & 1 \end{bmatrix} \tag{5.1}$$

The entries in the first n rows and n columns in Σ are correlations for baseline measurement and the entries in the last n rows and n columns are correlation between follow-up outcomes, where n is the cluster size. The other entries are the temporal correlation (same subject) or the autocorrelation (different subject) of baseline measurement and follow-up outcome. We assumed autocorrelation to be $r\rho$ because it should be weaker than both r and ρ. A higher autocorrelation results in simulated outcomes with higher cluster-level temporal correlation. We additionally consider higher autocorrelation as 0.9ρ for ordered category outcomes to simulate outcomes where cluster-level temporal correlation tends to be higher than individual-level temporal correlation.

Based on the value of WinP, continuous outcomes are generated from a multivariate normal distribution with unit standard deviation and the correlation matrix as described above. We assign the mean for control group as 0 and use the relation

$$\mathrm{WinP} = \Phi\left(\frac{u_2}{\sqrt{2}}\right)$$

to obtain the mean for the treatment group (u_2), which are 0.21, 0.51 and 0.78 corresponding to WinP = 0.56, 0.64 and 0.71, respectively. For binary outcomes we consider the event rate as 0.1 for control group and use the relation

$$\mathrm{WinP} = \frac{p - 0.1 + 1}{2}$$

to obtain the event rate p for the treatment group, which are 0.22, 0.38 and 0.52 corresponding to WinP = 0.56, 0.64 and 0.71, respectively.

We consider ordered category outcomes with five categories because Likert-like scales often have five or seven categories. To generate ordered category outcomes, we specify the probabilities for each category for the control arm with the binomial distribution (4,0.5), where the first category has probability of zero success from the binomial distribution, second category has probability of one success, and so on. The probabilities of each category for the treatment arm are specified by shifting some higher categories into the next category to obtain the desired WinP as shown in Table 5.2.

Table 5.2: Probabilities of each category to generate ordered category outcomes with five categories.

Category	1	2	3	4	5
Control (q)	0.0625	0.2500	0.3750	0.2500	0.0625
Treatment (p)					
WinP[a] = 0.56	0.0600	0.2500	0.2500	0.2560	0.1840
WinP = 0.64	0.0080	0.1000	0.4000	0.3560	0.1360
WinP = 0.71	0.0080	0.1000	0.2000	0.5080	0.1840

[a] The WinP is obtained by WinP= $q\Omega p'$, where q and p are the row vector of probabilities of each category for the control arm and treatment arm respectively, and Ω is the upper triangle matrix of ones but half on the diagonal.

5.2 Generation of ordered data

Correlated ordered category outcomes and correlated binary outcomes are generated by the mean mapping method (Kaiser *et al.*, 2011). The method generates correlated normal outcomes and transforms them into ordered category outcomes by cutting off the normal outcome with quantiles. As an example of the cut-off, suppose we intend to generate an ordered category outcome of three categories with the probability of each category as $(0.1, 0.3, 0.6)$. We first calculate the cut-offs on a standard normal distribution such that the interval between cut-offs forms an area with the corresponding probability. The cut-offs for probabilities $(0.1, 0.3, 0.6)$ are shown in Table 5.3.

Table 5.3: Example ordered category proportion and its cut-off points

Category	1	2	3
Proportion	0.1	0.3	0.6
Cumulative proportion	0.1	0.4	1
Cut-off	$\Phi^{-1}(0.1) = -1.28$	$\Phi^{-1}(0.4) = -0.25$	$\Phi^{-1}(1) = \infty$

The next step is to sample from a standard normal distribution and transform it into an integer corresponding to the ordered categories based on which interval of cut-offs it is in. For example, suppose -1 is randomly sampled from a standard normal distribution, instead of keeping the outcome as -1 it is converted into 2 because -1 lies within the interval of $(-1.28, -0.25)$.

To illustrate the mechanisms of the mean mapping method in generating correlated ordered categories outcomes, we first consider the case where only two outcomes are generated and we will extend it to the general case with any number of correlated outcomes. To generate ordered category outcomes X and Y of q categories that have correlation r, we first draw from a bivariate normal distribution with mean $\mathbf{0}$ and covariance matrix with ones on the diagonals and an arbitrary chosen value p on the off-diagonal entries. The next step is to transform them into ordered categories with the cut-offs and calculate the correlation of the ordered category outcomes. Since X and Y are generated from bivariate normal distribution, the joint probability can be calculated by the cumulative function ($\Phi(x, y, p)$) of the bivariate normal distribution by

$$\Pr(X = x, Y = y) = \Phi(q_x, q_y, p) - \Phi(q_{x-1}, q_y, p) - \Phi(q_x, q_{y-1}, p) + \Phi(q_{x-1}, q_{y-1}, p) \qquad (5.2)$$

where q_x denotes the cut-off quantile corresponding to category x. The correlation between X and Y are $r' = E[XY] - E[X]E[Y]$, where Kaiser *et al.* (2011) derived an approximation for $E(XY)$ given as

$$E(XY) = r' \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)} + E(X)E(Y) - q^2 + q \sum_{x=1}^{q-1} F_X(x) + q \sum_{y=1}^{q-1} F_Y(y), \qquad (5.3)$$

which depends only on r' and the cumulative distributions of X and Y denoted as $F_X(x)$ and $F_Y(x)$, respectively.

Combining the relation in equation (5.2) and (5.3) through

$$\begin{aligned} E[r'] &= E[XY] - E[X]E[Y] \\ &= \sum_{x=1}^{q} \sum_{y=1}^{q} xy\Pr(X = x, Y = y) - \left(\sum_{x=1}^{q} x\Pr(X = x) \right) \left(\sum_{y=1}^{q} y\Pr(Y = y) \right) \end{aligned}$$

implies that the correlation of the ordered category outcomes r' can be determined by p and the joint and marginal probabilities of X and Y. Due to the complexity of the relation, the mean mapping method calculates r' by using equation (5.2) and (5.3) on a grid of possible values of p from -1 to 1 incremented by 0.01. Out of the 200 values of p, the mean mapping method finds the one that provides r' closest to the desired r. The mean mapping method was showed to be a valid and reliable method to generate correlated ordered category outcomes by simulation studies (Kaiser *et al.*, 2011). Generating more than two correlated outcomes follows by sampling from a multivariate normal distribution instead of a bivariate normal distribution.

Generating ordered category outcomes for cluster randomization trials requires three correlation parameters: (i) the individual temporal correlation r, (ii) the intraclass correlation coefficient ρ, and (iii) the autocorrelation p. The mean mapping method solves correlation parameters one by one with the method we described above and then generates correlated ordered category outcomes with the desired correlation structure. Generating the baseline measurements and follow-up outcomes for a cluster of size *m* requires sampling from a 2*m*-dimension multivariate normal distribution with the covariance matrix shown in Equation (5.1).

5.3 Performance of interval estimation

We conducted a simulation study to evaluate the performance of interval estimation of WinP. Four variance estimators of the WinP adjusted baseline are used: three variance estimators are based on the weighted least square approach from equation (3.3 and 3.4), which are cluster size weighted estimator in equation (3.7), ICC weighted estimator in equation (3.6) and ratio estimator in equation (3.9), the fourth variance estimator is based on the mixed model approach (3.13), which regresses the follow-up win fractions on the baseline win fractions in a mixed model. Arsinh transformed confidence interval from equation (3.21) are constructed using the four different variance estimates. The logit transformed confidence interval from equation (3.19) have similar coverage to arsinh transformed intervals but are wider; hence, we will only present the results of the arsinh transformed intervals.

For confidence interval estimation, the most important performance measure in a simulation study is the empirical coverage as it measures the validity of the method. This criterion is crucial for the WinP because the range of restriction and skewness of sampling distribution can cause imbalanced tail errors. Another important performance measure for interval estimation is the average width of the interval. Shorter interval width implies the study has higher efficiency which is desirable for randomized controlled trials due to its high cost. Our proposed methods use covariance analysis of the follow-up and baseline to increase efficiency based on large-sample results (Teerenstra *et al.*, 2012). It is unclear whether the efficiency gain is similar for small samples. We will compare our confidence intervals to the intervals without baseline adjustment to calculate the efficiency gained from baseline adjustment. All confidence intervals in this section are calculated at 5% significant level. The simulation study is conducted with 1,825 replicates such that the empirical coverage rate is acceptable if it is between 94% and 96% because $0.95 \mp 1.96\sqrt{0.95 \times 0.05/1825} = (0.94, 0.96)$.

The simulation results for ordered categorical outcomes, continuous outcomes and binary outcomes are presented in the followings.

5.3.1 Ordered category outcomes

Table 5.5 presents the simulation results for ordered category outcomes with five clusters per arm. All four baseline adjusted confidence intervals showed acceptable coverage rates with very few entries outside the range of (94%, 96%). The average confidence interval width from the mixed model approach is the narrowest among the four baseline-adjusted confidence intervals. Baseline adjustment by the weighted least square approach only yielded narrower confidence intervals under medium correlation ($r = 0.5$).

Table 5.6 presents the simulation results for 15 clusters per arm. All four baseline-adjusted confidence intervals showed acceptable coverage rates. The mixed model approach yielded the narrowest confidence interval compared to the other three baseline adjusted intervals, but the difference is smaller compared to Table 5.5 with five clusters per arm. Baseline adjustment yielded efficiency regardless of the analysis method (mixed model or weighted least square) or the scenarios.

Table 5.7 presents the simulation results for 20 clusters in the control arm and 10 clusters in the treatment arm. Confidence intervals from the weighted least square approach are more likely to overshoot the coverage rate (> 96%), and intervals from the mixed model approach are more likely to undershoot the coverage rate (< 94%). However, most entries are still within the acceptable range of (94%, 96%). The mixed model approach yielded the most narrow confidence intervals compared to the intervals from the weighted least square approach. The average width of confidence intervals from the weighted least square approach is only narrower than the unadjusted intervals under medium correlation ($r = 0.5$). Comparing confidence interval widths between Table 5.7 and Table 5.6 shows that the imbalance design increases the confidence interval width.

Table 5.8 presents the simulation results for 10 clusters in the control arm and 20 clusters in the treatment arm. All four baseline adjusted confidence intervals showed satisfactory coverage rates, with almost no entries outside the range $(94\%, 96\%)$. Similar to previous findings, the mixed model approach yields confidence intervals with the narrowest width, and the weighted least square approach only gains efficiency under medium correlation ($r = 0.5$). Comparing the confidence interval width between Tables Table 5.6 to Table 5.8, we observe that a balanced design is the most efficient design, and having more clusters in the treatment arm is more efficient than having more clusters in the control arm.

Table 5.9 presents the simulation results for 15 clusters per arm with high cluster size variability (coefficient of variation as 0.3) and high autocorrelation (0.9ρ) or low autocorrelation ($r\rho$). All four baseline adjustment methods have slightly more entries of coverage outside the acceptable range $(94\%, 96\%)$, with the intervals from the size-weighted estimator performing the worst. The weighted least square approach has higher efficiency than the mixed model approach under high autocorrelation, and the mixed model approach has higher efficiency under low autocorrelation. This result is expected from the theoretical results in Chapter 3, where the efficiency gained from baseline adjustment depends only on cluster-level correlation for the weighted least square approach, but the efficiency depends on both cluster- and individual-level correlation for the mixed model approach.

We presented the efficiency of our confidence intervals adjusted by baseline measurement compared to unadjusted intervals for balanced design with different autocorrelation (resulting in different cluster-level correlation) in Figure 5.1. We can observe that the weighted least square approach gains more efficiency compared to the mixed model approach under high autocorrelation, or high cluster-level correlation. Otherwise, the mixed model approach usually gains more efficiency compared to the weighted least squared approach.

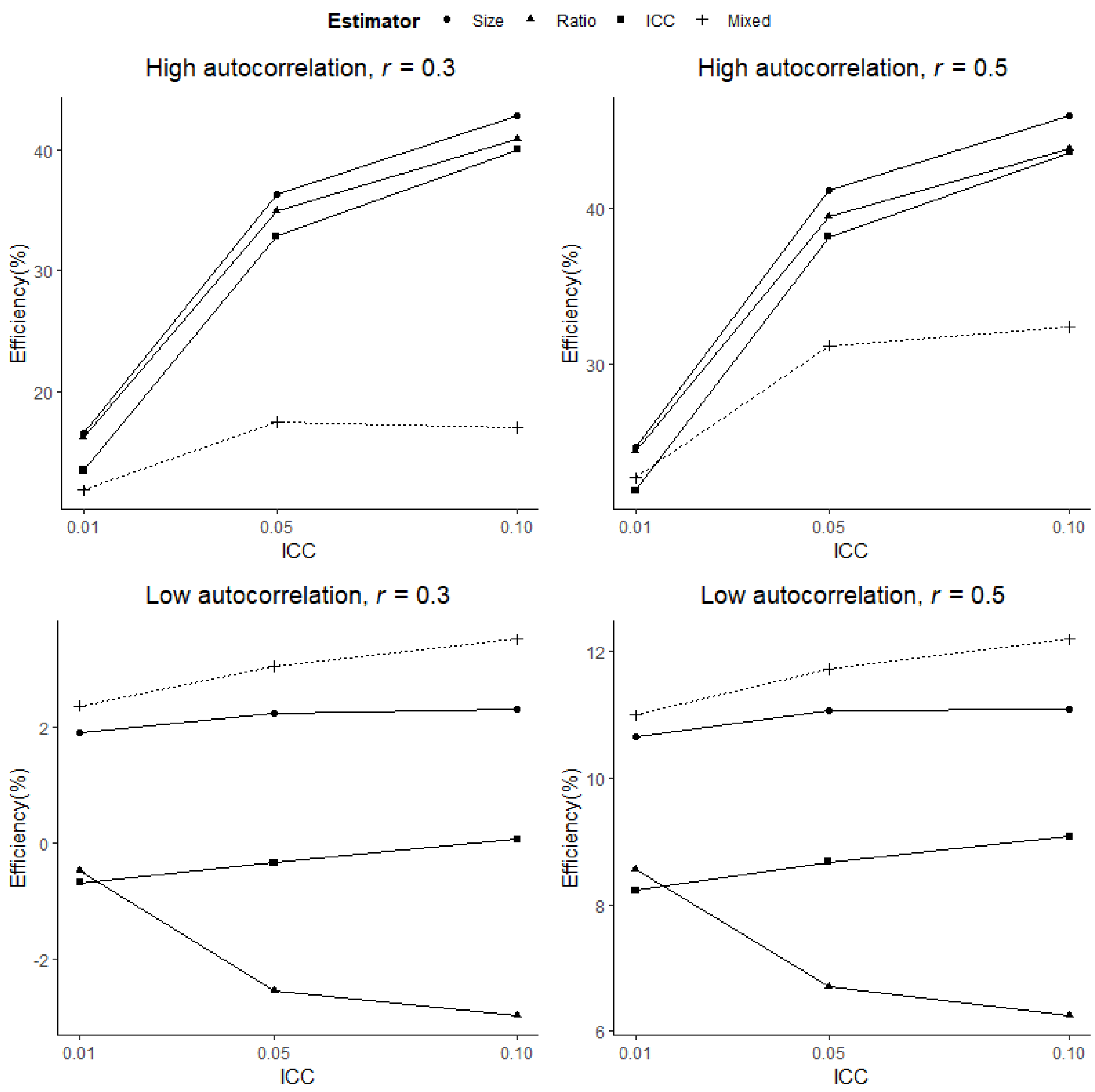

Figure 5.1: Average efficiency (% reduction in interval width) of the confidence intervals from the weighted least square approach (solid line) and the mixed model approach (dash line) compared to interval without baseline adjustment. There are 15 clusters for each arm, with the cluster size generated from uniform distribution (24,76).

5.3.2 Continuous outcomes

Table 5.10 presents the simulation results for continuous outcomes with five clusters per arm. All four baseline-adjusted confidence intervals showed acceptable coverage rates. The average confidence interval width from the mixed model approach is the narrowest among the four baseline-adjusted confidence intervals. Baseline adjustment by the weighted least square approach only yielded narrower confidence intervals under medium correlation ($r = 0.5$).

Table 5.11 presents the simulation results with 15 clusters per arm. All four baseline adjusted confidence intervals showed satisfactory coverage rates with no entries outside the range of $(94\%, 96\%)$. Similarly to previous results, the mixed model yields the most narrow confidence intervals. Baseline adjustment with the weighted least square approach yields narrower confidence intervals compared to unadjusted intervals regardless of the scenario.

Table 5.12 presents the simulation results for 20 clusters in the control arm and 10 clusters in the treatment arm. All four baseline-adjusted confidence intervals showed acceptable coverage rates. The mixed model approach yielded the most narrow confidence intervals, and baseline adjustment by the weighted least square approach only yields narrower confidence intervals under medium correlation ($r = 0.5$).

Table 5.13 presents the simulation results for 10 clusters in the control arm and 20 clusters in the treatment arm. All four baseline adjusted confidence intervals showed acceptable coverage rates, with a few entries of the mixed model approach falling below 94%. Similar to previous findings, the mixed model approach yields confidence intervals with the narrowest width, and the weighted least square approach only gains efficiency under medium correlation ($r = 0.5$). Comparing the confidence interval width between Tables 5.11 to 5.13, we observe that a balanced design is the most efficient design, and having more clusters in the treatment arm is more efficient than having more clusters in the control arm.

5.3.3 Binary outcomes

Table 5.14 presents the simulation results for binary outcomes with five clusters per arm. All four baseline adjusted confidence intervals and unadjusted intervals have a few entries of coverage falling below 94%. Under-coverage is more common for unadjusted intervals and adjusted intervals from the mixed model approach. The average confidence interval width from the mixed model approach is the narrowest among the four baseline-adjusted confidence intervals. Baseline adjustment by the weighted least square approach does not yield narrower confidence interval width, even for medium correlation (r = 0.5).

Table 5.15 presents the simulation results with 15 clusters per arm. All four baseline adjusted confidence intervals showed acceptable coverage rates, but the size-weighted estimator and the mixed model approach yielded a few more entries of coverage below 94%. Similarly to previous results, the mixed model yields the most narrow confidence intervals. Baseline adjustment with the weighted least square approach yields narrower confidence intervals compared to unadjusted intervals regardless of the scenario.

Table 5.16 presents the simulation results for 10 clusters in the control arm and 20 clusters in the treatment arm. All four baseline-adjusted confidence intervals showed acceptable coverage rates. The mixed model approach yielded the most narrow confidence intervals, and baseline adjustment by the weighted least square approach only yields narrower confidence intervals under medium correlation (r = 0.5).

Table 5.17 presents the simulation results for 20 clusters in the control arm and 10 clusters in the treatment arm. Unadjusted confidence intervals and intervals from the mixed model approach have more under-coverage entries compared to intervals from the weighted least square approach. Similar to previous findings, the mixed model approach yields confidence intervals with the narrowest width, and the weighted least square approach only gains efficiency under medium correlation (r = 0.5). Comparing the confidence interval width between Tables 5.15 to 5.17, we observe that a balanced design is the most efficient design, and having more clusters in the treatment arm is more efficient than having more clusters in the control arm.

5.4 Performance of sample size estimation

The performance of the sample size formula (equation 4.7) is evaluated using the empirical assurance probability (i.e., the lower limit of 2-sided 95% CI for WinP being above 0.5). We considered the nominal assurance probabilities of 80% and 90%. The performances of the weighted least square and mixed model approaches are different due to the difference between cluster-level and individual-level temporal correlation. We evaluate the sample size formulas using both approaches in terms of empirical coverage and assurance probability, where the ICC weighted variance estimator is used to construct arsinh transformed confidence interval.

In this section, the most critical performance measures is the empirical assurance probability, although we also consider the empirical coverage. All the confidence intervals in this section are calculated at 5% significant level. Each entries are calculated with 1,825 replicates such that the empirical coverage rate is acceptable if it is between 94% and 96% because $0.95 \mp 1.96\sqrt{0.95 \times 0.05/1825} = 0.94, 0.96$, and the empirical assurance probability is satisfactory if between 78.2% and 81.8% for 80% nominal assurance probability, and between 88.6% and 91.3% for 90% nominal assurance probability.

We assume the outcome has five ordered categories with the probabilities of each category presented in Table 5.4. We then use equation (4.7) to calculate the sample size. Since the temporal correlation of cluster means is often unknown in practice, we use the individual temporal correlation on the raw scale in calculating sample size. Similarly, we use the ICC of the raw scale in sample size calculation.

Although the number of clusters was suggested to be increased by two for studies with 5% level when 30 or fewer clusters are randomized, we do not include such adjustment in our simulation study to confirm whether or not such adjustment is necessary. Hence, we will additionally examine the performance of our sample size formula when less than 30 clusters are randomized.

Table 5.4: Probabilities of each categories used to plan for the sample size.

Category	1	2	3	4	5
Control (q)	0.0625	0.2500	0.3750	0.2500	0.0625
Treatment (p)					
WinP = 0.56	0.0430	0.1914	0.3627	0.3128	0.0901
WinP = 0.60	0.0332	0.1564	0.3415	0.3543	0.1146

[a] The WinP is obtained by WinP= $q\Omega p'$, where q and p are the row vector of probabilities of each categories for the control arm and treatment arm respectively, and Ω is the upper triangle matrix of ones but half on the diagonal.

Table 5.18 presents the results for 80% assurance probability for sample size estimation.

Both analysis methods maintained the assurance probability in most parameter combinations when there are more than 20 clusters in the trial. Using the mixed model approach to analyze the data results in a higher empirical assurance probability compared to the weighted least square approach. This is coherent with previous results that the mixed model approach tends to yield narrower confidence intervals.

Table 5.19 presents the results for 90% assurance probability for sample size estimation. Assurance probability are maintained in most entries as most scenario results in sample size more than 20 clusters. In Tables 5.18 and 5.19, we can see that our sample size formula generally has acceptable performance when over 30 clusters are randomized. The empirical assurance probability may be below the acceptable range when less than 30 clusters are randomized; hence, increasing the total number of clusters by two for studies with 95% level and by four for studies with 99% level could better maintain the assurance probability for studies under 30 clusters (Donner and Klar, 2000, p.67).

The empirical assurance probabilities under different number of clusters are graphically presented in Figure 5.2, where it can be seen that the mixed model approach generally has a higher empirical assurance probability. This may be explained by the consistency of efficiency gained for the mixed model approach under a small sample size, as observed previously.

5.5 Correlation on raw scale and win fractions

Our sample size formula was derived based on the Pearson correlation of the win fractions, but such information is often not available in practice. The correlation on the original scale may be used for the estimation of sample size to account for baseline adjustment and clustering. Although simulation results for the sample size formula were satisfactory, exploring the relationship between the two correlations is still beneficial for future research. It also gives the researcher more confidence to use the correlation of the original scale for the purpose of sample size planning.

