Title
Optimal Design of Cluster Randomized Trials with Binary Outcomes

Author
Wu, Sheng

Publication Date
2015

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Optimal Design of Cluster Randomized Trials with Binary Outcomes

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Sheng Wu

2015

Abstract of the Dissertation

Optimal Design of Cluster Randomized Trials with Binary Outcomes

by

Sheng Wu

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2015

Professor Weng Kee Wong, Co-chair

Professor Catherine Crespi-Chun, Co-chair

Cluster randomized trials (CRTs) are increasingly used in many fields including public health and medicine. We consider two-arm CRTs with binary outcomes with possibly unequal intraclass correlations coefficients (ICCs) in the two arms. The efficacy of the intervention may be measured in terms of the risk difference (RD), relative risk (RR) or odds ratio (OR). We define cost efficiency (CE) as the ratio of the precision of the efficacy measure to the study cost and develop optimal allocations to the two arms for maximizing CE. The optimal design, which is based on the optimal allocation, could be different for different measures. We define relative cost efficiency (RCE) of a design as the ratio of its CE to CE of the optimal design and use RCE to compare different designs. Because the optimal allocation can be highly sensitive to the unknown ICCs and anticipated success rates, we propose a Bayesian method and a maximin method to construct an efficient and robust design. We show that the RCE of the designs based on the Bayesian method or the maximin method is generally larger than the balanced design. Based on the optimal allocation, we derive optimal sample size formulas which satisfy the power requirement and minimize the total study cost. All the results above are based on the assumption of constant cluster size. When there is extreme variation in cluster size, the usually used sample size formula assuming a constant cluster size may result in a design with low power. Assuming a balanced design, we develop a sample

size formula for a two-arm CRT which obtains the desired power even though the cluster sizes are very different. This formula can be modified to incorporate optimal allocation consideration, hence it minimizes the study cost while satisfying the power requirement for a CRT with varying cluster sizes. Simulation is used to verify that our formulas can obtain the desired power.

The dissertation of Sheng Wu is approved.

Damla Senturk

May Wang

Catherine Crespi-Chun, Committee Co-chair

Weng Kee Wong, Committee Co-chair

University of California, Los Angeles

2015

To my family.

Table of Contents

List of Figures

List of Tables

Acknowledgments

My deepest gratitude goes first to my advisors Professor Catherine Crespi and Professor Weng Kee Wong. They introduced cluster randomized trials and optimal design theory to me. They helped me to generate the initial idea of this dissertation and gave me inspiring suggestions and consistent encouragement. Without their guidance and help, I would never have been able to complete this dissertation. In addition, I want to thank Professor May Wang and Professor Damla Senturk for kindly serving on my doctoral committee and spending time reviewing my manuscript and providing insightful comments. My thanks also go to Professor Thomas Belin, who helped me to obtain financial support during my PhD study. Finally, I thank all my friends for their support and encouragement in my difficult time.

Vita

2005	Bachelor of Medicine, Peking University, Beijing, P.R.China.
2008	M.S. (Physiology), University of Nebraska Medical Center, NE, USA.
2010	M.S. (Biostatistics), University of California, Los Angeles, CA, USA.
2010-Present	Graduate Student (Biostatistics), University of California, Los Angeles, CA, USA.

Publications

Crespi CM, Maxwell AE, Wu S. Cluster randomized trials of cancer screening interventions: are appropriate statistical methods being used? Contemporary Clinical Trials. 2011 Jul; 32(4):477-484.

Crespi CM, Wong WK, Wu S. A new dependence parameter approach to improve the design of cluster randomized trials with binary outcomes. Clinical Trials. 2011 Dec; 8(6):687-698.

Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemporary Clinical Trials. 2012 Sep; 33(5):869-880.

Braakenburg AM, Crespi CM, Holland GN, Wu S, Yu F, Rothova A. Recurrence rates of ocular toxoplasmosis during pregnancy. American Journal of Ophthalmology. 2014 Apr; 157(4):767-773.

Wu S, Wong WK, Crespi CM. Efficient design for cluster randomized trials with binary outcomes. International Chinese Statistical Association and Korean International Statistical Society Joint Applied Statistics Symposium, Portland, OR, 2014. (Oral presentation)

Wu S, Crespi CM, Wong WK. Comparison of methods for analyzing cluster randomized trials with binary outcomes when cluster size varies and intraclass correlation coefficients are unequal. Joint Statistical Meetings, Boston, MA, 2014. (Oral presentation)

Sverdlov O, Ryeznik Y, Wu S. Exact Bayesian inference comparing binomial proportions, with application to proof-of-concept clinical trials. Therapeutic Innovation & Regulatory Science. 2015 Jan; 49(1):163-174.

Lu N, Crespi CM, Liu N, Vu J, Ahmadieh Y, Wu S, Lynn S, Neiman D, Han S, Saab S, Durazo F, McClune A, Beaven S and French S. Quercetin bioflavonoids demonstrate safety as potential antiviral maintenance to lower hepatitis C relapse. Submitted.

Wu S, Wong WK, Crespi CM. Maximin optimal designs for cluster randomized trials with binary outcomes. Submitted.

CHAPTER 1

Background and Motivation

Cluster randomized trials (CRTs) are increasingly used in many fields including cancer control and prevention, public health and medicine [1, 2]. CRTs are experiments in which clusters of individuals rather than independent individuals are randomly allocated to intervention groups. CRTs are also called group randomized trials. All individuals in a cluster receive the same treatment, assigned to that cluster at random. Clusters can be churches, villages, medical practices, families or schools.

A key feature of CRTs is that outcomes of individuals within a cluster are correlated. The intraclass correlation coefficient (ICC), usually denoted by ρ, provides a quantitative measure of within-cluster correlation. The ICC is variously defined as the Pearson correlation between responses from two members in the same cluster or the proportion of the total variance in the outcome attributable to the variance between clusters. In the design and analysis of CRTs, ρ is an important parameter. Since the correlation increases the sampling error of estimating the intervention effect [3], CRTs are less efficient than individual randomized trials (IRTs), where individuals are randomized to study conditions. For the same total number of individuals in a study, CRTs have lower power to detect the intervention effect than IRTs. However, CRTs may have some advantages although they are less efficient. There are many reasons to use CRTs, including administrative convenience, ethical considerations, to avoid treatment group contamination and because the intervention is naturally applied at the cluster level [4].

The paper published in 1978 by Cornfield [5] attracted health professionals' attention to the appealing statistical features of CRTs. Since then, extensive development of methodology in CRTs has appeared in statistical and health science journals. There is a

lot of methodological research work on CRTs with continuous outcomes, but relatively little when the outcome is binary. Although a two-arm CRT with a binary outcome seems simple and its design problems have been discussed somewhat in the literature, such as in Donner and Klar [4], Hayes and Moulton [2] and Campbell and Walters [6], there are still design issues that need to be fully addressed.

In any study, investigators would prefer to expend minimal resources to obtain the most accurate estimate of an intervention effect. This is even more important when designing CRTs, because CRTs are less efficient than IRTs. In practice, investigators often assign the same number of clusters to each arm of a two-arm CRT. We call such a design a balanced design. Previous research in IRTs has showed that a balanced design may not be the most efficient when the outcome is binary. More discussions can be found in Walter [7], Meydrich [8], Morgenstern and Winn [9], Yanagawa and Bolt [10], Lubin [11] and Brittain and Schlesselman [12]. However, current research on design problems for CRTs largely consider equal allocation of clusters to different arms, and unequal allocation is rarely considered. It seems that investigators have not fully realized the possible low efficiency for a balanced design in CRTs. Hence some issues naturally arise. What are optimal proportions of clusters in the two arms? What factors determine the optimal allocation?

Cluster size in a CRT usually varies. However, investigators often assume the cluster size is constant. The most commonly used sample size formula for a two-arm CRT with binary outcomes is based on the assumption of a constant cluster size, see Donner and Klar [4]. Guittet et al. [13] and Ahn et al. [14] investigated the impact of cluster size variability on the power of CRTs and concluded that the desired power may not be obtained if large variability of cluster size is overlooked. Some research such as Kerry and Bland [15], Manatunga and Hudgens [16], Guittet et al. [13] considered sample size calculations for a CRT with varying cluster size. However, the outcome variable in their research is continuous. Jung et al. [17] provided a sample size formula for binary outcomes, but their sample size formula is for one-arm CRTs. More importantly, previous studies on sample size calculation for a two-arm CRT are based on the balanced

design; few people have addressed the sample size calculation for unequal allocation of clusters in the two arms. Therefore some issues naturally arise. For a two-CRT with binary outcomes and varying cluster sizes, how should we calculate the sample size in order to obtain the desired power? How many samples are needed in each arm to minimize the total cost in addition to satisfying the power requirement?

The main aim of this dissertation is to address these design issues more fully for investigators using a CRT in their research. Our focus is on applications in cancer control and prevention trials but methodology developed in the dissertation is applicable to any two-arm CRT with a binary response.

1.1 Terminology and Notation

1.1.1 Terminology

Here are some key terms we use in this dissertation.

Balanced design: A design that has the same number of clusters in the two arms.

Unbalanced design: A design that is not a balanced design.

Efficiency: This term refers to the precision of the estimated treatment effect in the responses between the two arms.

Cost efficiency (CE): The ratio of precision of the estimated treatment effect in the responses between the two arms to the study cost.

Optimal allocation: The fraction of the number of clusters in arm 1 to the total number of clusters that maximizes the CE of a CRT.

Optimal design: A design with the optimal allocation.

Relative cost efficiency (RCE): The ratio of CE of a design to CE of the optimal design. The design has higher efficiency if the RCE of a design is larger. The RCE is equal to 1 for the optimal design.

Efficient design: A design with high RCE. An efficient design may not be an optimal

design.

Robust design: A design whose RCE is not affected much by misspecified design parameters.

Optimal sample size: The sample size associated with the lowest cost among all that meet our power and type 1 error rate requirements.

1.1.2 Notation

Here are some key notations we use in this dissertation.

π_h, $h = 1, 2$: success rate in the hth arm.

ρ_h, $h = 1, 2$: intraclass correlation coefficient in the hth arm.

k: total number of clusters in the two-arm study.

$k_h, h = 1, 2$: number of clusters in the hth arm.

k_0: number of clusters in either arm for a balanced design, i.e., $k_0 = k_1 = k_2$. It is used in sample size calculation.

m: cluster size when all clusters have the same size.

m_{hi}: cluster size of the ith cluster in the hth arm.

γ: cost ratio of a cluster in arm 2 to a cluster in arm 1.

w_x^*: optimal proportion of clusters in arm 1 for the measure x with constant cluster size but without cost consideration.

w_x^{c*}: optimal proportion of clusters in arm 1 for the measure x with constant cluster size and cost consideration.

w_x^{v*}: optimal proportion of clusters in arm 1 for the measure x with varying cluster size but without cost consideration.

w_x^{vc*}: optimal proportion of clusters in arm 1 for the measure x with varying cluster size and cost consideration.

X_{hij}: the binary response of the jth individual in the ith cluster in the hth arm.

Ψ_x^{-1}: the variance of the estimator of the measure x.

CV: coefficient of variation of the cluster size.

1.2 Organization of this dissertation

In this dissertation, we only consider two-arm CRTs with a binary outcome. We consider design issues when cluster size is constant in Chapters 2-4 and when cluster size is varying in Chapter 5. We compare different analysis methods for CRTs with varying cluster sizes in Chapter 6.

In Chapter 2, we assume that the total number of clusters k is fixed. We consider three commonly used treatment effect measures, risk difference (RD), relative risk (RR) and odds ratio (OR), and find optimal allocations based on these measures. First we determine the optimal allocations which minimize the asymptotic variances of the estimated measures without cost consideration. In practice, the unit cost in a control arm may be very different from that in a treatment arm. We then define the concept of cost efficiency (CE), which combines statistical and economic considerations and adjust our optimal allocations to minimize CE. To compare designs, we define relative cost efficiency (RCE) and show how different allocations affect RCE. We apply the results to Samoan Women's Health Study, which was a CRT designed to increase rates of mammogram usage in women of Samoan ancestry. All designs in this chapter assume nominal values for the design parameters.

Chapter 3 explores how success rates π_1 and π_2, ICCs ρ_1 and ρ_2, cost ratio and cluster size affect the optimal allocations. When the allocation w is equal to the optimal allocation, the RCE reaches the maximal value 1. When w diverges from the optimal allocation, the RCE value decreases. Hence for any pre-selected value a of the RCE, there exists an interval of w, and any w in such interval makes the RCE larger than a. We explore the effects of different parameters on the width of this interval. To mitigate the effects of misspecifications of the nominal values of the parameters, a more robust design is desirable. We propose two methods to make the design less sensitive

to misspecifications to the parameters, a Bayesian method and a maximin method. We compare designs from the two methods with the balanced design. We end this chapter by application of the two methods to the Samoan Women's Health Study.

In Chapter 2 and 3 we assume the total number of clusters k is known and fixed. In Chapter 4, we revisit the sample size calculation for a CRT to satisfy the fixed type 1 error rate α and power $1-\beta$ requirements. The usual sample size formula for a CRT assumes RD as the outcome measure and a balanced design, for example, see Donner and Klar [1]. In this chapter, we first assume a balanced design and derive the sample size formulas for the measures RR and OR. Then costs are factored in and we show that the total cost of a balanced design is not necessarily the lowest and so we incorporate the optimal allocation approach from Chapter 2. Accordingly we modify the sample size formulas for the numbers of clusters in each treatment arm for different measures that guarantee our designs satisfy the power requirement and also minimize total cost. Finally, we consider non-inferiority CRTs and equivalence CRTs. We extend the results of sample size calculation to non-inferiority CRTs and equivalence CRTs.

The results in Chapter 2, 3 and 4 assume a constant cluster size. In Chapter 5, we consider varying cluster sizes and the outcome measure is RD. A common way to determine sample size that meets the specific type 1 error rate α and power $1-\beta$ requirements is to calculate the cluster size using mean cluster size. Guittet et al. [13] and Ahn et al. [14] investigated the effect of cluster size variability on the power of CRTs and concluded that extreme variation in cluster size may affect the power of CRTs. We review the sample size calculation formulas for a CRT with varying cluster sizes, but most of them deal with continuous responses or binary responses in a one-arm CRT. We focus on binary responses in a two-arm CRT. Our work in this chapter includes discussion on how to incorporate optimal allocation and cost considerations into the sample size calculation for CRTs with varying cluster size. At the end of this chapter, we apply our method to redesign Samoan Women's Health Study.

Chapter 6 discusses analysis methods for CRTs. When the outcome of a CRT is a binary variable, analysis methods for such a CRT are not as well established as when

the outcomes is continuous. Available analysis methods are either at the cluster level or the individual level. We first review the commonly used analysis methods including cluster-level t-test, weighted t-test, adjusted chi-square approach, generalized estimating equations approach (GEE) and mixed effect logistic regression model (MELR). We seek answers to the following questions: How does varying cluster sizes affect the performances of these analysis? How does having unequal ICCs in the two arms affect the performance of those methods if a constant ICC is assumed in the analysis? We use simulation to generate data for CRTs with a binary outcome and varying cluster sizes and use them to compare these methods in terms of their control for type 1 and 2 errors. We identify the methods that seem to outperform others under a broad set of scenarios.

CHAPTER 2

Optimal allocation of clusters when the total number of clusters is fixed and cluster sizes are equal

2.1 Introduction

When designing a two-arm CRT, investigators have to decide how many clusters or individuals are needed and what is the proportion of clusters or individuals to allocate to each treatment arm. More clusters or individuals per cluster in the study will detect the differences in responses rates between the two arms with larger power and obtain more accurate estimates of the treatment effect. However, the budget in a study is limited and hence investigators want to have an accurate estimate of treatment effect and save resources at the same time. This requires a carefully designed study.

In a CRT, the cluster size is the number of individuals in a cluster and this number may vary in different clusters. When designing a CRT, investigators often assume equal cluster size in the design stage and ignore cluster size variation. For simplicity, in this chapter we assume that the cluster size is constant. We will consider the case when cluster sizes are unequal in Chapter 5. In some CRTs, for example, when community or geographical zone is the cluster, investigators may recruit a sample of individuals from a cluster rather than all individuals in the cluster. Hence even the cluster sizes are different, investigators may recruit the same number of individuals from different clusters. That number is called sample size per cluster, see Hayes [2]. We don't distinguish sample size per cluster from cluster size in this thesis. That means cluster size can be the real total number of a cluster, or the total number recruited from a cluster to the study. If investigators recruit the same number of individuals from different clusters, we regard

cluster size is equal and the methods for CRTs with equal cluster size can be applied directly.

When the total number of cluster in a two-arm CRT is fixed, the ratio of the number of clusters in the two arms can influence the precision of estimation and the cost of the study. The precision is the inverse variance of an estimator of treatment effect measure. The precision, or the variance can be used to determine the efficiency of a design. When two designs are compared under the same measure, we say the one with higher precision (or lower variance) is more efficient. Investigators often assign the same number of clusters to each arm. In individual randomized trials, the design which assigns the same number of individuals to each arm is called balanced design. We also call a CRT design is a balanced design when the same number of clusters are assigned to each arm of a CRT. Note that when cluster size is equal, the total individuals in the two arms are also equal in a balanced design. For a fixed number of total clusters in a CRT, when the outcome follows a normal distribution, a balanced design is the most efficient. However, when the outcome is binary, a balanced design may cause the study to be less efficient.

When the randomized unit is an individual, research has showed that a balanced design may not be the most efficient. Walter [7] found that when success rates π_1 and π_2 are different, a balanced design results in some loss of efficiency. In his paper, he also defined cost efficiency (CE), which is the ratio of precision to study cost. CE combines statistical and cost considerations and the efficiency of a design is determined by CE. For the same measure of treatment effects, we say a design with higher CE is more efficient or have larger efficiency. Based on this concept of efficiency, Walter indicated that when unit cost ratio in one arm is very different from the other, a balanced design could have very low efficiency. The unit cost here means the cost for one individual in the study. An optimal allocation was also given in his paper. The design with an optimal allocation has the highest efficiency. Other authors also considered unbalanced design, such as Meydrich [8], Morgenstern and Winn [9], Yanagwa and Bolt [10], Lubin [11], Brittain and Schlesselman [12] and Gail et al.[18]. The discussed measures include risk difference (RD), relative risk (RR) and odds ratio (OR). All of these authors pointed out that

an unbalanced design might have higher efficiency than a balanced design. Dette [19] considered all three measures and gave different optimal allocation formulas for those measures. In addition, he pointed out that previous optimal designs are dependent on unknown probabilities, hence such design might not be robust when success rates are mis-specified. In his paper, he proposed a maximin method to construct an efficient and robust design in an individually randomized trial (IRT).

Our observation is that investigators frequently use balanced designs for CRTs and do not seem to realize the possible low efficiency of a balanced design. In this chapter, we extend previous work on optimal allocation in a two-arm individual randomized trials to a two-arm CRT. For a fixed total number of clusters, we will derive optimal allocation formulas for RD, RR and OR. The design with the optimal allocation is called optimal design.

2.2 Optimal allocation for RD, RR and OR

Our two-arm CRTs with a binary outcome are based on the common correlation model [20, 21]. Let X_{hij} denote the response of the jth individual in the ith cluster in the hth treatment arm. Let $X_{hij} = 1$ when the outcome of interest is present (success) and $X_{hij} = 0$ otherwise (failure). We assume the (unconditional) success rate $X_{hij} = 1$ for all individuals in all clusters in the h-th treatment arm is the same and equal to π_h, $h = 1, 2$. The responses of individuals from different clusters are assumed to be independent, while within each cluster, the correlation between any pair of responses is ρ_{hi}, the ICC, which takes value in [-1, 1]. In the common correlation model, we also assume the ICC for all clusters in the h-th treatment arm is the same. Hence the dependence on i in ρ_{hi} can be suppressed and the ICC in the hth arm is ρ_h. This ICC can be expressed as [22]:

$$\rho_h = \frac{Cov(X_{hij}, X_{hil})}{\sqrt{Var(X_{hij})Var(X_{hil})}} = \frac{Pr(X_{hij} = 1, X_{hil} = 1) - \pi_h^2}{\pi_h(1 - \pi_h)}, \forall j \neq l, h = 1, 2 \qquad (2.1)$$

Although many investigators assume a constant $\rho = \rho_1 = \rho_2$ across all the clusters,

ρ_1 is not necessarily equal to ρ_2 in real studies, especially when the outcome is binary. For example, in the Samoan Women's Health Study reported by Mishra et al. [23], the observed ICC in the control arm is much lower than that in the treatment arm. Through Chapter 2 to 4, we let m be constant cluster size in the trial.

To compare the success rates in the two arms, we consider three different measures RD, RR and OR. RD, RR and OR are defined as follows:

$RD = \pi_1 - \pi_2 \in (-1, 1)$,

$RR = \pi_1/\pi_2 \in (0, \infty)$,

and $OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \in (0, \infty)$.

RD is the most used measure when investigators design a CRT. RR is often used in randomized controlled trials and cohort studies and OR is suited for cross-sectional and case-control studies, see Sistrom et al. [24]. However, OR is also used in randomized controlled trials, see Knol [25]. Previous research on CRTs usually consider the measure RD. Since RR and OR can be used in randomized controlled trials, we discuss them too.

We consider the following CRT design problem. Suppose that the total number of clusters in the trial is predetermined and is equal to k. For a given measure, our goal is to determine the optimal proportion of clusters allocated to treatment arm 1, i.e. $w = \frac{k_1}{k}$ where k_1 and k_2 are the numbers of clusters in arm 1 and arm 2 such that $k = k_1 + k_2$. Our criterion for the optimal allocation $w \in (0, 1)$ is that the asymptotic variance of our estimated outcome measure is minimized. Note that minimizing the asymptotic variance is equal to maximizing the precision. The design with the optimal allocation is called optimal design. The three outcome measures of interest here are RD, RR and OR. We expect that they will have different optimal designs.

Recalling that m is the constant cluster size. Our derivations are based on the approximate normal distribution of the maximum-likelihood estimator (MLE) $(\hat{\pi}_1, \hat{\pi}_2)$:

$$\sqrt{k}\left\{\begin{pmatrix}\hat{\pi}_1\\ \hat{\pi}_2\end{pmatrix} - \begin{pmatrix}\pi_1\\ \pi_2\end{pmatrix}\right\} \xrightarrow{D} N\left\{\begin{pmatrix}0\\ 0\end{pmatrix}\begin{pmatrix}\frac{\pi_1(1-\pi_1)[1+(m-1)\rho_1]}{wm} & 0\\ 0 & \frac{\pi_2(1-\pi_2)[1+(m-1)\rho_2]}{(1-w)m}\end{pmatrix}\right\}.$$

From the above result, we derive asymptotic approximations of the distributions of

the estimates $\hat{RD} = \hat{\pi}_1 - \hat{\pi}_2$, $\hat{RR} = \hat{\pi}_1/\hat{\pi}_2$ and $\hat{OR} = \frac{\hat{\pi}_1(1-\hat{\pi}_2)}{\hat{\pi}_2(1-\hat{\pi}_1)}$. Using the delta method, we obtain the following formulas for the asymptotic variances:

For RD:

$$\Psi_{RD}^{-1} \propto \left[\frac{1}{w} + \frac{\pi_2(1-\pi_2)(1+(m-1)\rho_2)}{\pi_1(1-\pi_1)(1+(m-1)\rho_1)(1-w)}\right] \tag{2.2}$$

For RR:

$$\Psi_{RR}^{-1} \propto \left[\frac{1}{w} + \frac{\pi_1(1-\pi_2)(1+(m-1)\rho_2)}{\pi_2(1-\pi_1)(1+(m-1)\rho_1)(1-w)}\right] \tag{2.3}$$

For OR:

$$\Psi_{OR}^{-1} \propto \left[\frac{1}{w} + \frac{\pi_1(1-\pi_1)(1+(m-1)\rho_2)}{\pi_2(1-\pi_2)(1+(m-1)\rho_1)(1-w)}\right]. \tag{2.4}$$

For each of the above variances, we may treat the parameters π_1, π_2, ρ_1, ρ_2 and m as fixed and known so that the righthand side can be regarded as a function $f(w)$, where w is the only unknown variable. Different values of w result in different values of $f(w)$. The minimal value of $f(w)$ is $f(w^*)$ if the first derivative $f^{'}(w^*) = 0$ and the second derivative $f^{''}(w^*) > 0$. Using this approach, we obtain the optimal allocation $w_x^*(x = RD, RR, OR)$ for each measure to arm 1 as:

For RD:

$$w_{RD}^* = \frac{\sqrt{\pi_1(1-\pi_1)(1+(m-1)\rho_1)}}{\sqrt{\pi_1(1-\pi_1)(1+(m-1)\rho_1)} + \sqrt{\pi_2(1-\pi_2)(1+(m-1)\rho_2)}} \tag{2.5}$$

For RR:

$$w_{RR}^* = \frac{\sqrt{\pi_2(1-\pi_1)(1+(m-1)\rho_1)}}{\sqrt{\pi_2(1-\pi_1)(1+(m-1)\rho_1)} + \sqrt{\pi_1(1-\pi_2)(1+(m-1)\rho_2)}} \tag{2.6}$$

For OR:

$$w_{OR}^* = \frac{\sqrt{\pi_2(1-\pi_2)(1+(m-1)\rho_2)}}{\sqrt{\pi_2(1-\pi_2)(1+(m-1)\rho_2)} + \sqrt{\pi_1(1-\pi_1)(1+(m-1)\rho_1)}}. \tag{2.7}$$

Note that for the different measures, the optimal allocation w_x^*s are different. Also we note that $w_{RD}^* = 1 - w_{OR}^*$. Hence when investigators design a study, they should

consider which measure they plan to use to estimate the treatment difference between the two arms. One value of w may be the best for one measure, but not for the others.

In practice, we observe that RR and OR have skewed distributions, and the log transformation is often used to transform the measures to make them less skewed. However, the approximate variances of $log(RR)$ and $log(OR)$ are proportional to the corresponding variances of RR and OR. For example, by the delta method, we have $Var(log(RR)) \propto \frac{1}{RR} Var(RR)$. The term $\frac{1}{RR}$ does not involve w and so the value of w^*_{RR} will be the same, whether a log transformation is used or not. Hence, these optimal allocations apply regardless of whether or not a log transformation is used.