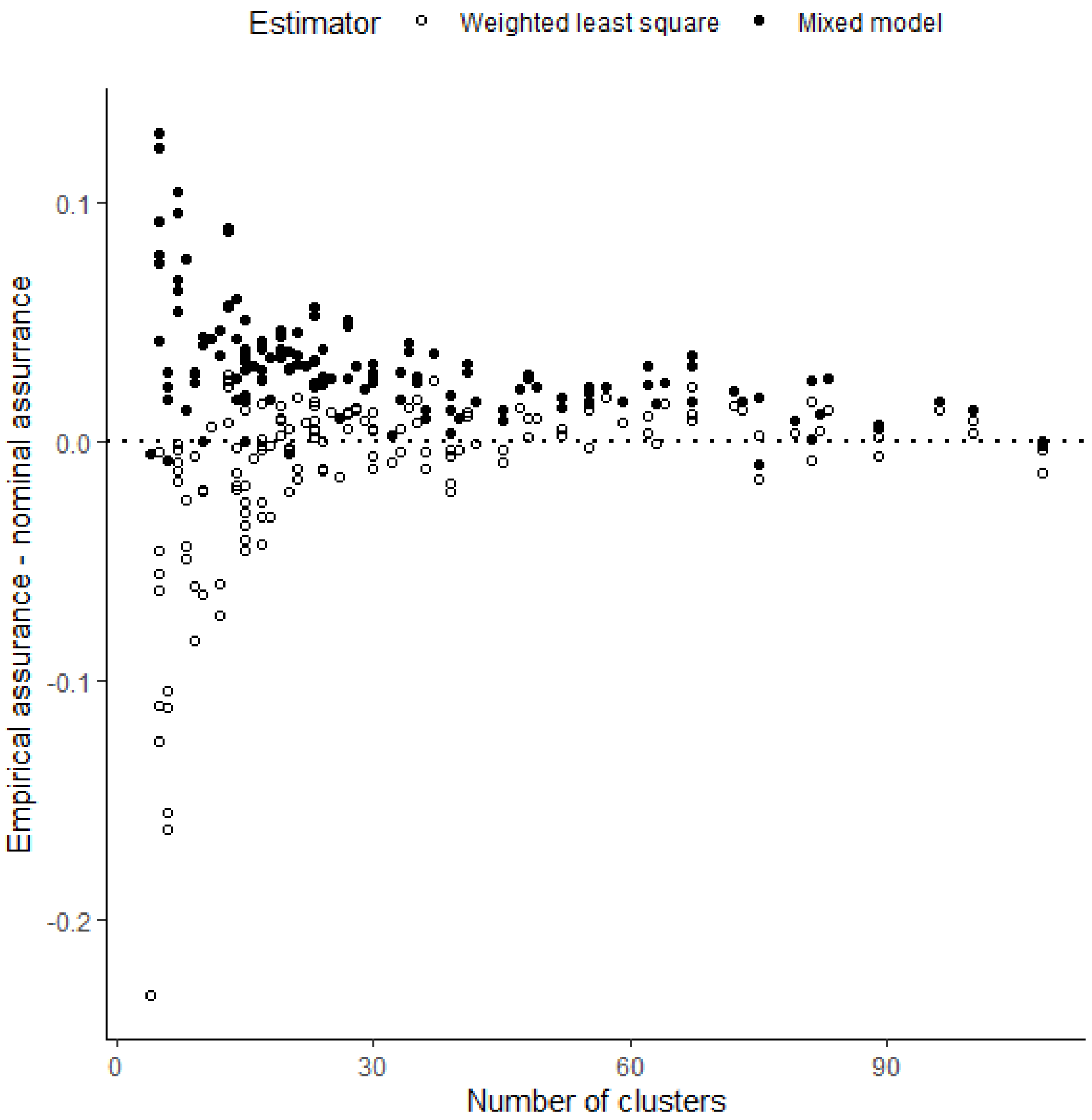

Figure 5.2: Empirical assurance probability under small to large sample size. The y-axis is the empirical assurance probability minus nominal assurance probability, hence >0 is more preferable than < 0 for small sample size.

We will focus on comparing the individual temporal correlation of the original scale to the individual temporal correlation of win fractions obtained from the same data. We will also examine the relationship between the individual temporal correlation of the original scale and the correlations of cluster means of win fractions since the latter are crucial in determining the efficiency gained from baseline adjustment. To be specific, the temporal correlation of cluster means of win fractions relates to the efficiency gained from baseline adjustment using the weighted least square approach, and the autocorrelation of cluster means of win fractions relates to the efficiency gained from the mixed model approach. The ICC of the original scale is also compared to the ICC of win fractions.

We generate data with two different correlation structures, where one has a lower autocorrelation of different individuals within the same cluster, and the other has a higher correlation. The temporal correlation of cluster means of win fractions is estimated by Pearson correlation weighted by cluster size given by

$$r_c = \frac{\sum_{i=1}^{2}\sum_{j=1}^{k_i} m_{ij}\overline{w}^X_{ij.}\overline{w}^Y_{ij.} - M\overline{w}^X_{...}\overline{w}^Y_{...}}{\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{k_i} m_{ij}(\overline{w}^X_{ij.})^2 - M(\overline{w}^X_{...})^2}\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{k_i} m_{ij}(\overline{w}^X_{ij.})^2 - M(\overline{w}^X_{...})^2}},$$

where $\overline{w}^X_{ij.}$ and $\overline{w}^Y_{ij.}$ are the cluster-specific mean of win fractions for baseline and follow-up, respectively, m_{ij} is the size of cluster j in arm i, and $M = \sum_{i=1}^{2}\sum_{j=1} m_{ij}$ is the total number of participants (Bland and Altman, 1995*b*). The individual temporal correlation r_i is estimated with repeated measure analysis of variance (ANOVA) to account for clustering (Bland and Altman, 1995*a*). The ICC ρ is estimated using the ANOVA estimator (Donner and Klar, 2000), and the autocorrelation of cluster means is estimated with

$$r_a = \frac{\overline{m}_{..}\rho}{1 + (\overline{m}_{..} - 1)\rho} r_c + \frac{1-\rho}{1 + (\overline{m}_{..} - 1)\rho} r_i,$$

where $\overline{m}_{..} = M/(k_1 + k_2)$ is the mean cluster size.

We generated 1,825 replicates with given individual temporal correlation (r) and ICC of the original scale and then estimated the individual temporal correlation, correlation of cluster means, the autocorrelation of cluster means and ICC of win fractions. We summarized the simulation results in Table 5.20, by presenting the mean of the correlation estimates and their standard errors.

The individual temporal correlation and ICC of win fractions are relatively close to its counterpart on the original scale compared to the temporal correlation of cluster means. The temporal correlation of cluster means of win fractions (r_c) is smaller than the individual temporal correlation of original scale r under low autocorrelation, where r_c could be larger than r under high autocorrelation. The differences between r_c and r increase as WinP increase. Estimating r_c yields a much higher standard error than the individual temporal correlation of win fractions. On the other hand, the autocorrelation (r_a) is closer to r and has a smaller standard error. This explains why the mixed model approach performs more reliably than the weighted least square approach in efficiency in our simulation study.

The difference between r and r_a or r_c and their standard errors increases as the treatment effect increases. This might explain why our sample size formula could be anticonservative when only a small number of clusters are required. It is because estimating the temporal correlation of cluster mean has a high standard error; hence, the required sample size varies more when the data is analyzed with the weighted least square approach. On the other hand, r_a has a smaller standard error which implies that the sample size required when the data is analyzed with the mixed model approach varies less.

5.6 Summary and discussion

We conducted simulation studies to evaluate the performance of estimating WinP for cluster randomization trials with baseline assessments by the weighted least square and the mixed model approach. Both methods of baseline adjustment have sufficient coverage from our simulation study. Baseline adjustment also reduced variance compared to the unadjusted estimates in all scenarios for the mixed model approach. The weighted least square approach could lose efficiency when only a few clusters are randomized because the temporal correlation of cluster means is unstable and could be much smaller for a small sample size. Hence, losing degrees of freedom for confidence interval construction may not make up for the efficiency gained in variance estimation from baseline adjustment using the weighted least square approach. This is also confirmed in the simulation study for the sample size formula, where we observed low assurance probability for the weighted least square approach when a small number of clusters are randomized. The low assurance probability is due to the increased width of the confidence interval from the weighted least square approach under a small sample size.

The mixed model approach has two advantages over the weighted least square approach. The mixed model approach often has higher efficiency than the weighted least square approach. The mixed model approach is better at maintaining the assurance probability with our sample size formula. Another reason is that the mixed model is commonly available in most statistical software without extra programming. One only needs to obtain the win fractions by ranks of the outcomes and baseline measurements and then use the mixed model to analyze the win fractions for effect estimation. In very rare cases, we did encounter simulated data sets that cannot be analyzed with the mixed model approach due to the singularity of the covariance matrix. The weighted least square approach could be an alternative choice in such cases.

The correlation of cluster means appears in the sample size formulas for cluster randomization trials, but it is often not known in practice. The approach we have taken in this thesis is to use only the individual temporal correlation and the ICC on the original scale to estimate sample size, although the formula was derived based on correlation and ICC on win fractions. Our simulation study shows that such an approach works well for moderate to large sample sizes. To maintain the assurance probability for small sample size, one can increase two clusters to our sample size formula when less than 30 clusters are randomized for $\alpha = 0.5$ and four clusters for $\alpha = 0.01$.

We evaluated the performance of our sample size formulas where the baseline measurement and follow-up outcome come from the same individual, i.e, the clusters consist of the same individuals over time. Our sample size formula is also applicable to cross-sectional designs, where the recruitment occurs at both follow-up and baseline. The sample size formula for cross-sectional cluster randomization trials is essentially the same for a trial where the same individuals are followed, except that the autocorrelation of cluster means is smaller because individual temporal correlation is close to zero for cross-sectional designs. We expect our formula would be liberal for cross-sectional designs if the individual temporal correlation were used in sample size estimation because the temporal correlation of cluster mean win fractions is often smaller than the individual temporal correlation as shown in our simulation study.

Table 5.5: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with ordered category outcomes. Each arm consists of 5 clusters. Entries are presented as coverage% $_{\text{confidence interval length} \times 100}$.

r^a	WinP^b	ICC^c	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
			Cluster size ~ Binomial(50,0.5)				
0.3	0.56	0.01	$94.63_{17.31}$	$95.18_{18.46}$	$95.29_{18.62}$	$94.85_{18.42}$	$94.63_{16.37}$
		0.05	$95.12_{22.55}$	$95.34_{23.98}$	$95.45_{24.13}$	$95.23_{23.95}$	$95.07_{21.38}$
		0.10	$95.51_{27.38}$	$95.23_{29.05}$	$95.07_{29.17}$	$95.34_{29.07}$	$94.90_{25.96}$
	0.64	0.01	$94.74_{15.99}$	$95.29_{17.11}$	$95.62_{17.25}$	$95.23_{17.07}$	$94.41_{15.29}$
		0.05	$96.05_{20.75}$	$94.79_{22.23}$	$95.07_{22.36}$	$94.74_{22.21}$	$95.56_{19.95}$
		0.10	$94.90_{25.59}$	$94.79_{27.18}$	$94.79_{27.32}$	$94.63_{27.17}$	$94.36_{24.52}$
	0.71	0.01	$95.89_{15.24}$	$95.29_{16.36}$	$95.18_{16.50}$	$95.18_{16.31}$	$96.27_{14.64}$
		0.05	$95.18_{19.58}$	$95.45_{20.97}$	$95.95_{21.10}$	$95.34_{20.95}$	$95.01_{18.86}$
		0.10	$94.74_{24.13}$	$95.18_{25.63}$	$95.45_{25.76}$	$95.18_{25.66}$	$94.74_{23.21}$
0.5	0.56	0.01	$95.18_{17.22}$	$95.51_{16.82}$	$95.56_{16.95}$	$95.45_{16.79}$	$95.23_{14.94}$
		0.05	$95.40_{22.69}$	$95.67_{22.17}$	$95.73_{22.31}$	$95.67_{22.15}$	$95.34_{19.69}$
		0.10	$93.97_{27.48}$	$95.40_{27.06}$	$95.45_{27.23}$	$95.34_{27.03}$	$94.68_{24.14}$
	0.64	0.01	$94.30_{15.96}$	$94.96_{15.61}$	$95.01_{15.75}$	$94.85_{15.57}$	$94.03_{13.91}$
		0.05	$95.62_{21.13}$	$95.84_{20.72}$	$96.00_{20.83}$	$95.78_{20.71}$	$95.84_{18.53}$
		0.10	$94.52_{25.65}$	$95.45_{25.10}$	$95.62_{25.25}$	$95.40_{25.09}$	$95.29_{22.56}$
	0.71	0.01	$95.78_{15.10}$	$95.62_{14.96}$	$95.62_{15.08}$	$95.45_{14.93}$	$95.56_{13.41}$
		0.05	$95.23_{19.58}$	$95.07_{19.27}$	$94.79_{19.44}$	$94.79_{19.23}$	$94.25_{17.31}$
		0.10	$95.51_{24.27}$	$95.29_{23.80}$	$95.45_{23.93}$	$95.40_{23.79}$	$94.74_{21.43}$
			Cluster size~ Binomial(100,0.5)				
0.3	0.56	0.01	$94.58_{13.22}$	$95.56_{14.13}$	$95.56_{14.19}$	$95.45_{14.11}$	$94.52_{12.47}$
		0.05	$95.18_{20.00}$	$95.12_{21.35}$	$95.34_{21.40}$	$95.12_{21.35}$	$94.52_{18.90}$
		0.10	$95.95_{25.62}$	$95.73_{27.31}$	$95.78_{27.39}$	$95.78_{27.29}$	$95.62_{24.29}$
	0.64	0.01	$94.41_{12.59}$	$95.62_{13.50}$	$95.62_{13.55}$	$95.78_{13.49}$	$94.74_{12.02}$
		0.05	$94.74_{18.63}$	$95.62_{19.92}$	$95.84_{19.97}$	$95.45_{19.92}$	$94.63_{17.78}$
		0.10	$95.73_{23.94}$	$95.84_{25.51}$	$95.95_{25.57}$	$95.78_{25.51}$	$96.05_{22.95}$
	0.71	0.01	$94.85_{11.70}$	$95.12_{12.49}$	$95.18_{12.54}$	$95.07_{12.48}$	$94.85_{11.24}$
		0.05	$94.52_{17.53}$	$95.01_{18.76}$	$94.96_{18.81}$	$95.01_{18.75}$	$94.90_{16.77}$
		0.10	$94.30_{22.40}$	$94.52_{23.98}$	$94.58_{24.05}$	$94.52_{23.97}$	$94.36_{21.51}$
0.5	0.56	0.01	$93.70_{13.25}$	$94.47_{13.02}$	$94.58_{13.07}$	$94.36_{13.00}$	$93.92_{11.47}$
		0.05	$94.58_{19.98}$	$94.52_{19.41}$	$94.63_{19.46}$	$94.58_{19.41}$	$94.96_{17.19}$
		0.10	$94.08_{25.81}$	$94.90_{25.07}$	$94.79_{25.14}$	$94.63_{25.06}$	$95.07_{22.22}$
	0.64	0.01	$95.40_{12.39}$	$95.56_{12.15}$	$95.67_{12.20}$	$95.40_{12.14}$	$95.01_{10.82}$
		0.05	$94.14_{18.51}$	$95.01_{18.16}$	$94.96_{18.21}$	$95.07_{18.17}$	$94.85_{16.24}$
		0.10	$94.90_{24.06}$	$95.29_{23.37}$	$95.29_{23.43}$	$95.07_{23.38}$	$95.73_{21.00}$
	0.71	0.01	$94.68_{11.70}$	$95.45_{11.58}$	$95.51_{11.63}$	$95.51_{11.57}$	$95.07_{10.41}$
		0.05	$95.62_{17.62}$	$95.51_{17.27}$	$95.34_{17.32}$	$95.62_{17.26}$	$95.12_{15.45}$
		0.10	$95.01_{22.38}$	$95.78_{22.01}$	$95.73_{22.07}$	$96.11_{22.00}$	$95.62_{19.82}$

[a] r: individual temporal correlation of the original scale. [b] WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. [c] ICC: intraclass correlation of coefficient.

Table 5.6: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with ordered category outcomes. Each arm consists of 15 clusters. Entries are presented as coverage% confidence interval length × 100.

r[a]	WinP[b]	ICC[c]	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
			Cluster size~ Binomial(50,0.5)				
0.3	0.56	0.01	94.90 9.13	94.85 8.99	94.90 9.02	95.07 8.99	94.58 8.69
		0.05	95.34 12.04	95.07 11.83	95.40 11.90	95.18 11.87	94.52 11.45
		0.10	95.56 14.87	95.45 14.61	95.51 14.71	95.56 14.68	95.01 14.15
	0.64	0.01	95.12 8.53	95.01 8.40	95.07 8.43	95.01 8.41	95.12 8.17
		0.05	95.34 11.22	95.67 11.03	95.84 11.09	95.84 11.07	95.01 10.74
		0.10	95.18 13.87	95.73 13.64	95.78 13.72	95.78 13.70	95.34 13.26
	0.71	0.01	95.29 8.04	94.90 7.93	95.01 7.96	95.12 7.94	94.90 7.74
		0.05	95.34 10.50	95.62 10.34	95.84 10.40	95.62 10.38	95.51 10.11
		0.10	95.07 12.95	95.07 12.74	95.01 12.83	94.96 12.80	95.23 12.44
0.5	0.56	0.01	95.07 9.12	95.07 8.20	95.07 8.23	95.01 8.21	94.41 7.93
		0.05	95.12 12.05	95.23 10.81	95.34 10.87	95.07 10.85	94.58 10.47
		0.10	95.45 14.88	94.79 13.34	94.90 13.42	94.90 13.41	94.58 12.92
	0.64	0.01	95.40 8.53	95.29 7.69	95.34 7.71	95.23 7.69	95.34 7.47
		0.05	95.45 11.24	95.67 10.10	95.84 10.15	95.89 10.14	95.62 9.83
		0.10	95.40 13.89	95.67 12.47	95.78 12.54	95.62 12.53	95.51 12.12
	0.71	0.01	95.01 8.02	95.29 7.28	95.29 7.30	95.34 7.28	94.90 7.10
		0.05	95.56 10.51	95.40 9.49	95.56 9.54	95.40 9.52	95.12 9.28
		0.10	95.07 12.96	95.40 11.67	95.56 11.74	95.56 11.73	95.45 11.41
			Cluster size~ Binomial(100,0.5)				
0.3	0.56	0.01	95.07 7.10	95.84 6.96	95.84 6.97	95.89 6.97	95.40 6.72
		0.05	95.18 10.70	96.11 10.50	96.16 10.54	96.11 10.53	95.51 10.13
		0.10	95.40 13.92	95.67 13.66	95.95 13.72	95.73 13.70	95.73 13.15
	0.64	0.01	96.11 6.61	95.89 6.50	95.89 6.51	95.78 6.50	95.56 6.32
		0.05	96.00 9.94	95.95 9.76	96.00 9.80	95.95 9.79	95.84 9.48
		0.10	96.05 12.92	95.89 12.69	95.95 12.75	95.78 12.73	96.11 12.31
	0.71	0.01	95.78 6.20	95.62 6.10	95.73 6.11	95.73 6.10	95.34 5.95
		0.05	95.89 9.26	96.05 9.10	96.11 9.13	95.95 9.12	95.78 8.87
		0.10	96.00 12.02	96.16 11.81	96.27 11.87	96.05 11.85	95.73 11.50
0.5	0.56	0.01	95.40 7.10	95.95 6.34	96.05 6.35	96.11 6.35	95.62 6.13
		0.05	95.45 10.71	95.67 9.56	95.84 9.59	95.73 9.58	95.95 9.23
		0.10	95.12 13.92	95.45 12.45	95.51 12.50	95.67 12.49	95.34 11.99
	0.64	0.01	95.62 6.61	95.56 5.91	95.62 5.92	95.34 5.92	95.51 5.75
		0.05	95.45 9.94	95.56 8.89	95.67 8.91	95.56 8.91	95.73 8.63
		0.10	95.34 12.91	95.51 11.56	95.62 11.61	95.51 11.59	95.40 11.21
	0.71	0.01	95.51 6.20	95.89 5.58	95.89 5.59	95.78 5.58	95.62 5.45
		0.05	95.12 9.26	95.95 8.31	95.89 8.33	95.89 8.33	95.89 8.10
		0.10	95.29 12.02	95.51 10.79	95.73 10.84	95.62 10.82	95.45 10.51

[a] r: individual temporal correlation of the original scale. [b] WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. [c] ICC: intraclass correlation of coefficient.

Table 5.7: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with ordered category outcomes. The control arm consists of 20 clusters and the treatment arm consists of 10 clusters. Entries are presented as coverage% $_{\text{confidence interval length} \times 100}$.

r^a	WinP^b	ICC^c	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
Cluster size~ Binomial(50,0.5)							
0.3	0.56	0.01	$93.97_{10.10}$	$94.30_{10.25}$	$94.74_{10.39}$	$94.58_{10.24}$	$93.86_{9.42}$
		0.05	$94.30_{13.25}$	$94.90_{13.40}$	$95.12_{13.56}$	$94.96_{13.43}$	$94.19_{12.34}$
		0.10	$94.25_{16.42}$	$94.85_{16.63}$	$94.79_{16.80}$	$94.85_{16.70}$	$94.47_{15.25}$
	0.64	0.01	$94.74_{8.77}$	$95.73_{8.93}$	$95.78_{8.96}$	$95.84_{8.93}$	$95.89_{8.54}$
		0.05	$95.34_{11.58}$	$95.56_{11.81}$	$95.84_{11.82}$	$95.51_{11.84}$	$96.16_{11.30}$
		0.10	$94.47_{14.22}$	$95.62_{14.49}$	$95.89_{14.47}$	$95.56_{14.56}$	$95.62_{13.91}$
	0.71	0.01	$94.85_{8.48}$	$95.40_{8.62}$	$95.45_{8.69}$	$95.40_{8.62}$	$95.78_{8.20}$
		0.05	$95.18_{11.10}$	$96.11_{11.33}$	$96.00_{11.39}$	$95.89_{11.35}$	$95.51_{10.71}$
		0.10	$94.52_{13.65}$	$95.12_{13.89}$	$95.29_{13.94}$	$95.18_{13.94}$	$95.23_{13.20}$
0.5	0.56	0.01	$93.86_{10.06}$	$93.26_{9.31}$	$93.48_{9.44}$	$92.88_{9.31}$	$92.60_{8.57}$
		0.05	$94.30_{13.35}$	$94.63_{12.36}$	$94.79_{12.49}$	$94.68_{12.38}$	$94.30_{11.31}$
		0.10	$93.97_{16.34}$	$94.14_{15.09}$	$94.41_{15.21}$	$94.19_{15.15}$	$93.21_{13.87}$
	0.64	0.01	$94.25_{8.75}$	$94.96_{8.16}$	$94.85_{8.18}$	$95.07_{8.16}$	$94.68_{7.78}$
		0.05	$93.92_{11.52}$	$95.01_{10.72}$	$94.90_{10.70}$	$95.01_{10.75}$	$94.74_{10.26}$
		0.10	$94.47_{14.28}$	$95.45_{13.27}$	$95.56_{13.24}$	$95.51_{13.34}$	$95.84_{12.72}$
	0.71	0.01	$94.74_{8.38}$	$95.62_{7.86}$	$95.51_{7.92}$	$95.51_{7.86}$	$94.90_{7.45}$
		0.05	$94.79_{11.09}$	$95.01_{10.38}$	$94.85_{10.41}$	$95.07_{10.41}$	$94.68_{9.83}$
		0.10	$94.58_{13.46}$	$94.96_{12.51}$	$94.74_{12.51}$	$95.18_{12.56}$	$95.45_{11.94}$
Cluster size~ Binomial(100,0.5)							
0.3	0.56	0.01	$94.41_{7.80}$	$94.41_{7.92}$	$94.36_{7.98}$	$94.36_{7.92}$	$94.03_{7.27}$
		0.05	$94.36_{11.68}$	$94.68_{11.84}$	$94.85_{11.92}$	$94.58_{11.86}$	$94.19_{10.81}$
		0.10	$94.52_{15.15}$	$94.79_{15.35}$	$95.01_{15.46}$	$94.68_{15.39}$	$93.75_{13.99}$
	0.64	0.01	$95.45_{6.82}$	$95.73_{6.96}$	$95.67_{6.95}$	$95.73_{6.96}$	$95.73_{6.64}$
		0.05	$94.90_{10.19}$	$94.74_{10.38}$	$94.36_{10.35}$	$94.79_{10.39}$	$95.07_{9.97}$
		0.10	$94.30_{13.32}$	$95.51_{13.53}$	$95.34_{13.50}$	$95.56_{13.56}$	$95.34_{13.06}$
	0.71	0.01	$95.23_{6.49}$	$95.89_{6.64}$	$95.84_{6.66}$	$95.89_{6.64}$	$95.40_{6.30}$
		0.05	$93.97_{9.67}$	$94.58_{9.84}$	$94.58_{9.85}$	$94.58_{9.86}$	$94.30_{9.31}$
		0.10	$94.47_{12.52}$	$95.40_{12.75}$	$95.18_{12.76}$	$95.29_{12.78}$	$95.51_{12.10}$
0.5	0.56	0.01	$94.47_{7.79}$	$94.58_{7.23}$	$94.79_{7.30}$	$94.52_{7.23}$	$94.25_{6.64}$
		0.05	$94.41_{11.72}$	$95.40_{10.87}$	$95.40_{10.94}$	$95.45_{10.88}$	$94.58_{9.91}$
		0.10	$94.36_{15.21}$	$95.51_{14.07}$	$95.67_{14.16}$	$95.34_{14.09}$	$94.52_{12.86}$
	0.64	0.01	$94.79_{6.75}$	$95.40_{6.29}$	$95.18_{6.27}$	$95.40_{6.29}$	$95.18_{6.00}$
		0.05	$94.79_{10.22}$	$96.16_{9.52}$	$96.05_{9.49}$	$96.05_{9.54}$	$96.27_{9.13}$
		0.10	$95.18_{13.29}$	$95.89_{12.40}$	$95.56_{12.33}$	$95.78_{12.43}$	$95.56_{11.91}$
	0.71	0.01	$94.36_{6.51}$	$96.05_{6.11}$	$95.95_{6.12}$	$96.22_{6.11}$	$96.00_{5.77}$
		0.05	$94.58_{9.71}$	$94.90_{9.03}$	$94.68_{9.03}$	$94.85_{9.04}$	$94.79_{8.57}$
		0.10	$94.68_{12.53}$	$95.45_{11.73}$	$95.34_{11.72}$	$95.23_{11.75}$	$94.74_{11.12}$

$^a r$: individual temporal correlation of the original scale. ^{b}WinP = Pr(Y_1 < Y_2) + 0.5Pr(Y_1 = Y_2). ^{c}ICC: intraclass correlation of coefficient.

Table 5.8: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with ordered category outcomes. The control arm consists of 10 clusters and the treatment arm consists of 20 clusters. Entries are presented as coverage% $_{\text{confidence interval length} \times 100}$.

r^a	WinP^b	ICC^c	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
			Cluster size~ Binomial(50,0.5)				
0.3	0.56	0.01	$93.97_{9.22}$	$94.96_{9.40}$	$94.96_{9.39}$	$94.58_{9.41}$	$95.51_{8.98}$
		0.05	$95.34_{12.15}$	$95.51_{12.38}$	$95.56_{12.33}$	$95.56_{12.41}$	$95.45_{11.86}$
		0.10	$95.62_{15.10}$	$95.84_{15.38}$	$95.45_{15.29}$	$95.89_{15.45}$	$95.89_{14.76}$
	0.64	0.01	$94.41_{9.24}$	$94.85_{9.39}$	$95.34_{9.51}$	$94.90_{9.40}$	$94.74_{8.72}$
		0.05	$94.85_{12.20}$	$95.01_{12.35}$	$95.12_{12.48}$	$94.85_{12.37}$	$94.58_{11.51}$
		0.10	$93.92_{14.97}$	$94.25_{15.16}$	$94.47_{15.29}$	$94.47_{15.21}$	$94.25_{14.11}$
	0.71	0.01	$94.30_{8.54}$	$95.51_{8.71}$	$95.89_{8.79}$	$95.34_{8.72}$	$95.29_{8.22}$
		0.05	$94.08_{11.11}$	$94.41_{11.31}$	$94.47_{11.39}$	$94.58_{11.34}$	$94.47_{10.71}$
		0.10	$94.41_{13.83}$	$95.23_{14.06}$	$95.45_{14.14}$	$95.29_{14.11}$	$95.12_{13.32}$
0.5	0.56	0.01	$94.30_{9.15}$	$95.67_{8.58}$	$95.40_{8.54}$	$95.84_{8.58}$	$95.34_{8.20}$
		0.05	$94.47_{12.18}$	$96.16_{11.36}$	$95.73_{11.28}$	$96.16_{11.40}$	$95.62_{10.87}$
		0.10	$94.41_{15.13}$	$95.18_{14.02}$	$94.74_{13.91}$	$95.01_{14.08}$	$94.52_{13.43}$
	0.64	0.01	$94.08_{9.21}$	$94.25_{8.56}$	$94.19_{8.67}$	$94.19_{8.56}$	$94.30_{7.97}$
		0.05	$93.53_{12.15}$	$94.03_{11.28}$	$94.14_{11.40}$	$94.03_{11.31}$	$94.47_{10.51}$
		0.10	$95.01_{14.98}$	$95.40_{13.84}$	$95.56_{13.95}$	$95.56_{13.90}$	$94.41_{12.88}$
	0.71	0.01	$94.36_{8.52}$	$95.29_{7.96}$	$95.45_{8.03}$	$95.18_{7.96}$	$95.34_{7.53}$
		0.05	$95.23_{11.28}$	$95.51_{10.48}$	$95.84_{10.56}$	$95.62_{10.51}$	$95.56_{9.93}$
		0.10	$94.68_{13.70}$	$95.95_{12.72}$	$95.84_{12.78}$	$95.73_{12.78}$	$95.84_{12.11}$
			Cluster size~ Binomial(100,0.5)				
0.3	0.56	0.01	$94.85_{7.13}$	$95.51_{7.26}$	$95.56_{7.21}$	$95.78_{7.27}$	$95.73_{6.95}$
		0.05	$95.18_{10.81}$	$95.73_{11.04}$	$95.45_{10.95}$	$95.56_{11.07}$	$95.51_{10.58}$
		0.10	$94.68_{13.97}$	$95.51_{14.26}$	$95.29_{14.14}$	$95.23_{14.29}$	$95.73_{13.79}$
	0.64	0.01	$93.97_{7.16}$	$95.12_{7.26}$	$95.12_{7.32}$	$95.07_{7.26}$	$94.03_{6.74}$
		0.05	$94.58_{10.67}$	$94.90_{10.82}$	$95.23_{10.90}$	$94.79_{10.84}$	$94.30_{10.06}$
		0.10	$94.68_{13.86}$	$94.30_{14.02}$	$94.74_{14.10}$	$94.52_{14.05}$	$94.96_{12.96}$
	0.71	0.01	$94.68_{6.55}$	$95.40_{6.67}$	$95.51_{6.71}$	$95.45_{6.67}$	$94.85_{6.30}$
		0.05	$94.52_{9.87}$	$95.18_{10.01}$	$95.12_{10.05}$	$95.12_{10.03}$	$94.90_{9.46}$
		0.10	$94.08_{12.67}$	$95.12_{12.89}$	$95.07_{12.94}$	$95.12_{12.92}$	$94.79_{12.14}$
0.5	0.56	0.01	$95.12_{7.14}$	$95.56_{6.67}$	$95.45_{6.62}$	$95.73_{6.68}$	$95.51_{6.37}$
		0.05	$94.30_{10.75}$	$94.79_{9.99}$	$94.47_{9.88}$	$94.90_{10.01}$	$95.18_{9.58}$
		0.10	$94.47_{14.03}$	$95.67_{13.08}$	$95.84_{12.95}$	$95.67_{13.12}$	$95.01_{12.64}$
	0.64	0.01	$94.79_{7.13}$	$94.96_{6.62}$	$94.90_{6.67}$	$94.85_{6.62}$	$94.68_{6.15}$
		0.05	$94.08_{10.71}$	$94.47_{9.90}$	$94.52_{9.96}$	$94.41_{9.92}$	$93.42_{9.16}$
		0.10	$93.70_{13.97}$	$95.40_{13.01}$	$95.62_{13.08}$	$95.45_{13.05}$	$94.19_{11.95}$
	0.71	0.01	$94.85_{6.58}$	$95.01_{6.17}$	$94.96_{6.20}$	$95.07_{6.17}$	$94.79_{5.83}$
		0.05	$94.90_{9.79}$	$95.45_{9.05}$	$95.51_{9.09}$	$95.29_{9.07}$	$95.62_{8.61}$
		0.10	$94.74_{12.86}$	$95.45_{11.93}$	$95.34_{11.98}$	$95.29_{11.95}$	$95.67_{11.20}$

$^a r$: individual temporal correlation of the original scale. ^{b}WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. ^{c}ICC: intraclass correlation of coefficient.

Table 5.9: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with ordered category outcomes. Each arm consists of 15 clusters with variable cluster size generated from discrete uniform distribution (24, 76). Entries are presented as coverage%$_{\text{confidence interval length} \times 100}$.

r^a	WinP^b	ICC^c	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
				High auto correlation^d			
0.3	0.56	0.01	94.30$_{5.77}$	94.85$_{4.81}$	95.23$_{4.99}$	94.52$_{4.83}$	94.68$_{5.07}$
		0.05	93.37$_{9.88}$	94.19$_{6.25}$	95.45$_{6.60}$	94.36$_{6.39}$	94.58$_{8.14}$
		0.10	92.93$_{13.30}$	93.75$_{7.54}$	95.34$_{7.92}$	93.97$_{7.80}$	94.79$_{11.04}$
	0.64	0.01	94.30$_{5.37}$	95.12$_{4.47}$	96.16$_{4.64}$	95.18$_{4.48}$	94.85$_{4.73}$
		0.05	93.32$_{9.16}$	94.08$_{5.80}$	95.34$_{6.13}$	94.41$_{5.93}$	94.74$_{7.56}$
		0.10	93.04$_{12.33}$	93.70$_{7.01}$	95.29$_{7.37}$	93.70$_{7.25}$	94.90$_{10.24}$
	0.71	0.01	94.03$_{5.03}$	95.18$_{4.22}$	95.84$_{4.38}$	94.79$_{4.24}$	95.12$_{4.46}$
		0.05	93.59$_{8.52}$	94.19$_{5.47}$	95.73$_{5.78}$	94.47$_{5.60}$	94.90$_{7.05}$
		0.10	93.26$_{11.47}$	94.08$_{6.63}$	95.34$_{6.96}$	93.97$_{6.86}$	94.85$_{9.54}$
0.5	0.56	0.01	94.14$_{5.78}$	94.58$_{4.32}$	95.23$_{4.49}$	94.14$_{4.34}$	94.30$_{4.43}$
		0.05	93.21$_{9.87}$	93.53$_{5.78}$	94.85$_{6.07}$	93.59$_{5.94}$	94.36$_{6.78}$
		0.10	93.26$_{13.30}$	93.59$_{7.14}$	94.85$_{7.46}$	94.08$_{7.42}$	94.25$_{8.99}$
	0.64	0.01	94.14$_{5.37}$	95.29$_{4.02}$	96.16$_{4.18}$	95.23$_{4.04}$	95.07$_{4.14}$
		0.05	93.32$_{9.15}$	93.70$_{5.36}$	95.12$_{5.63}$	93.97$_{5.50}$	94.36$_{6.28}$
		0.10	93.15$_{12.33}$	93.86$_{6.63}$	94.90$_{6.93}$	94.08$_{6.89}$	94.14$_{8.32}$
	0.71	0.01	94.36$_{5.03}$	95.18$_{3.83}$	95.89$_{3.97}$	95.12$_{3.84}$	95.45$_{3.93}$
		0.05	93.97$_{8.51}$	94.14$_{5.08}$	94.96$_{5.33}$	94.03$_{5.22}$	95.12$_{5.88}$
		0.10	93.26$_{11.47}$	93.97$_{6.30}$	94.96$_{6.57}$	94.30$_{6.54}$	94.63$_{7.77}$
				Low auto correlation			
0.3	0.56	0.01	94.08$_{5.76}$	94.14$_{5.65}$	94.74$_{5.80}$	94.08$_{5.78}$	94.52$_{5.59}$
		0.05	92.66$_{9.84}$	93.32$_{9.62}$	94.03$_{9.89}$	93.75$_{10.09}$	93.92$_{9.49}$
		0.10	92.49$_{13.24}$	92.88$_{12.93}$	93.86$_{13.27}$	93.86$_{13.63}$	93.97$_{12.71}$
	0.64	0.01	93.59$_{5.35}$	94.03$_{5.25}$	94.58$_{5.39}$	94.19$_{5.37}$	94.79$_{5.23}$
		0.05	92.71$_{9.12}$	93.04$_{8.91}$	93.97$_{9.16}$	94.14$_{9.35}$	94.25$_{8.85}$
		0.10	92.27$_{12.28}$	92.60$_{11.99}$	93.42$_{12.30}$	93.86$_{12.64}$	94.36$_{11.86}$
	0.71	0.01	93.81$_{5.01}$	94.63$_{4.92}$	95.23$_{5.05}$	94.41$_{5.04}$	94.36$_{4.92}$
		0.05	93.04$_{8.49}$	93.75$_{8.30}$	94.52$_{8.53}$	94.19$_{8.70}$	94.36$_{8.26}$
		0.10	92.33$_{11.41}$	92.93$_{11.14}$	93.75$_{11.44}$	94.58$_{11.75}$	94.41$_{11.05}$
0.5	0.56	0.01	93.92$_{5.76}$	94.47$_{5.14}$	94.74$_{5.28}$	94.25$_{5.26}$	94.41$_{5.09}$
		0.05	92.60$_{9.85}$	93.75$_{8.76}$	94.19$_{9.01}$	94.19$_{9.19}$	94.25$_{8.65}$
		0.10	92.27$_{13.25}$	93.26$_{11.77}$	94.08$_{12.07}$	93.97$_{12.41}$	94.25$_{11.56}$
	0.64	0.01	93.86$_{5.36}$	94.79$_{4.78}$	95.18$_{4.91}$	94.96$_{4.89}$	94.96$_{4.77}$
		0.05	92.77$_{9.13}$	93.64$_{8.11}$	94.52$_{8.34}$	94.30$_{8.50}$	94.74$_{8.06}$
		0.10	92.44$_{12.29}$	92.55$_{10.92}$	93.42$_{11.20}$	93.70$_{11.52}$	94.25$_{10.80}$
	0.71	0.01	93.97$_{5.02}$	94.52$_{4.50}$	95.12$_{4.63}$	94.58$_{4.61}$	94.96$_{4.50}$
		0.05	93.15$_{8.49}$	94.08$_{7.56}$	94.47$_{7.78}$	94.36$_{7.93}$	94.85$_{7.54}$
		0.10	92.38$_{11.42}$	93.26$_{10.17}$	94.03$_{10.43}$	94.03$_{10.73}$	94.68$_{10.09}$

[a] r: individual temporal correlation of the original scale. [b] WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. [c] ICC: intraclass correlation of coefficient.

Table 5.10: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with continuous outcomes generated from normal distributions. Each arm consists of 5 clusters. Entries are presented as coverage% × 100 $_{\text{confidence interval length} \times 100}$.

r^a	WinP[b]	ICC[c]	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
Cluster size~ Binomial(50,0.5)							
0.3	0.56	0.01	$95.07_{17.61}$	$95.73_{18.86}$	$95.56_{19.01}$	$95.89_{18.82}$	$95.34_{16.94}$
		0.05	$94.25_{23.09}$	$95.23_{24.68}$	$95.40_{24.84}$	$95.01_{24.65}$	$94.74_{22.33}$
		0.10	$94.58_{28.07}$	$94.85_{29.71}$	$95.23_{29.87}$	$95.01_{29.69}$	$94.85_{27.02}$
	0.64	0.01	$94.52_{16.93}$	$95.34_{18.17}$	$95.67_{18.32}$	$95.34_{18.12}$	$94.68_{16.26}$
		0.05	$95.45_{22.02}$	$96.05_{23.51}$	$96.00_{23.66}$	$95.95_{23.48}$	$95.73_{21.27}$
		0.10	$94.96_{27.04}$	$95.67_{28.64}$	$95.67_{28.79}$	$95.78_{28.63}$	$95.01_{26.03}$
	0.71	0.01	$95.84_{15.74}$	$95.73_{16.80}$	$96.00_{16.94}$	$95.45_{16.76}$	$95.12_{15.12}$
		0.05	$95.56_{20.22}$	$95.29_{21.66}$	$95.40_{21.79}$	$95.34_{21.65}$	$95.34_{19.55}$
		0.10	$94.47_{24.73}$	$95.78_{26.33}$	$96.05_{26.48}$	$95.84_{26.30}$	$95.23_{23.80}$
0.5	0.56	0.01	$95.12_{17.55}$	$95.62_{17.35}$	$95.67_{17.50}$	$95.67_{17.31}$	$95.56_{15.60}$
		0.05	$94.96_{22.95}$	$94.25_{22.45}$	$94.52_{22.61}$	$94.25_{22.41}$	$95.23_{20.27}$
		0.10	$94.85_{28.12}$	$95.67_{27.51}$	$95.95_{27.63}$	$95.62_{27.53}$	$95.62_{24.87}$
	0.64	0.01	$95.12_{16.95}$	$95.56_{16.63}$	$95.56_{16.76}$	$95.51_{16.61}$	$95.56_{14.98}$
		0.05	$95.40_{22.05}$	$95.84_{21.47}$	$96.00_{21.63}$	$95.56_{21.44}$	$95.34_{19.43}$
		0.10	$94.30_{26.68}$	$94.47_{26.09}$	$94.68_{26.25}$	$94.30_{26.07}$	$94.14_{23.67}$
	0.71	0.01	$95.18_{15.85}$	$95.01_{15.70}$	$95.07_{15.82}$	$95.07_{15.67}$	$95.12_{14.09}$
		0.05	$94.58_{20.30}$	$95.07_{20.12}$	$95.18_{20.26}$	$95.12_{20.10}$	$95.01_{18.11}$
		0.10	$95.01_{24.90}$	$95.23_{24.53}$	$95.29_{24.68}$	$94.90_{24.52}$	$94.85_{22.06}$
Cluster size~ Binomial(100,0.5)							
0.3	0.56	0.01	$95.73_{13.75}$	$95.62_{14.70}$	$95.78_{14.74}$	$95.62_{14.69}$	$95.29_{13.21}$
		0.05	$95.67_{20.51}$	$95.23_{21.91}$	$95.29_{21.97}$	$95.01_{21.91}$	$95.45_{19.70}$
		0.10	$95.45_{26.37}$	$95.78_{27.96}$	$95.62_{28.03}$	$95.67_{27.97}$	$96.11_{25.35}$
	0.64	0.01	$94.19_{12.96}$	$94.85_{13.85}$	$95.07_{13.91}$	$94.74_{13.84}$	$94.08_{12.48}$
		0.05	$94.68_{19.60}$	$94.85_{20.89}$	$95.01_{20.95}$	$94.79_{20.88}$	$94.58_{18.82}$
		0.10	$96.44_{25.24}$	$96.88_{26.92}$	$96.82_{26.99}$	$96.82_{26.91}$	$96.38_{24.24}$
	0.71	0.01	$95.23_{12.15}$	$95.12_{12.95}$	$95.12_{13.00}$	$95.18_{12.94}$	$95.23_{11.68}$
		0.05	$95.23_{17.96}$	$95.34_{19.20}$	$95.45_{19.26}$	$95.34_{19.18}$	$95.89_{17.28}$
		0.10	$94.96_{23.06}$	$95.01_{24.58}$	$95.01_{24.64}$	$95.01_{24.59}$	$94.85_{22.19}$
0.5	0.56	0.01	$95.07_{13.68}$	$94.63_{13.45}$	$94.68_{13.51}$	$94.47_{13.44}$	$94.47_{12.03}$
		0.05	$95.45_{20.28}$	$95.07_{19.81}$	$95.18_{19.87}$	$95.45_{19.81}$	$94.79_{17.85}$
		0.10	$94.36_{26.35}$	$94.68_{25.59}$	$94.63_{25.66}$	$94.52_{25.59}$	$94.52_{23.11}$
	0.64	0.01	$95.56_{13.21}$	$96.33_{12.93}$	$96.16_{12.98}$	$96.55_{12.91}$	$95.67_{11.61}$
		0.05	$94.90_{19.41}$	$94.90_{19.03}$	$95.07_{19.08}$	$94.68_{19.03}$	$95.01_{17.10}$
		0.10	$95.23_{25.26}$	$95.56_{24.62}$	$95.62_{24.67}$	$95.45_{24.64}$	$94.85_{22.26}$
	0.71	0.01	$95.40_{12.16}$	$95.51_{12.04}$	$95.51_{12.09}$	$95.51_{12.03}$	$95.62_{10.75}$
		0.05	$94.74_{17.85}$	$94.74_{17.50}$	$94.85_{17.56}$	$94.96_{17.49}$	$95.40_{15.67}$
		0.10	$95.29_{23.08}$	$95.01_{22.55}$	$95.29_{22.62}$	$94.96_{22.53}$	$94.47_{20.25}$

[a] r: individual temporal correlation of the original scale. [b] WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. [c] ICC: intraclass correlation of coefficient.

Table 5.11: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with continuous outcomes generated from normal distributions. Each arm consists of 15 clusters. Entries are presented as coverage% $_{\text{confidence interval length} \times 100}$.

r^a	WinP^b	ICC^c	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
Cluster size~ Binomial(50,0.5)							
0.3	0.56	0.01	95.40$_{9.35}$	95.51$_{9.21}$	95.51$_{9.24}$	95.67$_{9.21}$	95.29$_{9.00}$
		0.05	95.56$_{12.30}$	95.56$_{12.10}$	95.62$_{12.17}$	95.62$_{12.14}$	95.18$_{11.83}$
		0.10	95.62$_{15.18}$	95.23$_{14.92}$	95.40$_{15.02}$	95.29$_{14.99}$	95.18$_{14.58}$
	0.64	0.01	95.62$_{8.95}$	95.56$_{8.82}$	95.56$_{8.85}$	95.62$_{8.83}$	95.56$_{8.62}$
		0.05	95.62$_{11.74}$	95.45$_{11.55}$	95.67$_{11.61}$	95.67$_{11.58}$	95.18$_{11.29}$
		0.10	95.56$_{14.47}$	95.34$_{14.23}$	95.51$_{14.33}$	95.29$_{14.29}$	95.51$_{13.90}$
	0.71	0.01	95.67$_{8.31}$	95.78$_{8.21}$	95.78$_{8.24}$	95.73$_{8.21}$	95.62$_{8.01}$
		0.05	95.45$_{10.85}$	95.40$_{10.68}$	95.45$_{10.75}$	95.73$_{10.72}$	95.18$_{10.44}$
		0.10	95.29$_{13.35}$	95.34$_{13.13}$	95.62$_{13.23}$	95.40$_{13.19}$	95.29$_{12.83}$
0.5	0.56	0.01	95.56$_{9.35}$	95.67$_{8.43}$	95.67$_{8.45}$	95.73$_{8.43}$	95.07$_{8.23}$
		0.05	95.51$_{12.29}$	95.67$_{11.05}$	95.56$_{11.10}$	95.67$_{11.08}$	94.96$_{10.80}$
		0.10	95.56$_{15.17}$	95.51$_{13.62}$	95.67$_{13.71}$	95.67$_{13.68}$	95.51$_{13.31}$
	0.64	0.01	95.45$_{8.95}$	95.51$_{8.09}$	95.62$_{8.12}$	95.51$_{8.09}$	95.29$_{7.90}$
		0.05	95.40$_{11.73}$	95.23$_{10.56}$	95.23$_{10.62}$	95.45$_{10.59}$	95.07$_{10.32}$
		0.10	95.56$_{14.46}$	95.67$_{13.00}$	95.73$_{13.08}$	95.62$_{13.06}$	95.34$_{12.70}$
	0.71	0.01	95.29$_{8.31}$	95.40$_{7.55}$	95.45$_{7.58}$	95.56$_{7.55}$	95.56$_{7.37}$
		0.05	95.56$_{10.84}$	95.18$_{9.79}$	95.34$_{9.84}$	95.34$_{9.82}$	95.07$_{9.57}$
		0.10	95.23$_{13.35}$	95.51$_{12.02}$	95.51$_{12.10}$	95.34$_{12.07}$	95.29$_{11.74}$
Cluster size~ Binomial(100,0.5)							
0.3	0.56	0.01	95.29$_{7.25}$	95.07$_{7.12}$	95.18$_{7.13}$	95.07$_{7.12}$	95.12$_{6.96}$
		0.05	94.74$_{10.89}$	94.90$_{10.68}$	94.96$_{10.71}$	94.96$_{10.70}$	95.18$_{10.42}$
		0.10	94.85$_{14.15}$	94.68$_{13.87}$	95.18$_{13.93}$	94.90$_{13.91}$	95.23$_{13.53}$
	0.64	0.01	95.45$_{6.93}$	94.85$_{6.80}$	94.90$_{6.82}$	94.90$_{6.81}$	95.23$_{6.65}$
		0.05	94.63$_{10.37}$	94.96$_{10.17}$	94.90$_{10.20}$	94.90$_{10.20}$	95.34$_{9.93}$
		0.10	95.07$_{13.47}$	94.68$_{13.21}$	95.01$_{13.27}$	94.79$_{13.24}$	95.45$_{12.88}$
	0.71	0.01	95.51$_{6.42}$	94.68$_{6.31}$	94.68$_{6.32}$	94.63$_{6.32}$	95.18$_{6.17}$
		0.05	94.79$_{9.55}$	94.79$_{9.37}$	94.68$_{9.41}$	94.68$_{9.40}$	95.23$_{9.15}$
		0.10	95.12$_{12.40}$	94.63$_{12.16}$	94.74$_{12.22}$	94.68$_{12.19}$	95.29$_{11.86}$
0.5	0.56	0.01	95.51$_{7.25}$	95.07$_{6.49}$	95.07$_{6.51}$	95.01$_{6.50}$	95.01$_{6.35}$
		0.05	94.90$_{10.90}$	95.07$_{9.72}$	95.01$_{9.75}$	95.12$_{9.74}$	95.67$_{9.49}$
		0.10	94.63$_{14.16}$	95.01$_{12.62}$	95.01$_{12.67}$	94.90$_{12.66}$	95.45$_{12.32}$
	0.64	0.01	95.56$_{6.93}$	94.68$_{6.22}$	94.74$_{6.23}$	94.85$_{6.23}$	94.96$_{6.08}$
		0.05	94.85$_{10.38}$	95.01$_{9.26}$	95.01$_{9.29}$	95.07$_{9.29}$	95.45$_{9.05}$
		0.10	94.41$_{13.48}$	94.85$_{12.02}$	95.01$_{12.07}$	94.90$_{12.06}$	95.18$_{11.73}$
	0.71	0.01	95.51$_{6.42}$	94.90$_{5.79}$	94.90$_{5.80}$	95.07$_{5.79}$	95.23$_{5.66}$
		0.05	95.07$_{9.56}$	94.85$_{8.55}$	95.01$_{8.58}$	94.90$_{8.57}$	95.29$_{8.35}$
		0.10	94.90$_{12.40}$	94.52$_{11.08}$	94.85$_{11.12}$	94.63$_{11.11}$	95.18$_{10.81}$

[a] r: individual temporal correlation of the original scale. [b] WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. [c] ICC: intraclass correlation of coefficient.