Table 2.1 gives the optimal allocation w^* for RD, RR and OR under different combinations of π_1 and π_2 when $\rho_1 = \rho_2$. Note that when $\rho_1 = \rho_2$, the optimal allocation w^*_x does not depend on cluster size m and simplifies to:

For RD:

$$w^*_{RD} = \frac{\sqrt{\pi_1(1-\pi_1)}}{\sqrt{\pi_1(1-\pi_1)} + \sqrt{\pi_2(1-\pi_2)}} \tag{2.8}$$

For RR:

$$w^*_{RR} = \frac{\sqrt{\pi_2(1-\pi_1)}}{\sqrt{\pi_2(1-\pi_1)} + \sqrt{\pi_1(1-\pi_2)}} \tag{2.9}$$

For OR:

$$w^*_{OR} = \frac{\sqrt{\pi_2(1-\pi_2)}}{\sqrt{\pi_2(1-\pi_2)} + \sqrt{\pi_1(1-\pi_1)}}. \tag{2.10}$$

This table shows that under different combinations of π_1 and π_2, the optimal allocations for the different measures are different, and also $w^*_{RD} = 1 - w^*_{OR}$. For example, when $\pi_1 = 0.3$ and $\pi_2 = 0.1$, $w^*_{RD} = 0.60$, $w^*_{RR} = 0.34$, $w^*_{OR} = 0.40$. When $\pi_1 = \pi_2$, the optimal allocation for all three measures coincides and is equal to 0.5. However, trials are usually designed under the assumption that $\pi_1 \neq \pi_2$. The optimal allocations of RD and OR are symmetric around $\pi_1 = 0.5$ and $\pi_2 = 0.5$, and the optimal allocations of RR is symmetric around $\pi_1 = 1 - \pi_2$. For example, when π_1=0.4, π_2=0.2 we

have w^*_{RD}=0.55 and w^*_{OR}=0.45; and when π_1=0.6, π_2=0.2, we also have w^*_{RD}=0.55 and w^*_{OR}=0.45. When π_1=0.4, π_2=0.2 we have w^*_{RR}=0.38; and when π_1=0.8, π_2=0.6, we also have w^*_{RR}=0.38.

In this table, we let $\rho_1 = \rho_2$, so that the optimal allocation values are the same for different cluster size m. If $\rho_1 \neq \rho_2$, the optimal allocation values are different for different cluster size m. However, the symmetric patterns shown in Table 2.1 still hold.

2.3 Optimal allocation for RD, RR and OR with cost consideration

In Section2.2, we find the optimal allocation by minimizing the asymptotic variance of the estimator of the measure RD, RR or OR. In individual randomized trials, usually there is difference in cost per individual between the two arms in the study. For example, in a cancer and control prevention trial designed to increase a cancer screening, there are two intervention arms. The intervention in the first arm consists of take home print materials, a reminder letter and a letter to participants' providers. In the second arm, individuals additionally receive a free fecal occult blood test kit. Hence assigning more individuals to the second arm cost more money. Walter [7], Miettinen [26], Meydrich [8], Morgenstern and Winn [9] and Dette [19] considered combining cost and statistical considerations in one optimality criterion. Similar cost concerns also arise in two-arm CRTs but they have not been much considered. A cluster in one arm may cost more than that in another arm. Thus we consider designs in which the goal is to minimize total study costs while maximizing the precision of estimating when the number of clusters is fixed.

In CRTs individuals are nested in clusters, hence the study cost consists of cost spent on individuals and cost spent directly on clusters. Let the individual-level cost per individual be c_i and the cluster-level cost per cluster is e_i in the ith arm. For a fixed

Table 2.1: Optimal allocation w_x^* for estimating RD, RR and OR with a fixed number of clusters in the trial ($\rho_1 = \rho_2 = 0.1$)

π_1/π_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
a: RD									
0.1	0.50	0.43	0.40	0.38	0.38	0.38	0.40	0.43	0.50
0.2	0.57	0.50	0.47	0.45	0.44	0.45	0.47	0.50	0.57
0.3	0.60	0.53	0.50	0.48	0.48	0.48	0.50	0.53	0.60
0.4	0.62	0.55	0.52	0.50	0.49	0.50	0.52	0.55	0.62
0.5	0.62	0.56	0.52	0.51	0.50	0.51	0.52	0.56	0.62
0.6	0.62	0.55	0.52	0.50	0.49	0.50	0.52	0.55	0.62
0.7	0.60	0.53	0.50	0.48	0.48	0.48	0.50	0.53	0.60
0.8	0.57	0.50	0.47	0.45	0.44	0.45	0.47	0.50	0.57
0.9	0.50	0.43	0.40	0.38	0.37	0.38	0.40	0.43	0.50
b: RR									
0.1	0.50	0.60	0.66	0.71	0.75	0.79	0.82	0.86	0.90
0.2	0.40	0.50	0.57	0.62	0.67	0.71	0.75	0.80	0.86
0.3	0.34	0.43	0.50	0.56	0.60	0.65	0.70	0.75	0.82
0.4	0.29	0.38	0.44	0.50	0.55	0.60	0.65	0.71	0.79
0.5	0.25	0.33	0.40	0.45	0.50	0.55	0.60	0.67	0.75
0.6	0.21	0.29	0.35	0.40	0.45	0.50	0.56	0.62	0.71
0.7	0.18	0.25	0.30	0.35	0.40	0.44	0.50	0.57	0.66
0.8	0.14	0.20	0.25	0.29	0.33	0.38	0.43	0.50	0.60
0.9	0.10	0.14	0.18	0.21	0.25	0.29	0.34	0.40	0.50

cluster size m, the total cost function is:

$$k_1(mc_1 + e_1) + k_2(mc_2 + e_2) = k[w(mc_1 + e_1) + (1 - w)(mc_2 + e_2)] \tag{2.11}$$

where k_1 is the number of clusters in the arm 1, k_2 is the number of clusters in the arm 2 and $k = k_1 + k_2$ and $w = k_1/k_2$. Here m and k are known.

π_1/π_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
					c: OR				
0.1	0.50	0.57	0.60	0.62	0.62	0.62	0.60	0.57	0.50
0.2	0.43	0.50	0.53	0.55	0.56	0.55	0.53	0.50	0.43
0.3	0.40	0.47	0.50	0.52	0.52	0.52	0.50	0.47	0.40
0.4	0.38	0.45	0.48	0.50	0.51	0.50	0.48	0.45	0.38
0.5	0.38	0.44	0.48	0.49	0.50	0.49	0.48	0.44	0.37
0.6	0.38	0.45	0.48	0.50	0.51	0.50	0.48	0.45	0.38
0.7	0.40	0.47	0.50	0.52	0.52	0.52	0.50	0.47	0.40
0.8	0.43	0.50	0.53	0.55	0.56	0.55	0.53	0.50	0.43
0.9	0.50	0.57	0.60	0.62	0.62	0.62	0.60	0.57	0.50

Following Walter [7], we define cost efficiency (CE) as the ratio of the precision in estimating the effect (RD, RR, OR) to the total study cost. This is a natural way to combine statistical and cost considerations. The criterion for optimal allocation w_x^{c*} is to maximizes CE for measure $x, x \in (RD, RR, OR)$. Note that without considering cost, a design with higher efficiency often refers to the design with lower variance of estimating. However, when cost is included, we need define a design with higher efficiency as the design with higher CE. Hence efficiency here refers to CE when cost is considered and we may interchange using these two terms. The CE for each measure is:

For RD:

$$CE_{RD} = \frac{\Psi_{RD}}{k_1(mc_1 + e_1) + k_2(mc_2 + e_2)} \tag{2.12}$$

For RR:

$$CE_{RR} = \frac{\Psi_{RR}}{k_1(mc_1 + e_1) + k_2(mc_2 + e_2)} \tag{2.13}$$

For OR:

$$CE_{OR} = \frac{\Psi_{OR}}{k_1(mc_1 + e_1) + k_2(mc_2 + e_2)} \tag{2.14}$$

Let the cost ratio for a cluster in arm 1 versus arm 2 be defined as $\gamma = \frac{mc_1+e_1}{mc_2+e_2}$. Note that the cost for a cluster include both the cluster-level cost and the individual-level cost. Since in CRTs the randomized unit is cluster, we call γ unit cost ratio. It should be distinguished from unit cost in individual randomized trials, which refers to the ratio of cost per individual. Using the same method as in last section, we obtain the optimal allocation w_x^* which maximizes CE_{RD}, CE_{RR} or CE_{OR} for each measure:

For RD:

$$w_{RD}^{c*} = \frac{\sqrt{\pi_1(1-\pi_1)(1+(m-1)\rho_1)}}{\sqrt{\pi_1(1-\pi_1)(1+(m-1)\rho_1)} + \sqrt{\gamma\pi_2(1-\pi_2)(1+(m-1)\rho_2)}} \tag{2.15}$$

For RR:

$$w_{RR}^{c*} = \frac{\sqrt{\pi_2(1-\pi_1)(1+(m-1)\rho_1)}}{\sqrt{\pi_2(1-\pi_1)(1+(m-1)\rho_1)} + \sqrt{\gamma\pi_1(1-\pi_2)(1+(m-1)\rho_2)}} \tag{2.16}$$

For OR:

$$w_{OR}^{c*} = \frac{\sqrt{\pi_2(1-\pi_2)(1+(m-1)\rho_2)}}{\sqrt{\pi_2(1-\pi_2)(1+(m-1)\rho_2)} + \sqrt{\gamma\pi_1(1-\pi_1)(1+(m-1)\rho_1)}} \tag{2.17}$$

These formulas are similar to equations 2.5, 2.6 and 2.7 except for the presence of γ in each denominator. The optimal allocation w^{c*} is different for different measures. When costs in the two arms are not equal, e.g.,$\gamma \neq 1$, the relationship $w_{DR}^{c*} = 1 - w_{OR}^{c*}$ no longer holds, which is true for w^* when costs are not considered. If we assume $\rho_1 = \rho_2$, then all three w^{c*}s reduce to those in Dette [19]. If $\pi_1 = \pi_2$ and $\rho_1 = \rho_2$, the optimal allocation for all three measures coincides and is equal to $\frac{1}{1+\sqrt{\gamma}}$. In addition, if unit costs are the same ($\gamma = 1$), then w_x^{c*} is equal to $\frac{1}{2}$, the balanced design.

In Table 2.2, we give the values of optimal allocation w_x^{c*} for RD , RR and OR for different combinations of π_1 and π_1 when $\gamma = 5$, assuming $\rho_1 = \rho_2$.

The table illustrates that when the unit costs are different in the two arms, the relationship $w_{DR}^{c*} = 1 - w_{OR}^{c*}$ does not hold. For example, when $\pi_1 = 0.3$ and $\pi_2 = 0.1$, $w_{RD}^{c*} = 0.41$, and $w_{OR}^{c*} = 0.23$. When $\pi_1 = \pi_2$, all optimal allocations for RD, RR and OR coincide, and are equal to $\frac{1}{1+\sqrt{\gamma}} = \frac{1}{1+\sqrt{5}} = 0.31$ rather than 0.5. All the optimal

Table 2.2: Optimal allocation w_x^{c*} for estimating RD, RR and OR with a fixed number of clusters in the trial($\gamma = 5, \rho_1 = \rho_2 = 0.1$)

π_1/π_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
a: RD									
0.1	0.31	0.25	0.23	0.21	0.21	0.21	0.23	0.25	0.31
0.2	0.37	0.31	0.28	0.27	0.26	0.27	0.28	0.31	0.37
0.3	0.41	0.34	0.31	0.29	0.29	0.29	0.31	0.34	0.41
0.4	0.42	0.35	0.32	0.31	0.30	0.31	0.32	0.35	0.42
0.5	0.43	0.36	0.33	0.31	0.31	0.31	0.33	0.36	0.43
0.6	0.42	0.35	0.32	0.31	0.30	0.31	0.32	0.35	0.42
0.7	0.41	0.34	0.31	0.29	0.29	0.29	0.31	0.34	0.41
0.8	0.37	0.31	0.28	0.27	0.26	0.27	0.28	0.31	0.37
0.9	0.31	0.25	0.23	0.22	0.21	0.22	0.23	0.25	0.31
b: RR									
0.1	0.31	0.40	0.47	0.52	0.57	0.62	0.67	0.73	0.80
0.2	0.23	0.31	0.37	0.42	0.47	0.52	0.58	0.64	0.73
0.3	0.19	0.25	0.31	0.36	0.41	0.46	0.51	0.58	0.67
0.4	0.15	0.22	0.26	0.31	0.35	0.40	0.46	0.52	0.62
0.5	0.13	0.18	0.23	0.27	0.31	0.35	0.41	0.47	0.57
0.6	0.11	0.15	0.19	0.23	0.27	0.31	0.36	0.42	0.52
0.7	0.09	0.13	0.16	0.19	0.23	0.26	0.31	0.37	0.47
0.8	0.07	0.10	0.13	0.15	0.18	0.22	0.25	0.31	0.40
0.9	0.05	0.07	0.09	0.11	0.13	0.15	0.19	0.23	0.31

allocations under the situation $\gamma = 5$ are smaller than those in Table 2.1, that is, fewer units are allocated to the more expensive arm. Note that if $\rho_1 \neq \rho_2$, then when $\pi_1 = \pi_2$, the optimal allocations for RD and RR are same, but different from that for OR. This can be easily observed from formulas (2.15, 2.16, 2.15).

π_1/π_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
					c: OR				
0.1	0.31	0.37	0.41	0.42	0.43	0.42	0.41	0.37	0.31
0.2	0.25	0.31	0.34	0.35	0.36	0.35	0.34	0.31	0.25
0.3	0.23	0.28	0.31	0.32	0.33	0.32	0.31	0.28	0.23
0.4	0.21	0.27	0.29	0.31	0.31	0.31	0.29	0.27	0.21
0.5	0.21	0.26	0.29	0.30	0.31	0.30	0.29	0.26	0.21
0.6	0.21	0.27	0.29	0.31	0.31	0.31	0.29	0.27	0.21
0.7	0.23	0.28	0.31	0.32	0.33	0.32	0.31	0.28	0.23
0.8	0.25	0.31	0.34	0.35	0.36	0.35	0.34	0.31	0.25
0.9	0.31	0.37	0.41	0.42	0.43	0.42	0.41	0.37	0.31

2.4 Comparison of optimal allocation designs and balanced designs

We have seen that only under specific conditions the balanced design is the most efficient design. If the balanced design is not the most efficient design, we may ask how good it is? In a balanced design, we have the allocation $w = \frac{1}{2}$. In this section, we introduce the idea of relative efficiency and use it to compare efficient and balanced designs.

Recall that cost efficiency CE_x, $x \in \{RD, RR, OR\}$ is a function of $\pi_1, p_2, w, m, \rho_1, \rho_2, c_1$ and c_2. Given $w \in \{0, 1\}$, we can define relative cost efficiency (call it RCFFF)as:

$$RCE_x(\pi_1, \pi_2, w, m, \rho_1, \rho_2, c_1, c_2) = \frac{CE_x(\pi_1, \pi_2, w, m, \rho_1, \rho_2, c_1, c_2)}{max_{\eta \in \{0,1\}} CE_x(\pi_1, \pi_2, \eta, m, \rho_1, \rho_2, c_1, c_2)} \tag{2.18}$$

Given $\pi_1, \pi_2, m, \rho_1, \rho_2, c_1, c_2$, RCE is a function of w:

$$RCE_x(w) = \frac{CE_x(w)}{CE_x(w_x^{c*})} \tag{2.19}$$

Different values of w result in different values of RCE. The maximal value of RCE is 1, which is reached when w takes the optimal allocation value w_x^* for measure $x \in (RD, RR, OR)$. If RCE is close to 1, the design is a good design; if RCE is much lower

than 1, the design may not be preferred. For a balanced CRT design, $w = \frac{1}{2}$, and the RCE is expressed as $RCE_x(w = \frac{1}{2})$. If $RCE_x(w = \frac{1}{2})$ is close to 1, then the balanced design is good. Note that for different measures x, the values of $RCE_x(w = \frac{1}{2})$ is different. This means a balanced design may have high RCE for one measure but not another.

The following tables show the RCE of a balanced design for RD, RR and OR for different combinations of π_1 and π_2. Note that when ρ_1 and ρ_2 are assumed equal, RCE is independent of cluster size m.

Table 2.3 shows the RCE for estimating RD for different combinations of π_1 and π_2, comparing equal allocation to the optimal allocation. We observe a symmetrical pattern. The RCE is symmetrical about $\pi_1 = 0.5$ and about $\pi_2 = 0.5$. In Table 2.3a, the cost ratio γ is 5. The range of RCE is between 0.67 and 0.99, and in many scenarios it is larger than 0.8. The lowest RCE 0.67 occurs when $\pi_1 = 0.1$ or 0.9 and $\pi_2 = 0.5$. In Table 2.3b, the cost ratio γ is 10. When the cost ratio γ increases from 5 to 10, for all combinations of π_1 and π_2, the RCE decreases, meaning that fewer units are allocated to the more costly arm. The range of RCE for $\gamma = 10$ is between 0.56 and 0.95. The lowest RCE 0.56 also occurs when $\pi_1 = 0.1, 0.9$ and $\pi_2 = 0.5$. But in many scenarios, the RCE is lower than 0.8. Therefore, when costs in the two arms are very different, investigators should be careful about choosing a balanced design for estimating RD.

Table 2.4 shows the RCE for estimating RR under different combinations of π_1 and π_2. The RCE is symmetrical about the diagonal line $\pi_1 = 1 - \pi_2$. In Table 2.4a, the cost ratio γ is 5. The range of RCE is between 0.26 and 1, and in many scenarios it is smaller than 0.8. The smallest RCE, 0.26, occurs when $\pi_1 = 0.9$ and $\pi_2 = 0.1$. In Table 2.4b, the cost ratio γ is 10. The same symmetrical pattern as in Table 2.4 is observed. When the cost ratio γ increases to 10, for all combinations of π_1 and π_2, the RCE decreases. The range of RCE for $\gamma = 10$ is between 0.16 and 1. The smallest RCE 0.16 also occurs when $\pi_1 = 0.9$ and $\pi_2 = 0.1$. But in many other scenarios, the RCE is smaller than 0.8. Thus when cost in the two arms are very different and $\pi_1 > \pi_2$, we recommend investigators should be careful about choosing a balanced design for estimating RR.

Table 2.3: RCE of equal allocation $w = \frac{1}{2}$ for estimating RD with fixed number of clusters under different combinations of π_1 and π_2 ($\rho_1 = \rho_2$)

π_1/π_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
a: $\gamma = 5$									
0.1	0.87	0.76	0.71	0.68	0.67	0.68	0.71	0.76	0.87
0.2	0.95	0.87	0.82	0.80	0.79	0.80	0.82	0.87	0.95
0.3	0.97	0.91	0.87	0.85	0.84	0.85	0.87	0.91	0.97
0.4	0.98	0.93	0.89	0.87	0.87	0.87	0.89	0.93	0.98
0.5	0.99	0.94	0.90	0.88	0.87	0.88	0.90	0.94	0.99
0.6	0.98	0.93	0.89	0.87	0.87	0.87	0.89	0.93	0.98
0.7	0.97	0.91	0.87	0.85	0.84	0.85	0.87	0.91	0.97
0.8	0.95	0.87	0.82	0.80	0.79	0.80	0.82	0.87	0.95
0.9	0.87	0.76	0.71	0.68	0.67	0.68	0.71	0.76	0.87
b: $\gamma = 10$									
0.1	0.79	0.66	0.60	0.57	0.56	0.57	0.60	0.66	0.79
0.2	0.89	0.79	0.73	0.70	0.69	0.70	0.73	0.79	0.89
0.3	0.93	0.84	0.79	0.76	0.75	0.76	0.79	0.84	0.93
0.4	0.94	0.86	0.81	0.79	0.78	0.79	0.81	0.86	0.94
0.5	0.95	0.87	0.82	0.80	0.79	0.80	0.82	0.87	0.95
0.6	0.94	0.86	0.81	0.79	0.78	0.79	0.81	0.86	0.94
0.7	0.93	0.84	0.79	0.76	0.75	0.76	0.79	0.84	0.93
0.8	0.89	0.79	0.73	0.70	0.69	0.70	0.73	0.79	0.89
0.9	0.79	0.66	0.60	0.57	0.56	0.57	0.60	0.66	0.79

Investigators also should be aware that the lowest RCE for estimating RR is worse than that for estimating RD.

Table 2.5 shows the RCE for estimating OR for different combinations of π_1 and π_2. Compared to Tables 2.3, we can find the same symmetrical pattern. In Table 2.5a,

Table 2.4: RCE of equal allocation $w = \frac{1}{2}$ for estimating RR with fixed number of clusters under different combinations of π_1 and π_2 ($\rho_1 = \rho_2$)

π_1/π_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
a: $\gamma = 5$									
0.1	0.87	0.97	1.00	1.00	0.99	0.98	0.96	0.94	0.91
0.2	0.72	0.87	0.95	0.98	1.00	1.00	0.99	0.97	0.94
0.3	0.61	0.77	0.87	0.94	0.97	1.00	1.00	0.99	0.96
0.4	0.52	0.68	0.79	0.87	0.93	0.97	1.00	1.00	0.98
0.5	0.46	0.60	0.71	0.80	0.87	0.93	0.97	1.00	0.99
0.6	0.40	0.52	0.62	0.72	0.80	0.87	0.94	0.98	1.00
0.7	0.35	0.45	0.54	0.62	0.71	0.79	0.87	0.95	1.00
0.8	0.31	0.38	0.45	0.52	0.60	0.68	0.77	0.87	0.97
0.9	0.26	0.31	0.35	0.40	0.46	0.52	0.61	0.72	0.87
b: $\gamma = 10$									
0.1	0.79	0.92	0.97	0.99	1.00	1.00	0.99	0.98	0.96
0.2	0.61	0.79	0.89	0.94	0.98	0.99	1.00	1.00	0.98
0.3	0.49	0.67	0.79	0.87	0.93	0.97	0.99	1.00	0.99
0.4	0.41	0.57	0.69	0.79	0.86	0.92	0.97	0.99	1.00
0.5	0.35	0.48	0.60	0.70	0.79	0.86	0.93	0.98	1.00
0.6	0.29	0.41	0.51	0.61	0.70	0.79	0.87	0.94	0.99
0.7	0.25	0.34	0.43	0.51	0.60	0.69	0.79	0.89	0.97
0.8	0.21	0.27	0.34	0.41	0.48	0.57	0.67	0.79	0.92
0.9	0.16	0.21	0.25	0.29	0.35	0.41	0.49	0.61	0.79

the range of RCE is the same as in Table 2.3a, and the same is true comparing Table 2.5b to Table 2.3b. But the lowest RCE for estimating the OR occurs when $\pi_1 = 0.5$ and $\pi_2 = 0.1$ or 0.9 instead. Just as when estimating RD, when costs in the two arms are very different, investigators should be careful about choosing a balanced design for

Table 2.5: RCE of equal allocation $w = \frac{1}{2}$ for estimating OR with fixed number of clusters under different combinations of π_1 and π_2 ($\rho_1 = \rho_2$)

π_1/π_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
a: $\gamma = 5$									
0.1	0.87	0.95	0.97	0.98	0.99	0.98	0.97	0.95	0.87
0.2	0.76	0.87	0.91	0.93	0.94	0.93	0.91	0.87	0.76
0.3	0.71	0.82	0.87	0.89	0.90	0.89	0.87	0.82	0.71
0.4	0.68	0.80	0.85	0.87	0.88	0.87	0.85	0.80	0.68
0.5	0.67	0.79	0.84	0.87	0.87	0.87	0.84	0.79	0.67
0.6	0.68	0.80	0.85	0.87	0.88	0.87	0.85	0.80	0.68
0.7	0.71	0.82	0.87	0.89	0.90	0.89	0.87	0.82	0.71
0.8	0.76	0.87	0.91	0.93	0.94	0.93	0.91	0.87	0.76
0.9	0.87	0.95	0.97	0.98	0.99	0.98	0.97	0.95	0.87
b: $\gamma = 10$									
0.1	0.79	0.89	0.93	0.94	0.95	0.94	0.93	0.89	0.79
0.2	0.66	0.79	0.84	0.86	0.87	0.86	0.84	0.79	0.66
0.3	0.60	0.73	0.79	0.81	0.82	0.81	0.79	0.73	0.60
0.4	0.57	0.70	0.76	0.79	0.80	0.79	0.76	0.70	0.57
0.5	0.56	0.69	0.75	0.78	0.79	0.78	0.75	0.69	0.56
0.6	0.57	0.70	0.76	0.79	0.80	0.79	0.76	0.70	0.57
0.7	0.60	0.73	0.79	0.81	0.82	0.81	0.79	0.73	0.60
0.8	0.66	0.79	0.84	0.86	0.87	0.86	0.84	0.79	0.66
0.9	0.79	0.89	0.93	0.94	0.95	0.94	0.93	0.89	0.79

estimating OR.

2.5 Application

We show how to use the results in previous sections in the Samoan Women's Health study [23]. This study was a CRT designed to increase rates of mammogram usage in women of Samoan ancestry. In the trial, Samoan churches in southern California were randomized to intervention and control arms, and women at intervention churches participated in a culturally appropriate breast cancer education program. The control arm received usual care. The outcome was self-reported receipt of mammogram at follow-up.

Churches are clusters in this study and 55 churches were recruited. The actual cluster size varies and the mean cluster size is 14. But for illustration purpose, in this section we assume that the cluster size is constant and equal to 14. We suppose that the cost for an individual in the intervention is 10 times as the cost per individual in the control arm. The success proportions in the intervention and and control arm are 0.5 and 0.4, respectively, and the ICCs are 0.3 and 0.1.

If investigators want to estimate RD, equation (2.15) is used and the optimal allocation is $w^*_{RD} = 0.32$. This means that 55x0.32 = 17.6 clusters must be assigned to the intervention arm. Since the number of cluster should be an integer, 18 clusters should be assigned to the intervention arm and 37 clusters should be assigned to the control arm.

In the original design, 30 churches were assigned to the intervention arm and 25 churches were assigned to the control arm, hence $w = 0.55$. From the equation (2.18), the relative efficiency of the original design is 0.876. Therefore, if the investigators assign 18 clusters to the intervention arm instead of 30, the RCE increases about 14% for the measure of RD.

Similarly, if investigators want to estimate the RR, equation (2.16) is used and the optimal allocation $w^*_{RR} = 0.27$. Hence 15 clusters should be assigned to the intervention arm and 40 clusters should be assigned to the control arm. The RCE of the original design is 0.796. Therefore, if the investigators assign 15 clusters to the intervention arm instead of 30, the RCE increases about 26%.

Further, if the investigators want to estimate the OR, equation (2.17) is used and the optimal allocation $w^*_{OR} = 0.18$. Hence 10 clusters should be assigned to the intervention arm and 45 clusters should be assigned to control. The RCE of the original design is 0.940. Thus, the original design is close to the maximal RCE when the OR is the measure of interest.

Investigators sometimes are interested in estimating the success proportion in the intervention arm in addition to the difference of proportions between the two arms. Decreasing the number of clusters in the intervention arm leads to a less accurate estimate of the proportion in the intervention arm. Therefore if a large decrease of the number of clusters in the intervention arm only brings a small increase of RCE for estimating difference of proportions, we do not recommend making this adjustment.

Throughout we have assumed the true values of the success proportions and ICCs in the intervention arm and the control arm are available and use them to assign clusters to the two arms for estimating different measures. For our example, these numbers are actually estimates from the actual data after the Samoan study is completed. The question is how to determine the allocation when the true values of parameters are unknown. We will give some approaches in the next chapter.

2.6 Chapter summary and discussion

In this chapter, we answer the question how should investigators assign clusters in each arm efficiently to infer treatment effect when a fixed number of clusters is given. Without cost consideration, the criterion is to minimize the variance of the difference estimate. Then we consider the cost in the study. We define the CE, which is the ratio of precision of estimating to the study cost, and the criterion is to maximize the CE. The allocation scheme that corresponds to the maximal CE is the optimal allocation. The optimal design is dependent on the success proportions, ICCs, cluster size and cost ratio. If the individual-level cost and cluster-level cost in two arms are very different, the optimal allocation may be quite different from 0.5. We define RCE which is the ratio of CE in

two designs and it can be used to compare different designs. Current research on design problems for CRTs largely consider balanced design, which equally allocate clusters to different arms, and unequal allocation is rarely considered. We have shown that RCE may be very low when the unit cost is very different in two arms.

We showed that the optimal allocation values are very different for different measures. All of these measure, RD, RR and OR are used in randomized trials [25]. Hence all of these measures can be used in CRTs. RD is more discussed in literature, especially when sample size calculation is involved. Note that RD is absolute measure, but RR and OR are relative measures. Therefore, RR and OR tend to be insensitive to baseline risk and RD is sensitive to baseline risk. Investigators should decide which measure is appropriate for their CRT first before finding an optimal design.

As discussed in [12] and in last section, although the main objective of a CRT is to estimate the difference between arms, investigators may have the second objective in the intervention arm to increase the knowledge of the intervention population. From this perspective, assigning more clusters in the control arm is not efficient even though optimal allocation assigning more clusters to the control arm. This is because our optimal allocation is based on estimating the difference between two arms. If there is no preference among RD, RR or OD, investigators may consider using the measure with larger value of w_x^c which assign more clusters to the intervention arm.

In our optimal design, we assume all individuals in a cluster are recruited. We do not consider the problem how many individuals should be sampled from a cluster. For example, suppose the cluster in a CRT is the community with thousands of people, investigators may not recruit all individuals in a community to the study. Instead, investigators need to decide how many individuals recruited from a community and how many communities are needed in the study. Without cost consideration, if the total number of individuals is fixed, investigators may prefer to including more clusters with small sample size per cluster rather than including less clusters with large sample size per cluster. For example, see discussion in Donner [1]. However, if the cost is considered, the first choice is more expensive than the second. Therefore, the sample size per cluster

can be neither too small nor too big. Breukelen and Candel [27] answered how many individuals should be sampled from a cluster. The optimal sample size per cluster is given by:

$$m = \sqrt{\frac{(1-\rho)e}{\rho c}} \tag{2.20}$$

where e is the cluster-level cost per cluster and c is the individual-level cost per individual. This formula is also given in Raudenbush [28] in another form. If the total budget is B, then the number of cluster is $k = B/(e + mc)$. However, Breukelen and Candel did not consider the cost difference in the two arms and assumed a balanced design. Therefore, further research is needed to extend our optimal design to the design of CRTs in which sub-sample of individuals in a cluster is recruited.

Our optimal design is based on no-covariate case. In fact, when investigators analyze CRTs data to detect the difference between the two arms, they usually need adjust for individual-level and cluster-level covariates. Raudenbush [28] used the mixed linear regression model to explore the optimal design with covariates consideration, assuming the outcome following a normal distribution. Basically he dealt with the problem of the optimal sample size per cluster. Based on the mixed model, he gave a formula of the optimal sample size per cluster incorporating covariates consideration. However, he did not consider the cost difference in the two arms and did not consider the unbalanced design. Further research to incorporate covariates consideration in our optimal design is needed.

Note that in this chapter, we consider how to assign clusters to each arm when the total number of clusters is given. We do not consider the power of the trial and the sample size. That means, we do not answer the question how many clusters are required to satisfy the power requirement. We will consider the power and sample size problem in Chapter 4.

CHAPTER 3

Robust design when the total number of clusters is fixed and cluster sizes are equal

3.1 Introduction

In Chapter 2, we determined the optimal allocations w_x^* for a particular measure. We calculated RCE of balanced designs, the ratio of CE of balanced designs to CE of optimal allocation designs. Optimal allocation designs are always better than balanced designs. However, in practice true values of π_1, π_2, ρ_1 and ρ_2 in the two arms of a CRT are generally not available before the study. Hence investigators are likely to mis-specify the values of these parameters, and the misspecified parameters results in a design which is not really an optimal design.

In this chapter, we first explore the effects of success rates π_1 and π_2, ICCs ρ_1 and ρ_2, unit costs and cluster size on the optimal allocation. For an optimal design, the RCE is 1. If we set the RCE required to be at least a, then we can construct a range of values for w that will satisfy the requirement. We call that range of values for w the allocation interval. We next explore the effects of parameters on that interval.

In order to construct a robust design, we propose two methods: a Bayesian method and a maximin method. We will show how to use these two methods to construct a design when we do not know the true values of parameters. We end this chapter by comparing designs based on these two methods with a balanced design, in which equal number of clusters are randomly assigned to different arms.

3.2 Effects of parameters on optimal allocation

3.2.1 Effects of π_1 and π_2

The optimal allocation w_x^{c*} is dependent on the values of parameters π_1, π_2, ρ_1, ρ_2, m and γ. In this section, we briefly discuss how these parameters affect the value of w_x^{c*}. If a parameter has little effect on w_x^{c*}, even though that parameter is misspecified by investigators, the allocation w calculated may not be far away from w_x^{c*}, and the design should be relatively robust.

Let $\lambda = \frac{1+(m-1)\rho_2}{1+(m-1)\rho_1}$. The term $1+(m-1)\rho$ is design effect, for more discussion, see Donner and Klar [1]. Hence λ is the ratio of design effects in the two arms. Using an idea from [19], let $x \in \{RD, RR, OR\}$ and we define $\tilde{x}$:

$$\tilde{x} = \begin{cases} \frac{\pi_2(1-\pi_2)}{\pi_1(1-\pi_1)} & \text{if } x = RD \\ \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} & \text{if } x = RR \\ \frac{\pi_1(1-\pi_1)}{\pi_2(1-\pi_2)} & \text{if } x = OR \end{cases}$$

.

Then the optimal allocation w_x^{c*} for each $x \in \{RD, RR, OR\}$ given in formulas (2.15,2.16 and 2.17) can be expressed as:

$$w_x^{c*} = \frac{1}{1+\sqrt{\gamma\tilde{x}\lambda}}, \tag{3.1}$$

so that w_x^{c*} is determined by the combination effects of $\tilde{x}$, γ and λ. For different measures, $\tilde{x}$ is different even though π_1 and π_2 are the same. This is the reason why investigators should assign different numbers of clusters in the two arms when they use different measures.

Given γ and λ, we first discuss the effects of π_1 and π_2 on w_x^{c*} when they are misspecified. Suppose π_2 is fixed and we change the value of π_1. Clearly, as π_1 increases from 0 to 1, $1-\pi_1$ decreases from 1 to 0. Therefore when the measure is RR, increasing π_1 makes $\tilde{x}$ larger, hence w_{RR}^{c*} becomes smaller, and small number of clusters should be assigned to arm 1. Note that $\pi_1/(1-\pi_1)$ has no upper bound, so w_{RR}^{c*} can be very small if π_1 is very large.

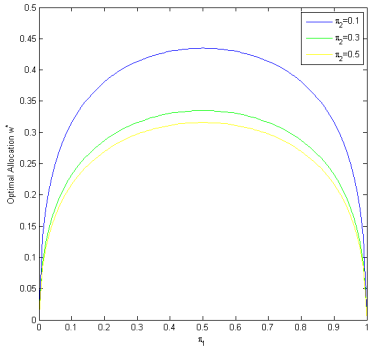
When the measure is RD and OR, π_1 and $1-\pi_1$ cancel each other's effect on $\tilde{x}$. Consequently, $\tilde{x}$ does not increase or decrease when π_1 increases. With π_1 increasing from 0 to 0.5, $\pi_1(1-\pi_1)$ increases from 0 to its maximal value 0.25, and with π_1 increasing from 0.5 to 1, $\pi_1(1-\pi_1)$ decreases from 0.25 to 0. Therefore, for the measure RD, increasing π_1 from 0 to 0.5 makes $\tilde{x}$ smaller, hence w_{RD}^{c*} becomes larger; and increasing π_1 from 0.5 to 1 makes $\tilde{x}$ larger, hence w_{RD}^{*} becomes smaller. Correspondingly, for the measure OR, w_{OR}^{c*} becomes smaller with π_1 increasing from 0 to 0.5 and becomes larger with π_1 increasing from 0.5 to 1.

The effect of π_2 can be studied in a similar way, but in a reverse way. When the measure is RR, increasing π_2 makes w_{RR}^{c*} larger. When the measure is RD, w_{RD}^{c*} becomes smaller when π_2 increases from 0 to 0.5 and becomes larger when π_2 increases from 0.5 to 1. When the measure is OR, w_{OR}^{c*} becomes larger when π_2 increases from 0 to 0.5 and becomes smaller when π_2 increases from 0.5 to 1.

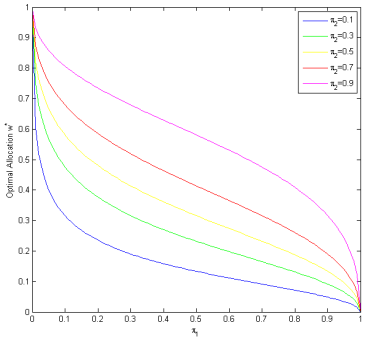
The effects of π_1 and π_2 on the optimal allocation w_x^{c*} are illustrated in Figure 3.1. In this figure, we let $\rho_1 = 0.3$, $\rho_2 = 0.1$ and $m = 14$, which are the parameter values in Samoan Women's Health Study.

Figure 3.1(a) shows how the optimal allocation value changes for different values of π_1 and π_2 for RD. We see that for a fixed value of π_2, w_{RD}^{c*} increases with π_1 increasing from 0 to 0.5 and then decreases with π_1 increasing from 0.5 to 1. When π_1 is close to 0 or close to 1, the change rate of w_{RD}^{c*} is very large. When π_1 is close to 0.5, the change rate of w_{RD}^{c*} is very small. For example, when $\pi_2 = 0.5$, w_{RD}^{c*} changes from 0 to about 0.23 with π_1 increasing from 0 to 0.1; in contrast, w_{RD}^{c*} almost does not change with π_1 increasing from 0 to 0.1. We see that the change rate of w_{RD}^{c*} is larger for $\pi_2 = 0.1$ than that for $\pi_2 = 0.5$. It indicates that if π_1 and π_2 are close to 0.5, for the measure RD, the optimal allocation value is robust. But if π_1 or π_2 is very close to 0 or 1, misspecifying π_1 or π_2 may result in a very different allocation value.

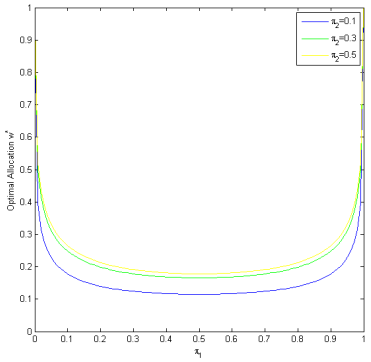
Figure 3.1(b) shows how the optimal allocation value changes for different values of π_1 and π_2 for RR. We see that for a fixed value of π_2, w_{RR}^{c*} keeps decreasing as π_1



(a) RD



(b) RR



(c) OR

Figure 3.1: Relationship between optimal allocation and π_1 for different π_2 when ρ_1=0.3, ρ_2=0.1, m=14, and γ=10)

increases from 0 to 1. When π_1 is close to 0 or close to 1, the change rate of w_{RR}^{c*} is very large. When π_1 is close to 0.5, the change rate of w_{RR}^{c*} is not small. It indicates that for the measure of RR, even though π_1 and π_2 are close to 0.5, the optimal allocation value is not very robust, and misspecifying π_1 or π_2 may result in a very different allocation value.

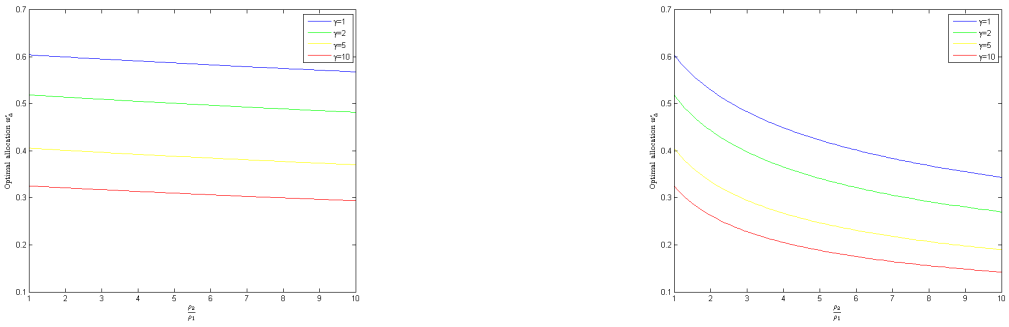
Figure 3.1(c) shows how optimal allocation value changes for different values of π_1 and π_2 for OR. For a fixed value of π_2, w_{OR}^{cc*} decrease with π_1 increasing from 0 to 0.5 and then increase with π_1 increasing from 0.5 to 1. When π_1 is close to 0 or close to 1, the change rate of w_{OR}^{c*} is very large. When π_1 is close to 0.5, the change rate of w_{OR}^{c*} is very small. The change rate of w_{OR}^{c*} is larger when $\pi_2 = 0.1$ than that when $\pi_2 = 0.5$. Hence for the measure OR, if π_1 and π_2 are close to 0.5, the optimal allocation value is robust. But if π_1 or π_2 is very close to 0 or 1, misspecifying π_1 or π_2 may result in a very different allocation value. The conclusion is similar to that for the measure RD.

In summary, when RR is used, the optimal allocation value is sensitive to different values of π_1 and π_2. When RD or OR is used, the optimal allocation value does not change a lot for different values of π_1 and π_2 when they are both close to 0.5. However, if π_1 or π_2 is close to 0 or 1, the optimal optimal allocation value can vary a lot for different values of π_1 and π_2.

3.2.2 Effects of cost ratio, ICCs and cluster size

Given π_1 and π_2, $\tilde{x}$ is fixed and the optimal allocation w_x^{c*} is determined by the product of $\gamma\lambda$. Recall that γ is the cost ratio and λ is the design effect ratio. As the value of $\gamma\lambda$ increases, w_x^{c*} decreases which means that fewer clusters need be assigned to arm 1; in contrast, as the value of $\gamma\lambda$ decreases, w_x^{c*} increases, and more clusters need be assigned to arm 1.

Recall that the unit cost ratio for a cluster in arm 1 versus arm 2 is $\gamma = \frac{mc_1+e_1}{mc_2+e_2}$. The parameters c_1, c_2 are individual-level cost and e_1, e_2 are cluster-level cost. To simplify the discussion in this chapter, we do not distinguish individual-level cost and cluster-level



(a) m=5 (b) m=500

Figure 3.2: Relationship between optimal allocation and ICC ratio for estimating RD with different cost ratios when π_1=0.3, π_2=0.1 and ρ_1=0.01

cost. Hence cost ratio is reduced to $\gamma = \frac{c_1}{c_2}$.

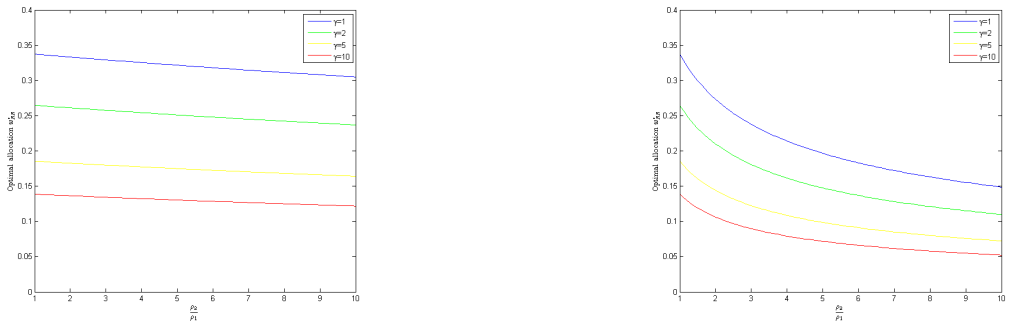
Given π_1 and π_2 we consider three more specific situations.

(1). $\rho_1 = \rho_2$: In this case, $w_x^{c*} = \frac{1}{1+\sqrt{\gamma\tilde{x}}}$. We see when ρ_1 and ρ_2 are equal, accurate values of cluster size of m, ρ_1 and ρ_2 are not needed to determine w_x^{c*}. The value of w_x^{c*} is determined by the cost ratio $\gamma = \frac{c_2}{c_1}$.

(2). $(m-1)\rho_1$,$(m-1)\rho_2$ are very small: $\lambda = \frac{1+(m-1)\rho_2}{1+(m-1)\rho_1} \approx 1$ and $w_x^* \approx \frac{1}{1+\sqrt{\gamma\tilde{x}}}$. We see that even though ρ_1 and ρ_2 are unequal, as long as ICCs are very small and cluster size m is small, accurate values of m, ρ_1 and ρ_2 are not needed to determine w_x^{c*}. The value of w_x^{c*} is determined by the cost ratio $\gamma = \frac{c_1}{c_2}$.

(3). Large cluster size $m \to \infty$: $w_x^{c*} \approx \frac{1}{1+\sqrt{\gamma\tilde{x}\frac{\rho_2}{\rho_1}}}$. When the cluster size is very large, an accurate value of m is not needed to determine w_x^{c*}. w_x^{c*} is determined by the cost ratio $\frac{c_2}{c_1}$ and the ICC ratio $\frac{\rho_2}{\rho_1}$, which is independent of cluster size.

From Figure 3.2, we further explore the effects of the cost ratio $\gamma = \frac{c_1}{c_2}$ and the ICC ratio $\frac{\rho_2}{\rho_1}$ on optimal allocation for different cluster size. In Figure 3.2(a) where cluster size=5, we see that for each value of γ, with an increase of $\frac{\rho_2}{\rho_1}$, the change in w_{RD}^{c*} is very small. For example, when $\gamma = 1$, with $\frac{\rho_2}{\rho_1}$ increasing 10-fold, w_{RD}^{c*} decreases from 0.62 to 0.59 and the absolute change in w_{RD}^{c*} is only about 0.03. When $\gamma = 10$, with $\frac{\rho_2}{\rho_1}$ increasing 10-fold, the absolute change in w_{RD}^{c*} is only about 0.04. However, for fixed



(a) m=5 (b) m=500

Figure 3.3: Relationship between optimal allocation and ICC ratio for estimating RR with different cost ratios when π_1=0.3, π_2=0.1 and ρ_1=0.01

values of $\frac{\rho_2}{\rho_1}$, with an increase of γ, the change in w_{RD}^{c*} is large. For example, when $\frac{\rho_2}{\rho_1} = 1$, when γ changes from 1 to 10, that means cost ratio increase 10-fold, w_{RD}^{c*} decrease from 0.62 to 0.34, and the absolute change in w_{RD}^{c*} is about 0.28. Therefore, when cluster size is small, w_{RD}^{c*} is mainly affected by cost ratio γ.

In Figure 3.2(b) where cluster size=500, we see that for each value of γ, with an increase of $\frac{\rho_2}{\rho_1}$, the change in w_{RD}^{c*} is much larger. For example, when $\gamma = 1$, with $\frac{\rho_2}{\rho_1}$ increasing 10-fold, w_{RD}^{c*} decreases from 0.76 to 0.52 and the absolute change in w_{RD}^{c*} is about 0.24. And for fixed values of $\frac{\rho_2}{\rho_1}$, with an increase in γ, the change in w_{RD}^{c*} is also much larger. For example, when $\frac{\rho_2}{\rho_1} = 1$, with γ changing from 1 to 10, w_{RD}^{c*} decreases from 0.76 to 0.50, and the absolute change in w_{RD}^{*} is about 0.26. Therefore, when cluster size is large, w_{RD}^{c*} is affected by both cost ratio γ and ICC ratio $\frac{\rho_2}{\rho_1}$.

Figures 3.3(a) and 3.3(b) show the relationship between optimal allocation for estimating RR with ICC ratio $\frac{\rho_2}{\rho_1}$ and cost ratio γ, and cluster size=5 in Figure 3.3(a) and cluster size =500 in Figure 3.3(b). Although for specific values of $\frac{\rho_2}{\rho_1}$ and γ, the value of w_{RR}^{c*} in Figure 3.3(a) with small cluster size is different from w_{RD}^{c*} in Figure 3.2(a), the pattern of curves are similar. For each value of γ, with an increase of $\frac{\rho_2}{\rho_1}$, the change in w_{RD}^{c*} is very small. Hence w_{RR}^{c*} is mainly affected by the cost ratio when cluster size is small. Further, the curves pattern in Figure 3.3(b) is similar with that in Figure 3.2(b).

When cluster size is large, w_{RR}^{c*} is affected by both cost ratio γ and ICC ratio $\frac{\rho_2}{\rho_1}$.

Interestingly, given π_1 and π_2, the effects of cluster size m, cost ratio γ and ICC ratio $\frac{\rho_2}{\rho_1}$ on optimal allocation w_x^{c*} for all the three measures ($x = RD, RR, OR$) are similar. This is not surprising. In equation (3.1), the optimal allocation w_x^{c*} for all three measures is determined by γ and $\lambda = \frac{1+(m-1)\rho_2}{1+(m-1)\rho_1}$ in the same way.

In summary, suppose π_1 and π_2 are given, when the cluster size is small, the optimal allocation w_x^{c*} is mainly determined by cost ratio, and is robust to ICCs. Even though the ICCs are mis-specified, the allocation w does not change much. Only when cluster size is very large, misspecifying ICCs may result in a wrong allocation value. The trend of w with ICC ratio change for the three measures are the same.

3.3 Effects of parameters on the allocation interval of w for a fixed RCE

3.3.1 Interval of w for a fixed RCE

In this section, we discussed how different parameters affect the optimal allocation value w for a measure. If parameters are misspecified in the study, the allocation value of w calculated may not be the optimal allocation w_x^{c*}. The value of w can be any number between 0 and 1 and determines the RCEFf of a design. When the value of w is equal to the optimal allocation w_x^{c*} for the measure $x \in (RD, RR, OR)$, the RCE for that measure x reaches the maximal value 1. When the value of w departs from w_x^{c*}, the RCE decreases.

However, in practice we may not need the RCE be 1. We may only need RCE be larger than a pre-determined value, e.g., 0.9. If we set the RCE required be at least a, say a=0.9, then the corresponding w falls in an interval of $(w_{xl}^{c*}, w_{xu}^{c*})$ which depends on a. Therefore even the parameters are mis-specified, as long as the allocation value w is in such an interval, the requirement of the RCE is still satisfied. We define that interval the allocation interval and we now address how parameters affect that interval.

In order to find the lower bound w_{xl}^{c*} and the upper bound w_{xu}^{c*} of that interval, we set the right hand of equation (2.19) to be a and solve the equation with respect to w. We obtain:

$$w_{xl}^{*} = \frac{2a + (1-a)(\dot{x}+a) + 2\sqrt{\dot{x}\gamma} - (\sqrt{\dot{x}}+\sqrt{\gamma})\sqrt{(1-a)[(1-a)(\dot{x}+\gamma) + 2(1+a)\sqrt{\dot{x}\gamma}]}}{2[a + (1-a)(\dot{x}+\gamma) + 2\sqrt{\dot{x}\gamma}) + a\dot{x}\gamma]} \tag{3.2}$$

$$w_{xu}^{*} = \frac{2a + (1-a)(\dot{x}+a) + 2\sqrt{\dot{x}\gamma} + (\sqrt{\dot{x}}+\sqrt{\gamma})\sqrt{(1-a)[(1-a)(\dot{x}+\gamma) + 2(1+a)\sqrt{\dot{x}\gamma}]}}{2[a + (1-a)(\dot{x}+\gamma) + 2\sqrt{\dot{x}\gamma}) + a\dot{x}\gamma]} \tag{3.3}$$

where $\dot{x}$ is a function of $\pi_1, \pi_2, \rho_1, \rho_2$ and m for a particular measure $x \in (RD, RR, OR)$:

$$\dot{x}(\pi_1, \pi_2, \rho_1, \rho_2, m) = \begin{cases} \frac{\pi_2(1-\pi_2)}{\pi_1(1-\pi_1)} \frac{1+(m-1)\rho_2}{1+(m-1)\rho_1} & \text{if } x = RD \\ \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} \frac{1+(m-1)\rho_2}{1+(m-1)\rho_1} & \text{if } x = RR \\ \frac{\pi_1(1-\pi_1)}{\pi_2(1-\pi_2)} \frac{1+(m-1)\rho_2}{1+(m-1)\rho_1} & \text{if } x = OR \end{cases}$$

If $\pi_1, \pi_2, \rho_1, \rho_2$ and m are given, both w_{xl}^{c*} and w_{xu}^{c*} are functions of a. The optimal allocation w_x^{c*} for a particular measure x is always within this allocation interval $(w_{xl}^{c*}, w_{xu}^{c*})$. When a=1, both w_{xl}^{c*} and w_{xu}^{c*} reduce to w_x^{c*}, the optimal allocation for measure x, and the interval reduces to one point. As a decreases, the length of the interval $(w_{xl}^{c*}, w_{xu}^{c*})$ increases.

For a CRT design, as long as w is chosen from the allocation interval $(w_{xl}^{c*}, w_{xu}^{c*})$, the RCE will be at least as large as a. When the length of this interval is larger, more values of w can be chosen from the interval to satisfy the RCE requirement. In other words, when the length of the interval is larger, the design is more robust. If investigators choose w outside the interval, the RCE of that design is smaller than a.