Table 5.12: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with continuous outcomes generated from normal distributions. The control arm consists of 20 clusters and the treatment arm consists of 10 clusters. Entries are presented as coverage%$_{\text{confidence interval length} \times 100}$.

r^a	WinP^b	ICC^c	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
Cluster size∼ Binomial(50,0.5)							
0.3	0.56	0.01	$94.68_{9.87}$	$95.40_{10.04}$	$95.40_{10.13}$	$95.12_{10.04}$	$94.79_{9.52}$
		0.05	$94.25_{13.01}$	$95.29_{13.22}$	$95.45_{13.29}$	$95.34_{13.26}$	$95.12_{12.57}$
		0.10	$94.96_{16.07}$	$95.95_{16.31}$	$95.89_{16.41}$	$96.00_{16.37}$	$95.84_{15.45}$
	0.64	0.01	$94.63_{9.46}$	$95.12_{9.62}$	$95.34_{9.71}$	$95.12_{9.62}$	$95.01_{9.16}$
		0.05	$94.03_{12.39}$	$94.41_{12.61}$	$94.79_{12.68}$	$94.63_{12.65}$	$94.68_{11.96}$
		0.10	$94.85_{15.24}$	$95.73_{15.51}$	$95.84_{15.57}$	$95.67_{15.57}$	$95.67_{14.75}$
	0.71	0.01	$95.01_{8.72}$	$95.56_{8.92}$	$95.95_{9.00}$	$95.45_{8.92}$	$94.79_{8.45}$
		0.05	$94.03_{11.46}$	$94.52_{11.65}$	$94.63_{11.73}$	$94.79_{11.67}$	$94.08_{11.06}$
		0.10	$94.96_{14.07}$	$95.89_{14.30}$	$95.73_{14.36}$	$95.95_{14.36}$	$95.40_{13.61}$
0.5	0.56	0.01	$94.74_{9.93}$	$94.58_{9.21}$	$94.68_{9.29}$	$94.68_{9.20}$	$94.85_{8.72}$
		0.05	$94.74_{13.02}$	$95.84_{12.14}$	$96.05_{12.20}$	$95.73_{12.17}$	$95.89_{11.51}$
		0.10	$94.74_{16.06}$	$94.96_{14.89}$	$95.07_{14.94}$	$94.90_{14.95}$	$94.96_{14.15}$
	0.64	0.01	$94.14_{9.41}$	$95.12_{8.82}$	$95.29_{8.89}$	$95.12_{8.81}$	$94.63_{8.37}$
		0.05	$94.58_{12.43}$	$94.74_{11.52}$	$94.52_{11.58}$	$94.58_{11.56}$	$94.25_{10.95}$
		0.10	$94.63_{15.30}$	$95.07_{14.17}$	$95.23_{14.22}$	$95.01_{14.22}$	$94.58_{13.49}$
	0.71	0.01	$94.03_{8.80}$	$95.40_{8.28}$	$95.45_{8.35}$	$94.96_{8.28}$	$95.07_{7.84}$
		0.05	$95.62_{11.58}$	$95.45_{10.79}$	$95.78_{10.86}$	$95.51_{10.82}$	$95.12_{10.22}$
		0.10	$94.74_{14.07}$	$95.89_{13.11}$	$95.95_{13.17}$	$96.05_{13.16}$	$95.73_{12.47}$
Cluster size∼ Binomial(100,0.5)							
0.3	0.56	0.01	$95.23_{7.70}$	$95.45_{7.85}$	$95.67_{7.88}$	$95.34_{7.86}$	$95.73_{7.41}$
		0.05	$94.08_{11.47}$	$95.45_{11.67}$	$95.34_{11.71}$	$95.56_{11.68}$	$95.12_{11.06}$
		0.10	$94.52_{14.93}$	$94.96_{15.14}$	$94.74_{15.17}$	$95.12_{15.17}$	$95.12_{14.37}$
	0.64	0.01	$95.29_{7.30}$	$95.62_{7.41}$	$95.62_{7.45}$	$95.62_{7.42}$	$95.01_{7.04}$
		0.05	$94.79_{10.91}$	$95.23_{11.11}$	$95.34_{11.14}$	$95.18_{11.13}$	$95.29_{10.49}$
		0.10	$95.34_{14.23}$	$95.56_{14.48}$	$95.51_{14.52}$	$95.62_{14.52}$	$94.90_{13.68}$
	0.71	0.01	$95.12_{6.77}$	$94.74_{6.90}$	$94.52_{6.93}$	$94.90_{6.90}$	$94.63_{6.53}$
		0.05	$94.14_{10.06}$	$94.41_{10.21}$	$94.58_{10.24}$	$94.36_{10.23}$	$94.63_{9.68}$
		0.10	$94.19_{13.08}$	$95.23_{13.29}$	$95.18_{13.33}$	$94.90_{13.32}$	$95.18_{12.56}$
0.5	0.56	0.01	$94.63_{7.68}$	$95.12_{7.13}$	$95.23_{7.15}$	$95.23_{7.13}$	$95.18_{6.74}$
		0.05	$93.86_{11.43}$	$95.51_{10.61}$	$95.84_{10.63}$	$95.40_{10.64}$	$95.56_{10.07}$
		0.10	$94.47_{14.86}$	$94.47_{13.80}$	$94.41_{13.82}$	$94.30_{13.83}$	$94.36_{13.09}$
	0.64	0.01	$93.70_{7.34}$	$95.62_{6.84}$	$95.45_{6.88}$	$95.84_{6.84}$	$94.96_{6.48}$
		0.05	$93.53_{10.90}$	$94.58_{10.10}$	$94.47_{10.11}$	$94.74_{10.11}$	$94.36_{9.53}$
		0.10	$94.36_{14.11}$	$95.89_{13.14}$	$96.00_{13.16}$	$95.73_{13.18}$	$95.62_{12.39}$
	0.71	0.01	$94.90_{6.79}$	$95.78_{6.36}$	$95.67_{6.38}$	$95.73_{6.36}$	$95.62_{6.02}$
		0.05	$93.64_{10.08}$	$94.25_{9.38}$	$94.30_{9.39}$	$94.30_{9.40}$	$93.97_{8.88}$
		0.10	$94.41_{13.04}$	$94.58_{12.09}$	$94.58_{12.11}$	$94.68_{12.11}$	$94.68_{11.47}$

$^a r$: individual temporal correlation of the original scale. ^{b}WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. ^{c}ICC: intraclass correlation of coefficient.

Table 5.13: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with continuous outcomes generated from normal distributions. The control arm consists of 10 clusters and the treatment arm consists of 20 clusters. Entries are presented as coverage% $_{\text{confidence interval length} \times 100}$.

r^a	WinP^b	ICC^c	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
			Cluster size~ Binomial(50,0.5)				
0.3	0.56	0.01	94.30$_{9.85}$	94.96$_{10.04}$	95.07$_{10.08}$	94.74$_{10.04}$	94.63$_{9.53}$
		0.05	94.25$_{12.93}$	94.52$_{13.15}$	94.63$_{13.24}$	94.68$_{13.18}$	94.25$_{12.52}$
		0.10	93.81$_{15.93}$	94.52$_{16.19}$	94.47$_{16.33}$	94.58$_{16.25}$	93.97$_{15.42}$
	0.64	0.01	94.25$_{9.43}$	94.58$_{9.61}$	94.63$_{9.66}$	94.41$_{9.61}$	94.79$_{9.12}$
		0.05	93.70$_{12.34}$	94.58$_{12.55}$	94.58$_{12.64}$	94.68$_{12.58}$	94.30$_{11.95}$
		0.10	93.86$_{15.19}$	94.25$_{15.44}$	94.30$_{15.57}$	94.41$_{15.50}$	93.97$_{14.71}$
	0.71	0.01	93.92$_{8.76}$	94.36$_{8.94}$	94.36$_{8.98}$	94.52$_{8.94}$	94.90$_{8.48}$
		0.05	93.32$_{11.41}$	94.30$_{11.61}$	94.41$_{11.69}$	94.52$_{11.64}$	94.19$_{11.05}$
		0.10	93.70$_{14.02}$	94.30$_{14.26}$	94.25$_{14.38}$	94.30$_{14.31}$	93.86$_{13.58}$
0.5	0.56	0.01	94.30$_{9.85}$	94.85$_{9.19}$	95.01$_{9.22}$	94.96$_{9.19}$	94.41$_{8.71}$
		0.05	93.86$_{12.92}$	94.52$_{12.02}$	94.63$_{12.09}$	94.74$_{12.05}$	94.63$_{11.43}$
		0.10	93.48$_{15.93}$	94.41$_{14.80}$	94.52$_{14.90}$	94.30$_{14.86}$	94.47$_{14.08}$
	0.64	0.01	94.30$_{9.43}$	94.85$_{8.83}$	94.79$_{8.85}$	94.96$_{8.82}$	94.58$_{8.36}$
		0.05	93.64$_{12.34}$	94.52$_{11.49}$	94.58$_{11.55}$	94.52$_{11.52}$	94.41$_{10.93}$
		0.10	93.21$_{15.19}$	94.25$_{14.12}$	94.25$_{14.22}$	94.19$_{14.18}$	94.14$_{13.44}$
	0.71	0.01	94.08$_{8.76}$	94.25$_{8.23}$	94.19$_{8.26}$	94.25$_{8.23}$	94.74$_{7.80}$
		0.05	93.32$_{11.40}$	94.63$_{10.65}$	94.63$_{10.71}$	94.58$_{10.67}$	94.41$_{10.13}$
		0.10	93.59$_{14.02}$	94.41$_{13.06}$	94.30$_{13.15}$	94.47$_{13.11}$	93.97$_{12.43}$
			Cluster size~ Binomial(100,0.5)				
0.3	0.56	0.01	94.74$_{7.66}$	94.96$_{7.77}$	94.90$_{7.80}$	94.85$_{7.78}$	95.18$_{7.38}$
		0.05	94.79$_{11.50}$	94.85$_{11.65}$	95.18$_{11.71}$	94.96$_{11.68}$	95.51$_{11.05}$
		0.10	94.74$_{14.93}$	95.07$_{15.13}$	95.23$_{15.23}$	95.12$_{15.16}$	95.78$_{14.34}$
	0.64	0.01	95.01$_{7.31}$	95.56$_{7.43}$	95.34$_{7.45}$	95.51$_{7.43}$	95.62$_{7.06}$
		0.05	95.07$_{10.95}$	95.07$_{11.10}$	95.34$_{11.16}$	95.07$_{11.12}$	95.73$_{10.53}$
		0.10	94.85$_{14.21}$	95.29$_{14.40}$	95.12$_{14.50}$	95.12$_{14.43}$	95.89$_{13.65}$
	0.71	0.01	95.18$_{6.77}$	95.89$_{6.89}$	95.78$_{6.91}$	96.00$_{6.89}$	95.62$_{6.54}$
		0.05	95.40$_{10.08}$	95.07$_{10.23}$	95.40$_{10.28}$	95.18$_{10.24}$	95.89$_{9.70}$
		0.10	95.18$_{13.07}$	95.34$_{13.25}$	95.29$_{13.35}$	95.45$_{13.28}$	95.89$_{12.56}$
0.5	0.56	0.01	94.96$_{7.66}$	94.79$_{7.10}$	94.96$_{7.12}$	94.90$_{7.11}$	95.34$_{6.74}$
		0.05	94.74$_{11.50}$	94.85$_{10.62}$	94.85$_{10.66}$	94.96$_{10.64}$	95.62$_{10.07}$
		0.10	94.63$_{14.94}$	94.85$_{13.78}$	95.07$_{13.85}$	95.12$_{13.81}$	95.67$_{13.06}$
	0.64	0.01	95.12$_{7.32}$	95.56$_{6.80}$	95.56$_{6.81}$	95.40$_{6.80}$	95.78$_{6.45}$
		0.05	95.12$_{10.95}$	95.40$_{10.12}$	95.34$_{10.16}$	95.23$_{10.14}$	95.62$_{9.59}$
		0.10	94.79$_{14.21}$	95.01$_{13.12}$	95.12$_{13.19}$	95.07$_{13.15}$	95.67$_{12.44}$
	0.71	0.01	95.23$_{6.78}$	96.11$_{6.32}$	96.16$_{6.34}$	96.00$_{6.32}$	96.22$_{6.00}$
		0.05	95.18$_{10.09}$	95.45$_{9.34}$	95.51$_{9.37}$	95.29$_{9.35}$	95.67$_{8.85}$
		0.10	94.90$_{13.08}$	95.18$_{12.09}$	95.23$_{12.15}$	95.12$_{12.11}$	95.73$_{11.46}$

$^a r$: individual temporal correlation of the original scale. ^{b}WinP = Pr($Y_1 < Y_2$) + 0.5Pr($Y_1 = Y_2$). ^{c}ICC: intraclass correlation of coefficient.

Table 5.14: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with binary outcomes. Each arm consists of 5 clusters. Entries are presented as coverage%$_{\text{confidence interval length} \times 100}$.

r^a	WinP^b	ICC^c	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
			Cluster size~ Binomial(50,0.5)				
0.3	0.56	0.01	95.73$_{11.07}$	96.11$_{12.00}$	96.22$_{12.09}$	96.05$_{11.98}$	95.78$_{10.56}$
		0.05	94.74$_{13.77}$	94.63$_{14.85}$	94.74$_{14.94}$	94.58$_{14.82}$	93.75$_{13.03}$
		0.10	95.12$_{16.62}$	95.01$_{17.81}$	95.01$_{17.89}$	95.01$_{17.80}$	94.79$_{15.61}$
	0.64	0.01	94.96$_{12.57}$	94.79$_{13.65}$	94.58$_{13.76}$	94.74$_{13.62}$	94.30$_{11.79}$
		0.05	94.36$_{16.52}$	94.90$_{17.83}$	94.96$_{17.92}$	94.79$_{17.81}$	93.70$_{15.28}$
		0.10	93.86$_{20.21}$	93.92$_{21.73}$	94.03$_{21.83}$	93.70$_{21.71}$	92.49$_{18.54}$
	0.71	0.01	95.45$_{12.82}$	94.63$_{13.99}$	94.63$_{14.09}$	94.63$_{13.95}$	94.52$_{12.07}$
		0.05	93.64$_{17.16}$	93.59$_{18.57}$	93.81$_{18.68}$	93.53$_{18.54}$	92.55$_{15.84}$
		0.10	93.48$_{20.98}$	93.75$_{22.70}$	93.75$_{22.78}$	93.59$_{22.72}$	92.71$_{19.26}$
0.5	0.56	0.01	95.23$_{10.90}$	94.90$_{10.99}$	95.12$_{11.08}$	95.01$_{10.96}$	94.25$_{9.62}$
		0.05	94.30$_{13.84}$	94.58$_{13.82}$	94.79$_{13.91}$	94.47$_{13.81}$	94.41$_{12.09}$
		0.10	95.18$_{16.78}$	94.58$_{16.64}$	94.79$_{16.73}$	94.41$_{16.64}$	93.64$_{14.54}$
	0.64	0.01	95.07$_{12.41}$	95.07$_{12.72}$	95.23$_{12.83}$	95.01$_{12.69}$	95.01$_{11.03}$
		0.05	93.75$_{16.54}$	93.97$_{16.71}$	94.14$_{16.82}$	93.92$_{16.69}$	92.71$_{14.35}$
		0.10	94.30$_{20.53}$	94.52$_{20.79}$	94.47$_{20.90}$	94.41$_{20.78}$	92.88$_{17.71}$
	0.71	0.01	94.63$_{12.96}$	94.68$_{13.51}$	95.01$_{13.63}$	94.74$_{13.48}$	93.64$_{11.61}$
		0.05	94.96$_{17.04}$	94.41$_{17.57}$	94.74$_{17.69}$	94.63$_{17.55}$	93.15$_{14.95}$
		0.10	93.70$_{20.94}$	94.19$_{21.30}$	94.58$_{21.42}$	93.92$_{21.28}$	92.38$_{18.12}$
			Cluster size~ Binomial(100,0.5)				
0.3	0.56	0.01	94.36$_{8.40}$	95.45$_{9.06}$	95.45$_{9.09}$	95.40$_{9.05}$	94.30$_{7.97}$
		0.05	94.19$_{11.97}$	94.96$_{12.83}$	94.96$_{12.87}$	94.79$_{12.83}$	94.14$_{11.24}$
		0.10	94.74$_{15.44}$	94.79$_{16.55}$	94.90$_{16.58}$	94.52$_{16.55}$	93.81$_{14.39}$
	0.64	0.01	94.79$_{9.76}$	95.18$_{10.65}$	95.07$_{10.69}$	95.29$_{10.63}$	94.36$_{9.09}$
		0.05	94.03$_{14.52}$	94.19$_{15.68}$	94.19$_{15.72}$	94.19$_{15.68}$	93.15$_{13.33}$
		0.10	93.42$_{18.86}$	93.81$_{20.29}$	93.75$_{20.33}$	93.75$_{20.30}$	91.95$_{17.13}$
	0.71	0.01	94.52$_{9.95}$	95.18$_{10.82}$	95.23$_{10.86}$	95.07$_{10.82}$	93.92$_{9.30}$
		0.05	93.92$_{15.23}$	94.03$_{16.47}$	94.19$_{16.52}$	94.03$_{16.47}$	93.26$_{13.91}$
		0.10	94.52$_{19.60}$	94.58$_{21.14}$	94.58$_{21.17}$	94.52$_{21.15}$	93.70$_{17.80}$
0.5	0.56	0.01	94.08$_{8.34}$	94.90$_{8.45}$	94.85$_{8.49}$	94.85$_{8.44}$	93.75$_{7.40}$
		0.05	94.30$_{11.95}$	94.47$_{11.92}$	94.30$_{11.95}$	94.63$_{11.92}$	94.08$_{10.40}$
		0.10	94.47$_{15.29}$	94.68$_{14.90}$	94.79$_{14.93}$	94.85$_{14.90}$	93.53$_{13.06}$
	0.64	0.01	93.75$_{9.72}$	93.75$_{9.90}$	93.86$_{9.95}$	93.86$_{9.89}$	93.04$_{8.49}$
		0.05	93.53$_{14.84}$	94.36$_{15.13}$	94.30$_{15.17}$	94.36$_{15.14}$	93.04$_{12.81}$
		0.10	93.59$_{19.14}$	94.03$_{19.24}$	94.03$_{19.29}$	93.97$_{19.25}$	92.88$_{16.33}$
	0.71	0.01	94.90$_{10.02}$	94.41$_{10.40}$	94.52$_{10.45}$	94.30$_{10.39}$	93.92$_{8.88}$
		0.05	92.82$_{15.33}$	93.10$_{15.59}$	93.21$_{15.63}$	93.21$_{15.60}$	91.78$_{13.19}$
		0.10	93.92$_{19.65}$	92.93$_{19.98}$	93.04$_{20.04}$	93.04$_{19.97}$	92.38$_{16.91}$

[a] r: individual temporal correlation of the original scale. [b] WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. [c] ICC: intraclass correlation of coefficient.

Table 5.15: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with binary outcomes. Each arm consists of 15 clusters. Entries are presented as coverage% $_{\text{confidence interval length} \times 100}$.

r^a	WinP^b	ICC^c	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
			Cluster size~ Binomial(50,0.5)				
0.3	0.56	0.01	94.96$_{5.99}$	94.47$_{5.87}$	94.63$_{5.88}$	94.47$_{5.87}$	94.58$_{5.65}$
		0.05	94.36$_{7.92}$	93.97$_{7.72}$	94.03$_{7.74}$	93.86$_{7.75}$	93.97$_{7.44}$
		0.10	94.63$_{9.78}$	94.03$_{9.49}$	94.30$_{9.53}$	93.97$_{9.54}$	93.53$_{9.14}$
	0.64	0.01	94.68$_{6.65}$	93.92$_{6.51}$	93.97$_{6.53}$	94.30$_{6.52}$	94.08$_{6.24}$
		0.05	94.79$_{8.84}$	94.63$_{8.62}$	94.41$_{8.65}$	94.74$_{8.65}$	94.08$_{8.25}$
		0.10	94.36$_{10.93}$	94.68$_{10.62}$	94.96$_{10.67}$	94.58$_{10.68}$	94.14$_{10.14}$
	0.71	0.01	95.62$_{6.80}$	95.45$_{6.65}$	95.40$_{6.67}$	95.62$_{6.65}$	95.34$_{6.36}$
		0.05	95.56$_{9.01}$	94.63$_{8.78}$	94.63$_{8.82}$	94.74$_{8.81}$	93.86$_{8.40}$
		0.10	95.12$_{11.18}$	94.85$_{10.88}$	94.74$_{10.95}$	94.79$_{10.94}$	93.92$_{10.37}$
0.5	0.56	0.01	94.68$_{5.98}$	94.68$_{5.31}$	94.63$_{5.32}$	94.58$_{5.31}$	94.74$_{5.13}$
		0.05	94.58$_{7.92}$	94.63$_{6.92}$	94.58$_{6.94}$	94.68$_{6.94}$	94.30$_{6.70}$
		0.10	94.47$_{9.78}$	94.52$_{8.50}$	94.52$_{8.53}$	94.63$_{8.54}$	93.53$_{8.24}$
	0.64	0.01	94.85$_{6.60}$	93.97$_{6.54}$	93.97$_{6.56}$	93.86$_{6.53}$	94.14$_{6.23}$
		0.05	94.30$_{8.78}$	94.25$_{8.33}$	94.30$_{8.37}$	94.19$_{8.34}$	94.03$_{8.00}$
		0.10	93.86$_{10.90}$	93.92$_{10.03}$	94.03$_{10.08}$	93.86$_{10.07}$	93.48$_{9.73}$
	0.71	0.01	94.79$_{6.76}$	94.63$_{6.71}$	94.68$_{6.73}$	94.63$_{6.71}$	93.75$_{6.38}$
		0.05	94.03$_{8.96}$	94.03$_{8.51}$	94.25$_{8.56}$	94.14$_{8.53}$	93.48$_{8.17}$
		0.10	94.41$_{11.11}$	94.30$_{10.24}$	94.47$_{10.31}$	94.41$_{10.29}$	94.03$_{9.92}$
			Cluster size~ Binomial(100,0.5)				
0.3	0.56	0.01	95.62$_{4.62}$	95.73$_{4.51}$	95.73$_{4.52}$	95.67$_{4.52}$	95.01$_{4.35}$
		0.05	95.56$_{7.02}$	95.56$_{6.81}$	95.62$_{6.83}$	95.62$_{6.83}$	95.12$_{6.55}$
		0.10	95.12$_{9.13}$	94.74$_{8.84}$	94.79$_{8.86}$	94.90$_{8.86}$	94.52$_{8.48}$
	0.64	0.01	95.51$_{5.15}$	95.23$_{5.03}$	95.23$_{5.04}$	94.96$_{5.04}$	95.18$_{4.82}$
		0.05	94.90$_{7.83}$	95.12$_{7.62}$	95.18$_{7.64}$	95.29$_{7.64}$	95.01$_{7.26}$
		0.10	95.23$_{10.21}$	94.90$_{9.91}$	94.90$_{9.94}$	94.90$_{9.93}$	94.03$_{9.41}$
	0.71	0.01	95.62$_{5.25}$	95.23$_{5.14}$	95.23$_{5.15}$	95.23$_{5.14}$	95.07$_{4.91}$
		0.05	95.01$_{7.99}$	94.85$_{7.78}$	94.96$_{7.80}$	95.01$_{7.80}$	93.92$_{7.41}$
		0.10	95.01$_{10.43}$	94.85$_{10.14}$	94.96$_{10.18}$	94.68$_{10.17}$	94.25$_{9.61}$
0.5	0.56	0.01	95.51$_{4.63}$	95.18$_{4.09}$	95.45$_{4.10}$	95.23$_{4.09}$	95.62$_{3.95}$
		0.05	95.34$_{6.38}$	95.23$_{5.83}$	95.23$_{5.84}$	95.34$_{5.84}$	94.41$_{5.57}$
		0.10	94.85$_{8.25}$	94.74$_{7.51}$	94.68$_{7.52}$	94.79$_{7.53}$	94.14$_{7.15}$
	0.64	0.01	95.56$_{5.17}$	95.34$_{5.07}$	95.40$_{5.08}$	95.40$_{5.07}$	95.40$_{4.81}$
		0.05	95.01$_{7.85}$	94.58$_{7.29}$	94.63$_{7.31}$	94.63$_{7.31}$	93.86$_{6.77}$
		0.10	94.58$_{10.24}$	94.58$_{9.49}$	94.58$_{9.52}$	94.52$_{9.52}$	93.59$_{8.75}$
	0.71	0.01	95.34$_{5.25}$	95.23$_{5.15}$	95.18$_{5.16}$	95.40$_{5.15}$	95.07$_{4.90}$
		0.05	95.34$_{8.01}$	95.18$_{7.40}$	95.23$_{7.42}$	95.23$_{7.41}$	94.58$_{7.17}$
		0.10	95.51$_{10.46}$	95.73$_{9.39}$	95.73$_{9.42}$	95.78$_{9.41}$	95.12$_{9.20}$

[a] r: individual temporal correlation of the original scale. [b] WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. [c] ICC: intraclass correlation of coefficient.

Table 5.16: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with binary outcomes. The control arm consists of 10 clusters and the treatment arm consists of 20 clusters. Entries are presented as coverage% × 100 $_{\text{confidence interval length} \times 100}$.

r^a	WinP^b	ICC^c	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
			Cluster size∼ Binomial(50,0.5)				
0.3	0.56	0.01	$95.45_{5.81}$	$96.11_{6.03}$	$96.05_{6.01}$	$95.95_{6.02}$	$95.34_{5.77}$
		0.05	$95.01_{7.23}$	$95.62_{7.46}$	$95.12_{7.39}$	$95.73_{7.48}$	$95.56_{7.25}$
		0.10	$96.00_{8.84}$	$96.05_{9.08}$	$95.95_{8.97}$	$96.05_{9.12}$	$95.89_{8.86}$
	0.64	0.01	$94.36_{6.44}$	$95.73_{6.69}$	$95.23_{6.59}$	$95.78_{6.69}$	$94.74_{6.38}$
		0.05	$95.34_{8.45}$	$95.67_{8.72}$	$95.01_{8.53}$	$95.95_{8.75}$	$95.45_{8.48}$
		0.10	$94.41_{10.39}$	$95.18_{10.69}$	$94.47_{10.45}$	$95.45_{10.74}$	$95.18_{10.49}$
	0.71	0.01	$93.64_{6.56}$	$95.23_{6.82}$	$94.96_{6.71}$	$95.12_{6.83}$	$94.30_{6.51}$
		0.05	$95.45_{8.60}$	$96.00_{8.89}$	$95.45_{8.69}$	$96.05_{8.92}$	$95.95_{8.60}$
		0.10	$94.52_{10.61}$	$95.95_{10.94}$	$95.67_{10.66}$	$95.73_{10.99}$	$95.73_{10.67}$
0.5	0.56	0.01	$94.85_{5.78}$	$95.23_{5.59}$	$95.07_{5.56}$	$95.34_{5.59}$	$95.18_{5.36}$
		0.05	$95.34_{7.28}$	$95.12_{6.95}$	$94.74_{6.87}$	$95.23_{6.97}$	$95.34_{6.76}$
		0.10	$94.79_{8.86}$	$95.95_{8.42}$	$95.67_{8.31}$	$96.16_{8.45}$	$95.67_{8.25}$
	0.64	0.01	$94.90_{6.44}$	$95.23_{6.24}$	$95.01_{6.16}$	$95.12_{6.24}$	$94.68_{5.98}$
		0.05	$94.25_{8.44}$	$95.51_{8.13}$	$95.07_{7.96}$	$95.45_{8.15}$	$95.29_{7.93}$
		0.10	$95.89_{10.48}$	$96.11_{10.03}$	$95.89_{9.80}$	$96.11_{10.08}$	$96.05_{9.88}$
	0.71	0.01	$95.29_{6.54}$	$95.45_{6.48}$	$95.18_{6.39}$	$95.40_{6.48}$	$95.45_{6.18}$
		0.05	$95.40_{8.70}$	$96.49_{8.50}$	$95.89_{8.31}$	$96.55_{8.53}$	$96.05_{8.24}$
		0.10	$95.01_{10.68}$	$95.84_{10.32}$	$95.34_{10.09}$	$96.00_{10.36}$	$96.00_{10.11}$
			Cluster size∼ Binomial(100,0.5)				
0.3	0.56	0.01	$95.29_{4.37}$	$95.78_{4.52}$	$95.62_{4.48}$	$95.84_{4.53}$	$95.45_{4.36}$
		0.05	$94.90_{6.22}$	$95.73_{6.39}$	$95.12_{6.30}$	$95.95_{6.40}$	$95.51_{6.26}$
		0.10	$94.96_{8.01}$	$95.23_{8.20}$	$94.90_{8.07}$	$95.29_{8.22}$	$95.67_{8.13}$
	0.64	0.01	$95.78_{4.96}$	$96.44_{5.14}$	$96.11_{5.04}$	$96.44_{5.14}$	$96.00_{4.93}$
		0.05	$95.01_{7.42}$	$95.95_{7.64}$	$95.62_{7.44}$	$95.95_{7.65}$	$96.05_{7.45}$
		0.10	$95.34_{9.65}$	$96.11_{9.90}$	$95.67_{9.64}$	$96.16_{9.93}$	$96.11_{9.77}$
	0.71	0.01	$94.96_{5.07}$	$95.78_{5.28}$	$95.23_{5.16}$	$95.62_{5.29}$	$95.23_{5.06}$
		0.05	$95.51_{7.65}$	$95.89_{7.90}$	$95.23_{7.67}$	$95.95_{7.92}$	$95.12_{7.67}$
		0.10	$94.08_{9.90}$	$95.01_{10.18}$	$94.52_{9.91}$	$95.12_{10.21}$	$94.96_{9.97}$
0.5	0.56	0.01	$94.85_{4.39}$	$95.34_{4.23}$	$95.18_{4.18}$	$95.23_{4.23}$	$95.29_{4.06}$
		0.05	$94.90_{6.26}$	$95.84_{5.93}$	$95.51_{5.84}$	$95.73_{5.94}$	$96.11_{5.80}$
		0.10	$95.95_{8.11}$	$95.95_{7.63}$	$95.23_{7.50}$	$95.73_{7.64}$	$96.38_{7.57}$
	0.64	0.01	$94.03_{4.99}$	$95.07_{4.83}$	$94.41_{4.73}$	$95.01_{4.83}$	$95.29_{4.65}$
		0.05	$95.89_{7.42}$	$96.27_{7.13}$	$95.84_{6.95}$	$96.27_{7.15}$	$96.49_{7.00}$
		0.10	$95.67_{9.58}$	$96.05_{9.18}$	$95.62_{8.95}$	$96.11_{9.20}$	$96.27_{9.10}$
	0.71	0.01	$94.52_{5.06}$	$94.90_{4.99}$	$94.41_{4.89}$	$94.85_{5.00}$	$94.68_{4.78}$
		0.05	$93.86_{7.65}$	$95.62_{7.40}$	$95.01_{7.22}$	$95.78_{7.42}$	$95.56_{7.23}$
		0.10	$94.14_{9.82}$	$95.07_{9.44}$	$94.19_{9.19}$	$95.18_{9.47}$	$94.90_{9.29}$

$^a r$: individual temporal correlation of the original scale. ^{b}WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. ^{c}ICC: intraclass correlation of coefficient.

Table 5.17: Performance of two-sided 95% arsinh-transformed confidence intervals for WinP using four variance estimators in cluster randomization trials with binary outcomes. The control arm consists of 20 clusters and the treatment arm consists of 10 clusters. Entries are presented as coverage% $\times$ 100$_{\text{confidence interval length} \times 100}$.

r^a	WinP^b	ICC^c	Unadjusted	Size weight	ICC weight	Ratio estimator	Mixed model
			Cluster size~ Binomial(50,0.5)				
0.3	0.56	0.01	94.79$_{6.64}$	95.07$_{6.73}$	95.07$_{6.78}$	95.18$_{6.73}$	94.79$_{6.13}$
		0.05	94.41$_{8.79}$	94.74$_{8.84}$	94.96$_{8.94}$	94.58$_{8.87}$	93.42$_{8.06}$
		0.10	93.32$_{10.81}$	94.47$_{10.79}$	94.52$_{10.93}$	94.47$_{10.83}$	93.32$_{9.86}$
	0.64	0.01	95.07$_{7.51}$	95.12$_{7.62}$	95.40$_{7.69}$	95.07$_{7.62}$	93.97$_{6.88}$
		0.05	93.97$_{10.01}$	94.08$_{10.07}$	94.30$_{10.19}$	94.25$_{10.10}$	92.66$_{9.07}$
		0.10	93.26$_{12.36}$	94.25$_{12.37}$	94.79$_{12.55}$	94.30$_{12.43}$	92.66$_{11.11}$
	0.71	0.01	93.70$_{7.69}$	94.74$_{7.77}$	94.79$_{7.84}$	94.47$_{7.77}$	94.03$_{7.02}$
		0.05	93.32$_{10.21}$	94.74$_{10.25}$	94.79$_{10.38}$	94.58$_{10.29}$	92.88$_{9.25}$
		0.10	93.26$_{12.67}$	93.97$_{12.70}$	94.19$_{12.89}$	94.14$_{12.76}$	92.99$_{11.39}$
0.5	0.56	0.01	94.08$_{6.63}$	95.18$_{6.09}$	95.56$_{6.13}$	95.23$_{6.09}$	94.63$_{5.59}$
		0.05	93.70$_{8.78}$	93.81$_{7.89}$	94.25$_{7.97}$	94.14$_{7.92}$	93.59$_{7.28}$
		0.10	94.14$_{10.83}$	94.63$_{9.67}$	94.63$_{9.78}$	94.52$_{9.71}$	93.81$_{8.93}$
	0.64	0.01	94.30$_{7.43}$	94.74$_{7.70}$	94.96$_{7.77}$	94.79$_{7.70}$	93.59$_{6.98}$
		0.05	93.70$_{9.89}$	94.74$_{9.79}$	94.85$_{9.89}$	94.85$_{9.81}$	92.88$_{8.89}$
		0.10	93.26$_{12.28}$	93.86$_{11.75}$	94.19$_{11.88}$	93.92$_{11.78}$	92.71$_{10.79}$
	0.71	0.01	93.48$_{7.61}$	95.01$_{7.91}$	95.29$_{7.98}$	94.63$_{7.90}$	93.64$_{7.15}$
		0.05	93.53$_{10.12}$	94.47$_{10.04}$	94.68$_{10.14}$	94.36$_{10.06}$	92.99$_{9.10}$
		0.10	93.42$_{12.55}$	93.59$_{12.03}$	93.92$_{12.17}$	93.59$_{12.07}$	92.66$_{11.03}$
			Cluster size~ Binomial(100,0.5)				
0.3	0.56	0.01	94.25$_{5.13}$	95.29$_{5.16}$	95.67$_{5.20}$	95.40$_{5.16}$	94.47$_{4.72}$
		0.05	94.58$_{7.77}$	95.01$_{7.77}$	95.07$_{7.85}$	95.07$_{7.78}$	94.03$_{7.06}$
		0.10	93.97$_{10.08}$	94.36$_{10.03}$	94.79$_{10.16}$	94.36$_{10.05}$	93.75$_{9.10}$
	0.64	0.01	93.97$_{5.81}$	95.12$_{5.86}$	95.18$_{5.91}$	95.12$_{5.86}$	93.86$_{5.29}$
		0.05	93.92$_{8.83}$	94.36$_{8.86}$	94.68$_{8.96}$	94.30$_{8.87}$	92.71$_{7.96}$
		0.10	93.70$_{11.52}$	94.85$_{11.52}$	95.12$_{11.67}$	94.96$_{11.54}$	92.38$_{10.29}$
	0.71	0.01	95.23$_{5.94}$	95.18$_{6.01}$	95.29$_{6.06}$	95.23$_{6.01}$	94.30$_{5.42}$
		0.05	94.58$_{9.04}$	94.79$_{9.08}$	95.29$_{9.18}$	94.85$_{9.09}$	93.75$_{8.14}$
		0.10	94.25$_{11.81}$	94.96$_{11.84}$	95.34$_{12.00}$	94.85$_{11.86}$	93.37$_{10.53}$
0.5	0.56	0.01	94.58$_{5.14}$	95.51$_{4.70}$	95.62$_{4.72}$	95.67$_{4.70}$	95.07$_{4.29}$
		0.05	94.08$_{7.77}$	95.07$_{6.96}$	95.18$_{7.02}$	94.85$_{6.97}$	94.30$_{6.39}$
		0.10	94.36$_{10.08}$	94.14$_{8.97}$	94.25$_{9.06}$	94.08$_{8.99}$	94.08$_{8.24}$
	0.64	0.01	94.30$_{5.85}$	94.90$_{6.01}$	95.01$_{6.05}$	94.90$_{6.01}$	93.42$_{5.39}$
		0.05	94.14$_{8.86}$	95.12$_{8.51}$	95.51$_{8.58}$	95.18$_{8.52}$	93.97$_{7.80}$
		0.10	93.15$_{11.56}$	94.79$_{10.74}$	95.01$_{10.84}$	94.96$_{10.76}$	93.48$_{9.98}$
	0.71	0.01	93.64$_{5.95}$	94.52$_{6.11}$	94.74$_{6.15}$	94.25$_{6.11}$	93.70$_{5.50}$
		0.05	93.42$_{9.10}$	94.79$_{8.77}$	94.74$_{8.84}$	95.07$_{8.78}$	93.59$_{8.00}$
		0.10	93.48$_{11.86}$	95.23$_{11.07}$	95.45$_{11.17}$	95.34$_{11.09}$	93.70$_{10.24}$

[a] r: individual temporal correlation of the original scale. [b] WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. [c] ICC: intraclass correlation of coefficient.

Table 5.18: Performance of sample size formula to ensure that the lower limit of a two-sided 95% confidence interval is no less than 0.5 with 80% assurance probability for ordered category outcomes. The WinP is estimated with with the weighted least square approach (WLS) or the mixed model approach.

			WinP[a] = 0.56					WinP = 0.60				
				WLS		Mixed			WLS		Mixed	
r[b]	ICC[c]	s[d]	k	EAP[e]	ECP[f]	EAP	ECP	k	EAP	ECP	EAP	ECP
						Cluster size=25						
0.3	0.01	0.5	35	78.48	94.91	80.56	94.69	14	70.87	95.13	79.74	95.07
		1	32	78.09	93.48	79.35	93.81	12	71.14	95.51	78.04	95.73
		2	35	75.74	95.18	79.35	95.07	14	70.43	95.84	79.08	95.02
	0.05	0.5	62	81.38	95.02	83.52	94.74	23	74.26	95.02	79.35	94.80
		1	54	79.63	94.25	80.94	94.14	20	73.71	94.85	77.16	94.91
		2	62	80.12	95.56	81.87	95.35	23	75.68	94.80	80.45	94.52
	0.1	0.5	93	79.85	95.02	81.22	95.29	35	79.68	95.07	82.09	94.41
		1	84	78.53	95.40	79.35	95.40	30	76.67	94.47	78.59	93.98
		2	93	79.63	94.96	81.11	94.91	35	78.20	95.67	82.04	95.07
0.5	0.01	0.5	29	75.47	95.18	79.68	94.85	12	63.53	96.22	79.24	95.45
		1	26	75.90	95.13	78.15	94.91	10	65.61	95.35	74.64	95.07
		2	29	76.18	95.13	80.56	95.89	12	65.55	96.17	80.28	95.13
	0.05	0.5	51	76.67	94.63	79.13	94.85	18	66.37	95.56	75.74	94.25
		1	44	76.12	94.91	77.71	94.36	16	70.97	94.69	74.92	94.58
		2	51	77.55	94.74	79.52	94.69	18	66.59	95.62	75.79	95.13
	0.1	0.5	78	78.86	95.07	80.39	95.40	29	77.38	94.74	81.33	94.91
		1	68	79.46	95.78	80.45	95.89	26	77.33	94.52	79.63	94.30
		2	78	79.03	94.96	80.34	94.85	29	77.98	94.69	82.09	95.18
						Cluster size=50						
0.3	0.01	0.5	23	78.42	94.96	84.50	95.13	11	75.03	95.24	87.57	94.80
		1	22	81.49	95.24	83.84	95.45	10	78.70	95.02	86.69	95.07
		2	23	78.86	94.19	83.52	94.69	11	77.55	96.00	87.57	95.45
	0.05	0.5	49	80.94	94.85	82.58	94.58	20	77.44	96.00	84.17	95.51
		1	44	79.90	94.91	81.60	94.96	18	79.30	95.35	83.13	94.96
		2	49	80.12	95.67	82.75	95.40	20	76.78	96.44	83.46	95.67
	0.1	0.5	82	81.65	96.17	82.48	95.62	32	78.81	94.30	82.80	94.47
		1	72	81.49	95.18	82.04	95.24	28	78.48	95.07	80.89	95.18
		2	82	79.19	95.29	80.07	95.13	32	79.35	95.13	83.24	94.74
0.5	0.01	0.5	20	76.78	95.84	82.91	95.18	10	79.57	95.24	89.49	95.13
		1	18	78.15	94.96	81.93	94.63	8	68.84	95.45	81.71	95.45
		2	20	75.68	95.67	82.48	95.45	10	79.90	95.78	90.42	94.63
	0.05	0.5	41	78.20	94.91	81.87	95.13	17	77.38	95.35	85.05	94.85
		1	38	80.72	94.74	82.42	94.80	16	80.72	96.28	85.54	95.62
		2	41	77.82	95.45	80.34	95.95	17	76.45	94.91	83.57	93.70
	0.1	0.5	69	81.11	95.78	83.13	96.00	26	78.86	95.24	82.69	95.29
		1	60	80.72	94.91	81.65	94.91	24	80.72	95.84	83.13	95.78
		2	69	82.20	96.00	83.52	95.51	26	78.70	95.62	83.84	94.96

[a]WinP = Pr($Y_1 < Y_2$) + 0.5Pr($Y_1 = Y_2$). [b]r: individual temporal correlation of the original scale. [c]ICC: intraclass correlation of coefficient. [d]s: randomization ratio of treatment over control. [e]EAP:empirical assurance probability. [f] ECP: empirical coverage rate.

Table 5.19: Performance of sample size formula to ensure that the lower limit of a two-sided 95% confidence interval is no less than 0.5 with 90% assurance probability for ordered category outcomes. The WinP is estimated with with the weighted least square approach (WLS) or the mixed model approach.

			WinP[a] = 0.56					WinP = 0.60				
				WLS		Mixed			WLS		Mixed	
r[b]	ICC[c]	s[d]	k	EAP[e]	ECP[f]	EAP	ECP	k	EAP	ECP	EAP	ECP
Cluster size=25												
0.3	0.01	0.5	47	89.21	94.69	91.57	94.69	18	82.80	95.29	89.76	94.25
		1	42	88.99	94.74	89.92	95.07	16	85.98	94.58	89.21	94.74
		2	47	89.38	94.63	90.47	94.96	18	82.58	95.84	88.77	95.40
	0.05	0.5	81	89.05	94.80	90.31	94.58	30	86.53	94.58	89.98	94.36
		1	72	89.54	95.73	89.59	95.29	26	86.64	94.69	88.06	94.58
		2	81	89.76	95.02	90.31	95.24	30	85.16	94.52	88.94	93.48
	0.1	0.5	126	89.92	94.69	90.58	95.02	45	86.14	94.69	88.34	93.98
		1	112	89.65	94.63	90.09	94.80	40	88.17	94.69	89.27	94.41
		2	126	90.74	94.91	91.18	94.74	45	87.40	94.41	89.54	94.30
0.5	0.01	0.5	39	88.50	95.45	91.02	95.35	15	77.88	96.11	88.44	95.13
		1	34	87.46	94.36	88.44	94.30	12	78.42	96.55	84.72	95.45
		2	39	87.95	95.18	90.14	95.13	15	78.64	94.91	88.94	94.80
	0.05	0.5	68	88.72	94.96	90.20	95.51	24	84.01	94.19	88.94	94.47
		1	60	89.49	94.36	89.59	94.30	22	86.80	95.35	88.44	95.18
		2	68	90.58	96.00	91.46	96.00	24	83.02	94.69	87.40	94.36
	0.1	0.5	104	89.10	94.58	90.47	94.25	38	89.27	95.35	91.35	95.07
		1	92	90.80	94.85	91.13	94.74	34	88.06	93.70	89.43	93.15
		2	104	90.85	95.40	91.79	95.13	38	88.17	94.63	90.31	94.41
Cluster size=50												
0.3	0.01	0.5	31	91.29	94.41	93.10	94.36	13	87.84	96.22	94.30	95.84
		1	28	91.18	95.40	92.55	95.56	12	89.38	96.22	92.88	95.18
		2	31	91.35	94.63	93.15	94.69	13	87.95	95.62	94.03	95.13
	0.05	0.5	64	89.87	95.29	91.57	95.62	26	90.53	95.51	92.44	95.18
		1	60	91.84	95.24	92.22	95.13	24	91.84	96.00	93.21	95.67
		2	66	91.51	95.13	92.44	95.07	26	90.85	95.56	93.26	95.13
	0.1	0.5	108	88.66	95.29	89.81	95.67	41	89.65	94.74	91.29	94.80
		1	96	91.24	95.84	91.62	95.67	38	91.73	95.18	92.72	95.02
		2	108	89.65	95.07	89.92	95.29	41	89.38	95.67	91.24	95.56
0.5	0.01	0.5	26	90.14	94.80	92.28	94.96	11	81.65	95.29	92.44	94.63
		1	22	87.84	95.24	89.43	95.29	10	85.60	95.40	91.29	94.69
		2	26	90.36	95.35	93.37	94.69	11	83.95	95.89	92.77	94.74
	0.05	0.5	54	90.20	95.35	91.79	94.69	22	90.91	95.56	94.30	96.06
		1	50	91.35	95.95	92.17	95.67	20	89.65	95.62	92.55	95.40
		2	54	90.53	95.45	91.40	94.74	22	90.25	95.18	93.48	95.07
	0.1	0.5	90	89.38	95.07	90.58	95.02	35	90.53	96.00	92.83	95.51
		1	80	90.31	94.96	90.85	95.13	32	90.85	94.80	92.11	94.03
		2	90	89.38	95.84	90.53	95.78	35	89.49	95.40	91.68	94.85

[a] WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. [b] r: individual temporal correlation of the original scale. [c] ICC: intraclass correlation of coefficient. [d] s: randomization ratio of treatment over control. [e] EAP:empirical assurance probability. [f] ECP: empirical coverage rate.