If the RCE of a is given, the values of w_{xl}^{c*} and w_{xu}^{c*} are determined by the cost ratio γ and $\dot{x}$. Further, $\dot{x}$ is determined by the measure $x \in (RD, RR, OR)$ and the values of parameters π_1, π_2, ρ_1, ρ_2 and m. We are interested in how those parameters affect the interval $(w_{xl}^{c*}, w_{xu}^{c*})$ for a particular measure x.

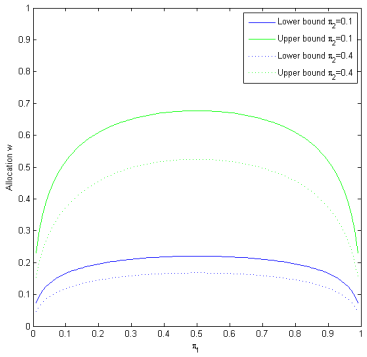
3.3.2 Effects of π_1 and π_2

We first explore the relationship between the allocation interval of w and π_1 and π_2 in Figure . In Figures 3.4(a), 3.4(b) and 3.4(c), the measures are RD, RR and OR respectively. We let $\rho_1 = 0.3, \rho_2 = 0.1$ and cluster size $m = 14$, which are the same values in Samoan women's health study. Assume that we want a CRT with at least 0.9 RCE.

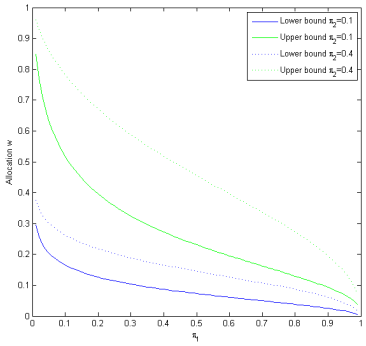
Figure 3.4(a) shows that for RD, given a specific value of π_2, both upper bound and lower bound increase with π_1 increasing from 0 to 0.5 and shifts the location of the interval towards 1. The upper bound and the lower bound decrease as π_1 increases from 0.5 to 1, and shifts the the interval towards 0. The length of the interval changes too. When π_1 increases from 0 to 0.5, the length of interval increases correspondingly, and obtain the maximal length when $\pi_1 = 0.5$. When π_1 increases from 0.5 to 1, the length of interval decreases. However, when π_1 is close to 0.5, the change of length is very small.

However, if the measure changes from RD to RR, we observe the change in the pattern of the behavior of the interval is different. Figure 3.4(b) shows that both the upper bound and the lower bound continue to decrease as π_1 increases from 0 to 1. Hence the location of the interval always moves towards 0 as π_1 increases from 0 to 1. In addition, the length of the interval continues to decrease. The implication is that for a large value of π_1, the interval is less likely to maintain the desired RCE.

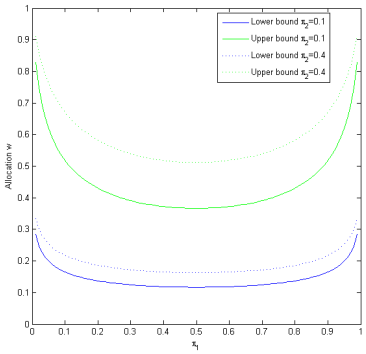
Figure 3.4(c) shows how the interval changes for different values of π_1 and π_2 for estimating OR. In contrast to the case with RD, both the upper bound and the lower bound decrease as π_1 increases from 0 to 0.5. Hence the location of the interval moves towards 0. When π_1 increases from 0.5 to 1, both the upper bound and the lower bound increase, and the location of the interval moves towards 1. The length of the interval decreases as π_1 increases from 0 to 0.5 and increases as π_1 increases from 0.5 to 1. Although the length of the interval achieves its smallest value at $\pi_1 = 0.5$, the change of the length of the interval is very gradual as π_1 gets closer to 0.5.



(a) RD



(b) RR



(c) OR

Figure 3.4: Relationship between the allocation interval of w and π_1 and π_2 with at least 0.9 RCE when $\rho_1 = 0.3, \rho_2 = 0.1, m = 14, \gamma = 10$.

3.3.3 Effects of cost ratio

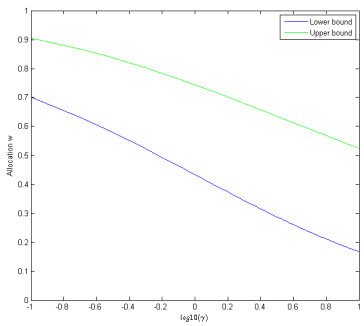
We continue to explore the relationship between the allocation interval of w and the cost ratio γ. Figure 3.5 shows the patterns of the upper bound and the lower bound for all three measures and they appear to be similar. With $log_{10}(\gamma)$ increasing from -1 to 1 (γ increase from 0.1 to 10), both w_{xl}^{c*} and w_{xu}^{c*} decrease, regardless of which measure is used. The location of interval $(w_{xl}^{c*}, w_{xu}^{c*})$ moves towards 0 as γ increases. This is reasonable since larger values of γ mean more resources need to be spent on individuals in arm 1, hence less individuals in arm 1, as expected.

When $log_{10}(\gamma) = 0$(i.e., $\gamma = 1$), Figure 3.5(a) shows the allocation interval of w with at least 0.9 RCE is (0.38,0.69) for estimating RD. When γ increases 10-fold, the interval is (0.18, 0.52). If the allocation in a CRT design is w=0.6, with equal costs in the two arms, the RCE of that design is at least 0.9, but when cost ratio γ increases to 10, the RCE is smaller than 0.9. From Figure 3.5(b), for estimating RR, the length of the interval slightly increases as γ increases. But from Figure 3.5(a) and 3.5(c), for estimating RD and OR, the length of the interval increases from 0.2 to 0.34 with γ increasing from 0.1 to 10. It indicates that when γ is larger, it is more likely to keep the RCE as large as the value required for RD and OR, but not for RR.

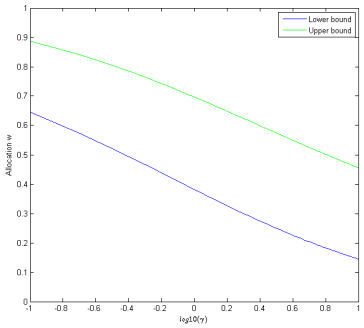
3.3.4 Effects of cluster size

We next explore the relationship between the interval of w and the cluster size m. We continue to assume that the true design parameters values are π_1=0.5, π_2=0.4, ρ_1=0.3 and ρ_2=0.1, the same parameter values in Samoan Women's Health Study. The cost ratio is $\gamma = 1$. In Samoan Women's Health Study, the mean cluster size is 14. To explore the relationship between the interval of w and cluster size m, we let m change from 1 to 200.

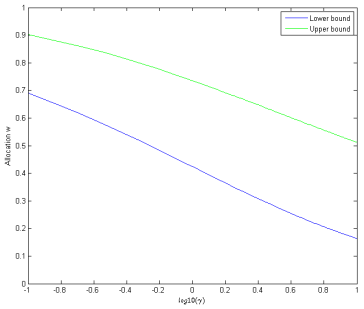
In Figures 3.6(a), 3.6(b) and 3.6(c), the patterns of the upper and lower bounds are similar. As m increases from 1 to 20, both w_{xl}^{c*} and w_{xu}^{c*} increase a bit, but as m increases from 20 to 200, there is little change in w_{xl}^{*} and w_{xu}^{*}. Therefore, the location of



(a) RD

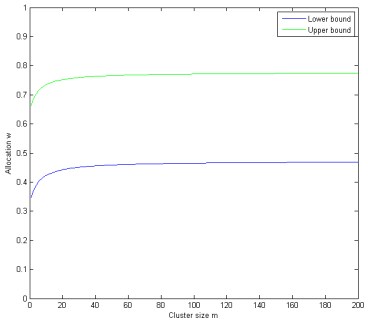


(b) RR

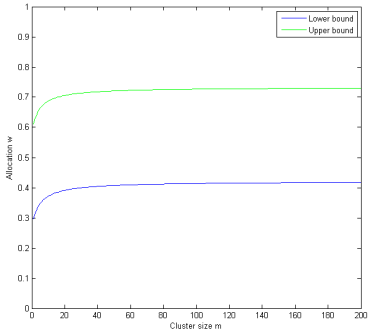


(c) OR

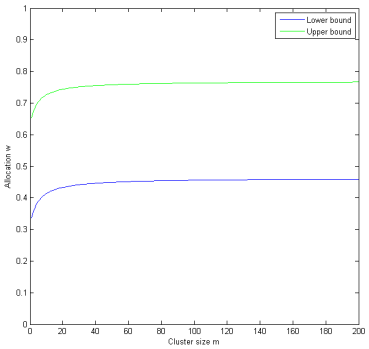
Figure 3.5: Relationship between the allocation interval of w and cost ratio γ with at least 0.9 RCE when $\pi_1 = 0.5, \pi_2 = 0.4, \rho_1 = 0.3, \rho_2 = 0.1, m = 14$.



(a) RD



(b) RR

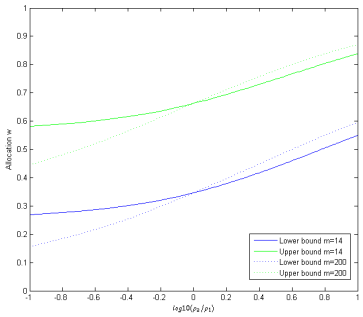


(c) OR

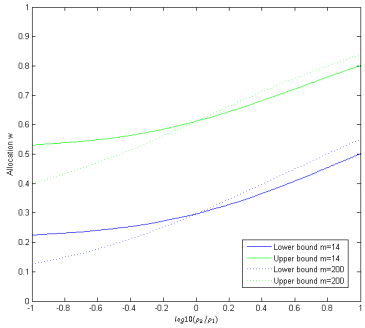
Figure 3.6: Relationship between the allocation interval of w and cluster size m with at least 0.9 RCE when $\pi_1 = 0.5$, $\pi_2 = 0.4$, $\rho_1 = 0.3$, $\rho_2 = 0.1$, $\gamma = 1$.

interval $(w_{xl}^{c*}, w_{xu}^{c*})$ moves toward 1 a bit as m increases from 1 to 20, and then remains approximately the same after that. From Figure 3.6(a) where RD is estimated, with $m = 1$, the interval with at least 0.9 RCE is (0.35,0.67); as m increases to 20, the interval is (0.44,0.75); as m increases from 20 to 200, there is almost no change of the interval. From Figure 3.6(a), 3.6(b) and 3.6(c), if allocation $w = 0.5$ in a CRT design, the RCE of that design which is the balance design stays at least 0.9, no matter how m changes. The length of the interval stays about 0.3, no matter how m changes.

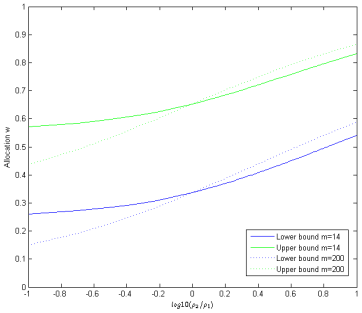
In summary, the allocation interval of w is insensitive to cluster size. The length of the interval is almost the same as cluster size increases. The location of the interval has little change with cluster size increasing when cluster size is larger than 20. This conclusion is the same for all three measures. Therefore, it seams reasonable to use the



(a) RD



(b) RR



(c) OR

Figure 3.7: Relationship between the allocation interval of w and ICC ratio with at least 0.9 RCE when $\pi_1 = 0.5$, $\pi_2 = 0.4$, $\rho_1 = 0.3$, $\rho_2 = 0.1$, $\gamma = 1$.

mean of cluster size to calculate the allocation value when cluster size is varying.

3.3.5 Effects of ICCs

In this subsection, we explore the relationship between the interval of w and the ICC ratio $\frac{\rho_1}{\rho_2}$. We choose $\frac{\rho_1}{\rho_2}$ rather than $\rho_1 - \rho_2$ is because the former may approximate design effect ratio. Suppose we have π_1=0.5, π_2=0.4, ρ_2=0.1, m=14, the same as in Samoan Women's Health Study and $\gamma = 1$. To explore the relationship between the interval of w and the ICC ratio, we let ρ_1 changes from 0.01 to 1 and fix ρ_2 as 0.1. In addition, we also consider the situation in which $m = 200$, which stands for the case when we have a large cluster size.

In Figures 3.7(a), 3.7(b) and 3.7(c), the patterns of upper and lower bounds are

similar. As $log_{10}(\rho_1/\rho_2)$ increases from -1 to 1 (ρ_1/ρ_2 increase from 0.1 to 10), both w_{xl}^{c*} and w_{xu}^{c*} increase. Therefore, the location of interval $(w_{xl}^{c*}, w_{xu}^{c*})$ moves toward 1 as ICC ratio ρ_1/ρ_2 increases. The rates of increase in w_{xl}^{c*} and w_{xu}^{c*} when $m = 200$ are faster than the rates of increase when $m = 14$. Hence the location of the interval is more robust to ICC ratio changes when the cluster size is small. In addition, the lengths of the interval for all three measures stay about 0.3 as the ICC ratio increases. That means ICC change may affect the location of the interval, but has little effects on the length of the interval.

3.4 Bayesian methods for the efficient and robust design

When the total cluster number is fixed, to determine the optimal allocation w_x^{c*} in a CRT, investigators need to know the values of π_1, π_2, ρ_1, ρ_2 and m. While m may be known in advance, investigators often do not know accurate values of π_1, π_2, ρ_1 and ρ_2 before the study has been conducted. Investigators usually look for previously reported similar studies to guess what values those parameters are. In particular, accurately predicting parameters in the treatment arm is more difficult than predicting in the control arm, since the treatment is often novel, while the control condition is often usual care. If those parameters values are incorrect, the calculated w_x^{c*} may be incorrect, resulting in a low CE of the design, especially when the cost ratio of two arms is very different from unity.

We want to find an efficient and robust design. An efficient design means that the design has relatively high RCE, although it may not have the maximal RCE 1. A design is robust if the design is not affected very much by the misspecified parameters. Bayesian methods can help to construct an efficient and robust design when parameter values are uncertain. Although investigators may not able to obtain accurate point estimates of parameters before the study, they may obtain some information about the values of those parameters from previous similar studies. Based on such prior information, they may be able to construct prior distribution for those parameters.

The basic idea of constructing an efficient and robust design using Bayesian methods

is the same as discussed in previous sections: to find an optimal allocation w which maximizes CE. The criterion for Bayesian methods is to maximize the expected precision or to minimize the expected variance with respect to the prior information of $(\pi_1, \pi_2) \in \Pi$ and $(\rho_1, \rho_2) \in R$ over the total cost.

In Section 3.3 we defined $\dot{x}$:

$$\dot{x}(\pi_1, \pi_2, \rho_1, \rho_2, m) = \begin{cases} \frac{\pi_2(1-\pi_2)}{\pi_1(1-\pi_1)} \frac{1+(m-1)\rho_2}{1+(m-1)\rho_1} & \text{if } x = RD \\ \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} \frac{1+(m-1)\rho_2}{1+(m-1)\rho_1} & \text{if } x = RR \\ \frac{\pi_1(1-\pi_1)}{\pi_2(1-\pi_2)} \frac{1+(m-1)\rho_2}{1+(m-1)\rho_1} & \text{if } x = OR \end{cases}$$

Using this notation, the variances for all three measures in formulas (2.2, 2.3 and 2.4) can be written in a unified format:

$$\Psi_x^{-1} \propto \left(\frac{1}{w} + \frac{\dot{x}(\pi_1, \pi_2, \rho_1, \rho_2, m)}{1-w} \right). \tag{3.4}$$

From formulas (2.12, 2.13 and 2.14), CE for all three measures can be written in a unified format:

$$CE_x \propto \left(\frac{1}{w} + \frac{\dot{x}(\pi_1, \pi_2, \rho_1, \rho_2, m)}{1-w} \right)^{-1} (\gamma w + (1-w))^{-1}. \tag{3.5}$$

Therefore, the criterion for Bayesian methods is to minimize $\Gamma(w)_x$:

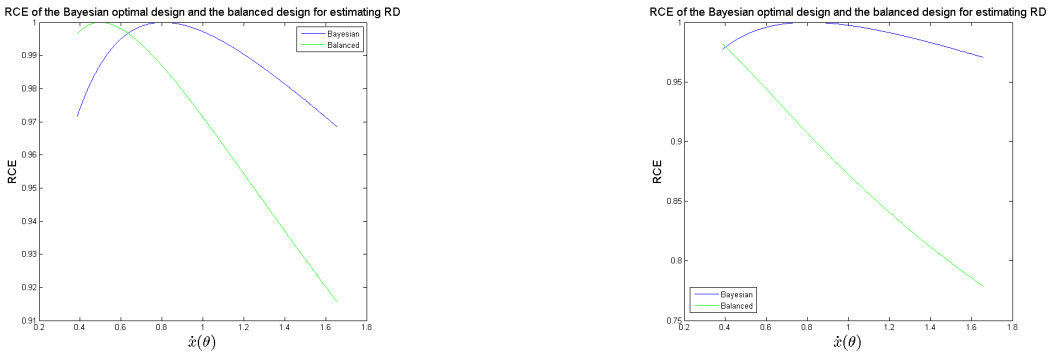
$$\Gamma(w)_x \propto \int_R \int_\Pi [\gamma w + (1-w)] \left\{ \frac{1}{w} + \frac{\dot{x}(\pi_1, \pi_2, \rho_1, \rho_2, m)}{1-w} \right\} dP(\pi_1, \pi_2) dP(\rho_1, \rho_2), \tag{3.6}$$

A direct calculation shows that the optimal allocation $w_{\Phi x}^{c*}$ for a particular measure $x \in (RD, RR, OR)$ is given by

$$w_{\Gamma x}^{c*} = \frac{1}{1 + \sqrt{\gamma E[\dot{x}(\pi_1, \pi_2, \rho_1, \rho_2, m)]}}. \tag{3.7}$$

We observe that $w_{\Phi x}^{c*}$ is determined by the expectation of $\dot{x}$, which depends on the prior distributions of π_1, π_2, ρ_1 and ρ_2. Investigators need to specify the prior distributions of those parameters before the Bayesian design can be carried out. We let $\theta = (\pi_1, \pi_2, \rho_1, \rho_2)$.

Now we compare the Bayesian design to the balanced design. In Figure 3.8, π_1, π_2, ρ_1 and ρ_2 follow uniform distributions on the intervals (0.3,0.5), (0.2,0.3), (0.05, 0.1)

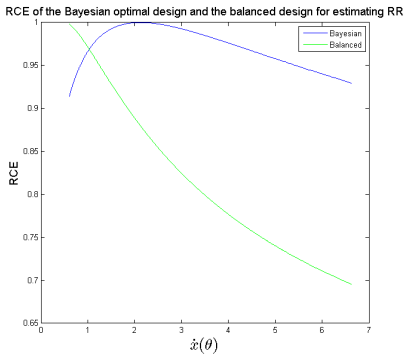


(a) $\gamma = 2$ (b) $\gamma = 5$

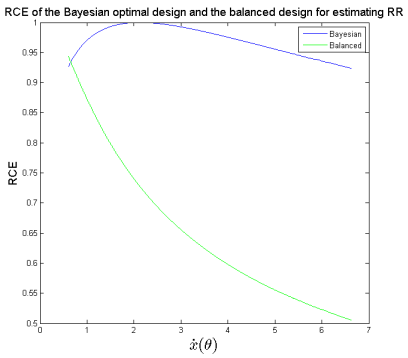
Figure 3.8: RCE of the Bayesian design and the balanced design for estimating RD for different cost ratios γ when $\pi_1 \sim u(0.3, 0.5), \pi_2 \sim u(0.2, 0.3), \rho_1 \sim u(0.1, 0.2), \rho_2 \sim u(0.1, 0.2)$ and all cluster sizes have $m = 20$ subjects.

and (0.05, 0.1), respectively. Figure 3.8(a) shows RCEs of the Bayesian design and the balanced design for estimating RD when the cost ratio is 2. We see that when $\dot{x}(\theta) > 0.62$, the RCE of the Bayesian design is larger than that of the balanced design; when $\dot{x}(\theta) < 0.62$, the RCE of the Bayesian design is smaller than that of the balanced design. But over the whole range of $\dot{x}(\theta)$, the RCE for the Bayesian design is at least 0.97, while the lowest RCE of the balanced design is less than 0.92. Figure 3.8(b) shows results for cost ratio $\gamma = 5$. We see that over the whole range of $\dot{x}(\theta)$, the RCE of the Bayesian design is larger than 0.97, while the lowest RCE of the balanced design is less than 0.78. In addition, in almost the whole range of $\dot{x}(\theta)$, the RCE of the Bayesian design is larger than that of the balanced design.

Figure 3.9(a) shows RCE of the Bayesian design and the balanced design for estimating RR when the cost ratio is 2. We see that for a large range of $\dot{x}(\theta)$, the RCE of the Bayesian design is larger than that of the balanced design. Over the whole range of $\dot{x}(\theta)$, the RCE for the Bayesian design is at least 0.92, and the lowest RCE of the balanced design is less than 0.7. Figure 3.9(b) shows results for cost ratio $\gamma = 5$. We see that in almost the whole range of $\dot{x}(\theta)$, the RCE of the Bayesian design is larger than that of the balanced design. The RCE of the Bayesain design is at least 0.92, while the lowest RCE of the balanced design reduces to 0.5.



(a) $\gamma = 2$



(b) $\gamma = 5$

Figure 3.9: RCE of the Bayesian design and the balanced design for estimating RR for different cost ratios γ when $\pi_1 \sim u(0.3, 0.5), \pi_2 \sim u(0.2, 0.3), \rho_1 \sim u(0.1, 0.2), \rho_2 \sim u(0.1, 0.2)$ and all cluster sizes have $m = 20$ subjects.

3.5 Maximin methods for the efficient and robust design

Another method to construct an efficient and robust design is to use the maximin method. We briefly explain the logic behind the maximin method. For an arbitrary value of $w \in (0, 1)$, there exists at least one combination of π_1, π_2, ρ_1 and ρ_2 from their respective ranges for which the RCE is the smallest. Let the smallest RCE be $min(RCE)$. In all of those associated pairs of $(w, min(RCE))$, we can find one specific w which maximizes $min(RCE)$ among all designs (i.e. all values of w). In practice, we do not know accurate values of $\pi_1, \pi_2, \rho_1, \rho_2$, but only ranges of values for those parameters from previous similar studies and experiences. However, such information is enough to find w. Let parameter vector $\theta = (\pi_1, \pi_2, \rho_1, \rho_2)$. The criterion is to find a design that maximizes $min(RCE_x(\theta, w, m, \gamma)|\theta)$, where, $x \in$ RD, RR or OR.

Using $\dot{x}$ introduced before, the optimal allocation w_{RD}^{c*}, w_{RR}^{c*} and w_{OR}^{c*} in formulas (2.15, 2.16, 2.17) can all be written as:

$$w_x^{c*} = \frac{1}{1 + \sqrt{\gamma \dot{x}}}. \tag{3.8}$$

Let $\underline{x} = min(\dot{x}(\theta))$ and $\overline{x} = max(\dot{x}(\theta))$, where $\theta = (\pi_1, \pi_2, \rho_1, \rho_2)$. Based on our criterion to find w_{maxmin}^{c*}, we have the following theorem.

Theorem 3.1:

The maximin design for estimating a measure of difference of success rates in a two-arm CRT allocates w^{c*}_{maxmin} of clusters to the arm 1, and w^{c*}_{maxmin} is:

$$w^{c*}_{maxmin} = \frac{(\sqrt{\gamma}+\sqrt{\underline{x}})^2 - (\sqrt{\gamma}+\sqrt{\overline{x}})^2}{(\sqrt{\gamma}+\sqrt{\overline{x}})^2(\underline{x}-1) - (\sqrt{\gamma}+\sqrt{\underline{x}})^2(\overline{x}-1)} \tag{3.9}$$

Proof:

The proof follows Dette [19] who considered the maximin method for IRTs with binary outcomes. Recall that the optimal allocation $w^{c*} = \frac{1}{1+\sqrt{\gamma\dot{x}}}$ and the RCE from equation (2.18) is

$$RCE(\dot{x},\gamma,m) = \frac{(1+\sqrt{\dot{x}/\gamma})^2}{(w+(1-w)/\gamma)(1/w+\dot{x}/(1-w))}.$$

Based on our definitions of $\underline{x}$ and $\overline{x}$, we have:

$$min(RCE(w,\theta,\gamma)) = min_{\dot{x}\in[\underline{x},\overline{x}]}RCE(\dot{x},\gamma,w).$$

Further, calculus shows that there is only one solution to $\frac{\partial}{\partial\dot{x}}RCE(\dot{x},\gamma,w)$=0, and is given by $\dot{x}^* = \frac{1-w}{w\sqrt{\gamma}}$. We also have $\frac{\partial^2}{\partial^2\dot{x}}RCE(\dot{x},\gamma,w)|\dot{x}* = \frac{1-w}{w\sqrt{\gamma}} < 0$ and so $RCE(\dot{x},\gamma,w)$ has a local maximum in the interval $[\underline{x},\overline{x}]$. It follows that:

$$min_{\dot{x}\in[\underline{x},\overline{x}]}RCE(\dot{x},\gamma,w) = min(RCE(\underline{x},\gamma,w), RCE(\overline{x},\gamma,w)). \tag{3.10}$$

To maximize the right-hand of equation (3.10) by choice of w, we first assume the optimal choice is

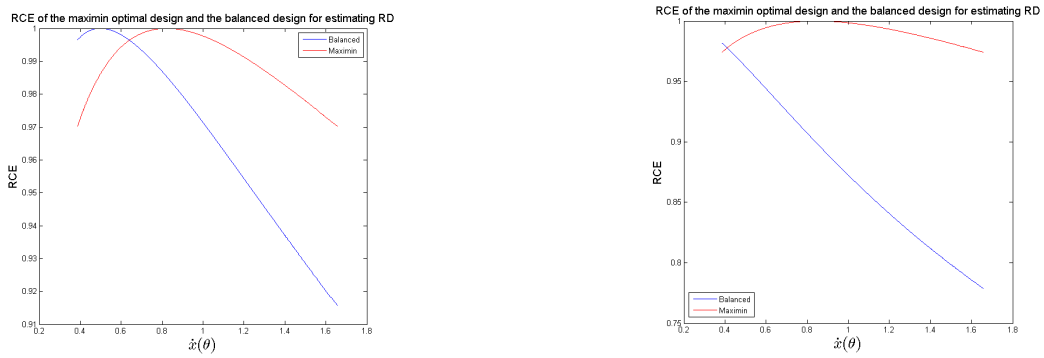
$$w^{c*}_{maximin} \in M_1 = \{w\in[0,1]|RCE(\underline{x},\gamma,w) < RCE(\overline{x},\gamma,w)\},$$

whereupon one obtains $w^*_{maximin} = \frac{1}{1+\sqrt{\gamma\underline{x}}}$. By the definition of M_1, we have

$$RCE(\underline{x},\gamma,\frac{1}{1+\sqrt{\gamma\underline{x}}}) < RCE(\overline{x},\gamma,\frac{1}{1+\sqrt{\gamma\underline{x}}}),$$

implying that $(\sqrt{\underline{x}}-\sqrt{\overline{x}})^2/\sqrt{\gamma} < 0$, and hence a contradiction. Therefore $w^{c*}_{maximin}$ is not in the set M_1. Similarly, if we assume

$$w^{c*}_{maximin} \in M_2 = \{w\in[0,1]|RCE(\underline{x},\gamma,w) > RCE(\overline{x},\gamma,w)\},$$



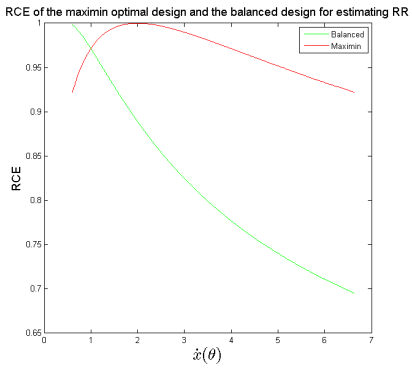
(a) $\gamma = 2$ (b) $\gamma = 5$

Figure 3.10: RCE of the maximin design and the balanced design for estimating the RD for different cost ratios γ when $\pi_1 \in [0.3, 0.5], \pi_2 \in [0.2, 0.3], \rho_1 \in [0.1, 0.2], \rho_2 \in [0.1, 0.2]$ and all cluster sizes have $m = 20$ subjects.

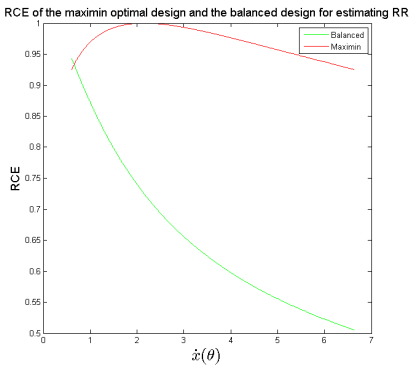
a similar argument also yields a contradiction and $w^{c*}_{maximin}$ is not in the set M_2. It follows that $w^{c*}_{maximin} \in M_3$=$\{w \in [0,1] | RCE(\underline{x}, \gamma, w)$=$RCE(\overline{x}, \gamma, w)\}$ and solving this equation $RCE(\underline{x}, \gamma, w) = RCE(\overline{x}, \gamma, w)$ yields the desired value of $w^{c*}_{maximin}$ in our result.

Let us compare the maximin design to the balanced design for estimating RD, RR and OR. We first consider the RD. Figure 3.10(a) shows RCEs of the maximin design and the balanced design when the cost ratio is 2, which is relatively small. We see that over the whole range of $\dot{x}(\theta)$, the RCE for the maximin design is at least 0.97, while the lowest RCE of the balanced design is less than 0.92. In addition, for a larger range of $\dot{x}(\theta)$, the RCE of the maximin design is larger than that of the balanced design. Figure 3.10(b) shows results for cost ratio $\gamma = 5$. We see that over the whole range of $\dot{x}(\theta)$, the RCE of the maximin design is larger than 0.96, while the lowest RCE of the balanced design is less than 0.78. In addition, in almost the whole range of $\dot{x}(\theta)$, the RCE of the maximin design is larger than that of the balanced design. Hence the maximin design is much better than the balanced design when the cost ratio is 5.

We next consider the RR. Figure 3.11(a) is the graph for cost ratio $\gamma = 2$ and Figure 3.11(b) is the corresponding graph for cost ratio $\gamma = 5$. From both figures, we see that the maximin design is better than the balanced design. Note that the lowest RCEs of



(a) $\gamma = 2$



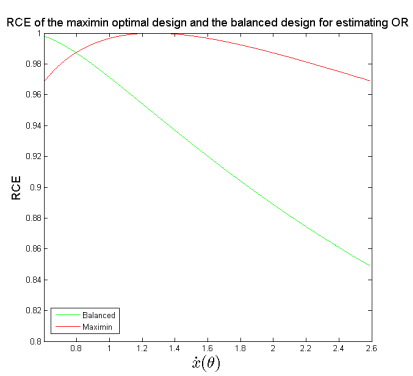
(b) $\gamma = 5$

Figure 3.11: RCE of the maximin design and the balanced design for estimating RR for different cost ratios γ when $\pi_1 \in [0.3, 0.5], \pi_2 \in [0.2, 0.3], \rho_1 \in [0.1, 0.2], \rho_2 \in [0.1, 0.2]$ and all cluster sizes have $m = 20$ subjects.

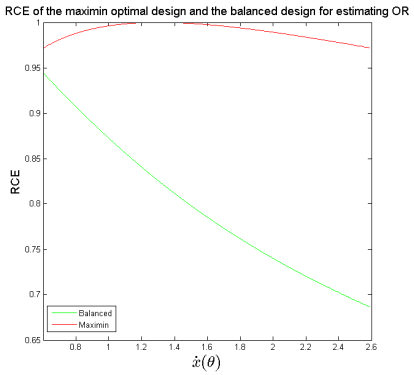
the balanced designs are less than 0.70 and 0.52 for $\gamma = 2$ and $\gamma = 5$ respectively in Figure 3.11, and both are lower than those in Figure 3.10. Hence if the measure is RR rather than RD, the balanced design is more sensitive to misspecified parameters and the maximin design is more helpful to avoid low RCE.

We finally consider the OR. Figure 3.12(a) is the graph for cost ratio $\gamma = 2$ and Figure 3.12(b) is the corresponding graph for cost ratio $\gamma = 5$. The lowest RCEs of the maximin design are larger than 0.96, obviously better than the balanced design. Note that the lowest RCEs of the balanced design are about 0.85 and 0.68 for $\gamma = 2$ and $\gamma = 5$ respectively in Figure 3.12, and both are lower than those in Figure 3.10 but larger than those in Figure 3.11. Hence estimating OR is less sensitive than estimating RR but more sensitive than estimating RD in the balanced design. The maximin design is helpful to avoid a low RCE.

It is interesting to compare the Bayesian design with the maximin design. Figure 3.13 shows that the RCE curves of the Bayesian design almost coincide with those from the maximin design, no matter the measure is RD or RR. It is very hard to tell which is better. But both are much better than the balanced design. Note that in this figure, for the Bayesian design, π_1, π_2, ρ_1 and ρ_2 are assumed to follow uniform distributions.

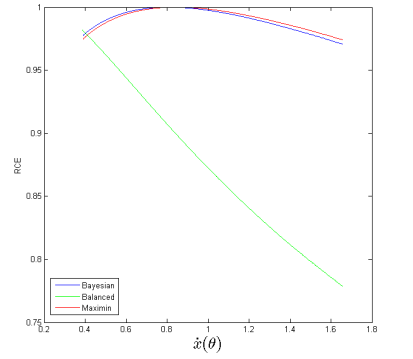


(a) $\gamma = 2$

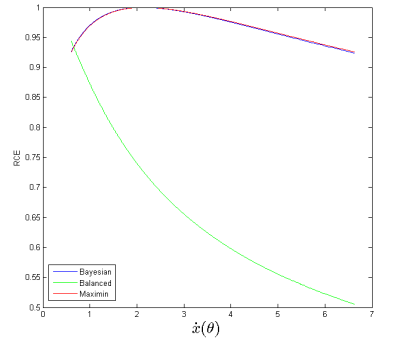


(b) $\gamma = 5$

Figure 3.12: RCE of the maximin design and the balanced design for estimating OR for different cost ratios γ when $\pi_1 \in [0.3, 0.5], \pi_2 \in [0.2, 0.3], \rho_1 \in [0.1, 0.2], \rho_2 \in [0.1, 0.2]$ and all cluster sizes have $m = 20$ subjects.



(a) RD



(b) RR

Figure 3.13: RCE of the Bayesian, the maximin and the balanced designs for RD and RR when $\pi_1 \sim u(0.3, 0.5), \pi_2 \sim u(0.2, 0.3), \rho_1 \sim u(0.1, 0.2), \rho_2 \sim u(0.1, 0.2)$, cost ratio $\gamma = 5$ and all cluster sizes have $m = 20$ subjects.

That is the reason why the Bayesian design and the maximin design have comparable RCE. The general comparison between the two designs is difficult [19, 29].

3.6 Application

Now we consider re-designing the Samoan women's health study using the results in previous sections. We first need determine the distribution or the range of π_1, π_2, ρ_1 and ρ_2. π_2 is the probability of mammography use for Samoan women without intervention. Mishra et al. reported the probability is 0.224 and 0.244 in Hawaii and Los Angeles [23]. Hence we assume that π_2 follows uniform distribution in a range of 0.2 and 0.3. Now we consider the possible value for π_1 in the intervention arm. First, we think the intervention will increase the mammography use. Therefore, the smallers value of π_1 should be higher than the largest possible value of π_2. Second, we have less certainty about the intervention arm, so we specify a larger variance of π_1. Considering these issues, we assume that π_1 follow a uniform distribution in a range of 0.3 and 0.6. Hade et al. [30] reported ICCs for cancer screening CRTs ranging from 0.05 to 0.3. Not all the clusters in those CRTs were churches and not all CRTs were about mammography use. However, these values provide reasonable ranges of ρ_1 and ρ_2. We assume that ρ_1 and ρ_2 follow a uniform distribution with range of 0.05 and 0.3. The intervention consisted of three components: specially developed English and Samoan language breast cancer educational booklets; skill building and behavioral exercises; and interactive group discussion sessions. In the control arm, women were only provided with the breast cancer educational materials. Hence we have the high cost ratio $\gamma = 5$.

From equation (3.7), the Bayesian allocation value is 0.32, hence we need assign 55x0.32=18 clusters to arm 1. From equation (3.9), the maximin allocation value is 0.30, hence we need assign 55x0.30=17 clusters to arm 1. The RCE comparison is illustrated in Figure 3.14. It is clear that either assigning 18 or 17 clusters to arm 1 has higher RCE than the original design in which 30 clusters are assigned to arm 1.

3.7 Chapter summary and discussion

In this chapter, we first explored how different parameters affect the optimal allocation. The effects of π_1 and π_2 are different for different measures. For RD and OR, when both π_1 and π_2 are close to 0.5, misspecifying π_1 and π_2 tend to yield an allocation that is similar with the optimal allocation. When either π_1 or π_2 is close to 0 or 1, misspecifying π_1 and π_2 can yield an allocation that is very different from the optimal allocation. This means that if π_1 and π_2 are close to 0.5, the optimal allocation is robust. For RR, when both π_1 and π_2 are close to 0.5, misspecifying π_1 and π_2 can yield an allocation that is very different from the optimal allocation. Therefore, RR is more sensitive to misspecifying π_1 and π_2. This result suggests that investigator may want to use RD and OR rather than RR when there is great uncertainty about π_1 and π_2. The effects of ρ_1 and ρ_2 depend on cluster size. If the cluster size is very small, then the ratio of ρ_1 and ρ_2 has little impact on the optimal allocation value. When the cluster size increases, the effects of misspecification of ρ_1 and ρ_2 become larger. However, their impact on allocation value is not as large as misspecification of π_1 and π_2.

If we set the RCE required be at least a, the corresponding w falls in an interval which depends on a. We called it the allocation interval and explored the effects of parameters on that interval. It seems that ρ_1 and ρ_2 and cluster size m have little impact on the length of such interval. When cluster size is larger than 20, the location of the interval is almost the same. The effects of ρ_1, ρ_2 and cluster size m on the interval are similar for different measures. However, the impact of π_1 and π_2 on the interval depends on different measures. For RD and OD, for given π_2, the length of the interval is largest when π_1 equals to 0.5 and becomes smaller when π_1 is close to 0 or 1. For RR, the length of the interval becomes shorter with increasing of π_1. For a given π_2, it seems that the length of the interval is larger for RD than RR unless π_1 is very small. It suggests that RD may be preferred than RR unless π_1 is very small.

We have seen that misspecifying parameters may not yield the optimal allocation and may result in a design with low RCE. But true values of π_1, π_2, ρ_1 and ρ_2 are generally

not known before the study. Hence in order to construct a robust design, we proposed two methods: a Bayesian method and a maximin method. For the Bayesian method, we assumed uniform and beta distributions for π_1 and π_2 and uniform distribution for ρ_1 and ρ_2. We assumed that these distributions are obtained from historical data and they are informative priors. One alternative way is to assume non-informative priors and create a likelihood model. Then we can use the historical data to obtain the posterior distribution of these parameters. For the success rates, a non-informative conjugate Beta(1/3, 1/3) is suggested by Kerman [31]. For the ICCs, a non-informative uniform [0,1] is often assumed [32, 33]. The likelihood is based on the distributional assumption that have been used to construct confidence interval for ρ. One possible method is to assume that observed $\hat{\rho}$ follows a normal distribution around the true ρ [32].

In summary, the maximin method does not need distribution of π_1, π_2, ρ_1 and ρ_2 but only need the range of those parameters. The designs based on the Bayesian method and the maximin method often have the similar RCE. Therefore, we prefer the maximin method since it avoids requirement of the knowledge of the distribution of those parameters. We showed that designs based on the two methods often have higher RCE than a balanced design, in which cluster are equally assigned to the two arms. Current research often use the balanced design in CRTs, and we recommend investigators to consider using the maximin method or the Bayesian method to construct a robust design.

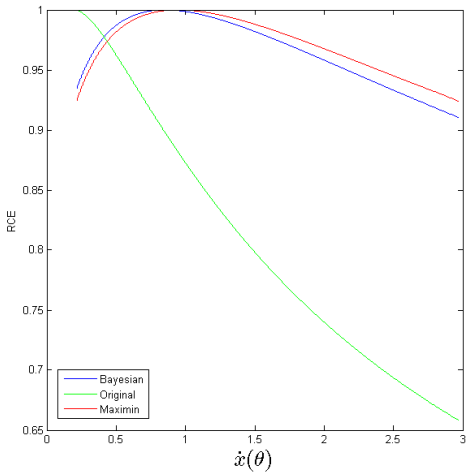


Figure 3.14: RCE of Bayesian design, maximin design and original design for estimating RD in Samoan study when $\pi_1 \sim u(0.3, 0.6)$, $\pi_2 \sim u(0.2, 0.3)$, $\rho_1 \sim u(0.05, 0.3)$, $\rho_2 \sim u(0.05, 0.3)$, cluster size m=14 and cost ratio $\gamma = 5$.

CHAPTER 4

Optimal sample size when cluster sizes are equal

4.1 Introduction

In previous chapters, we investigated the design problems for a two-arm CRT when the total number of clusters is pre-determined and equal to k and all cluster sizes are the same. We derived the optimal allocation w_x^{c*}, which is a function of success rates π_1 and π_2, ICCs ρ_1 and ρ_2, cost ratio γ and the constant cluster size equal to m. For different outcome measures x, $x \in (RD, RR, OR)$, the optimal allocation w_x^{c*} was found to be different. If the unit costs are very different in the two arms, a balanced design which assigns the same number of clusters in the two arms could be inefficient. We showed how to assign clusters to the two arms so that investigators can use the smallest amount of resources to obtain the most accurate estimate of the difference in response rates between the two arms.

Suppose now investigators want to attain a specific power to detect a given difference in response rates between the two arms in a CRT and control type 1 error rate as well. How many cluster are needed in the trial and how should they be distributed between the two arms? In this chapter, we derive the sample size formulas to satisfy the fixed type 1 error rate α and power $1-\beta$ requirements. The usually used sample size formula in a CRT with binary outcomes is given in equation (5.5) in Donner and Klar [4]. It says in each arm, the number of individuals needed is:

$$n = \frac{(Z_{\alpha/2} + Z_{\beta})^2(\pi_1(1-\pi) + \pi_2(1-\pi_2))(1+(m-1)\rho)}{m(\pi_1 - \pi_2)^2}.$$

However, that formula assumes a balanced design and RD is the outcome measure. In this chapter, we derive the sample size formulas for RR and OR as well.

We further show that, although a balanced design in a CRT can satisfy the power and type 1 rate requirements, the total cost of the balanced design is not always minimal. Accordingly, we incorporate the optimal allocation w_x^{c*} and derive modified formulas for the numbers of clusters in each arm of a CRT for different measures. Although the total clusters or numbers of individuals by these sample size formulas are not minimal, the total cost in the study is guaranteed to be minimal.

We next consider non-inferiority CRT study. Non-inferiority trials are often used in medical studies. The hypothesis of interest in non-inferiority trials is that the new treatment is at worst inferior to the standard treatment (control) by a defined margin. There is a lot of research on design problems for non-inferiority IRTs, but few discuss non-inferiority CRTs. We extend the results of a non-inferiority IRT to a non-inferiority balanced CRT study. Further, we extend the results to equivalence CRTs.

4.2 Sample size for a balanced CRT

We consider sample size calculation for a two-arm balanced CRT with a common cluster size. Let X_{hij} denote the binary response of the jth individual in the ith cluster in the hth arm, where $j = 1, ..., m$, $i = k_1, k_2$ and $h = 1, 2$. The success rate in the hth arm is π_h and its ICC is ρ_h, $h = 1, 2$.

The estimator of π_{hi} in the ith cluster in the hth arm is $\hat{\pi}_{hi} = \frac{\sum_{j=1}^{m} X_{hij}}{m}$, where m is the cluster size. The unbiased estimator of π_h is $\hat{\pi}_h = \frac{\sum_{i=1}^{k_h} \hat{\pi}_{hi}}{k_h} = \frac{\sum_{i=1}^{k_h} \sum_{j=1}^{m} X_{hij}}{k_h m}$, where k_h is the number of clusters in the hth arm. It is easily shown that the variance of $\hat{\pi}_h$ is $Var(\hat{\pi}_h) = \pi_h(1-\pi_h)\frac{1+(m-1)\rho_h}{k_h m}$.

We introduced the common correlation model in Section 2.2. In that model, responses from different clusters are independent. Hence direct calculation shows that the variance of $\hat{\pi}_1$-$\hat{\pi}_2$ is:

$$Var(\hat{\pi}_1 - \hat{\pi}_2) = \pi_1(1-\pi_1)\frac{1+(m-1)\rho_1}{k_1 m} + \pi_2(1-\pi_2)\frac{1+(m-1)\rho_2}{k_2 m} \qquad (4.1)$$

We consider the null hypothesis $H_0 : \pi_1 - \pi_2 = 0$ and the alternative hypothesis

$H_1 : \pi_1 - \pi_2 \neq 0$. Our test statistic is $\frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{Var(\hat{\pi}_1 - \hat{\pi}_2)}}$. From the central limit theorem, this statistic approximately follows a normal distribution. Hence we reject the null hypothesis at the α level of significance if $\left| \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{Var(\hat{\pi}_1 - \hat{\pi}_2)}} \right| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution.

Let C be the critical cut-off point. Under $H_0 : \pi_1 - \pi_2 = 0$, we have

$$Pr(|\hat{\pi}_1 - \hat{\pi}_2| > C) = \alpha \Rightarrow Pr(\hat{\pi}_1 - \hat{\pi}_2 > C) = \alpha/2$$

$$\Rightarrow Pr\left(\frac{(\hat{\pi}_1 - \hat{\pi}_2) - 0}{\sqrt{Var(\hat{\pi}_1 - \hat{\pi}_2)}} > \frac{C - 0}{\sqrt{Var(\hat{\pi}_1 - \hat{\pi}_2)}} \right) = \alpha/2$$

$$\Rightarrow C = Z_{1-\alpha/2} \sqrt{Var(\hat{\pi}_1 - \hat{\pi}_2)}.$$

Under the alternative hypothesis $H_1 : \pi_1 - \pi_2 = \epsilon \neq 0$, we will have power equal to $1 - \beta$ if

$$Pr(|(\hat{\pi}_1 - \hat{\pi}_2) - \epsilon| > C) = 1 - \beta \Rightarrow Pr((\hat{\pi}_1 - \hat{\pi}_2) - \epsilon > C) \approx 1 - \beta$$

$$\Rightarrow Pr\left(\frac{(\hat{\pi}_1 - \hat{\pi}_2) - \epsilon}{\sqrt{Var(\hat{\pi}_1 - \hat{\pi}_2)}} > \frac{C - \epsilon}{\sqrt{Var(\hat{\pi}_1 - \hat{\pi}_2)}} \right) = 1 - \beta$$

$$\Rightarrow C = Z_\beta \sqrt{Var(\hat{\pi}_1 - \hat{\pi}_2)} + \epsilon.$$

Therefore, in order to achieve power of $1 - \beta$, we set:

$$\frac{|\epsilon|}{\sqrt{\pi_1(1 - \pi_1)\frac{1+(m-1)\rho_1}{k_1 m} + \pi_2(1 - \pi_2)\frac{1+(m-1)\rho_2}{k_2 m}}} = Z_{\alpha/2} + Z_\beta. \tag{4.2}$$

When $m = 1$, this formula reduces to that for an IRT.

In most studies, investigators use a balanced design and assign the same number of clusters to the two arms. Putting $k_1 = k_2 = k_0$ in equation (4.2) and solving the equation for k_0, the number of clusters required in each arm is given by:

$$k_0 = \frac{(Z_{\alpha/2} + Z_\beta)^2 [\pi_1(1 - \pi_1)(1 + (m-1)\rho_1) + \pi_2(1 - \pi_2)(1 + (m-1)\rho_2)]}{(\pi_1 - \pi_2)^2 m} \tag{4.3}$$

If $\rho_1 = \rho_2 = \rho$, we multiply both sides by m, and equation (4.3) is exactly the same with equation (5.5) in Donner and Klar [1]. This equation is the common sample size formula for CRTs design based on the measure RD.

Next, we consider the measure RR. In this case, the null hypothesis is $H_0 : \pi_1/\pi_2 = 1$. This is equivalent to $H_0 : log(\pi_1/\pi_2) = 0$. Since $\hat{\pi}_1/\hat{\pi}_2$ is highly skewed, the log transformation $log(\hat{\pi}_1/\hat{\pi}_2)$ is preferred, which is less skewed than $\hat{\pi}_1/\hat{\pi}_2$ and more likely to be normally distributed.

By using the delta method, the variance of $log(\hat{\pi}_1/\hat{\pi}_2)$ is given by:

$$Var(log(\hat{\pi}_1/\hat{\pi}_2)) = \frac{1-\pi_1}{\pi_1}\frac{1+(m-1)\rho_1}{k_1 m} + \frac{1-\pi_2}{\pi_2}\frac{1+(m-1)\rho_2}{k_2 m}. \tag{4.4}$$

The following procedure to obtain the sample size formula is similar to earlier work. A direct calculation shows the required number of clusters in each arm is:

$$k_0 = \frac{(Z_{\alpha/2}+Z_\beta)^2[\frac{(1-\pi_1)(1+(m-1)\rho_1)}{\pi_1} + \frac{(1-\pi_2)(1+(m-1)\rho_2]}{\pi_2})}{(\log(\pi_1/\pi_2))^2 m}. \tag{4.5}$$

Finally in this section, we consider the measure OR. The corresponding null hypothesis is $H_0 : OR = 1$ where $OR = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$. Like RR, the estimated OR is also highly skewed. We take a log transformation of $\hat{OR}$ and our hypothesis is $H_0 : log(OR) = 0$.

By using the delta method, the variance of $log(\hat{OR})$ is given by:

$$Var(log(\hat{OR}) = \frac{1+(m-1)\rho_1}{\pi_1(1-\pi_1)k_1 m} + \frac{1+(m-1)\rho_2}{\pi_2(1-\pi_2)k_2 m}. \tag{4.6}$$

Hence the number of clusters required in each arm is:

$$k_0 = \frac{(Z_{\alpha/2}+Z_\beta)^2[\frac{1+(m-1)\rho_1}{\pi_1(1-\pi_1)} + \frac{1+(m-1)\rho_2}{\pi_2(1-\pi_2)}]}{(\log(OR))^2 m}. \tag{4.7}$$

4.3 Optimal sample size

In the previous section, we assumed each arm had the same number of clusters and we derived the sample size formulas for different measures in a two-arm CRT. However, when the unit costs are very different in the two arms, assigning equal number of clusters

in the two arms may not be cost efficient. Hence although a balanced design can satisfy the power and type 1 rate requirements in a study, it may cost more. Investigators usually want to use minimal resources to obtain the desired power. Hence the balanced design while convenient may not be the preferred choice.

Our optimal sample size by definition is the cheapest sample size in terms of cost among all those that meet our power and type 1 error rate requirements. In this section, we will derive the optimal sample size formulas which meet pre-selected power $1-\beta$ and type 1 rate α requirements for all of the three measures.

First we consider the measure RD. We wish to test the null hypothesis $H_0 : \pi_1 - \pi_2 = 0$. From equation (4.2), we know that for any combination of k_1 and k_2, as long as they satisfy the equation, they satisfy the power and type 1 rate requirements. A balanced design, in which $k_1 = k_2$, is a special case. Suppose that the number of clusters in the arm 1 is s times that in the arm 2. We plug $k_1 = sk_2$ in equation (4.2) and solve this equation for k_1 and obtain:

$$k_1 = \frac{(Z_{\alpha/2} + Z_\beta)^2[\pi_1(1-\pi_1)(1+(m-1)\rho_1) + s\pi_2(1-\pi_2)(1+(m-1)\rho_2)]}{(\pi_1 - \pi_2)^2 m}. \quad (4.8)$$

If π_1, π_2, ρ_1 and ρ_2 are known, for any s, we can obtain corresponding k_1 through the above equation for any s. All combinations of s and k_1 obtained from the above equation satisfy power and type 1 error requirements. The value of k_2 is determined from s and k_1 through the equation $k_2 = k_1/s$. The total cost is expressed by the cost function (2.11), $k_1 m c_1 + k_2 m c_2 = k_1 m(c_1 + c_2/s)$, and different combinations of s and k_1 result in different total costs. In all combinations of s and k_1, there is one pair of s and k_1 that minimizes the total cost.