Table 5.20: Relation between correlation parameters on the original scale and win fractions. Correlation parameters of win fractions are displayed with a tilde symbol. The columns with SE are the standard error of the correlation parameter estimated with the win fractions.

r^a	WinP^a	ICC^c	$\tilde{r}^d$	SE. $\tilde{r}$	$\tilde{r}_c^e$	SE. $\tilde{r}_c$	$\tilde{r}_a^f$	SE. $\tilde{r}_a$	$\widetilde{\text{ICC}}$	SE. $\widetilde{\text{ICC}}^g$
Low autocorrelation^h										
0.3	0.56	0.01	0.29	0.15	0.19	0.52	0.26	0.09	0.01	0.09
		0.05	0.29	0.16	0.23	0.49	0.25	0.17	0.05	0.13
		0.10	0.29	0.17	0.25	0.47	0.26	0.19	0.10	0.17
	0.64	0.01	0.29	0.16	0.09	0.57	0.23	0.11	0.01	0.09
		0.05	0.29	0.16	0.13	0.55	0.18	0.21	0.05	0.13
		0.10	0.29	0.17	0.16	0.54	0.19	0.24	0.09	0.17
	0.71	0.01	0.28	0.16	0.06	0.58	0.22	0.11	0.01	0.09
		0.05	0.28	0.16	0.09	0.57	0.15	0.22	0.05	0.13
		0.10	0.28	0.17	0.11	0.56	0.14	0.26	0.09	0.17
0.5	0.56	0.01	0.48	0.14	0.31	0.51	0.44	0.08	0.01	0.09
		0.05	0.48	0.15	0.38	0.47	0.42	0.15	0.05	0.13
		0.10	0.48	0.15	0.42	0.45	0.43	0.17	0.10	0.17
	0.64	0.01	0.48	0.14	0.15	0.56	0.39	0.11	0.01	0.09
		0.05	0.48	0.15	0.22	0.55	0.30	0.20	0.05	0.13
		0.10	0.48	0.16	0.27	0.53	0.31	0.23	0.09	0.17
	0.71	0.01	0.47	0.14	0.10	0.57	0.37	0.12	0.01	0.09
		0.05	0.47	0.15	0.14	0.57	0.25	0.22	0.05	0.13
		0.10	0.47	0.16	0.18	0.56	0.23	0.25	0.09	0.17
High autocorrelation										
0.3	0.56	0.01	0.29	0.02	0.31	0.26	0.30	0.09	0.01	0.01
		0.05	0.29	0.02	0.56	0.18	0.48	0.14	0.05	0.02
		0.10	0.28	0.03	0.67	0.14	0.61	0.13	0.10	0.03
	0.64	0.01	0.29	0.02	0.15	0.32	0.25	0.11	0.01	0.01
		0.05	0.29	0.03	0.31	0.28	0.31	0.19	0.05	0.02
		0.10	0.28	0.03	0.42	0.24	0.40	0.20	0.09	0.03
	0.71	0.01	0.28	0.02	0.10	0.33	0.23	0.11	0.01	0.01
		0.05	0.28	0.02	0.20	0.31	0.23	0.21	0.05	0.02
		0.10	0.28	0.03	0.28	0.29	0.28	0.23	0.09	0.03
0.5	0.56	0.01	0.48	0.02	0.39	0.24	0.46	0.08	0.01	0.01
		0.05	0.48	0.02	0.60	0.17	0.57	0.12	0.05	0.02
		0.10	0.48	0.02	0.70	0.13	0.67	0.11	0.10	0.03
	0.64	0.01	0.48	0.02	0.19	0.31	0.40	0.11	0.01	0.01
		0.05	0.48	0.02	0.33	0.27	0.38	0.18	0.05	0.02
		0.10	0.48	0.02	0.44	0.24	0.45	0.19	0.09	0.03
	0.71	0.01	0.47	0.02	0.12	0.33	0.37	0.11	0.01	0.01
		0.05	0.47	0.02	0.22	0.31	0.30	0.21	0.05	0.02
		0.10	0.47	0.02	0.30	0.28	0.33	0.23	0.09	0.03

[a] r: individual temporal correlation of the original scale. [b] WinP = $\Pr(Y_1 < Y_2) + 0.5\Pr(Y_1 = Y_2)$. [c] ICC: intraclass correlation of coefficient of the original scale. [d] $\tilde{r}$: individual temporal correlation of win fractions. [e] $\tilde{r}_c$: temporal correlation of cluster-mean win fractions. [f] $\tilde{r}_a$: autocorrelation of cluster mean win fractions. [g] $\widetilde{\text{ICC}}$: intraclass correlation of coefficient of win fractions. [h] autocorrelation of individuals, used for data generation.

Chapter 6

Illustrative examples

6.1 Introduction

We studied the finite sample property of the proposed methods in Chapter 3 and Chapter 4 through simulation studies in Chapter 5. The confidence intervals for the adjusted win probability (WinP) maintain the coverage rate even for small samples (five clusters for each arm). Confidence intervals based on the mixed model approach are consistently narrower than unadjusted intervals. Confidence intervals from the weighted least square approach are narrower than the adjusted intervals only when at least 30 clusters are randomized in the trial and at least medium strength of correlation between baseline and follow-up ($r > 0.5$).

We illustrate the statistical methods proposed in Chapter 3 and Chapter 4 with two published cluster randomization trials with ordinal outcomes. We illustrate how they can be analyzed with our methods. Sample sizes for future trials are estimated using different models to derive the outcome distribution and variance components for sample size estimation. The analyses in this chapter are only for illustration purposes of statistical methods, not for drawing clinical conclusions. Before analyzing both trials, we briefly review the methods in the next section.

An example `SAS` code of analyzing the data with our methods is provided in Appendix.

6.2 Illustrating methods

Our methods of confidence interval estimation for WinP adjusted for baseline measurement can be summarized in three steps:

1. Denote the jth outcome in arm i as Y_{ij}, $i = 1, 2$ and $j = 1, \cdots N_i$. Rank the outcome Y_{ij} within its own arm (r_{ij}) and the whole sample (R_{ij}), and obtain the win fraction for the outcome by $w_{ij} = (R_{ij} - r_{ij})/(N - N_i)$ where n_i is the size of arm i and $N = n_1 + n_2$ is the size of the trial. The win fraction w_{ij}^X can be obtained similarly for the baseline measurements.

2a. For the mixed model approach, regress the win fractions for the outcome by the win fractions for the baseline measurement and the treatment indicator in a random intercept model. The adjusted WinP is then obtained by dividing the regression coefficient of the treatment indicator by two and then adding 0.5, whereas the variance of adjusted WinP is obtained from the variance of the regression coefficient of the treatment indicator.

2b. For the weighted least square approach, calculate cluster-specific summaries (mean and summation) of win fractions for the outcome and baseline measurement first and apply one of the three (co)variance estimators in Chapter 3 with the weighted least square approach to obtain adjusted WinP and its variance.

3. Obtain the arsinh-transformed interval with equations (3.20) and (3.21). The degrees of freedom for the weighted least square approach are based on the Satterthwaite approximation, and the degrees of freedom for the mixed model approach are the same as the degrees of freedom for the regression coefficient.

To estimate the sample size required for the lower limit of WinP exceeding a certain threshold, denoted by WinP$_l$, the variance of estimating WinP can be obtained from hypothetical distributions or pilot data. To be specific the sample size can be obtained in three steps:

1. Obtain the win fractions of the outcomes from either pilot data or hypothetical distribution. For ordered categorical outcomes, the win fractions for each category can be obtained by summing the probability of inferior categories in the other arm plus half of the probability of the same category in the other arm. If prior knowledge of treatment effect is based on the common odds ratio, the probabilities of each category for the treatment arm can be obtained from the common odds ratio and the probabilities of each categories for the control arm, additional details can be found in Subsection 4.1.1.3.

2. Calculate the variance of win fractions in each arm and estimate the sample size required for an individually randomized trial focusing on estimating the $(1-\alpha)100\%$ confidence interval of WinP, with the lower limit exceeding WinP_l with $1-\beta$ probability. The formula is given by

$$N = \left(1 + \frac{1}{s}\right)\left\{\frac{z_{\alpha/2} + z_\beta}{[\mathrm{logit}(\mathrm{WinP}) - \mathrm{logit}(\mathrm{WinP}_l)]}\right\}^2 \frac{s\phi_1^2 + \phi_2^2}{\mathrm{WinP}^2(1-\mathrm{WinP})^2}, \qquad (6.1)$$

   where s is the randomization ratio of treatment over control and ϕ_i^2 is the variance of win fractions for arm i.

3. Adjust the sample size in (6.1) by the designing features of the trial. For cluster randomization trials, N should be multiplied by $\{1 + (m-1)\rho\}$, where m is the mean cluster size, and ρ is the intraclass correlation coefficient for the outcome. When baseline measurement is included in the analysis, the sample size is decreased by multiplying N by $1 - r^2$, where r is the correlation between baseline measurement and follow-up outcome.

Simulation results in Chapter 5 indicate that we safely use r and ρ of the original scale in our sample size formula, even though the formula is derived for correlation parameters of win fractions.

6.3 A trial for evaluating the treatment of Crohn's disease

As the first example, we consider the randomised evaluation of an algorithm for Crohn's treatment (REACT) trial (Khanna *et al.*, 2015), which is a cluster randomization trial aiming to evaluate the effectiveness of early combined immunosuppression (ECI). The trial randomizes practices to provide conventional care or ECI to their patients. Crohn's disease is a type of inflammatory bowel disease that can lead to abdominal pain, diarrhea and other complications. The exact cause of Crohn's disease is unknown, but genetic and immune responses are known associated factors.

The conventional care of Crohn's disease usually starts with the use of corticosteroids, followed by the use of antimetabolites and TNF antagonists if the symptoms are not controlled. Although such step-care avoids overtreatment of low risk patients, it also delays highly effective treatments for patients at greater risk of complications. Additionally, long-term use of corticosteroids is associated with infection and mortality; hence, ECI that applies the antimetabolites and TNF antagonists earlier could be a more appealing treatment strategy.

The severity of Crohn's disease was assessed by the Harvey-Bradshaw index (HBI), which is the score of summing five disease-related components (Harvey and Bradshaw, 1980). The five components are status of well-being (from 0=very well to 4=terrible), the severity of abdominal pain (from 0=none to 3=severe), number of liquid stools per day, presence of abdominal mass (from 0=none to 3=definite and tender), and the number of related complications (from 0 to 8). There is no clear interpretation of an one-unit increase or decrease of HBI, except a higher score indicates more disease activity. The trial used a cut-off of four points to dichotomize HBI and compared the proportion between the ECI group and the conventional management group because no widely accepted minimal clinical difference was established for HBI. The lack of meaningful units for the HBI left investigators no choice but to dichotomize the index even though statistical power could be compromised.

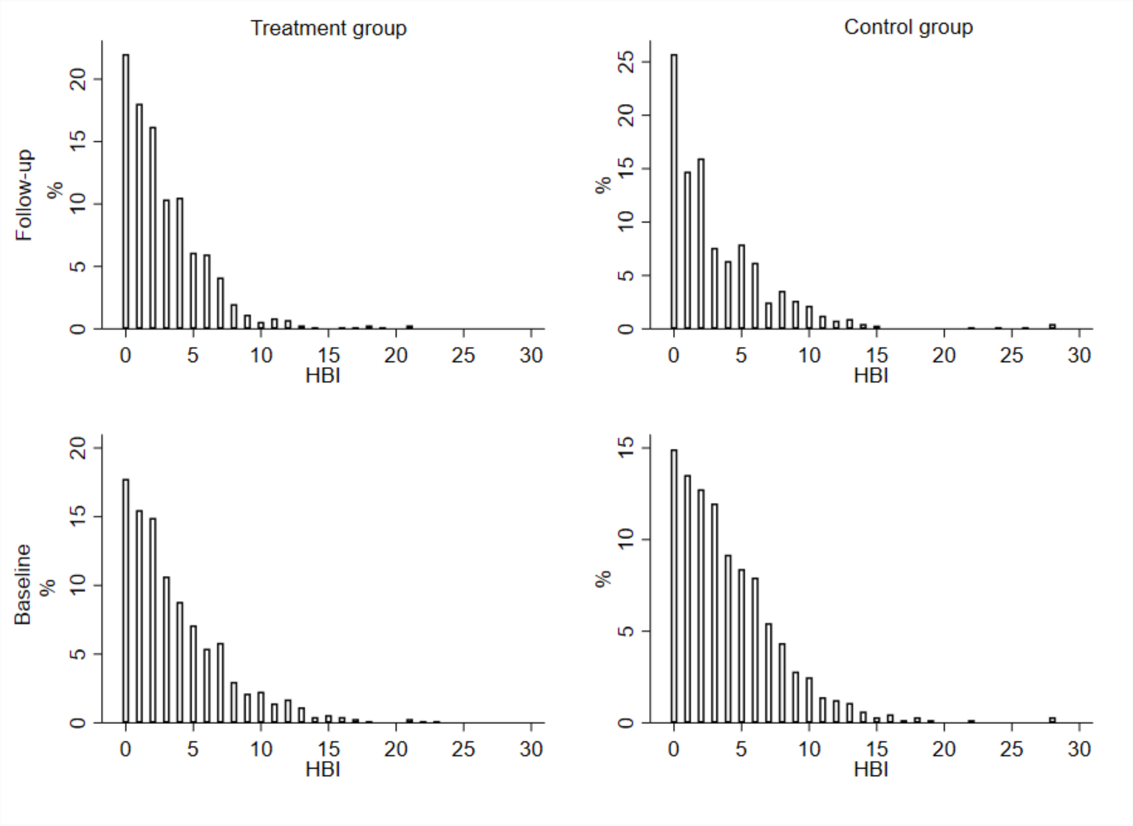

Figure 6.1: HBI score from the REACT trial. A higher HBI score indicates more severe Crohn's disease.

6.3.1 Estimation of the win probability

Zou (2021) has used this data set to illustrate the method for follow-up outcome only. We use the HBI at a two-year follow-up for this illustration and only include individuals with complete records for baseline and follow-up HBI. A total of 17 practices were assigned to the conventional management group, and 19 practices were assigned to the ECI group. There is a substantial variation of the cluster sizes (number of patients under a practice) due to loss of follow-up, with the smallest cluster having 7 patients and the largest having 58 patients. The HBI of both baseline and follow-up are displayed in Figure 6.1 where the distribution of both arms is highly overlapped.

After using step one in Section 6.1 to obtain the win fractions for the outcome from the ranks, we use a mixed model with the win fractions as the dependent variable and treatment indicator as the independent variable. We obtained the unadjusted WinP from dividing the regression coefficient by two and adding 0.5, which yields 0.49; this approach was used by Zou (2021) to illustrate estimating unadjusted WinP for this data set. To obtain adjusted WinP from the mixed model approach, we added the baseline win fraction as a covariate to the mixed model, yielding the regression coefficient for treatment indicator $\hat{\beta}_1 = 0.013$ and its standard error $\text{SE}(\hat{\beta}_1) = 0.023$, yielding $\widehat{\text{WinP}} = 0.507$ and its standard error as 0.023. The 95% arsinh-transformed confidence interval is obtained by first calculate with equation (3.20),

$$\begin{aligned}(l, u) &= \ln\frac{0.507}{1-0.507} \mp 2\text{arsinh}\left[t_{0.025,34}\frac{0.023}{2\times 0.507(1-0.507)}\right]\\ &= (-0.160, 0.212)\end{aligned}$$

where the degrees of freedom is obtained from the mixed model output. We did not consider small-sample adjustments on the degrees of freedom here because our data consists of 36 clusters. We then apply the inverse logit function to l and u yielding,

$$(L, U) = (0.460, 0.533) \tag{6.2}$$

and the p-value = 0.568 is obtained from the p-value of the regression coefficient.

To obtain adjusted WinP from the weighted least square approach, we first obtain the cluster-specific means and summation of follow-up and baseline win fractions, respectively. A list of the cluster-level summary statistics that are required in the weighted least square method is shown for four clusters in Table 6.1 for illustration. The size-weighted WinP and its variance are obtained from equations (3.3) and (3.4), respectively. The size-weighted (co)variances are estimated from equation (3.7) using cluster size, follow-up mean cluster-specific win fractions and baseline mean cluster-specific win fractions (columns 3,4 and 6 in Table 6.1). The ICC weighted estimator can be obtained similarly with the (co)variances estimated from equation (3.6), where the ICC is obtained from the analysis of variance of follow-up win fractions as

Table 6.1: Cluster-level summary statistics of four clusters for the REACT trial.

Cluster id	arm	size	Mean($w^Y_{ij.}$)[a]	Sum($w^Y_{ij.}$)[b]	Mean($w^X_{ij.}$)[c]	Sum($w^X_{ij.}$)[d]
1	0	42	0.475	19.954	0.451	18.942
2	0	47	0.669	31.425	0.735	34.545
3	1	51	0.456	23.273	0.361	18.391
4	1	24	0.465	11.148	0.378	9.074

[a] The cluster-specific mean of follow-up win fractions. [b] The cluster-specific summation of follow-up win fractions. [c] The cluster-specific mean of baseline win fractions. [d] The cluster-specific summation of baseline win fractions.

0.076. The ratio estimator is obtained from the cluster-specific summation of win fractions (columns 5 and 7 in Table 6.1) with the (co)variance estimated from equation (3.9). The P-values of testing WinP = 0.5 from these three estimators are obtained from t-tests where the degrees of freedom are obtained from Satterthwaite approximation in equation (3.17). The results without baseline adjustment and with baseline adjustment by the weighted least square approach and mixed model approach are presented in Table 6.2.

The adjusted analysis showed treatment effect in the opposite direction from the unadjusted results, with all of the adjusted WinP > 0.5. However, all the analyses showed non-significant effects. The proportion of HBI scores smaller or equal to four points at one-year follow-up also did not show a significant difference (Khanna *et al, 2015*).

The individual temporal correlation is 0.481 for the original scale and 0.486 based on win fractions, whereas the intraclass correlation coefficients are 0.050 and 0.076 based on the original scale and win fractions, respectively. The temporal cluster correlation of the small difference in ICC and individual temporal correlation between the original scale and win fractions confirms the analysis based on the original scale or win fractions should not have a big difference in the conclusion regarding the treatment effect.

6.3.2 Sample size estimation

The WinP in this trial is too small to be realistic for a future trial to plan sample size based on such WinP. Hence, we consider a hypothetical scenario where the pilot data have a total of

Table 6.2: Treatment effect estimation of the REACT trial.

Method	Estimate θ (95%CI)	p-value
Unadjusted	0.491 (0.429, 0.554)	0.784
Size weighted	0.519 (0.473, 0.565)	0.405
Ratio	0.521 (0.475, 0.568)	0.347
Size+ICC weighted	0.517 (0.469, 0.565)	0.470
Mixed model	0.507 (0.460, 0.553)	0.568

17 participants in the conventional group and 19 participants in the ECI group. The pilot data is generated from sampling one individual from each cluster, and the HBI in the ECI group is reduced to yield a WinP of 0.64. The variance of win fractions is 0.088 for the conventional group and 0.092 for the ECI group. We calculate the required sample size for an individual trial with WinP = 0.64 to have 80% chance to have the 95% confidence interval excluding WinP$_l$ = 0.56 corresponding to a small effect size using equation (6.1),

$$N = (1+1)(0.088+0.092)\left\{\frac{z_{0.2}+z_{0.025}}{[\text{logit}(0.64)-\text{logit}(0.56)]\,0.64(1-0.64)}\right\}^2 = 476.6\,,$$

yielding 478 individuals in a balanced trial.

For a cluster randomization trial with a mean cluster size of 40 and an ICC of 0.01, the required sample would then be increased to $N[1+(40-1)\times 0.01] = 662.4$, yielding a total of 663 individuals. However, since one only uses intact clusters of size 40, this means 9 clusters in each arm for a total of 720 individuals are required if no missing data occurs.

6.4 A trial for evaluating educational interventions of smoking prevention

As another example, we consider the television, school, and family smoking prevention and cessation project (TVSFP), which is a cluster randomization trial aiming to evaluate effectiveness of educational programs at preventing smoking for students (Flay *et al.*, 1995). Smoking was the leading preventable cause of mortality and morbidity in the United States; thus, delaying the onset of tobacco use or lower the prevalence of tobacco use was a major public health issue. The aim of the trial was to evaluate the effectiveness of social resistance curriculum and mass media for students on altering smoking behavior or prevent from experimental tobacco use. Previous studies mostly focused the effect of school-based intervention along, where TVSFP considers the complement effect from school-based and media-based intervention.

A total of 47 schools from Los Angeles and San Diego were randomized in a factorial design with two interventions: (i) television intervention and (ii) social-resistance classroom curriculum. For this illustration, we use a subset of the data that contains 28 schools from Los Angeles, where the data is available online at `https://content.sph.harvard.edu/fitzmaur/ala/tvsfp.txt` (accessed on August 18 2022).

The outcome is the knowledge score obtained from the number of questions correctly answered in the questionnaire, which ranges from 0 to 7. The data was analyzed by Hedeker and Gibbons (1994) with a mixed model comparing the mean difference of scores for different intervention arms. Mean comparison on the knowledge score can take on non-integer values, which leads to difficulty in interpreting mean differences. Additionally, the difficulty of each question is not the same. It is hard to argue whether a student with a score of 3 at baseline and 4 at follow-up shows the same improvement as a student with 6 at baseline and 7 at follow-up. The data was also analyzed by categorizing the score into four categories to make the categories have a more uniform proportion and fitted with a mixed logistic model (Raman and Hedeker, 2005), where the treatment effect is estimated with odds ratio. However, the categorization is rather arbitrary and could potentially lose information.

6.4.1 Win probability estimation

We consider only the social-resistance classroom curriculum as the intervention (14 schools) in this illustration, any school that did not receive the curriculum is in the control arm (14 schools). The sizes of the schools vary substantially from 18 students to 136 students. The proportion of the score at baseline and follow-up is displayed in Figure 6.2, where we can see that students

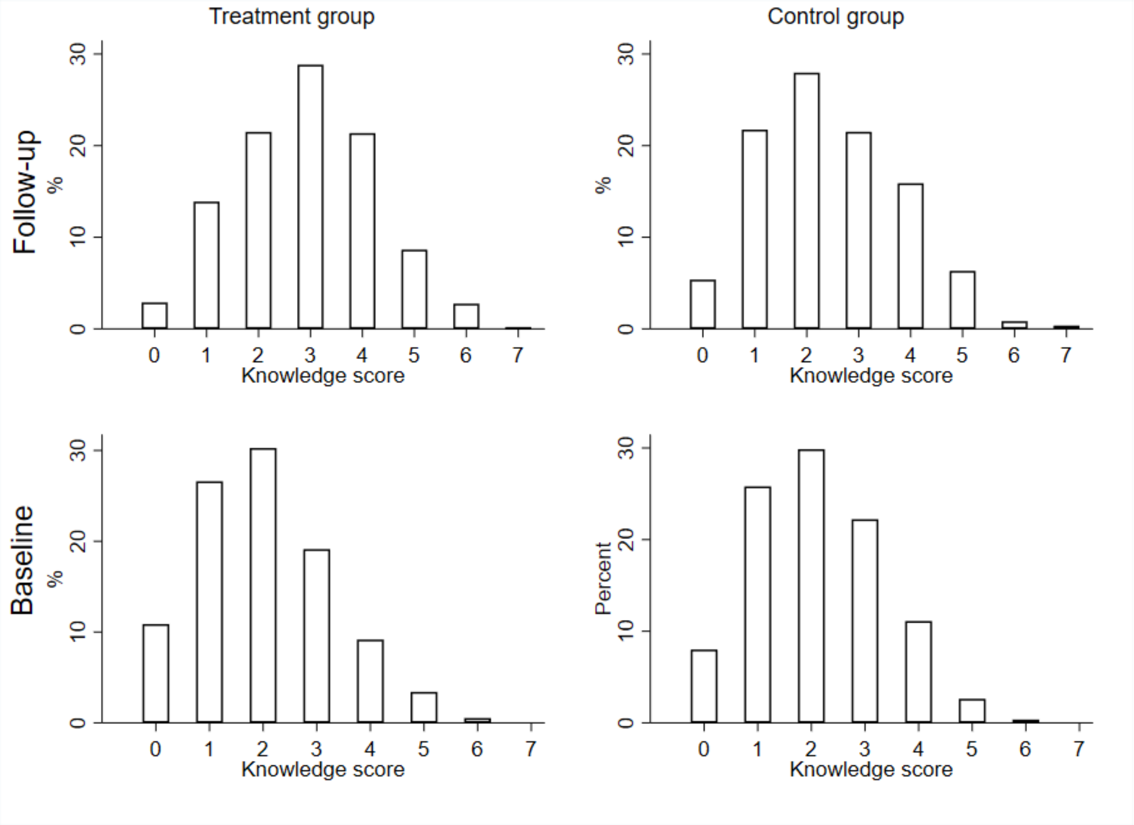

Figure 6.2: Knowledge score from the number of questions answered correctly regarding tobacco and health from the TVSFP study.

who received the curriculum tend to have higher knowledge scores. After obtaining the win fractions for the outcome with the steps in Section 6.1, we fitted a mixed model with the win fractions as the dependent variable and treatment indicator as the independent variable to obtain unadjusted WinP from dividing the regression coefficient by two plus 0.5, which is 0.59. To obtain adjusted WinP from the mixed model approach, we add in the baseline win fraction as a covariate to the mixed model, yielding the regression coefficient for treatment indicator $\hat{\beta}_1 = 0.200$ and its standard error $\text{SE}(\hat{\beta}_1) = 0.024$, yielding $\widehat{\text{WinP}} = 0.600$ and its standard error as 0.024. The 95% arsinh-transformed confidence interval is then obtained from equations (3.20) and (3.21) as $(0.551, 0.645)$, and the p-value < 0.001 is obtained from the p-value of the regression coefficient.

We follow similar steps in our illustration of the weighted least square approach for the REACT trial to obtain cluster-specific summaries of win fractions and use them with equations

(3.3) and (3.4) to obtain the adjusted WinP and its variance for three different (co)variance estimators, respectively. We presented the results without baseline adjustment and with baseline adjustment by the weighted least square approach and the mixed model approach in Table 6.3. All five estimates do not include 0.5 in the confidence interval, and adjusted analyses showed a reduction of confidence interval width around 15%. The individual temporal correlation is 0.3 for the original scale and 0.27 for win fractions, whereas the ICC is 0.04 and 0.05 for the original scale and win fractions, respectively. The data from TVSFP also showed small differences in ICC and individual temporal correlation between the original scale and win fractions.

Table 6.3: Treatment effect estimation of the TVFSP trial.

Method	Estimate θ (95%CI)	p-value
Unadjusted	0.593 (0.539, 0.646)	<0.001
Size weighted	0.612 (0.569, 0.653)	<0.001
Ratio	0.611 (0.568, 0.652)	<0.001
Size+ICC weighted	0.612 (0.567, 0.656)	<0.001
Mixed model	0.600 (0.551, 0.648)	<0.001

6.4.2 Sample size estimation

For illustrating obtaining win fractions for sample size estimation with different methods, we assume we do not have the pilot data for both arm, but we assume the treatment effect from prior studies suggested a common odds ratio of 3.5 and use the distribution of knowledge score of the no curriculum group to obtain the distribution for the treatment arm by equation (4.10). The win fractions and the proportions are listed in Table 6.4, which yields WinP = 0.61. The probabilities of each categories for both group in Table 6.4 can be used to obtain the variance of win fractions from equation (4.12) as 0.073 for the control arm and 0.072 for the treatment arm.

Suppose the aim of the trial is to determine whether or not the curriculum has at least a small effect size, which means whether or not the lower limit $\text{WinP}_l > 0.56$ is the interest of the trial. Additionally, suppose the stakeholders of the trial intend to have more schools receiving the curriculum, implying a randomization ratio of 2:1 for treatment to control ($s = 2$). We calculate the sample size for an individual trial to have 80% chance to have the lower limit of 95% confidence interval exceeding 0.56 using equation (6.1),

$$N = (1 + 1/2)\,(2 \times 0.073 + 0.072)\left\{\frac{z_{0.2} + z_{0.025}}{[\text{logit}(0.61) - \text{logit}(0.56)]\,0.61(1 - 0.61)}\right\}^2 = 1065.9\,,$$

yielding 1,066 individuals, or 356 individuals to the control arm and 712 individuals to the treatment arm.

The average school size of the data set is around 60 and the ICC is 0.04. Hence, a future trial randomizing schools from the same area would then need a total of 3,582 individuals ($N[1 + (60 - 1) \times 0.04] = 3581.4$), or 20 schools for the control arm and 40 schools for the treatment arm. If we include the baseline knowledge score in the analysis with a correlation with the follow-up as 0.3, the sample size would be $N[1 + (60 - 1) \times 0.04](1 - 0.3^2) = 3259.1$, yielding 3,260 individuals, or 18 schools for the control arm and 36 schools for the treatment arm if no missing data occurs.

Table 6.4: Proportions of the knowledge scores from pilot data corresponding to a common odds ratio of 3.5.

Knowledge score	0	1	2	3	4	5	6	7
Control	0.05	0.22	0.28	0.22	0.16	0.05	0.01	0.01
Treatment	0.01	0.08	0.16	0.23	0.30	0.14	0.04	0.04

6.5 Discussion

We have illustrated the estimation of WinP while adjusting for baseline with data from two published cluster randomization trials. The distributions of the outcomes in these two trials are different, resulting in different effect measures and different analysis methods in the literature. For example, the HBI score is dichotomized at 4 points in the REACT trial, which is a popular approach for ordinal outcomes when a meaningful effect measure is hard to define on the raw scale. However, dichotomizing outcomes could lead to loss of power (Altman and Royston, 2006). Additionally, such an approach complicates baseline adjustment because the non-linear link in the logistic or proportional odds model results in noncollapsibility. In contrast, the adjusted WinP from our method has the same interpretation as the unadjusted WinP.

The efficiency gained from baseline adjustment for the mixed model approach is similar to the weighted least square approach in both examples, which could be explained by both examples having sufficient number of clusters (around 30) for large sample theory to be relevant. Our simulation results in Chapter 5 also showed that both approaches have similar efficiency when 30 or more clusters are randomized.

We also illustrated sample size estimation of trials focusing on estimating WinP from pilot data or information derived from the two published cluster randomization trials. The key is to use the pilot data or information regarding the distribution of the outcomes to obtain the variance of win fractions and estimate the sample size as if individuals were randomized. The next step is to adjust the sample size by design considerations such as cluster randomization or baseline adjustment. Our method treats win fractions as continuous, making adjustments for design considerations more straightforward.

We have illustrated our statistical methods in designing and analyzing cluster randomization trials with baseline adjustment from two published trials. We will summarize this thesis and discuss the potential impact and future research in the next chapter.

Chapter 7

Summary and discussion

7.1 Summary

Ordinal outcomes are ubiquitous in randomized trials because the intervention effect often cannot be measured directly from physical phenomenons. By ordinal we mean both categorical and continuous data. It can appear as a single ordinal item such as the Rankin scale (Rankin, 1957), a summation of multiple ordinal items such as the Harvey-Bradshaw index (Harvey and Bradshaw, 1980), or a score derived from a questionnaire such as the number of correctly answered questions.

Ordinal outcomes do not have a clear clinical interpretation of a one-unit increase. Hence, comparing the mean of such outcomes is difficult to interpret as a treatment effect. Additionally, non-integer mean differences could be difficult to interpret. A better way to quantify the treatment effects for ordinal outcomes is the probability that a randomly chosen individual in the treatment arm wins over a randomly chosen individual in the control arm, i.e. the win probability (WinP). Estimation of WinP can be done using rank information.

We reviewed the literature on analyzing and sample size planning of WinP in Chapter 2 and categorized them into three approaches based on different variance estimators. The parametric approach derives the variance estimator based on distributional assumptions, such as normal distribution or other distributions from the exponential family. However, such distributional assumptions often are violated for ordinal outcomes due to the range restriction and skewness.

One nonparametric approach is based on all the pairwise comparisons between outcomes, where the pairwise comparison results in one if the treated outcome wins over the untreated outcome, half if the treated outcome is tied to the untreated outcome and zero otherwise. The pairwise comparisons are incorporated in a regression framework by the probabilistic index model (Thas *et al.*, 2012). It is computationally demanding as it regresses on $N_1 \times N_2$ pairs of pairwise comparisons, where N_i is the sample size of arm i. Moreover, the pairwise comparisons are correlated even for independent outcomes; hence, complicating the variance estimation and impeding the extension to clustered outcomes.

We have used an alternative approach that is based on the win fractions for each participant, which is the probability of such a participant winning over all the participants in the other arm. The win fractions are asymptotically independent of each other and can be analyzed with statistical models for continuous outcomes (Zou *et al.*, 2023), making it more feasible to extend the methods to correlated outcomes in cluster trials. Although the win fractions can be obtained from averaging the pairwise comparisons between outcomes, using the arm-specific ranks and overall ranks of outcomes to obtain the win fractions reduces the calculation time and complexity because only $N_1 + N_2$ win fractions are used for regression.

This thesis developed methods for confidence interval estimation and sample size estimation of WinP while adjusting for baseline measurements in cluster randomization trials by extending methods for continuous outcomes into a nonparametric framework with win fractions. Our methods for confidence interval estimation of WinP in Chapter 3 can be summarized in three steps: (i) obtaining the win fractions for the outcomes and the baseline measurements with ranks. The win fraction is the number of wins over the number of comparisons, where the number of wins of an outcome is obtained by the overall rank minus the rank in its own arm (ii) estimating the adjusted WinP and the variance with the mixed model approach or weighted least square approach to adjust for baseline measurements and clustering, and (iii) constructing confidence intervals based on logit-transformation or arsinh-transformation to improve coverage and balance of tail errors.

The extended weighted least square approach of Koch *et al.* (1998) uses the cluster-specific mean of win fractions to obtain point and variance estimates, whereas the mixed model approach regresses an individual's follow-up win fraction by baseline win fraction and treatment indicator. The main difference between the two approaches is that the efficiency gained for the weighted least square approach depends only on the temporal correlation of cluster-specific mean of win fractions, whereas the efficiency of the mixed model approach additionally depends on individual temporal correlation of win fractions.

We developed our methods based on asymptotic results; hence, we examined the finite sample properties by simulation studies. To be specific, we examined the coverage and width of confidence intervals constructed with our estimators under small sample sizes, with 10 or 30 clusters in total and an average of 25 or 50 individuals in each cluster. We considered small ($r = 0.3$) and medium ($r = 0.5$) individual temporal correlations and weak and strong autocorrelations resulting in weak or strong cluster-level temporal correlation.

Our simulation results showed that the mixed model approach often outperforms or is at least similar to the weighted least square approach in terms of efficiency because the mixed model depends on both individual and cluster temporal correlations; hence, the efficiency gained is less affected by the cluster temporal correlation, which is unreliably estimated with few number of clusters.

One advantage of the mixed model approach is that it is already available in most statistical software without writing additional scripts. It only needs to rank the outcomes to obtain the win fractions and then regress them with the mixed model while specifying variance heterogeneity. The WinP is obtained by dividing the regression coefficient of the treatment indicator by two plus 0.5, where the variance of the estimated WinP is directly obtained as the variance of the regression coefficient.

We developed sample size formulas in Chapter 4 by extending the sample size formula for continuous outcomes into a nonparametric framework by using individual win fractions as the analysis unit. Our method can be summarized into two steps: (i) estimate the sample size with the variance of win fractions as if independent individuals are randomized, and (ii) increase the sample size to account for the design effect, such as randomizing clusters or adjusting for baseline imbalance.

We illustrated different strategies to obtain the variance of win fractions from pilot studies or expert knowledge. Sample size formulas in the literature mostly focused on hypothesis testing, making them less useful for trials focusing on effect estimation. Our sample size formula is developed focusing on effect estimation; however, it can also be used for trials focusing on hypothesis testing because assurance probability is equivalent to power when the lower limit of WinP is 0.5.

Although, in theory, our sample size formula will have the highest precision when correlation and ICC of win fractions are known, our simulation results in Chapter 5 showed acceptable performance even if we use correlation and ICC of the original scale. We proved that the Spearman correlation of the original scale and the Pearson correlation of the win fractions are equal when WinP = 0.5. This might explain why the correlation of win fractions differs to the correlations of the original scale more when WinP is larger. However, it should not be a concern since studies with such a large effect are rare, and they only require a small sample size.

In summary, this thesis has developed statistical methods for interval estimation and sample size planning of WinP with baseline adjustments in cluster randomization trials. The validity of our methods was proven asymptotically and examined for finite samples with simulation studies.

7.2 Discussions

Ordinal continuous outcomes in the literature are often analyzed by mean comparisons followed by reporting the Cohen's effect size, which assumes the normality of outcomes. The normality assumption is often violated for ordinal outcomes due to the range restriction and skewness of the distribution. A simulation study by Zou *et al.* (2023) showed that the confidence intervals developed from the normality assumption could have under coverage when the normality assumption is violated. They investigated the confidence interval of WinP from an analysis of (co)variance model on the original scale, where WinP = $\Phi(\text{ES}/\sqrt{2})$ and ES is the Cohen's effect size obtained from the analysis of (co)variance model.

Note that the relationship between WinP and Cohen's effect size only holds for outcomes following normal distributions, which is often untenable in practice. Another problem with such an approach is that the mean difference of the original scale will always have decreased variance from baseline adjustment, leading to a bigger WinP. Although it can be avoided by using the variance of the unadjusted mean difference, it requires fitting two models to obtain the adjusted WinP.

The proportional odds model is an alternative option for ordinal outcomes, where the model assumes a common odds ratio across the categories of the outcome. However, such an assumption is often violated in practice and could lead to misleading results when it is violated. Furthermore, the noncollapsibility of the odds ratio complicates baseline adjustment and meta-analysis because adjusted odds ratios are different from the population-averaged odds ratio. In addition, the number of parameters needed to be estimated increases as the range of the outcome increases, making the proportional odds assumption more unplausible.

Ordinal outcomes with a wide range is not uncommon, as the HBI at two-year follow-up in the REACT trial ranged from 0 to 28 and at baseline ranged from 0 to 52. Although it is possible to collapse the outcome into fewer categories to avoid violation of the proportional odds assumption and make the model more parsimonious, the cutoffs to define the new categories are arbitrary, and different cutoffs could have different results. In addition, it makes comparing results from multiple studies complicated or impossible.

The problems associated with the parametric approach and proportional odds model for ordinal outcomes do not have simple solutions. It was until recently Zou (2021) proposed a rank-based method in estimating WinP focusing on the follow-up outcome, where the method accommodates any distributions as long as the outcomes can be ranked. We extended his method to include baseline measurements in the analysis. Cluster randomization trials often take additional measurements of the outcome at baseline before assigning intervention to the clusters, and it is desirable to include baseline measurement in the analysis to increase efficiency. Additionally, cluster randomization trials are more prone to baseline imbalance due to the recruitment of participants occurring after interventions are assigned to clusters.

We also derived sample size formulas for cluster randomization trials with ordinal outcomes, where only formulas for individually randomized trials (Zou *et al.*, 2022) are available prior to our work. Applying their formula to cluster randomization trials will result in underpowered studies.

Our methods build on the general ideal of Zou *et al.* (2023) that applies regression methods to win fractions of each outcome. The win fractions are obtained from the ranks of the outcome, making it applicable to any outcome as long as it can be ranked. Most other rank-based methods in the literature focused on hypothesis testing (Akritas, 1990; Akritas *et al.*, 1997) that cannot be used for confidence interval estimation. The potential of ranks in quantifying the treatment effect with confidence intervals has only been utilized recently (Zou, 2021; Zou *et al.*, 2023).

We note that ranking an outcome is inherently calculating the number of wins of such an outcome compared to others. Although such a link was pointed out long ago by Hoeffding (1948), it was only recently utilized by Zou (2021) to estimate WinP for cluster randomization trials. In fact, the WinP could be the only sensible effect measure for ranks because the mean rank difference is linearly related to WinP by the formula, mean rank difference = $N(\text{WinP} - 0.5)$.

A competing method in the literature to our method would be the probabilistic index model (PIM) by Thas *et al.* (2012), which regresses on the pairwise comparison of outcomes by covariates. The pairwise comparisons are correlated even for individually randomized trials because a participant can be compared with all the other participants in the other arm, respectively. Therefore, it is unclear how to extend the PIM for correlated outcomes. Additionally, the PIM could require a logit or a probit link function to produce estimates within the $[0, 1]$ range, where the link function can result in noncollapsibility, complicating the interpretation of the treatment effect after adjusting for baseline.

Estimating WinP with win fractions only requires an identity link in the regression model; hence, it is free from non-collapsibility, therefore, the baseline-adjusted WinP estimates the same underlying parameter as the unadjusted WinP. Another advantage is that it can be extended to meta-analysis with classical methods, as discussed by Zou *et al.* (2022). Finally, the PIM is computationally demanding as it regresses on $N_1 \times N_2$ pairwise comparisons. On the other hand, our method regresses on $N_1 + N_2$ win fractions, which is more suitable for cluster randomization trials as they usually have more participants.

Methods based on placement values (Hanley and Hajian-Tilaki, 1997) are similar to our methods as the placement values are the same as win fractions in the treatment arm and one minus win fractions in the control arm. However, placement values are difficult or impossible to be used in a regression model because the mean placement values are the same for both arms. On the other hand, win fractions can be easily used with linear regression models.

We have focused on our methods in cohort design, where the same individual is being followed through the trial. However, our method is also applicable to cross-sectional designs, where different individuals are assessed at different time points. The most intuitive way to do so is to apply the weighted least square approach because only the cluster means of win fractions are required.

7.3 Future research

We only developed methods for baseline adjustment by analysis of covariance (ANCOVA) and ignored changes from baseline analysis since it has lower power compared to ANCOVA in randomized studies (Van Breukelen, 2006). Change from baseline could be more useful for non-randomized studies by adjusting for preexisting differences (Van Breukelen, 2006). However, it is not clear whether the win fractions can be analyzed with a change from the baseline model and if the treatment parameter in such a model is meaningful. It is hard to argue if a change from baseline is comparable for two individuals measured by an ordinal scale. A more feasible research topic for this is the analysis of covariance with win fractions for non-randomized studies. We did not discuss this because treatment effects have different interpretations for randomized and non-randomized studies.

Our simulation study evaluated the validity of our methods and the efficiency gained from baseline adjustment. However, baseline adjustment can also increase the credibility of the results by making the two arms more comparable. When a considerable magnitude of imbalance at baseline occurs, it is hard to argue whether or not the treatment effect is confounded by the baseline imbalance. Previous research in the context of individually randomized trials has found that baseline adjustment improves the coverage of the confidence interval of WinP (Zou *et al.*, 2023), which could be due to the bias reduction of adjusted WinP.

We have focused our attention on cluster randomization trials with two intervention arms. However, trials with more than two intervention arms are not uncommon. For example, the TVSFP trial has four intervention arms due to the block design, where we only used the curriculum variable as an intervention for illustration. Extending our method to trials with more than two intervention arms can be one of the future research topics. For trials that compare several different interventions to one control, the WinP for each intervention compared to the control can be estimated with our method.

The WinP comparing different arms could be correlated since the control arm consists of the same participants. The work by DeLong *et al.* (1988) that compares multiple correlated areas under the receiver operating curve might be suitable to compare these correlated WinPs with the use of our (co)variance estimators based on cluster means of win fractions. Such an approach may be more appropriate for post hoc analyses as it is essentially pairwise comparing intervention arms.

Another approach to analyzing trials with more than two intervention arms is to define WinP as the probability of a participant winning to any outcomes in other arms, similar to the relative treatment effect with respect to all distributions in the trial defined by Brunner *et al.* (2017), except they also compare within-arm outcomes for factorial hypothesis testing purposes. In such an approach, the win fractions are obtained by calculating the wins of an outcome over all other arms, which can be obtained from the overall ranks minus the arm-specific ranks. This approach could be more useful in identifying the most effective combination of interventions in the study.

In Chapter 6, we illustrated our methods with the REACT trial, which randomizes gastroenterology practices to provide standard step-care or early combined immunosuprression to their patients with Crohn's disease. The primary outcome, Harvey-Bradhaw Index (HBI), was measured at baseline and once every six months, and we analyzed patients who completed the assessment at the two-year follow-up as an illustration. However, there were 32% of the participants had missing outcome at two-year follow-up, where the information of these participants could be partially recovered from the previous assessments. In addition, our analysis is only free from bias if we assume the missingness of an outcome is independent of both the value of the outcome and the baseline measurements Rubin (1976), i.e., missing completely at random; however, such an assumption is often unrealistic. For example, participants in the control arm that have worse outcomes could be less encouraged to show up in the trial, resulting in an underestimate of the treatment effect.

Extending our methods to incorporate missing data is especially useful as missing data is ubiquitous in medical research. It is common to assume the data is missing at random, where the missingness can be fully accounted by other observed values. Hence, the distribution of the missing values can be estimated from the observed values by multiple imputations, which often assume normality (Sterne *et al.*, 2009), making it inappropriate for ordinal outcomes. Other popular methods are also based on parametric assumptions, where the likelihood function needs to be specified (Carpenter and Smuk, 2021). One way to avoid parametric assumptions could be trying to use information from previous assessments to estimate the win fractions and include them in a mixed model as it provides unbiased estimates when the data are missing at random (Albert, 1999).

Since most trials make regular measurements of the primary outcome over time, the treatment effect could be quantified over time instead of just focusing on one time point of measurement. For example, the group difference of HBI over time was also tested by Khanna *et al.* (2015) in the REACT trial to see if the treatment benefits patients over time. Similarly, a trialist could ask if the WinP for treated participants increases through time, which can be answered by a mixed model of win fractions that includes treatment by time interaction (Albert, 1999). A positive coefficient of the interaction term indicates an increase in WinP over time, indicating an improvement through time.

Finally, comparing multiple endpoints is also an important future research topic because the efficacy of the intervention cannot be sufficiently measured with one variable in most cases. For example, one of the objectives of the REACT trial was to evaluate whether combining multiple drugs at once will increase drug-related complications compared to gradually administrating the drugs. Consequently, they compared adverse outcomes such as surgery, hospital admission and disease or drug-related complications, respectively.

Multiple endpoints are commonly analyzed by error-rate controlling methods (Sankoh *et al.*, 1997) that aim at controlling the overall type I error from multiple testing by assigning a smaller significance level for each individual test. However, Sankoh *et al.* (1997) showed in simulation studies that such an approach did not maintain type I error when the number of multiple testing increases and the correlation between the endpoints strengthens. Another approach is to consider summarizing the endpoints, where O'Brien (1984) proposed a rank-based test that could provide some insights. To be specific, he proposed to rank participants by different endpoints and add up all the ranks for each participant. His test then compares the rank-sum between intervention arms. Hence, adding up all the wins for each participant could potentially be a method to extend our method for multiple endpoints. Extensions of methods of O'Brien (1984) and Wei and Lachin (1984) to cluster randomization trials could be a fruitful research topic (Zou and Zou, 2023).

Bibliography

Agresti, A. (1999). Modelling ordered categorical data: recent advances and future challenges. *Statistics in Medicine* **18** (17-18), 2191–2207. doi:10.1002/(SICI)1097-0258(19990915/30)18:17/18<2191::AID-SIM249>3.0.CO;2-M.

Akritas, M. G. (1990). The rank transform method in some two-factor designs. *Journal of the American Statistical Association* **85** (409), 73–78. doi:10.2307/2289527.

Akritas, M. G., Arnold, S. F. and Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association* **92** (437), 258–265. doi:10.2307/2291470.