Recall that in Section 2.3, we had the same cost function and when the total number of clusters k is fixed, the optimal allocation w_{RD}^{c*} in equation (2.15) maximizes the precision of estimation per cost unit. This in turns implies that the power for the same total cost is maximized. In other words, the total cost for the optimal allocation w_{RD}^{c*} is minimized in all different w that satisfy the power requirement.

Since $w = k_1/(k_1 + k_2)$ and $s = k_1/k_2$, we have $s = w/(1-w)$. Hence k_1 which

is associated with $s^* = w_{RD}^{c*}/(1 - w_{RD}^{c*})$ is the optimal sample size we need. Plugging $s^* = w_{RD}^{c*}/(1 - w_{RD}^{c*})$ in equation (4.8) and solving for k_1 gives the required number of clusters in arm 1:

$$k_1 = \frac{w_{RD}^{c*}(Z_{\alpha/2} + Z_\beta)^2[\frac{\pi_1(1-\pi_1)(1+(m-1)\rho_1)}{w_{RD}^{c*}} + \frac{\pi_2(1-\pi_2)(1+(m-1)\rho_2)}{1-w_{RD}^{c*}}]}{(\pi_1 - \pi_2)^2 m}. \tag{4.9}$$

The corresponding number of clusters needed in arm 2 is:

$$k_2 = \frac{(1 - w_{RD}^{c*})(Z_{\alpha/2} + Z_\beta)^2[\frac{\pi_1(1-\pi_1)(1+(m-1)\rho_1)}{w_{RD}^{c*}} + \frac{\pi_2(1-\pi_2)(1+(m-1)\rho_2)}{1-w_{RD}^{c*}}]}{(\pi_1 - \pi_2)^2 m}. \tag{4.10}$$

Let us summarize the steps to obtain the optimal sample size for the measure RD.

Step 1: Determine the parameters π_1, π_2, ρ_1, ρ_2, cost ratio γ and cluster size m.

Step 2: Calculate the optimal allocation w_{RD}^{c*} from equation (2.15).

Step 3: Determine the power and type 1 error requirements, and calculate k_1 and k_2 from equations (4.9) and (4.10).

The steps are similar when the measure is RR or OR. In step 2, w_{RR}^{c*} or w_{OR}^{c*} is used instead. In step 3, equations (4.11) and (4.12) or equations (4.13) and (4.14) are used.

The total number of clusters needed in a study calculated by the optimal sample size formulas is usually larger than the total number of clusters needed for a balanced design, unless the cost per individual in the two arms are the same. However, the total cost for the former is usually minimal. For example, Table 4.1 shows the numbers of clusters, individuals and total cost for estimating RD in a two-arm CRT. In this CRT, π_1=0.1, π_2=0.3, ρ_1=ρ_2=0.1 and the cluster size is $m = 20$. We assume the unit cost is 5 in the arm 1 and 1 in the arm 2. The desired power is 0.8 and type 1 error rate is 0.05. The first row in Table 4.1 has different allocation values of $w = k_1/(k_1 + k_2)$, the proportion of clusters assigned to the arm 1. The last value of 0.23 is the optimal allocation value. The fifth and the sixth rows show the numbers of total individuals needed and the total cost in the study. We observe that the optimal design with $w = 0.23$ has a total cost 800. This cost is the minimal among all different values of w. The total numbers of clusters and individuals for optimal allocation are 20 and 400 respectively, versus 18 and 360, for the balanced design (with $w = 0.5$).

Table 4.1: Cost differentials for estimating RD from different allocation schemes versus optimal ($w_{RD}^{c*} = 0.23$) when CRT can have different number of clusters ($\pi_1 = 0.1, \pi_2 = 0.3, m = 20, \gamma = 5, \rho_1 = \rho_2 = 0.1, 1 - \beta = 0.8$).

w	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	$0.23(w_{RD}^{c*})$
k_1	4	5	6	7	9	12	17	27	57	5
k_2	29	17	12	10	9	8	8	7	7	15
k	33	22	18	17	18	20	25	34	64	20
N	660	440	360	340	360	400	500	680	1280	400
Cost	980	840	840	900	1080	1360	1860	2840	5840	800

For the measure RR, the null hypothesis is $H_0 : log(\pi_1/\pi_2) = 0$. The procedure to derive the optimal sample size formula for estimating RR is the same with that for estimating RD. Hence we omit the details. A direct calculation shows the number of clusters for the measure RR in arm 1 is:

$$k_1 = \frac{w_{RR}^{c*}(Z_{\alpha/2} + Z_\beta)^2[\frac{(1-\pi_1)(1+(m-1)\rho_1)}{\pi_1 w_{RR}^{c*}} + \frac{(1-\pi_2)(1+(m-1)\rho_2)}{\pi_2(1-w_{RR}^{c*})}]}{(\log(\pi_1/\pi_2))^2 m}. \tag{4.11}$$

The corresponding number of clusters needed in arm 2 is:

$$k_2 = \frac{(1 - w_{RR}^{c*})(Z_{\alpha/2} + Z_\beta)^2[\frac{(1-\pi_1)(1+(m-1)\rho_1)}{\pi_1 w_{RR}^{c*}} + \frac{(1-\pi_2)(1+(m-1)\rho_2)}{\pi_2(1-w_{RR}^{c*})}]}{(\log(\pi_1/\pi_2))^2 m}. \tag{4.12}$$

Table 4.2 shows the numbers of clusters, individuals and total cost for estimating RR in a two-arm CRT. Note that the total cost for the optimal size is not the minimal. When $w = 0.4$, $k_1 = 10$ and $k_2 = 15$, the total cost is minimal. The reason is that for different values of w, the calculated values of k_1 and k_2 are not integers; but k_1 and k_2 have to be integers. Hence the calculated values of k_1 and k_2 are rounded to integers which are at least as large as the original values of k_1 and k_2.

For the measure OR, the null hypothesis is $H_0 : log(OR) = 0$. The procedure to derive the optimal sample size formula for estimating OR is the same with that for estimating RD and RR. Again we omit the details. A direct calculation shows the

Table 4.2: Cost differentials for estimating RR from different allocation schemes versus optimal ($w_{RR}^{c*} = 0.47$) when CRT can have different number of clusters ($\pi_1 = 0.1, \pi_2 = 0.3, m = 20, \gamma = 5, \rho_1 = \rho_2 = 0.1, 1 - \beta = 0.8$).

w	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	$0.47(w_{RR}^{c*})$
k_1	9	10	10	10	11	12	14	18	29	11
k_2	79	37	22	15	11	8	6	5	4	12
k	88	47	32	25	22	20	20	23	33	23
N	1760	940	640	500	440	400	400	460	660	460
Cost	2480	1740	1440	1300	1320	1360	1520	1900	2980	1340

Table 4.3: Cost differentials for estimating OR from different allocation schemes versus optimal ($w_{OR}^{c*} = 0.41$) when CRT can have different number of clusters ($\pi_1 = 0.1, \pi_2 = 0.3, m = 20, \gamma = 5, \rho_1 = \rho_2 = 0.1, 1 - \beta = 0.8$).

w	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	$0.41(w_{OR}^{c*})$
k_1	8	8	9	9	10	12	14	19	34	9
k_2	66	31	20	14	10	8	6	5	4	13
k	74	39	29	23	20	20	20	24	38	22
N	1480	780	580	460	400	400	400	480	760	440
Cost	2120	1420	1300	1180	1200	1360	1520	2000	3480	1160

optimal number of clusters for the measure OR in arm 1 is:

$$k_1 = \frac{w_{OR}^{c*}(Z_{\alpha/2} + Z_\beta)^2[\frac{1+(m-1)\rho_1}{\pi_1(1-\pi_1)w_{OR}^{c*}} + \frac{1+(m-1)\rho_2}{\pi_2(1-\pi_2)(1-w_{OR}^{c*})}]}{(\log(OR))^2 m}. \tag{4.13}$$

The corresponding number of clusters in arm 2:

$$k_2 = \frac{(1-w_{OR}^{c*})(Z_{\alpha/2} + Z_\beta)^2[\frac{1+(m-1)\rho_1}{\pi_1(1-\pi_1)w_{OR}^{c*}} + \frac{1+(m-1)\rho_2}{\pi_2(1-\pi_2)(1-w_{OR}^{c*})}]}{(\log(OR))^2 m}. \tag{4.14}$$

Table 4.3 shows the numbers of clusters, individuals and total cost for estimating OR in a two-arm CRT. The optimal sample size is $k_1 = 9$ and $k_2 = 13$ and the total cost is

1160. From Table 4.1, 4.2 and 4.3, we observe different measures have different optimal sample sizes and their corresponding total costs are different too. If investigators have no specific reason to choose a particular measure, they may consider choosing the measure with the minimal total cost for its optimal sample size. For the example discussed in this section, the minimal total cost is 800 for estimating RD, hence investigators may design the trial using the measure RD.

4.4 Sample size for a non-inferiority trial

In previous sections we consider the problems of sample size calculation to detect the difference between two binary responses in the two arms in a CRT. The aim for the design is to determine whether responses rate from a new intervention is different from that in the control intervention.

Sometimes, investigators are interested in whether a new intervention is nearly as effective as the standard intervention. For this purpose, what desired is a non-inferiority trial. The non-inferiority trial seeks to determine whether a new intervention is not worse than a reference intervention by more than an acceptable amount, which is also called the non-inferiority margin. The selection of the non-inferiority margin is key to design the non-inferiority trial. However, we won't discuss how to choose this margin here. Throughout, we assume that the non-inferiority margin has been determined. The new intervention is potentially less toxic, less costly, or easier to administer than a standard intervention and hence may be preferred. There are many papers to discuss non-inferiority trials, for example, see D'Agostino et al. [34] and Ch16 in Crowley [35]. Hilton [36] discussed the non-inferiority trial designs for RD and OR.

Almost all research on non-inferiority trials assumes IRTs. One exception is Dixon et. al. [37] who reported a non-inferiority CRT. That trial was to investigate whether triage nurses in the emergency department can safely reduce radial-head subluxation at rates that are not substantially lower than those of emergency department physicians. However, although it was a CRT, the authors did not include variance inflation factor

in the sample size calculation. More work is needed for non-inferiority CRTs. In this section, we will derive the sample size formula for non-inferiority CRTs assuming the measure is RD.

Let $\epsilon = \pi_1 - \pi_2$ be predetermined. We wish to test the null and alternative hypotheses: $H_0 : \epsilon \leq \delta$ and $H_0 : \epsilon > \delta$, where $\delta < 0$ is the non-inferiority margin. The variance of $(\pi_1 - \pi_2 - \delta)$ is the identical to the variance of $(\pi_1 - \pi_2)$, as shown in equation (4.1). We reject the null hypothesis at the α level of significance if $\frac{\hat{\pi}_1 - \hat{\pi}_2 - \delta}{\sqrt{Var(\hat{\pi}_1 - \hat{\pi}_2)}} > Z_\alpha$ where Z_α is the $100(1-\alpha)$ percentile of the standard normal distribution.

Under the alternative hypothesis that $\epsilon > \delta$, the power of the above test is approximately

$$\Phi\left(\frac{\epsilon - \delta}{\sqrt{\pi_1(1-\pi_1)\frac{1+(m-1)\rho_1}{k_1 m} + \pi_2(1-\pi_2)\frac{1+(m-1)\rho_2}{k_2 m}}} - Z_\alpha\right). \tag{4.15}$$

Therefore, in order to achieve the power $1 - \beta$, we have the equation:

$$\frac{\epsilon - \delta}{\sqrt{\pi_1(1-\pi_1)\frac{1+(m-1)\rho_1}{k_1 m} + \pi_2(1-\pi_2)\frac{1+(m-1)\rho_2}{k_2 m}}} = Z_\alpha + Z_\beta. \tag{4.16}$$

Suppose that the number of clusters in the arm 1 is as s times as that in the arm 2, $k_1 = sk_2$. Note $s = 1$ $(k_1 = k_2)$ is a special case, corresponding to a balanced design. Plugging $k_1 = sk_2$ in equation (4.16), we solve the equation for k_1:

$$k_1 = \frac{(Z_\alpha + Z_\beta)^2[\pi_1(1-\pi_1)(1+(m-1)\rho_1) + s\pi_2(1-\pi_2)(1+(m-1)\rho_2)]}{(\pi_1 - \pi_2 - \delta)^2 m}. \tag{4.17}$$

Since $k_1 = sk_2$, the number of clusters in the arm 2 is:

$$k_2 = \frac{(Z_\alpha + Z_\beta)^2[\frac{1}{s}\pi_1(1-\pi_1)(1+(m-1)\rho_1) + \pi_2(1-\pi_2)(1+(m-1)\rho_2)]}{(\pi_1 - \pi_2 - \delta)^2 m}. \tag{4.18}$$

For the balanced design, in which $k_1 = k_2$, we have the number of clusters in each arm is:

$$k_0 = \frac{(Z_\alpha + Z_\beta)^2[\pi_1(1-\pi_1)(1+(m-1)\rho_1) + \pi_2(1-\pi_2)(1+(m-1)\rho_2)]}{(\pi_1 - \pi_2 - \delta)^2 m}. \tag{4.19}$$

We consider the optimal sample size calculation for a non-inferiority trial. From the last section, we know that in order to obtain the desired power we need satisfy this equation $s* = w^*_{RD}/(1 - w^*_{RD})$, where w^*_{RD} is the optimal allocation. However, the optimal allocation w^*_{RD} for a non-inferiority trial may not be exactly the value given in equation (2.15). This is because in a non-inferiority trial, the success rate π_1 in the intervention arm is not a fixed value, but in a range.

4.5 Sample size for an equivalence trial

In this section, we extend the sample size formula for non-inferiority CRTs to equivalence CRTs. Equivalence trials aim to determine whether the new intervention is similar to the standard existing treatment.

Let $\epsilon = \pi_1 - \pi_2$. We wish to test the null and alternative hypotheses: $H_0 : |\epsilon| \geq \delta$ and $H_0 : |\epsilon| < \delta$.

We reject the null hypothesis at the α level of significance if $\frac{\hat{\pi}_1 - \hat{\pi}_2 - \delta}{\sqrt{\hat{Var}(\hat{\pi}_1 - \hat{\pi}_2)}} < -Z_\alpha$ and $\frac{\hat{\pi}_1 - \hat{\pi}_2 + \delta}{\sqrt{\hat{Var}(\hat{\pi}_1 - \hat{\pi}_2)}} > Z_\alpha$.

Under the alternative hypothesis that $|\epsilon| < \delta$, the power of the above test is approximately

$$2\Phi\left(\frac{\delta - |\epsilon|}{\sqrt{\pi_1(1-\pi_1)\frac{1+(m-1)\rho_1}{k_1 m} + \pi_2(1-\pi_2)\frac{1+(m-1)\rho_2}{k_2 m}}} - Z_\alpha\right) - 1. \tag{4.20}$$

Therefore, in order to achieve the power $1 - \beta$, we have the equation:

$$\frac{\delta - |\epsilon|}{\sqrt{\pi_1(1-\pi_1)\frac{1+(m-1)\rho_1}{k_1 m} + \pi_2(1-\pi_2)\frac{1+(m-1)\rho_2}{k_2 m}}} = Z_\alpha + Z_{\beta/2} \tag{4.21}$$

where Z_β is the $100(1-\beta)$ percentile of the standard normal distribution.

Suppose that the number of clusters in the arm 1 is as s times as that in the arm 2, $k_1 = sk_2$. Note $s = 1$ ($k_1 = k_2$) is a special case, corresponding to a balanced design. Plugging $k_1 = sk_2$ in equation (4.21), we solve the equation for k_1:

$$k_1 = \frac{(Z_\alpha + Z_{\beta/2})^2[\pi_1(1-\pi_1)(1+(m-1)\rho_1) + s\pi_2(1-\pi_2)(1+(m-1)\rho_2)]}{(\delta - |\pi_1 - \pi_2|)^2 m}. \tag{4.22}$$

Since $k_1 = sk_2$, the number of clusters in the arm 2 is:

$$k_2 = \frac{(Z_\alpha + Z_{\beta/2})^2[\frac{1}{s}\pi_1(1-\pi_1)(1+(m-1)\rho_1) + \pi_2(1-\pi_2)(1+(m-1)\rho_2)]}{(\delta - |\pi_1 - \pi_2|)^2 m}. \quad (4.23)$$

For the balanced design, in which $k_1 = k_2$, we have the number of clusters in each arm is:

$$k_0 = \frac{(Z_\alpha + Z_{\beta/2})^2[\pi_1(1-\pi_1)(1+(m-1)\rho_1) + \pi_2(1-\pi_2)(1+(m-1)\rho_2)]}{(\delta - |\pi_1 - \pi_2|)^2 m}. \quad (4.24)$$

For the optimal sample size calculation, we need satisfy this equation $s^* = w_{RD}^{c*}/(1 - w_{RD}^{c*})$. However, the optimal allocation w_{RD}^{c*} may not exactly be the same as in equation (2.15), since π_1 in the intervention arm is not assumed be a fixed value.

4.6 Chapter summary and discussion

In this chapter, we consider sample size calculation problem for a CRT. The common used sample size formula assumes the measure is RD and ICCs in the two arms are the same. In this chapter, we first derive the sample size formula including arm-specific ρ_1 and ρ_2. We also derive the sample size formulas for RR and OR.

Investigators often use the balanced design. We define the optimal sample size and derive the optimal sample size formulas for all three measures. In all the designs satisfying the power requirement, the total study cost for the optimal sample size is minimal. We show the results in Tabes 4.1, 4.2 and 4.3. However, the total number of clusters and individuals are not minimal in optimal sample size. When the number of clusters is large, investigators do not worry about the sample is not enough but more care about the cost in the study, hence they may prefer to use optimal sample size. However, if the number of clusters is very limited, investigators may use the balanced design, and the price is increasing the total study cost.

We also consider non-inferiority CRTs and equivalence CRTs. The former is to determine whether a new intervention is not worse than a standard intervention by more than an acceptable amount, and the latter is to determine whether the new intervention

is similar to the standard existing treatment. Very few literature discuss this topic. Dixon et al. [37] reported a non-inferiority CRT but they did not consider correlation in a cluster when they calculated the sample size. We derive sample size formulas for non-inferiority CRTs and equivalence CRTs respectively. We only consider the measure RD because the hypothesis to define the non-inferiority and equivalence uses RD scale. Also, we only give explicit sample size formulas for the balanced design. More work is needed for the optimal sample size in non-inferiority CRTs and equivalence CRTs.

In this chapter, we assume that cluster size is a constant. We will deal with a CRT with varying cluster size in Chapter 5.

CHAPTER 5

Optimal allocation and sample size when cluster size varies

5.1 Introduction

In Chapter 2, we investigated efficient design of a two-arm CRT with a fixed total number of clusters. We showed how to assign clusters to the two arms to obtain the most precise estimate of the difference in response rates between the two arms while minimizing the study cost. In Chapter 4, we derived the sample size formulas for optimal allocation for a two-arm CRT to achieve specific type 1 error rate and power requirements. In those two chapters, we assumed that cluster size was constant.

Constant cluster size can occur naturally, when couples (constant cluster size of 2) are randomized to conditions, for example. Sometimes, the investigator can control cluster size, even though the actual size of clusters is varying. For example, investigators may recruit a fixed number of individuals from each cluster, which has been termed *sample size per cluster*; see Hayes [2]. In this case, we can regard sample size per cluster as cluster size and results of Chapter 2 and 4 can be applied directly. More commonly, CRTs exhibit variation in cluster size. The variation may reflect natural variation in the actual size of the clusters, or variation in recruitment rates or loss to follow-up rates among equal-sized clusters. For example, in the Samoan women's health study reported by Mishra [23], Samoan churches in southern California were randomized to intervention and control arms. Cluster sizes ranged from 2-42 with a mean of 14. Another example is a CRT reported by Bastani [38] and Glenn [39]. In that trial, colorectal cancer cases were identified through the California Cancer Registry and relatives within the same

family composed clusters, which were randomized to intervention or control arms. The cluster sizes ranged from 1-7 with a mean of 1.6.

Guittet et al. [13] investigated the impact of unequal cluster size on power in CRTs with continuous outcomes. In their simulation studies, beside the constant cluster size, three types of dispersion were considered: a moderate imbalance, a Pareto imbalance and a Poisson imbalance. In the moderate imbalance, an individual has the same probability to be assigned to any cluster. Therefore, all clusters have the same expected cluster size. In Pareto imbalance, within each treatment arm, two strata were defined, larger clusters (20% of all clusters) and smaller clusters (80% of all clusters). 80% individuals were assigned to the larger cluster stratum and 20% were assigned to the smaller cluster stratum. Within each stratum, an individual has the same probability to be assigned to any cluster. Hence in each stratum, all clusters have the same expected cluster size. But the expected cluster sizes in the two strata are very different, with expected cluster size in the larger cluster stratum 16 times higher than that in the smaller cluster stratum. Hence they regarded Pareto imbalance as severe dispersion. In Poisson imbalance, they let cluster size follow a Poisson distribution. They compared the power of these three distributions to a constant cluster size distribution. The power is slightly reduced in the moderate imbalance and Poisson imbalance, but is greatly reduced in Pareto imbalance. They concluded that varying cluster size greatly reduces power in the case of severe dispersion, particularly if the number of clusters is low and/or the ICC is high.

Ahn et al. [14] investigated the effect of cluster size variability on the power of CRTs with a binary outcome. Cluster sizes were generated using a negative binomial distribution truncated below 1. They defined an *imbalance parameter* $\kappa = 1/(1+\sigma^2/\mu^2)$, where σ and μ are the standard deviation and mean of cluster size. The degree of imbalance can be measured by the imbalance parameter. When $\kappa = 1$, all clusters have the same cluster size. As κ decreases, the imbalance of cluster size increases. They considered the situations when $\kappa = 1$, $\kappa = 0.8$ and $\kappa = 0.6$. Their simulation studies showed that empirical power levels are not close to the nominal power when there is severe ($\kappa = 0.6$) imbalance in cluster size and the number of clusters is as small as

10. Note that coefficient of variation (CV) of cluster size is a function of the imbalance parameter κ: $CV = \sqrt{\frac{1-\kappa}{\kappa}}$. When $\kappa = 0.6$ and $\kappa = 0.8$, then we have $CV = 0.82$ and $CV = 0.5$, respectively.

Although severe imbalance in cluster size affects power, especially when number of clusters is small and ICC is large, investigators seldom account for varying cluster size in the trial design stage. Rather, they typically use average cluster size in formulas intended for use in trials with equal size clusters. Only a few authors, such as Kerry and Bland [15], Mantunga and Hudgen [16] and Guittet et al.[13], discuss sample size determination in the case of unequal cluster sizes. Fewer authors deal with a binary outcome rather than a continuous outcome.

In this chapter, we extend the results obtained in Chapters 2 and 4 to the situation where a CRT has unequal cluster sizes. In the first part of this chapter, we fix the total number of clusters and investigate the optimal design problem of how to allocate clusters to each arm to achieve the most precise estimate with minimal cost. In Chapter 2, the cluster size is assumed equal to m. Recall that w is the fraction of clusters assigned to arm 1. Since all clusters have the same number of individuals, w is also the fraction of individuals assigned to the arm 1. For measure $x, x \in (RD, RR, OR)$, optimal allocation w_x^{c*} is a function of m, the fixed cluster size. However, when cluster size is varying, this does not apply. Therefore, we need further consideration of optimal allocation when cluster size is varying. We address the optimal allocation problem in the first part of this chapter.

In Chapter 4, we derived optimal sample size formulas for CRTs with constant cluster size. In the second part of this chapter, we extend the results to CRTs with varying cluster size. The usual sample size calculation approach is to assign the same number of clusters to each arm and ignore the cluster size variation. However, when cluster size variation is large, the desired power may not be achieved, and equal allocation of clusters may not be efficient when costs on the individual level and the cluster level are very different in the two arms. Hence we first derive sample size formula for equal allocation of clusters, considering cluster size variation, to satisfy the desired type 1 error

rate and power requirements. Then as in Chapter 4, we derive a optimal sample size formula, which satisfies the power requirement but guarantees that total cost is minimal. For optimal sample size, the number of clusters in the two arms are usually unequal.

In this chapter, we confine attention to estimating RD. Results for RR and OR can be derived in the same manner.

5.2 Weighted estimation of proportions

The sample size requirement will be based on the estimator of the risk difference $\pi_1 - \pi_2$. Hence we consider the problem of estimating the success rate π_h in the hth arm of a CRT in this section.

Our approach is based on the common correlation model, which was described in Chapter 2 and has been discussed in the literature, for example, see Crespi et al. [40]. Recall X_{hij} denotes the response of the jth individual in the ith cluster in the hth arm. The success rate in the ith cluster in the hth arm is π_{hi}. The estimator of π_{hi} is $\hat{\pi}_{hi}=\frac{\sum_{j=1}^{m_{hi}} X_{hij}}{m_{hi}}$, where m_{hi} is the cluster size of the ith cluster in the hth arm.

We want to estimate the success rate π_h in the hth arm. Several different estimators for π_h could be considered. One estimator using π_{hi} is the simple average of the proportions over clusters, $\hat{\pi}_h^C = \frac{\sum_{i=1}^{k_h} \hat{\pi}_{hi}}{k_h}$, where k_h is the number of clusters in the hth arm. Under the common correlation model, this estimator is unbiased:

$$E(\hat{\pi}_h^C) = \frac{\sum_{i=1}^{k_h} E(\hat{\pi}_{hi})}{k_h} = \frac{k_h E(\hat{\pi}_{hi})}{k_h} = \pi_h.$$

Another estimator of π_h is $\hat{\pi}_h^I = \frac{\sum_{i=1}^{k_h}\sum_{j=1}^{m_{hi}} X_{hij}}{\sum_{i=1}^{k_h} m_{hi}}$, which averages over individuals; see Donner and Klar [41], for example. This estimator is also unbiased:

$$E(\hat{\pi}_h^I) = \frac{\sum_{i=1}^{k_h}\sum_{j=1}^{m_{hi}} E(X_{hij})}{\sum_{i=1}^{k_h} m_{hi}} = \frac{\sum_{i=1}^{k_h} m_{hi} E(X_{hij})}{\sum_{i=1}^{k_h} m_{hi}} = \pi_h.$$

When cluster size is the same for all clusters ($m_{hi} = m$), the two estimators are identical. However, when cluster size is varying, the two estimators can yield different estimates.

To generalize, we can express these estimators using a weighting scheme with weight for each cluster. Let b_{hi} be the weight assigned to ith cluster in the hth arm. Then the estimators can be expressed as:

$$\hat{\pi}_h = \sum_{i=1}^{k_h} b_{hi}\hat{\pi}_{hi} \qquad \text{subject} \qquad \text{to:} \qquad \sum_{i=1}^{k_h} b_{hi} = 1. \tag{5.1}$$

Note that the constraint $\sum_{i=1}^{k_h} b_{hi} = 1$ ensures that the weighted estimators are unbiased:

$$E(\hat{\pi}_h) = \sum_{i=1}^{k_h} b_{hi}E(\hat{\pi}_{hi}) = \pi_h \sum_{i=1}^{k_h} b_{hi} = \pi_h.$$

For the estimator $\hat{\pi}_h^C = \frac{\sum_{i=1}^{k_h}\hat{\pi}_{hi}}{k_h}$, the weight is:

$$b_{hi}^C = \frac{1}{k_h} \tag{5.2}$$

We refer to these weights as *cluster weights*. They give the same weight to each cluster regardless of cluster size. Hence these weights give more weight to individuals in smaller clusters than to individuals in larger clusters.

For the estimator $\hat{\pi}_h^I = \frac{\sum_{i=1}^{k_h}\sum_{j=1}^{m_{hi}} X_{hij}}{\sum_{i=1}^{k_h} m_{hi}}$, the weight is:

$$b_{hi}^I = \frac{m_{hi}}{\sum_{i=1}^{k_h} m_{hi}}. \tag{5.3}$$

We refer to these weights as *individual weights*. They give weight to a cluster according to its cluster size, i.e., the number of individuals in the cluster. Hence these weights give more weight to larger clusters.

Kerry and Bland [15] and Jung et al. [17] proposed another weighting scheme:

$$b_{hi}^{MV} = \frac{m_{hi}(1+(m_{hi}-1)\rho_h)^{-1}}{\sum_{i=1}^{k_h} m_{hi}(1+(m_{hi}-1)\rho_h)^{-1}} \tag{5.4}$$

"MV" stands for minimal variance, as we will explain. This weight depends on ρ_h, the value of the ICC in the hth arm. When $\rho_h = 1$, this weight reduces to cluster weight, and when $\rho_h = 0$, this weight reduces to individual weight. When $0 < \rho_h < 1$, the value of this weight is intermediate between the cluster weight and individual weight. We denote the corresponding estimator of success rate in the hth arm as $\hat{\pi}_h^{MV}$.

Note that the variance of $\hat{\pi}_{hi}$ is $\frac{1}{m_{hi}}\pi_h(1-\pi_h)[1+(m_{hi}-1)\rho_h]$. From equation (5.1), direct calculation shows that the variance of $\hat{\pi}_h^{MV}$ is:

$$Var(\hat{\pi}_h^{MV}) = \sum_{i=1}^{k_h} b_{hi}^{MV2} \frac{1}{m_{hi}} \pi_h(1-\pi_h)[1+(m_{hi}-1)\rho_h]. \tag{5.5}$$

If the number of clusters in the hth arm k_h is fixed, then for all weights satisfying $\sum_{i=1}^{k_h} b_{hi} = 1$, the weights in equation (5.4) make the variance of $\hat{\pi}_h^{MV}$ minimal. Therefore, we call b_{hi}^{MV} *minimal variance weight*.

All of these weight schemes have been used in data analysis and sample size calculation. For example, with respect to data analysis, cluster weights are used in Lee and Dubin [42], individual weights are used in Rao and Scott [43], and minimal variance weights are discussed in Jung and Ahn [44]; with respect to sample size calculation, these weights schemes are used in Kerry et al.[15] and Guittet et al.[13], although Kerry and Guittet used other terminology for these weight schemes.

5.3 Optimal allocation when the total number of clusters is fixed but cluster size varies

In Chapter 2, we derived results for the optimal allocation w_x^{c*} for different measures x. In that chapter, we assumed a constant cluster size m. Now we consider the situation in which the cluster size is varying.

An unbiased estimator of the risk difference $RD = \pi_1 - \pi_2$ is

$$\hat{RD} = \hat{\pi}_1 - \hat{\pi}_2 = \sum_{i=1}^{k_1} b_{1i}\hat{\pi}_{1i} - \sum_{i=1}^{k_2} b_{2i}\hat{\pi}_{2i}, \tag{5.6}$$

where the weights b_{hi} are as defined in the last section. The variance of $\hat{\pi}_h$ is $\sum_{i=1}^{k_h} b_{hi}^2 \frac{1}{m_{hi}}\pi_h(1-\pi_h)[1+(m_{hi}-1)\rho_h]$. The variance of $\hat{RD}$ is the sum of the variances of $\hat{\pi}_1$ and $\hat{\pi}_2$, given by:

$$\Psi_{RD}^{-1} = \pi_1(1-\pi_1)\sum_{i=1}^{k_1} b_{1i}^2 \frac{1+(m_{1i}-1)\rho_1}{m_{1i}} + \pi_2(1-\pi_2)\sum_{i=1}^{k_2} b_{2i}^2 \frac{1+(m_{2i}-1)\rho_2}{m_{2i}}. \tag{5.7}$$

As in Chapter 2, the objective for the optimal design is to allocate a total fixed number of clusters k to each arm so as to minimize Ψ_{RD}^{-1}, the variance of $\hat{RD}$. In Chapter 2, the cluster size is a constant m, and Ψ_{RD}^{-1} is a function of $w = \frac{k_1}{k_1+k_2}$, the fraction of clusters allocated to arm 1. Hence the optimal allocation problem involved finding the value of w that makes the variance minimal.

When cluster size varies, the situation is more complicated and there are additional considerations. Investigators not only consider how many clusters are assigned to each arm, but also consider which cluster is assigned to each arm. Investigators usually want the distribution of cluster size to be similar in the two arms. If an interaction exists between intervention and cluster size, this strategy can prevent bias in the estimation of the intervention effect. Note that we use w to denote the fraction of clusters assigned to arm 1. If the distribution of cluster size is the same in the two arms, w also stands for the fraction of individuals assigned to arm 1, even though cluster size is varying.

Since $\hat{\pi}_1$ and $\hat{\pi}_2$ may be based on various weighting schemes, we also need to consider which weighting approach to use. The total number of clusters is a constant k. Since our goal is to minimize the variance of $\hat{RD}$, we select the minimal variance weights. To simplify notation, we drop the superscript MV in b_{hi}^{MV} and $\hat{\pi}_h^{MV}$ through out the remainder of this chapter.

Using the minimal variance weights in equation (5.7), we obtain:

$$\Psi_{RD}^{-1} = \pi_1(1-\pi_1)\frac{1}{\sum_{i=1}^{k_1}\frac{m_{1i}}{1+(m_{1i}-1)\rho_1}} + \pi_2(1-\pi_2)\frac{1}{\sum_{i=1}^{k_2}\frac{m_{2i}}{1+(m_{2i}-1)\rho_2}}. \tag{5.8}$$

We now need to determine $k_1 = kw, k_2 = k(1-w)$, which are the numbers of clusters assigned to each arm. Note that equation (5.8) requires values for the size of each cluster, m_{1i}, $i = 1, ..., k_1$ and m_{2i}, $i = 1, ..., k_2$. To proceed, we can specify the cluster size distribution. Suppose cluster size has a discrete distribution with d different cluster sizes: $n_1, n_2, ..., n_d, d \leq k$, and the proportion of clusters with cluster size n_i is f_i with $\sum_{i=1}^{d} f_i = 1$. Hence the numbers of clusters of size $n_1, n_2, ..., n_d$ are $kf_1, kf_2, ..., kf_d$, respectively.

Since we have

$$\sum_{i=1}^{k_1} \frac{m_{1i}}{1+(m_{1i}-1)\rho_1} = kw \sum_{i=1}^{d} \frac{f_i n_i}{1+(n_i-1)\rho_1},$$

$$\sum_{i=1}^{k_2} \frac{m_{1i}}{1+(m_{1i}-1)\rho_2} = k(1-w) \sum_{i=1}^{d} \frac{f_i n_i}{1+(n_i-1)\rho_2},$$

we can rewrite Ψ_{RD}^{-1} (up to a factor $1/k$) as:

$$\Psi_{RD}^{-1} = \pi_1(1-\pi_1)\left[\frac{1}{w} + \frac{1}{(1-w)} \frac{\pi_2(1-\pi_2)\sum_{i=1}^{d} \frac{f_i n_i}{1+(n_i-1)\rho_1}}{\pi_1(1-\pi_1)\sum_{i=1}^{d} \frac{f_i n_i}{1+(n_i-1)\rho_2}}\right]. \tag{5.9}$$

When π_1, π_2, ρ_1, ρ_2, $f_i (i = 1, ..., d)$, and $n_i (i = 1, ..., d)$ are known, the variance is a function with one unknown variable w. Using the same method as in Chapter 2, we obtain the optimal allocation:

$$w_{RD}^{v*} = \frac{\sqrt{\pi_1(1-\pi_1)\sum_{i=1}^{d} \frac{f_i n_i}{1+(n_i-1)\rho_2}}}{\sqrt{\pi_1(1-\pi_1)\sum_{i=1}^{d} \frac{f_i n_i}{1+(n_i-1)\rho_2}} + \sqrt{\pi_2(1-\pi_2)\sum_{i=1}^{d} \frac{f_i n_i}{1+(n_i-1)\rho_1}}} \tag{5.10}$$

We use w_{RD}^{v*} to distinguish this allocation from w_{RD}^{*}. The superscript v indicates varying cluster size. Note that if $\rho_1 = \rho_2$, this equation reduces to $w_{RD}^{v*} = \frac{\sqrt{\pi_1(1-\pi_1)}}{\sqrt{\pi_1(1-\pi_1)}+\sqrt{\pi_2(1-\pi_2)}}$. Hence in the case of equal ICCs and same distribution of cluster sizes, the optimal allocation is determined by values of π_1 and π_2 only; we then have the same optimal allocation for CRTs with constant cluster size and CRTs with varying cluster size.

5.4 Optimal allocation when the total number of clusters is fixed and cost is considered

In Section 2.3, we derived optimal allocation w_x^{c*} of a fixed number of clusters with equal cluster sizes when cost is considered in the allocation. Here we extend that result to the case of varying cluster size.

We consider both cluster-level and individual-level costs. Suppose that the cluster-level cost per cluster in the hth arm is $e_h, h = 1, 2$, and cost per individual in the hth arm is $c_h, h = 1, 2$. The total cost associated with a particular allocation is $\sum_{i=1}^{k_1}(m_{1i}c_1 +$

$e_1) + \sum_{i=1}^{k_2}(m_{2i}c_2 + e_2)$. Cost efficiency, CE, the ratio of the precision to the total study cost is:

$$CE_{RD} = \frac{\Psi_{RD}}{\sum_{i=1}^{k_1}(m_{1i}c_1 + e_1) + \sum_{i=1}^{k_2}(m_{2i}c_2 + e_2)}. \tag{5.11}$$

The optimal allocation w_{RD}^{c*} will be the allocation that maximizes CE for measure RD.

Recall that Ψ_{RD} is the precision, the inverse of the variance of the estimator. Since Ψ_{RD} is the inverse of the right side of equation (5.9), CE can be expressed (up to a constant) as:

$$CE_{RD} = \left\{\pi_1(1-\pi_1)\left[\frac{1}{w} + \frac{1}{(1-w)}\frac{\pi_2(1-\pi_2)q_1}{\pi_1(1-\pi_1)q_2}\right][wq_3 + (1-w)q_4]\right\}^{-1}, \tag{5.12}$$

where $q_1 = \sum_{i=1}^{d}\frac{f_i n_i}{1+(n_i-1)\rho_1}$, $q_2 = \sum_{i=1}^{d}\frac{f_i n_i}{1+(n_i-1)\rho_2}$, $q_3 = c_1\sum_{i=1}^{d} f_i n_i + e_1$ and $q_4 = c_2\sum_{i=1}^{d} f_i n_i + e_2$.

Again, if cluster size distribution (f_i and n_i) is known, the righthand side of the equation is a function with one unknown variable w. Using the same method as in Chapter 2, we obtain the optimal allocation w_{RD}^{vc*}:

$$w_{RD}^{vc*} = \frac{\sqrt{\pi_1(1-\pi_1)q_2}}{\sqrt{\pi_1(1-\pi_1)q_2} + \sqrt{\pi_2(1-\pi_2)q_1\gamma}}, \tag{5.13}$$

where $\gamma = \frac{(c_1\sum_{i=1}^{d} f_i n_i + e_1)}{(c_2\sum_{i=1}^{d} f_i n_i + e_2)}$. When cost per cluster in the two arms (e_1, e_2) and cost per individual in the two arms (c_1, c_2) are given, this γ is a constant. As in Chapter 2, we call γ the cost ratio.

Note that due to the discrete nature of cluster size, investigators may need do some adjustments. Recall that the proportions of clusters of size n_i in the k clusters are $f_i, i = 1, 2, ..., d$ respectively and the total numbers of clusters of sizes $n_1, n_2, ..., n_d$ are $kf_1, kf_2, ..., kf_d$, respectively. However, the values of $w_{RD}^{vc*}kf_1, w_{RD}^{vc*}kf_2, ..., w_{RD}^{vc*}kf_d$ may not be integers. Therefore, investigators will typically need to round values to integers such that $k_1 + k_2 = k$.

In Chapter 2, we derived w_{RD}^{c*} when the cluster size is a constant m. One could ask,

suppose we calculate w_{RD}^{c*} using $\bar{n}$, mean cluster size, instead of m in equation (2.15). How different are w_{RD}^{vc*} and w_{RD}^{c*} for different cluster size distributions?

We first consider a special case. If ICCs are the same in both arms, $\rho_1 = \rho_2$, then $q_1 = q_2$ and w_{RD}^{vc*} reduces to $w_{RD}^{vc*} = \frac{\sqrt{\pi_1(1-\pi_1)}}{\sqrt{\pi_1(1-\pi_1)}+\sqrt{\pi_2(1-\pi_2)\gamma}}$, where $\gamma = \frac{(c_1\bar{n}+e_1)}{(c_2\bar{n}+e_2)}$. The parameters $c1, c2, e1, e2$ are individual and cluster level costs per unit in each arm and are known constants; the value of $\bar{n}$ is also known. The parameter γ is a constant. If we use constant cluster size m to substitute $\bar{n}$, w_{RD}^{vc*} is identical to w_{RD}^{c*} in the case of $\rho_1 = \rho_2$ and we do not need to know the exact distribution of cluster size. Rather, we only need the mean of cluster size and can directly calculate w_{RD}^{c*}. Compared to the exact distribution of cluster size, the mean of cluster size is easier to obtain.

Next we consider more generalized situations in which the ICCs are different in each arm. Suppose the total number of clusters is 100 and the mean of cluster size is 20. We consider several different cluster size distributions with increasing dispersion of cluster size distribution.

We use coefficient of variation (CV) of cluster size to measure the degree of dispersion. Recall that CV=σ/μ, where σ is the standard deviation and μ is the mean of cluster size. When the value of CV is larger, there is more dispersion in cluster size. Ahn et al. [14] defined and used an imbalance parameter to measure the dispersion and that imbalance parameter can be expressed as $1/(1 + 1/CV^2)$. But CV is more often used in research on CRTs with varying cluster size; for example, see Eldridge et al. [22] and Breukelen et al. [45].

We consider the following distributions:

Distribution 1 (Discrete uniform distribution): The cluster sizes are 10, 15, 20, 25 and 30. There are 20 clusters of each size. The corresponding CV is 0.35.

Distribution 2: The cluster sizes are 10 and 30. There are 50 clusters of each size. Distribution 2 has the same maximal, minimal and mean values of cluster size as Distribution 1. It has only two different values of cluster size, so it has more imbalance of cluster size, with CV value of 0.50.

Distribution 3: The cluster sizes are 10 and 60. There are 80 clusters of size 10 and 20 clusters for cluster size 60. Distribution 3 is a skewed distribution and has more imbalance in cluster size, with CV value of 1.

Distribution 4: The cluster sizes are 10 and 110. There are 90 clusters of size 10 and 10 clusters for cluster size 110. Distribution 4 is a more skewed distribution and more severe imbalance in cluster size, with CV value of 1.5.

We assume $\rho_1 = 0.3$, $\rho_2 = 0.1$ and $\gamma = 5$. The values of w_{RD}^{vc*} and w_{RD}^{c*} for different combinations of π_1 and π_2 are summarized in Table 5.1.

Table 5.1 shows that in Distribution 1, the discrete uniform distribution with CV of 0.35, w_{RD}^{vc*} and w_{RD}^{c*} are identical for almost all combinations of π_1 and π_2. A few small discrepancies between w_{RD}^{vc*} and w_{RD}^{c*} exist for Distribution 2 with w_{RD}^{c*} slightly higher than w_{RD}^{vc*}. We see more discrepancies for Distribution 3. However, these discrepancies are very small. The maximal value is just 0.02. The maximal value of discrepancies between w_{RD}^{vc*} and w_{RD}^{c*} is 0.03. Thus there is a pattern of more imbalance in cluster size leading to larger discrepancy between w_{RD}^{vc*} and w_{RD}^{c*}. However, even though the CV value is quite high at 1.5, the maximal value of discrepancies between w_{RD}^{vc*} and w_{RD}^{c*} is not large. Hence using w_{RD}^{c*} with mean cluster size may be a good approximation of w_{RD}^{vc*}.

In practice, CV of cluster size is typically less than 1. For example, in the Samoan Women's Health study [23], which was a CRT designed to increase rates of mammogram usage in women of Samoan ancestry, the CV of cluster size is 0.63. Another CRT study, the High Risk Colon Study, was designed to increase colorectal cancer (CRC) screening among high-risk individuals [38]. In that study, CRC cases were identified through the California Cancer Registry and relatives within the same family composed clusters. The CV of cluster size in that study is 0.58. Carter [46] reviewed four CRTs and CV values were between 0.30 and 0.51. Eldridge et al. [22] reported CVs for six CRTs and found values are between 0.42 to 0.75. We recommend investigators to estimate the CV of cluster size for their study. If the CV is not large, we may directly use w_{RD}^{c*} with mean

Table 5.1: Comparison of w_{RD}^{vc*} and w_{RD}^{c*} (in parentheses) values under different combinations of π_1 and π_2 for different cluster size distributions ($\rho_1 = 0.3, \rho_2 = 0.1, \gamma = 5$)

.

π_1/π_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Distribution 1									
0.1	.40(.40)	.34(.34)	.31(.31)	.29(.29)	.29(.29)	.29(.29)	.31(.31)	.34(.34)	.40(.40)
0.2	.47(.48)	.40(.40)	.37(.37)	.36(.36)	.35(.35)	.36(.36)	.37(.37)	.40(.40)	.47(.48)
0.3	.51(.51)	.44(.44)	.40(.40)	.39(.39)	.38(.38)	.39(.39)	.40(.40)	.44(.44)	.51(.51)
0.4	.52(.53)	.45(.45)	.42(.42)	.40(.40)	.40(.40)	.40(.40)	.42(.42)	.45(.45)	.52(.53)
0.5	.53(.53)	.46(.46)	.42(.43)	.41(.41)	.40(.40)	.41(.41)	.42(.43)	.46(.46)	.53(.53)
0.6	.52(.53)	.45(.45)	.42(.42)	.40(.40)	.40(.40)	.40(.40)	.42(.42)	.45(.45)	.52(.53)
0.7	.51(.51)	.44(.44)	.40(.40)	.39(.39)	.38(.38)	.39(.39)	.40(.40)	.44(.44)	.51(.51)
0.8	.47(.48)	.40(.40)	.37(.37)	.36(.36)	.35(.35)	.36(.36)	.37(.37)	.40(.40)	.47(.48)
0.9	.40(.40)	.34(.34)	.31(.31)	.29(.29)	.29(.29)	.29(.29)	.31(.31)	.34(.34)	.40(.40)
Distribution 2									
0.1	.40(.40)	.33(.34)	.30(.31)	.29(.29)	.29(.29)	.29(.29)	.30(.31)	.33(.34)	.40(.40)
0.2	.47(.48)	.40(.40)	.37(.37)	.35(.36)	.35(.35)	.35(.36)	.37(.37)	.40(.40)	.47(.48)
0.3	.51(.51)	.43(.44)	.40(.40)	.38(.39)	.38(.38)	.38(.39)	.40(.40)	.43(.44)	.51(.51)
0.4	.52(.53)	.45(.45)	.42(.42)	.40(.40)	.40(.40)	.40(.40)	.42(.42)	.45(.45)	.52(.53)
0.5	.53(.53)	.46(.46)	.42(.43)	.41(.41)	.40(.40)	.41(.41)	.42(.43)	.46(.46)	.53(.53)
0.6	.52(.53)	.45(.45)	.42(.42)	.40(.40)	.40(.40)	.40(.40)	.42(.42)	.45(.45)	.52(.53)
0.7	.51(.51)	.43(.44)	.40(.40)	.38(.39)	.38(.38)	.38(.39)	.40(.40)	.43(.44)	.51(.51)
0.8	.47(.48)	.40(.40)	.37(.37)	.35(.36)	.35(.35)	.35(.36)	.37(.37)	.40(.40)	.47(.48)
0.9	.40(.40)	.33(.34)	.30(.31)	.29(.29)	.29(.29)	.29(.29)	.30(.31)	.33(.34)	.40(.40)

cluster size even though cluster size is varying.

5.5 Sample size for balanced designs when cluster size varies

In Chapter 4, we derived the sample size formulas for estimating RD, RR and OR when the cluster size in CRTs is constant. In practice, the cluster size often varies. When the cluster size is varying, a common way to calculate the sample size is to use average

π_1/π_2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Distribution 3									
0.1	.39(.40)	.33(.34)	.30(.31)	.29(.29)	.29(.29)	.29(.29)	.30(.31)	.33(.34)	.39(.40)
0.2	.46(.48)	.39(.40)	.36(.37)	.35(.36)	.34(.35)	.35(.36)	.37(.37)	.39(.40)	.46(.48)
0.3	.50(.51)	.43(.44)	.39(.40)	.38(.39)	.37(.38)	.38(.39)	.39(.40)	.43(.44)	.50(.51)
0.4	.52(.53)	.44(.45)	.41(.42)	.39(.40)	.39(.40)	.39(.40)	.41(.42)	.44(.45)	.52(.53)
0.5	.52(.53)	.45(.46)	.42(.43)	.40(.41)	.39(.40)	.40(.41)	.42(.43)	.45(.46)	.52(.53)
0.6	.52(.53)	.44(.45)	.41(.42)	.39(.40)	.39(.40)	.39(.40)	.41(.42)	.44(.45)	.52(.53)
0.7	.50(.51)	.43(.44)	.39(.40)	.38(.39)	.37(.38)	.38(.39)	.39(.40)	.43(.44)	.50(.51)
0.8	.46(.48)	.39(.40)	.36(.37)	.35(.36)	.34(.35)	.35(.36)	.37(.37)	.39(.40)	.46(.48)
0.9	.39(.40)	.33(.34)	.30(.31)	.29(.29)	.29(.29)	.29(.29)	.30(.31)	.33(.34)	.39(.40)
Distribution 4									
0.1	.43(.40)	.36(.34)	.33(.31)	.32(.29)	.31(.29)	.32(.29)	.33(.31)	.36(.34)	.43(.40)
0.2	.50(.48)	.43(.40)	.40(.37)	.38(.36)	.38(.35)	.38(.36)	.40(.37)	.43(.40)	.50(.48)
0.3	.53(.51)	.46(.44)	.43(.40)	.41(.39)	.41(.38)	.41(.39)	.43(.40)	.46(.44)	.53(.51)
0.4	.55(.53)	.48(.45)	.45(.42)	.43(.40)	.42(.40)	.43(.40)	.45(.42)	.48(.45)	.55(.53)
0.5	.56(.53)	.48(.46)	.45(.43)	.43(.41)	.43(.40)	.43(.41)	.45(.43)	.48(.46)	.56(.53)
0.6	.55(.53)	.48(.45)	.45(.42)	.43(.40)	.42(.40)	.43(.40)	.45(.42)	.48(.45)	.55(.53)
0.7	.53(.51)	.46(.44)	.43(.40)	.41(.39)	.41(.38)	.41(.39)	.43(.40)	.46(.44)	.53(.51)
0.8	.50(.48)	.43(.40)	.40(.37)	.38(.36)	.38(.35)	.38(.36)	.40(.37)	.43(.40)	.50(.48)
0.9	.43(.40)	.36(.34)	.33(.31)	.32(.29)	.31(.29)	.32(.29)	.33(.31)	.36(.34)	.43(.40)

cluster size instead of the constant cluster size in sample size formulas designed for CRTs with constant cluster size. For example, if the goal is to estimate RD using equal allocation, under this approach we would use average cluster size $\bar{m}$ instead of m in equation (4.3) to calculate the number of clusters needed in each arm. However, the sample size calculated by this way may not reach the desired power; see Guittet [13], for example.

Our goal is to derive a sample size formula that will archive the desired power when cluster size varies. Our approach is to regard cluster size as a random variable denoted N. We denote the pdf of cluster size as $f(N)$.

Kerry and Bland [15] consider this sample size problem in a two-arm CRT with continuous outcomes. Recall that the design effect (DE) is the ratio of the sample size required for a CRT to the sample size required for an IRT with the same power, see Donner and Klar [1], for example. When cluster size is constant and equal to m, the design effect is $1 + (m-1)\rho$. Hence the sample size formula in a CRT can be obtained by multiplying the usual sample size formula by the design effect term. Suppose one arm of a CRT has k clusters with varying size, and the ith cluster has size m_i and its corresponding summary statistic is given a weight b_i. Kerry and Bland [15] show that the design effect can be expressed as:

$$DE = \frac{\bar{m}k \sum_{i=1}^{k} \frac{b_i^2}{m_i}(1 + (m_i - 1)\rho)}{(\sum_{i=1}^{k} b_i)^2} \tag{5.14}$$

When all cluster sizes are equal, $m_i = m$ and $b_i = 1/k$, and the design effect reduces to $1 + (m-1)\rho$.

Kerry and Bland consider the three different weights as we present in Section 5.2, *cluster weight*, *individual weight* and *minimal variance weight*. They refer to the first two as equal weight and cluster size weight. When clusters are given equal weight, the design effect is:

$$DE = \frac{\bar{m} \sum_{i=1}^{k} 1/m_i}{k}(1 - \rho) + \bar{m}\rho.$$

When clusters are weighted by their size, the design effect is given by:

$$DE = 1 + \left(\frac{\sum_{i=1}^{k} m_i^2}{\sum_{i=1}^{k} m_i} - 1 \right) \rho$$

Kerry and Bland [15] indicated the minimum variance mean of the cluster means is found by weighting the individual cluster means by the inverse of their variances. Weights proportional to the inverse of the variances become $m_i/(1 + (m_i - 1)\rho)$. The design effect is given by:

$$DE = \frac{\bar{m}k}{\sum_{i=1}^{k} \frac{m_i}{1+(m_i-1)\rho}}.$$

Although Kerry and Bland did not provide the explicit sample size formulas, the sample size required can be obtained by multiplying the usual sample size formula for an individual randomized trial by the design effect. Based on simulation results, Kerry and Bland conclude that minimum variance weights are always best, but equal weights may be acceptable for trials with large clusters and cluster size weights for trials with small clusters.