Albert, P. S. (1999). Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in Medicine* **18** (13), 1707–1732. doi:10.1002/(SICI)1097-0258(19990715)18:13<1707::AID-SIM138>3.0.CO;2-H.

Altman, D. G. and Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ* **332** (7549), 1080. doi:10.1136/bmj.332.7549.1080 .

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12** (4), 387–415. doi:10.1016/0022-2496(75)90001-2.

Bartlett, J. W. (2020). Robustness of ancova in randomized trials with unequal randomization. *Biometrics* **76** (3), 1036–1038. doi:10.1111/biom.13184.

Beal, S. L. (1989). Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics* **45** (3), 969–977. doi:10.2307/2531696.

Bland, J. M. and Altman, D. G. (1995*a*). Statistics notes: calculating correlation coefficients with repeated observations: part 1—correlation within subjects. *BMJ* **310** (6977), 446. doi:10.1136/bmj.310.6977.446.

Bland, J. M. and Altman, D. G. (1995*b*). Calculating correlation coefficients with repeated observations: part 2—correlation between subjects. *BMJ* **310** (6980), 633. doi:10.1136/bmj.310.6980.633.

Broderick, J. P., Adeoye, O. and Elm, J. (2017). Evolution of the modified rankin scale and its use in future stroke trials. *Stroke* **48** (7), 2007–2012. doi:10.1161/STROKEAHA.117.017866.

Brunner, E., Konietschke, F., Pauly, M. and Puri, M. L. (2017). Rank-based procedures in factorial designs: hypotheses about non-parametric treatment effects. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **79** (5), 1463–1485. link:http://www.jstor.org/stable/44682537.

Brunner, E. and Munzel, U. (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biometrical Journal* **42** (1), 17–25. doi:10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U.

Carpenter, J. R. and Smuk, M. (2021). Missing data: a statistical framework for practice. *Biometrical Journal* **63** (5), 915–947. doi:10.1002/bimj.202000196.

Casella, G. and Berger, R. L. (2001). *Statistical inference. 2nd Edition.* Cengage Learning.

Chakraborti, S., Hong, B. and van de Wiel, M. A. (2006). A note on sample size determination for a nonparametric test of location. *Technometrics* **48** (1), 88–94. doi:10.1198/004017005000000193.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10** (4), 417–451. doi:10.2307/3001616 .

Cochran, W. G. (1976). *Sampling Techniques. 3rd Edition.* Wiley & Sons.

Cohen, R. L. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Routledge, New York. doi:10.4324/9780203771587.

Colditz, G. A., Miller, J. N. and Mosteller, F. (1988). Measuring gain in the evaluation of medical technology the probability of a better outcome. *International Journal of Technology Assessment in Health Care* **4** (4), 637–642. doi: 10.1017/s0266462300007728.

D'Agostino, R. B., Campbell, M. and Greenhouse, J. (2006). The Mann–Whitney statistic: continuous use and discovery. *Statistics in Medicine* **25** (4), 541–542. doi:10.1002/sim.2508.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44** (3), 837–845. doi:10.2307/2531595.

Donner, A., Birkett, N. and Buck, C. (1981). Randomization by cluster: sample size requirements and analysis. *American Journal of Epidemiology* **114** (6), 906–914. doi:10.1093/oxfordjournals.aje.a113261.

Donner, A. and Klar, N. (1993). Confidence interval construction for effect measures arising from cluster randomization trials. *Journal of Clinical Epidemiology* **46** (2), 123–131. doi:10.1016/0895-4356(93)90050-B.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research.* Arnold, New York.

Dutta, S. and Datta, S. (2016). A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. *Biometrics* **72** (2), 432–440. doi:10.1111/biom.12447.

Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R. and Ukoumunne, O. C. (2004). Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clinical Trials* **1** (1), 80–90. doi:10.1191/1740774504cn006rr.

Emerson, J. D. and Moses, L. E. (1985). A note on the Wilcoxon-Mann-Whitney test for 2 × k ordered tables. *Biometrics* **41** (1), 303–309. doi:10.2307/2530667.

Fay, M. P. and Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys* **4**, 1–39. doi:10.1214/09-SS051.

Flay, B. R., Miller, T. Q., Hedeker, D., Siddiqui, O., Britton, C. F., Brannon, B. R., Johnson, C. A., Hansen, W., Sussman, S. and Dent, C. (1995). The television, school, and family smoking prevention and cessation project: VIII. student outcomes and mediating variables. *Preventive Medicine* **24** (1), 29–40. doi:10.1006/pmed.1995.1005.

Forrest, M. and Andersen, B. (1986). Ordinal scale and statistics in medical research. *BMJ* **292** (6519), 537–538. doi:10.1136/bmj.292.6519.537.

Gao, F., Earnest, A., Matchar, D. B., Campbell, M. J. and Machin, D. (2015). Sample size calculations for the design of cluster randomized trials: a summary of methodology. *Contemporary Clinical Trials* **42**, 41–50. doi:10.1016/j.cct.2015.02.011.

Greenland, S. (1988). On sample size and power calculations for studies using confidence intervals. *American Journal of Epidemiology* **128** (1), 231–237. doi:10.1093/oxfordjournals.aje.a114945.

Hanley, J. A. and Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. *Academic Radiology* **4** (1), 49–58. doi:10.1016/S1076-6332(97)80161-4.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** (1), 29–36. doi:10.1148/radiology.143.1.7063747.

Happ, M., Bathke, A. C. and Brunner, E. (2019). Optimal sample size planning for the Wilcoxon-Mann-Whitney test. *Statistics in Medicine* **38** (3), 363–375. doi:10.1002/sim.7983.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15** (4), 361–387. $doi : 10.1002/(SICI)1097 - 0258(19960229)15 : 4 < 361 :: AID - SIM168 > 3.0.CO; 2 - 4$.

Harvey, R. F. and Bradshaw, J. M. (1980). A simple index of Crohn's disease activity. *The Lancet (British edition)* **1** (8167), 514–514. doi:10.1016/s0140-6736(80)92767-1.

Hayes, R. J. and Moulton, L. H. (2017). *Cluster Randomised Trials, 2nd Edition*. Chapman and Hall/CRC, New York. doi:10.4324/9781315370286.

Hayter, A. J. (2012). Win-probabilities for regression models. *Statistical Methodology* **9** (5), 520–527. doi:10.1016/j.stamet.2012.02.002.

Hedeker, D. and Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50** (4), 933–944. doi:10.2307/2533433.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* **19** (3), 293–325. link:https://www.jstor.org/stable/2235637.

Hooper, R., Forbes, A., Hemming, K., Takeda, A. and Beresford, L. (2018). Analysis of cluster randomised trials with an assessment of outcome at baseline. *BMJ* **360:k1121**. doi:10.1136/bmj.k1121.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Satariano, S. A., Jewell, N., Bruckner, T. and Satariano, W. A. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology* **21** (4), 467–474. doi: 10.1097/EDE.0b013e3181caeb90.

Huskisson, E. (1974). Measurement of pain. *The Lancet* **304** (7889), 1127–1131. doi:10.1016/s0140-6736(74)90884-8.

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education* **38** (12), 1217–1218. doi:10.1111/j.1365-2929.2004.02012.x.

Kaiser, S., Träger, D. and Leisch, F. (2011). Generating correlated ordinal random values. Technical report University of Munich, Department of Statistics. url:https://epub.ub.uni-muenchen.de/12157/.

Kaner, E., Bland, M., Cassidy, P., Coulton, S., Dale, V., Deluca, P., Gilvarry, E., Godfrey, C., Heather, N., Myles, J., Newbury-Birch, D., Oyefeso, A., Parrott, S., Perryman, K., Phillips, T., Shepherd, J. and Drummond, C. (2013). Effectiveness of screening and brief alcohol intervention in primary care (sips trial): pragmatic cluster randomised controlled trial. *BMJ* **346:e8501**. doi:10.1136/bmj.e8501.

Kerry, S. M. and Bland, J. M. (2001). Unequal cluster sizes for trials in english and welsh general practice: implications for sample size calculations. *Statistics in Medicine* **20** (3), 377–390. doi:10.1002/1097-0258(20010215)20:3<377::AID-SIM799>3.0.CO;2-N.

Khanna, R., Bressler, B., Levesque, B. G., Zou, G. Y., Stitt, L., Greenberg, G. R., Panaccione, R., Bitton, A. and Pare, P. (2015). A cluster randomization trial of early combined immunosuppression for the management of Crohn's disease. *The Lancet* **386** (10006), 1825–1834. doi:10.1016/S0140-6736(15)00068-9.

Klar, N. and Darlington, G. (2004). Methods for modelling change in cluster randomization trials. *Statistics in Medicine* **23** (15), 2341–2357. doi:10.1002/sim.1858.

Koch, G. G., Tangen, C. M., Jung, J. W. and Amara, I. A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine* **17** (15–16), 1863–1892. doi:10.1002/(SICI)1097-0258(19980815/30)17:15/16<1863::AID-SIM989>3.0.CO;2-M.

Kupper, L. L. and Hafner, K. B. (1989). How appropriate are popular sample size formulas? *The American Statistician* **43** (2), 101–105. doi:10.2307/2527916.

Kuzon, W. M., Urbanchek, M. G. and McCabe, S. (1996). The seven deadly sins of statistical analysis. *Annals of plastic surgery* **37** (3), 265—272. doi:10.1097/00000637-199609000-00006.

Lachin, J. M. (2011). Power and sample size evaluation for the Cochran–Mantel–Haenszel mean score (Wilcoxon rank sum) test and the Cochran–Armitage test for trend. *Statistics in Medicine* **30** (25), 3057–3066. doi:10.1002/sim.4330.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** (4), 963–974. doi:10.2307/2529876.

Laupacis, A., Sackett, D. L. and Roberts, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine* **318** (26), 1728–1733. doi:10.1056/NEJM198806303182605.

Lee, A. J. (1990). *U-statistics: Theory and Practice (1st ed.).* Routledge, New York. doi:10.1201/9780203734520.

Lesaffre, E. and Senn, S. (2003). A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials. *Statistics in Medicine* **22** (23), 3583–3596. doi:10.1002/sim.1583.

Leyrat, C., Morgan, K. E., Leurent, B. and Kahan, B. C. (2018). Cluster randomized trials with a small number of clusters: which analyses should be used? *International Journal of Epidemiology* **47** (3), 1012–1012. doi:10.1093/ije/dyy057.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** (1), 13–22. doi:10.1093/biomet/73.1.13.

Likert, R. (1932). *A technique for the measurement of attitudes.* Archives of psychology ; No. 140, New York.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* **18** (1), 50 – 60. doi:10.1214/aoms/1177730491.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the mantel-haenszel procedure. *Journal of the American Statistical Association* **58** (303), 690–700. doi:10.2307/2282717 .

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI: Journal of the National Cancer Institute* **22** (4), 719–748. doi:10.1093/jnci/22.4.719 .

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M. and Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* **340:c869**. doi:10.1136/bmj.c869.

Moses, L. E., Emerson, J. D. and Hosseini, H. (1984). Analyzing data from ordered categories. *New England Journal of Medicine* **311**, 442–448. doi:10.1056/NEJM198408163110705.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17** (8), 857–872. doi:10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E.

Newcombe, R. G. (2001). Logit confidence intervals and the inverse sinh transformation. *The American Statistician* **55** (3), 200–202. doi:10.1198/000313001317098167.

Newcombe, R. G. (2006). Confidence intervals for an effect size measure based on the Mann–Whitney statistic. part 1: general issues and tail-area-based methods. *Statistics in Medicine* **25** (4), 543–557. doi:10.1002/sim.2323.

Noether, G. E. (1987). Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association* **82** (398), 645–647. doi:10.1080/01621459.1987.10478478.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40** (4), 1079–1087. doi:10.2307/2531158.

Obuchowski, N. A. (1997). Nonparametric analysis of clustered ROC curve data. *Biometrics* **53** (2), 567–578. doi:10.2307/2533958.

Perme, M. P. and Manevski, D. D. (2019). Confidence intervals for the Mann–Whitney test. *Statistical Methods in Medical Research* **28** (12), 3755–3768. doi:10.1177/0962280218814556.

Pocock, S. J. (2013). *Clinical Trials: a Practical Approach*. John Wiley & Sons, Ltd, New York. doi:10.1002/9781118793916.

Putter, J. (1955). The treatment of ties in some nonparametric tests. *The Annals of Mathematical Statistics* **26** (3), 368 – 386. doi:10.1214/aoms/1177728485.

Rahardja, D., Zhao, Y. D. and Y, Q. (2009). Sample size determinations for the Wilcoxon-Mann-Whiteny test: a comprehensive review. *Statistics in Biopharmaceutical Research* **1** (3), 317–322. doi:10.1198/sbr.2009.0016.

Raman, R. and Hedeker, D. (2005). A mixed-effects regression model for three-level ordinal response data. *Statistics in Medicine* **24** (21), 3331–3345. doi:10.1002/sim.2186.

Rankin, J. (1957). Cerebral vascular accidents in patients over the age of 60: ii. prognosis. *Scottish Medical Journal* **2** (5), 200–215. PMID: 13432835.

Reinert, D. F. and Allen, J. P. (2007). The alcohol use disorders identification test: an update of research findings. *Alcohol: Clinical and Experimental Research* **31** (2), 185–199. doi:10.1111/j.1530-0277.2006.00295.x.

Robinson, L. D. and Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review / Revue Internationale de Statistique* **59** (2), 227–240. doi:10.2307/1403444.

Rosner, B. and Glynn, R. J. (2009). Power and sample size estimation for the wilcoxon rank sum test with application to comparisons of c statistics from alternative prediction models. *Biometrics* **65** (1), 188–197.

Rosner, B. and Glynn, R. J. (2011). Power and sample size estimation for the clustered wilcoxon test. *Biometrics* **67** (2), 646–653. doi:10.1111/j.1541-0420.2010.01488.x.

Rosner, B. and Grove, D. (1999). Use of the Mann–Whitney U-test for clustered data. *Statistics in Medicine* **18** (11), 1387–1400. doi:10.1002/(SICI)1097-0258(19990615)18:11<1387::AID-SIM126>3.0.CO;2-V.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63** (3), 581–592. doi:10.2307/2335739.

Rutterford, C., Copas, A. and Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology* **44** (3), 1051–1067. doi:10.1093/ije/dyv113.

Sankoh, A. J., Huque, M. F. and Dubey, S. D. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine* **16** (22), 2529–2542. doi:10.1002/(SICI)1097-0258(19971130)16:22<2529::AID-SIM692>3.0.CO;2-J.

Schacht, A., Bogaerts, K., Bluhmki, E. and Lesaffre, E. (2008). A new nonparametric approach for baseline covariate adjustment for two-group comparative studies. *Biometrics* **64** (4), 1110–1116. doi:10.1111/j.1541-0420.2008.00994.x.

Sen, P. K. (1967). A note on asymptotically distribution-free confidence bounds for P(X < Y), based on two independent samples. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **29** (1), 95–102.

Shieh, G., Jan, S. and Randles, R. H. (2006). On power and sample size determinations for the Wilcoxon–Mann–Whitney test. *Journal of Nonparametric Statistics* **18** (1), 33–43. doi:10.1080/10485250500473099.

Simpson, J. M., Klar, N. and Donnor, A. (1995). Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *American Journal of Public Health* **85** (10), 1378–1383. doi:10.2105/AJPH.85.10.1378.

Snedecor, G. W. and Cochran, W. G. (1989). *Statistical Methods. 8th Edition.* Iowa State University Press.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M. and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338:b2393**. doi:10.1136/bmj.b2393.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science* **103** (2684), 677–680.

Svensson, E. (2000). Comparison of the quality of assessments using continuous and discrete ordinal rating scales. *Biometrical Journal* **42** (4), 417–434. doi:10.1002/1521-4036(200008)42:4<417::AID-BIMJ417>3.0.CO;2-Z.

Teerenstra, S., Eldridge, S., Graff, M., de Hoop, E. and Borm, G. F. (2012). A simple sample size formula for analysis of covariance in cluster randomized trials. *Statistics in Medicine* **31** (20), 2169–2178. doi:10.1002/sim.5352.

Thas, O., De Neve, J., Clement, L. and Ottoy, J. (2012). Probabilistic index models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **74** (4), 623–671. doi:10.1111/j.1467-9868.2011.01020.x.

Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline: more power in randomized studies, more bias in nonrandomized studies. *Journal of clinical epidemiology* **59** (9), 920—925. doi:10.1016/j.jclinepi.2006.02.007.

Varnell, S. P., Murray, D. M., Janega, J. B. and Blitstein, J. L. (2004). Design and analysis of group randomized trials: a review of recent practices. *American Journal of Public Health* **94** (3), 393–399. doi:10.2105/AJPH.94.3.393.

Vickers, A. J. and Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. *BMJ* **323:1123**, 1123–1124. doi:10.1136/bmj.323.7321.1123.

Wang, B., Ogburn, E. L. and Rosenblum, M. (2019). Analysis of covariance in randomized trials: more precision and valid confidence intervals, without model assumptions. *Biometrics* **75** (4), 1391–1400. doi:10.1111/biom.13062.

Wei, L. J. and Lachin, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association* **79** (387), 653–661. doi:10.2307/2288413 .

Wellek, S. (2017). A U-statistics based approach to sample size planning of two-arm trials with discrete outcome criterion aiming to establish either superiority or noninferiority. *Statistics in Medicine* **36** (5), 799–812. doi:10.1002/sim.7183.

Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in Medicine* **12** (24), 2257–2271. doi:10.1002/sim.4780122404.

Yang, L. and Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician* **55** (4), 314–321. doi:10.1198/000313001753272466.

Zhao, Y. D., Rahardja, D. and Qu, Y. (2008). Sample size calculation for the Wilcoxon–Mann–Whitney test adjusting for ties. *Statistics in Medicine* **27** (3), 462–468. doi:10.1002/sim.5466.

Zimmermann, H. and Rahlfs, V. W. (2014). Comments on number-needed-to-treat derived from ordinal scales. *Statistical Methods in Medical Research* **23** (1), 107–110. doi:10.1177/0962280212469202.

Zou, G. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in Medicine* **31** (29), 3972–3981. doi:10.1002/sim.5466.

Zou, G. (2021). Confidence interval estimation for treatment effects in cluster randomization trials based on ranks. *Statistics in Medicine* **40** (14), 3227–3250. doi:10.1002/sim.8918.

Zou, G., Donner, A. and Klar, N. (2005). Group sequential methods for cluster randomization trials with binary outcomes. *Clinical Trials* **2** (6), 479–487. doi:10.1191/1740774505cn126oa.

Zou, G., Smith, E. and Jairath, V. (2022). A nonparametric approach to confidence intervals for concordance index and difference between correlated indices. *Journal of Biopharmaceutical Statistics* **32** (5), 740–767. doi:10.1080/10543406.2022.2030747.

Zou, G., Smith, E. J., Zou, L., Qiu, S. and Shu, D. (2023). A rank-based approach to design and analysis of pretest-posttest randomized trials, with application to COVID-19 ordinal scale data. *Contemporary Clinical Trials* **126**. doi:10.1016/j.cct.2023.107085.

Zou, G. and Zou, L. (2023). Nonparametric methods for randomized controlled trials with multiple endpoints: Beyond O'Brien-Wei-Lachin. In: European Meeting of Statisticians 2023 Warsaw 3–7 July, 2023 Book of Abstracts (p. 272).

Zou, G., Zou, L. and Choi, Y. (2022). Distribution-free approach to the design and analysis of randomized stroke trials with the modified rankin scale. *Stroke* **53** (10), 3025–3031. doi:10.1161/STROKEAHA.121.037744.

Zou, G., Zou, L. and Qiu, S.-f. (2023). Parametric and nonparametric methods for confidence intervals and sample size planning for win probability in parallel-group randomized trials with likert item and likert scale data. *Pharmaceutical Statistics* **22** (3), 418–439. *Pharmaceutical Statistics*. doi:10.1002/pst.2280.

Appendix A

SAS code for WinP estimation of TVSFP trial

```
/* Import TVSFP data
x : baseline Kscore
y: follow-up Kscore
s_i : treatment indicator 1=treatment 0=control
schoolID: ID for cluster
*/
/*1. convert data to win fractions */
proc sort data=tvsfp; by descending s_i;
/*obtain overall rank */
proc rank data= tvsfp out=overall;
var x y;
ranks overallx overally; run;
/*obtain within group rank */
proc rank data=tvsfp out=group; by  descending s_i;
var x y;
```

```
ranks groupx groupy; run;
ods listing close;
/*obtain size for the other arm (N-N_i) */
proc freq data=tvsfp;
tables s_i/out=size4other(drop=percent); run;
data size4other;
set size4other;
s_i = 1 - s_i; run;
data WinF;
merge overall group size4other; by descending s_i;
winf =(overally - groupy)/count;
winf_base =(overallx - groupx)/count;
keep schoolID s_i winf_base winf x y; run;
proc print data=WinF (obs=10);
run;
/*2. Regression on win fractions*/
proc mixed data= WinF method = reml;
class schoolID s_i/ref=FIRST;
model winf = s_i winf_base / ddfm = betwithin;
random intercept/subject= schoolID(s_i) type=cs;
lsmeans s_i/diff cl;
ods output Diffs= est (keep = Estimate StdErr DF);
run;
/*3. Obtain confidence interval and point estiamtes */
data results;
merge est;
alpha=0.05;
```

```
point = (Estimate+1)/2;
lgtPoint = log(point/(1-point));
crit = tinv(1- alpha/2, Df);
se1 = StdErr/(point*(1-point));
l1 = lgtPoint - crit * se1;
u1 = lgtPoint + crit * se1;
logitlower = logistic(l1);
logitupper = logistic(u1);
l2 = lgtPoint - 2*arsinh( crit/2* se1 );
u2 = lgtPoint + 2*arsinh( crit/2* se1 );
asinelower = logistic(l2);
asineupper = logistic(u2);
ttest= (point -.5)/StdErr;
pvalue = 2*(1-probt(abs(ttest), DF));
run;
proc print data=results;
var  point logitlower logitupper asinelower asineupper pvalue;
run;
```

Curriculum Vitae

Chengchun (Edward) Yu

Education

Western University Sept 2017 – Sept 2023
PhD, Epidemiology and Biostatistics *Thesis topic: cluster randomization trials, nonparametric effect estimation*

National Taiwan University Sept 2014 – Aug 2016
MSc, Biostatistics *Thesis topic: recurrent survival analysis, informative censoring, correlation analysis*

National Chengchi University Sept 2010 – Aug 2014
BSc, Double major in mathematics and statistics

Professional experience

Research assistant Jan 2018 – Jan 2021
Western University *London, ON*

- Reported results of Alzheimer research in publishable format including tables, listing and graphs
- Analyzed the predictive value of novel cognitive markers on Alzheimer's disease and developed prediction model while optimizing multiple prediction performance indices including C-statistics, IDI and NRI, achieving 83% prediction accuracy

Teaching assistant Sept 2020 – Apr 2021
Western University *London, ON*

- Graded assignments and exams regarding statistical theories and practice in clinical research including hypothesis testing, confidence interval estimation and sample size estimation
- Taught sessions of `SAS` and `R` for graphical presentation and statistical analysis from clinical trials and observational studies

Research assistant May 2016 - May 2017
National Taiwan University *Taipei, Taiwan*

- Designed and conducted simulation studies of survival data with recurrent events for new statistical methods with Fortran/C
- Conducted literature review, proposed new statistical methodologies and collaborated with physicians and faculty members

Financial modelling intern Jan 2015 – Jan 2017
KPMG, accounting firm *Taipei, Taiwan*

- Conducted statistical and graphical analysis of default risk data from multiple banks using `R`, evaluating over 1.5 billion assets
- Pioneered model selection with stepwise model selection, and cross-validated prediction models, reaching accuracy rates above 90% and outperforming the market practices
- Collaborated with experts in financial modelling and developers to develop applicants for loan approval with `Python`, automating 80% of modelling processes

Specialized Skills

Statistical analysis: generalized linear models, hypothesis testing, nonparametric covariate adjustment, imputation of missing values, model selection, cross-validation, model calibration.
Statistical programming: R, SAS (macro,IML), SPSS, STATA.
Word processing programs: LaTeX, MS word.
Other programs: C, Fortran, SQL, Python, MS Excel.