Mantunga and Hudgen [16] derived a sample size formula for a continuous outcome for a two-arm CRT while accounting for variability due in cluster size. It is shown that the proposed formula can be obtained by adding a correction term to the traditional formula which uses the average cluster size. Their derivation is based on an estimator equivalent to the estimator using the individual weighting scheme. Hence their sample size formula can be derived using Kerry and Bland's method too, although Kerry and Bland did not give an explicit formula. The additional correction term in Mantunga and Hudgen's paper is essentially the design effect term.

Mantunga and Hudgen's work dealt with a continuous outcome rather than a binary outcome. Kang et al. [47] extend Mantunga's work to a CRT with a binary outcome. The sample size (number of clusters per arm) is given by:

$$k_0 = \frac{(z_{\alpha/2} + z_\beta)^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{(\pi_1 - \pi_2)^2}[(1-\rho)\frac{1}{m} + \rho + \rho CV^2], \tag{5.15}$$

where $m = E(N)$, $\sigma^2 = Var(N)$ and $CV = \sigma/m$, the coefficient of variation of cluster size. Note that the term $\frac{(z_{\alpha/2}+z_\beta)^2[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}{(\pi_1-\pi_2)^2}\rho CV^2$ is a correction term and represents the additional number of clusters needed for a CRT with varying cluster sizes and average cluster size m compared to a CRT with equal cluster sizes in order to achieve the same power $1-\beta$. When CV is equal to 0, which corresponds to constant cluster size, this correction term equals 0, and equation (5.15) is reduced to the traditional formula; see equation (5.5) in Donner [41]. With larger CV, which means more dispersion of cluster sizes, more clusters are needed.

In Mantunga and Hudgen's derivation for the sample size formula, individual weights are used. In that paper, the authors also mention minimal variance weight, but they

argue that the sample size formula based on individual weight may be a little conservative but provides reasonable values in practice. Note that equation (5.15) can also be written as:

$$k_0 = \frac{(z_{\alpha/2} + z_\beta)^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{(\pi_1 - \pi_2)^2} \left[\frac{E(N(1+(N-1)\rho))}{E(N)^2}\right]. \tag{5.16}$$

Jung et al.[17] discussed sample size calculation for a binary outcome in CRTs. They derived a sample size formula for a one-arm CRT for the hypothesis test $H_0 : \pi = \pi_0$. They considered cluster weight, individual weight and minimal variance weight in estimating the success rate. They showed that the minimal variance weight is the best among these three weights. It requires equal or smaller sample sizes and is more robust to misspecification of an input parameter than those assigning equal weights to individuals or clusters. The sample size formula based on the minimal variance weight for a one-arm CRT is given by:

$$k_0 = \frac{(z_{\alpha/2} + z_\beta)^2 \pi_1(1-\pi_1)}{(\pi_0 - \pi_1)^2} \frac{1}{E[\frac{N}{1+(N-1)\rho}]}. \tag{5.17}$$

Jung's formula is for one-arm CRTs with binary outcomes, and other formulas mentioned above focus on two-arm CRTs but assume that both arms have the same ICC. Since this assumption may not hold, we relax it and allow each arm to have its own ICC, denoted ρ_1 and ρ_2.

Our approach is based on the $\hat{RD} = \hat{\pi}_1 - \hat{\pi}_2$ and its variance as given in equation (5.7). As we have seen, using different weights in this equation leads to different variance estimates. Here we use the minimal variance weight, so the variance of the test statistic is given in equation (5.8).

Let m_{1i} and m_{2i} be the size of the i-th cluster in arm 1 and arm 2, respectively. We regard them as independent and identically distributed random variables. By the law of large numbers, as $k_1 \to \infty$ and $k_2 \to \infty$, we have $\sum_{i=1}^{k_1} \frac{m_{1i}}{1+(m_{1i}-1)\rho_1}/k_1 \xrightarrow{P} E\left(\frac{N}{1+(N-1)\rho_1}\right)$ and $\sum_{i=1}^{k_2} \frac{m_{2i}}{1+(m_{2i}-1)\rho_2}/k_2 \xrightarrow{P} E\left(\frac{N}{1+(N-1)\rho_2}\right)$.

By equation (5.8), we have:

$$Var(\hat{\pi}_1 - \hat{\pi}_2) \xrightarrow{P} \pi_1(1-\pi_1)\frac{1}{k_1 E\left(\frac{N}{1+(N-1)\rho_1}\right)} + \pi_2(1-\pi_2)\frac{1}{k_2 E\left(\frac{N}{1+(N-1)\rho_2}\right)}. \quad (5.18)$$

Our hypotheses are $H_0 : \pi_1 = \pi_2$ and $H_1 : \pi_1 \neq \pi_2$, and the test statistic is $\frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{Var(\hat{\pi}_1)+Var(\hat{\pi}_2)}}$. The two arms have equal numbers of clusters, so $k_1 = k_2 = k_0$. Using the same method as in Section 4.2, we can obtain the number of clusters required in each arm:

$$k_0 = \frac{(z_{\alpha/2} + z_\beta)^2 \left[\pi_1(1-\pi_1)\frac{1}{q_1} + \pi_2(1-\pi_2)\frac{1}{q_2}\right]}{(\pi_1 - \pi_2)^2}, \quad (5.19)$$

where $q_1 = E\left(\frac{N}{(1+(N-1)\rho_1)}\right)$ and $q_2 = E\left(\frac{N}{(1+(N-1)\rho_2)}\right)$.

Note that when we say cluster size is a random variable N following a particular probability distribution and the pdf is $f(N)$, we refer to the population distribution of cluster size. If in a study, the clusters are randomly chosen by investigators, then the population distribution of cluster size can be used to calculate the sample size; $E\left(\frac{N}{(1+(N-1)\rho_1)}\right)$ is calculated using the pdf $f(N)$. However, if investigators do not randomly select clusters, the distribution of cluster size will not reflect the population distribution. For example, investigators may attempt to choose smaller size clusters. Recall that the design effect $DE = 1 + (m-1)\rho$, the smaller value of m, the smaller value of DE. Hence the smaller sample size is required to obtain the same power. Therefore, selecting smaller size clusters is more efficient than selecting larger size clusters. In any case, the pdf $f(N)$ should reflect the realized probability distribution of cluster size.

5.6 Sample size for optimal allocation designs when cluster size varies

In Section 5.4, assuming each arm has the same number of clusters, we derived the sample size formula for a two-arm CRT with varying cluster size. From previous chapters, we know that when the cost per cluster and the cost per individual are very different in the two arms, assigning equal numbers of clusters in the two arms may not be efficient.

Suppose that the number of clusters in arm 1 is s times as the number in arm 2, $k_1 = sk_2$. Again using the same method in Section 4.2, after direct calculation we obtain the number of clusters in arm 1 and arm 2 to satisfy the power and type 1 error rate requirements:

$$k_1 = \frac{(z_{\alpha/2} + z_\beta)^2 \left[\pi_1(1-\pi_1)\frac{1}{q_1} + s\pi_2(1-\pi_2)\frac{1}{q_2}\right]}{(\pi_1 - \pi_2)^2}, \tag{5.20}$$

$$k_2 = \frac{(z_{\alpha/2} + z_\beta)^2 \left[\pi_1(1-\pi_1)\frac{1}{sq_1} + \pi_2(1-\pi_2)\frac{1}{q_2}\right]}{(\pi_1 - \pi_2)^2}, \tag{5.21}$$

where $q_1 = E\left(\frac{N}{(1+(N-1)\rho_1)}\right)$ and $q_2 = E\left(\frac{N}{(1+(N-1)\rho_2)}\right)$.

Given π_1, π_2, ρ_1 and ρ_2 and the pdf of cluster size N, for any s we can obtain corresponding k_1 and k_2. All combinations of s, k_1 and k_2 satisfying the above equations can satisfy power and type 1 error requirement. But different s, k_1 and k_2 are associated with different costs.

In Section 5.3, when the total number of clusters is fixed, we derived equation (5.13), the optimal allocation w_{RD}^{vc*}, which maximizes the precision of the estimator per unit cost, thereby maximizing the power for the total cost. In other words, the total cost for w_{RD}^{vc*} is minimized for all different w satisfying the same power requirement.

In contrast to the problem in Section 5.3, in which the total number of clusters is fixed, in this section we consider the sample size requirement problem. The number of clusters in the study is not pre-determined and our role is to find it. When k_1 and k_2 are large, we have

$$w_{RD}^{vc*} \xrightarrow{P} \frac{\sqrt{\pi_1(1-\pi_1)q_2}}{\sqrt{\pi_1(1-\pi_1)q_2} + \sqrt{\pi_2(1-\pi_2)q_1\gamma}} \tag{5.22}$$

where $q_1 = E\left(\frac{N}{(1+(N-1)\rho_1)}\right)$, $q_2 = E\left(\frac{N}{(1+(N-1)\rho_2)}\right)$ and $\gamma = \frac{(c_1E(N)+e_1)}{(c_2E(N)+e_2)}$. This w_{RD}^{vc*} can be used in sample size calculations.

Since $s = k_1/k_2$ and $w_{RD}^{vc*} = k1/(k1+k2)$, we obtain $s^* = \frac{w_{RD}^{vc*}}{1-w_{RD}^{vc*}}$. The value of $s*$ makes the total cost minimal for the study. Plugging $s^* = \frac{w_{RD}^{vc*}}{1-w_{RD}^{vc*}}$ into equations (5.20)

and (5.21), we obtain:

$$k_1 = \frac{(Z_{\alpha/2} + Z_\beta)^2 \left[\pi_1(1-\pi_1)\frac{1}{q_1} + \frac{w_{RD}^{vc*}}{1-w_{RD}^{c*}}\pi_2(1-\pi_2)\frac{1}{q_2}\right]}{(\pi_1 - \pi_2)^2}, \tag{5.23}$$

$$k_2 = \frac{(Z_{\alpha/2} + Z_\beta)^2 \left[\pi_1(1-\pi_1)\frac{1}{\frac{w_{RD}^{vc*}}{1-w_{RD}^{c*}}q_1} + \pi_2(1-\pi_2)\frac{1}{q_2}\right]}{(\pi_1 - \pi_2)^2}, \tag{5.24}$$

where $q_1 = E\left(\frac{N}{(1+(N-1)\rho_1)}\right)$, $q_2 = E\left(\frac{N}{(1+(N-1)\rho_2)}\right)$ and w_{RD}^{vc*} is obtained from equation (5.22).

We assess the performance of our sample size formula (5.19), which corresponds to equal sample sizes and formulas (5.23, 5.24) which correspond to the optimal sample size. We also compare them to the usually used sample size formula (4.3) in which mean cluster size replaces constant cluster size. Suppose the distribution of cluster size is one of following distributions:

Distribution 1: All clusters have size 5, with CV equal to 0;

Distribution 2: One fourth of clusters have size 2, 4, 6 and 8, respectively, with CV equal to 0.45;

Distribution 3: Half of clusters have size 2 and half have size 8, with CV equal to 0.6;

Distribution 4: Four fifths of clusters have size 2 and one fifth of clusters have size 17, with CV equal to 1.2.

The mean cluster size in the four cluster size distributions is equal to 5. However, the imbalance of cluster size increases sequentially from Distribution 1 to Distribution 4.

Our null hypothesis is $\pi_1 = \pi_2$ and we want to obtain 0.8 power with type 1 error rate of 0.05. We assume that the success rates in the two arms are 0.5 and 0.3 and the ICC in the arm 1 takes values 0.05, 0.1, 0.2 and 0.3 while the ICC in the arm 2 is a constant 0.1. We assume cost ratio $\gamma = 5$. We simulate data with cluster size following one of

Table 5.2: Empirical power and sample size (in parentheses) for different cluster size distributions ($\pi_1 = 0.5, \pi_2 = 0.3$, desired power 80%)

(ρ_1, ρ_2)	Dist 1	Dist 2	Dist 3	Dist 4
Based on formula (4.3)				
(0.05, 0.1)	78.9(*24, 24*)	76.6(*24, 24*)	74.8(*24, 24*)	67.4(*24, 24*)
(0.1, 0.1)	78.4(*26, 26*)	76.7(*26, 26*)	72.9(*26, 26*)	67.4(*26, 26*)
(0.2, 0.1)	78.8(*30, 30*)	75.6(*30, 30*)	74.4(*30, 30*)	68.0(*30, 30*)
(0.3, 0.1)	81.2(*34, 34*)	79.2(*34, 34*)	77.4(*34, 34*)	68.4(*34, 34*)
Based on formula (5.19)				
(0.05, 0.1)	78.9(*24, 24*)	76.9(*25, 25*)	77.8(*26, 26*)	78.9(*30, 30*)
(0.1, 0.1)	78.4(*26, 26*)	77.4(*27, 27*)	77.0(*28, 28*)	78.6(*33, 33*)
(0.2, 0.1)	78.8(*30, 30*)	77.4(*31, 31*)	80.9(*33, 33*)	80.8(*38, 38*)
(0.3, 0.1)	81.2(*34, 34*)	79.0(*35, 35*)	79.8(*37, 37*)	79.3(*42, 42*)
Based on formulas (5.23, 5.24)				
(0.05, 0.1)	80.3(*17, 38*)	78.5(*18, 40*)	80.2(*19, 41*)	80.0 (*22, 47*)
(0.1, 0.1)	79.8(*20, 40*)	80.8(*21, 42*)	82.6(*22, 44*)	81.2 (*26, 52*)
(0.2, 0.1)	81.8(*25, 44*)	82.6(*26, 46*)	82.3(*24, 25*)	82.1(*32, 57*)
(0.3, 0.1)	81.7(*29, 47*)	84.8(*31, 50*)	82.7(*32, 52*)	81.2(*37, 59*)

above distributions. The details of our simulation method are described in Section 6.3. The power and sample size under different formulas are summarized in Table 5.2.

We see that for all formulas, the sample sizes required increase with the increasing ρ_1. This is reasonable since when ICC is larger, the design effect is larger, leading to a larger sample size to obtain the same power.

When formula (4.3) is used, we see that for different distributions, the calculated sample sizes are the same. Recall that in formula (4.3), the mean cluster size is used. The mean cluster sizes are the same in all of the 4 distributions, hence we obtain the same sample sizes. For Distribution 1, the power levels under all ρ_1 and ρ_2 combinations

are around 0.8, which is the desired power. For Distribution 2, the power levels are decreased, but only slightly. For Distribution 3 and Distribution 4, the power levels are obviously smaller than 0.8.

We next assess our derived sample size formula (5.19). For this formula, we still assume that the two arms have the same number of clusters. However, we used the minimal variance weights scheme to incorporate the cluster distribution information in the sample size calculation. From Distribution 1 to Distribution 4, as CV of cluster size increases, the sample size required also increases. For example, when $\rho_1 = 0.3$ and $\rho_2 = 0.1$, the total number of clusters required for Distribution 1 is 48, and the total number of individuals required is 68x5=340; in contrast, the total number of clusters required for Distribution 4 is 84, and the total number of individuals required is 84x0.8x2+84x0.2x17=84x5=420. For various ρ_1 and ρ_2, the power levels for all 4 different distributions are around the desired power 0.8.

Finally we assess our derived optimal sample size formulas (5.23, 5.24), in which we allow the number of clusters in the two arms to be different. For those formula, besides incorporating cluster distribution information, we also incorporate the optimal allocation information. Similar with formula (5.19), from Distribution 1 to Distribution 4, with CV increasing, we see that the calculated sample size increases for various ρ_1 and ρ_2, the power levels for all 4 different distributions are around the desired power 0.8. In fact, in most cases, the power levels are slightly larger than 0.8. The difference from formula (5.19) is that the total sample size required by formulas (5.23, 5.24) is larger but the total cost is less. For example, when $\rho_1 = 0.3$ and $\rho_2 = 0.1$, for Distribution 3, the total number of clusters required is 74 by formula (5.19) and 84 by (5.23, 5.24), respectively. Since the cost ratio γ is 5, the total cost is 37x5+37=222 by formula (5.19) and 32x5+52=212 by formula (5.19). This suggests that the balanced design costs more than the design based on optimal sample size.

5.7 Application

We show how to use the results in this chapter using real studies. Recall the Samoan Women's Health study [23], which was a CRT designed to increase rates of mammogram usage in women of Samoan ancestry. In the trial, Samoan churches in southern California were randomized to intervention and control arms, and women at intervention churches participated in a culturally appropriate breast cancer education program. The control arm received usual care. The outcome was self-reported receipt of mammogram at follow-up. In previous chapters, we assume the cluster size in this CRT is constant. Now we abandon such assumption.

In this study, 55 churches were recruited. The actual cluster size varies, ranging from 2 to 42 participants with a mean cluster size of 14. A histogram of cluster size distribution is shown in Figure 5.1.

These 55 churches have 24 distinctive cluster sizes. The frequencies of churches for some cluster sizes are very small. For example, only 1 church with cluster size 3 exists. No matter what the optimal allocation is, we can only assign this church to one arm. For practical purpose, we group churches into several cluster size strata and randomly assign churches in each stratum to arm 1 or arm 2 according to w_{RD}^{vc*}. In order to calculate w_{RD}^{vc*}, n_i in equation (5.13) is the mean cluster size of churches in a stratum. Since the CV within each stratum will be small, the w_{RD}^{vc*} will be essentially the same for each stratum. We have seen that the difference between using w_{RD}^{c*} with mean cluster size and w_{RD}^{vc*} is small. Hence we do not need many strata.

From Figure 5.1, we can classify churches into 4 groups with mean cluster size 7, 17, 27 and 37 respectively. The corresponding density 0.45, 0.37, 0.1 and 0.08.

In this study, the cost per church is \$1000. For churches in the intervention arm (arm 1), there is an additional cost of \$4000 per church. The cost of materials for each individual is \$10 in both arms. Hence $\gamma = \frac{5000+14*10}{1000+14*10} = 4.51$. In the original Samoan Women's Health study, the observed success proportions in the intervention and and control arm were 0.5 and 0.4. Even though all Samoan churches available were used, a

significant difference was not found. Here we assume the success proportions in the two arms are more different, say, 0.5 and 0.3, respectively. The ICCs are assumed to be 0.3 and 0.1.

From equation (5.13), after calculation, we obtain $w_{RD}^{vc*} = 0.41$. Hence for clusters with size 2-12, $55 * 0.45 * 0.41 = 10$ clusters are assigned to arm 1 and 15 are assigned to arm 2. For clusters with size 12-22, 8 clusters are assigned to arm 1 and 12 are assigned to arm 2. For the rest clusters, 4 are assigned to arm 1 and 6 are assigned to arm 2.

Now we consider using equation (2.15). The mean of cluster size is 14. After calculation, we obtain $w_{RD}^{c*} = 0.41$, the same as $w_{RD}^{vc*} = 0.41$. Hence for this study, although the cluster size is varying, directly using w_{RD}^{c*} is good.

Now suppose we want to test the hypotheses $\pi_1 = \pi_2$ with power 0.8 and type 1 error 0.05. The optimal sample size can be directly calculated from equations (5.23) and (5.23). After calculation, $k_1 = 93.4$ and $k_2 = 134.5$. Hence we need 94 clusters in arm 1 and 135 clusters in arm 2.

5.8 Chapter summary and discussion

In Chapter 2 through 4, we assumed the cluster size is constant. In this chapter, we deal with a CRT with varying cluster size. When the cluster size is varying, there are different ways that we can assign weights to different clusters to obtain an unbiased estimator of RD. We reviewed three different weighting approaches: cluster weight, individual weight and minimal variance weight, which were introduced in Kerry and Bland [15]. For the same precision, the number of clusters needed is the smallest for minimal variance weights. Hence our work is based on minimal variance weights.

We considered the optimal allocation problem assuming the cluster sizes are known. We regarded cluster size as a random variable that follows a particular probability distribution. We let clusters in the two arms have similar distributions of cluster size and derived the optimal allocation w_{RD}^{vc*}, which makes the CE maximal among all allocation

values w. The optimal allocation w_{RD}^{vc*} is dependent on the cluster size distribution. However, if $\rho_1 = \rho_2$, the optimal allocation does not involve the distribution of cluster size; it only involves the mean cluster size. Even though $\rho_1 \neq \rho_2$, w_{RD}^{c*} derived for constant cluster size could be used to approximate w_{RD}^{vc*}. The constant cluster size is substituted by mean of cluster size to calculate w_{RD}^{c*}. The approximation is generally good. As shown in Table 5.1, even when the CV of cluster size is as large as 1.5, the difference between w_{RD}^{vc*} and w_{RD}^{c*} is negligible. Recall that we need the number of clusters to be integers, therefore we often obtain the same number of clusters after rounding kw_{RD}^{vc*} and kw_{RD}^{c*} to integers, as illustrated by the example in section 5.7.

We also consider the sample size calculation problem for a CRT with varying cluster sizes. We review recent research on sample size calculation for a CRT with varying cluster size. Previous research mainly focused on a continuous outcome. Besides the literature reviewed in Section 5.5, some other work on sample size calculation can be found in Breukelen et al. [48] and Candel et al. [49], in which they used CV to calculate how many more clusters are required for varying cluster size compared to equal cluster size. Kang et al. [47] extended the work of Mantunga and Hudgen [16] to a two-arm CRT with binary outcomes. Although they did not directly specify, they used individual weights. Jung et al.[17] provided sample size formulas for a one-arm CRT with a binary outcome using all three weighting approaches. In this chapter, we derived a sample size formula (5.19) for a two-arm CRT with a binary outcome using the minimal variance weights. The minimal variance weights can naturally connect to the concept of maximizing cost efficiency (CE), which was introduced in Chapter 2. In addition, previous studies assumed a common ρ in the two arms. In our sample size formula, we considered arm-specific ρ_1 and ρ_2.

All sample size formulas given in previous literature are for balanced designs, in which the same number of clusters are assigned in the two arms. As we showed in Chapter 4, among all designs satisfying the power requirement, the total cost of a balanced design may not be minimal. In this chapter, we modify our sample size formula (5.19) for varying cluster size using the optimal allocation w_{RD}^{vc*}, as we did in Chapter 4. The

modified sample size formulas (5.23, 5.24), which are called optimal sample size formula, include cost consideration and guarantee the total cost of the study is minimal. However, the total number of individuals in the study by optimal sample size formula is generally larger than that of a balanced design. From the simulation study shown in Table 5.2, it seems that formulas (5.23, 5.24) tend to give a conservative sample size. Note that the distribution of cluster size should be specified in the sample size calculation.

Optimal allocation and optimal sample size for varying cluster size in this chapter are based on the assumption that the cluster size follows a probability distribution and that distribution is similar in the two arms. This assumption is realistic. In practice, investigators usually want their intervention to work for clusters with various sizes, not for only smaller clusters or larger clusters. More research is needed if this assumption does not hold.

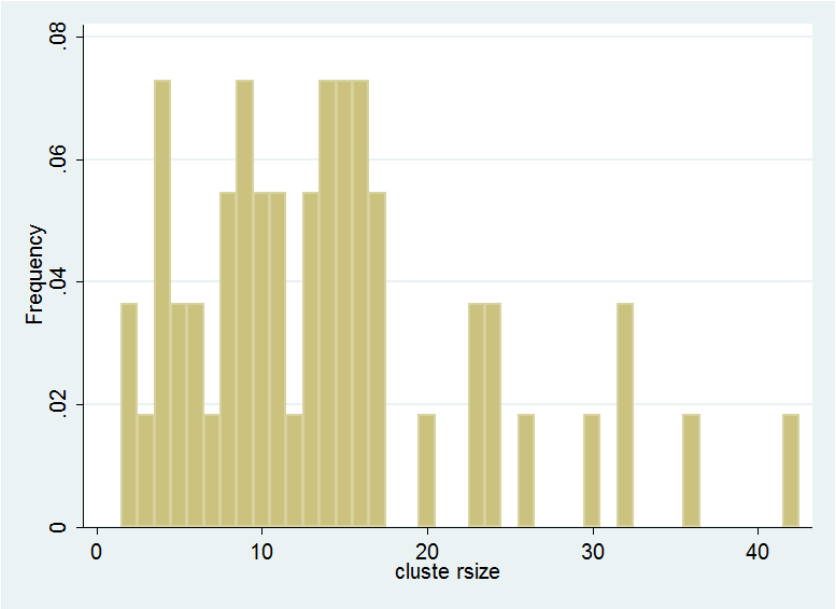


Figure 5.1: Histogram of cluster size distribution in Samoan Women's Health study.

CHAPTER 6

Comparison of analysis methods when cluster size varies

6.1 Introduction

In CRTs, clusters rather than individuals are randomized to different treatment arms and observations are typically made at the individual level. When observations are available at the individual level, there is a choice to be made as to whether to conduct analyses at the cluster level or on the individual level. In a cluster-level analysis, cluster-level summary measures are used. In individual-level analyses, individual-level measures are used, and correlated data methods are used to account for clustering.

Cluster-level analysis is conceptually straightforward; cluster-level summary measures are calculated and treated as independent. According to Donner [1] and Hayes [2], cluster-level analysis is more robust than individual-level analysis, especially when the number of clusters is small. However, cluster-level analysis does not use the individual level information and hence may not be fully efficient. Compared to the cluster-level analysis, the individual-level analysis can be more complex but may offer advantages. For example, when a multilevel regression model is used, covariates can be included in the model.

There is currently limited information on the comparative performance of different analysis methods for CRTs with binary outcomes in terms of type 1 error rate and statistical power, and even more limited information in the case of varying cluster size. Here we review relevant literature.

Heo and Leon [50] compared the performance of different methods for analysis of clustered binary outcomes. They considered the mixed effect logistic regression model (MELR), also called the "random effects logistic regression model" and generalized estimating equation method (GEE). For MELR, they used SAS PROC NLMIXED and the SAS GLIMMIX macro. The former corresponds to a full likelihood method and the latter corresponds to a penalized quasi-likelihood (PQL) method. They also included the ordinary logistic regression model as a comparison, although it is not a correct method to analyze CRTs data. Type I error rate, power, bias, and standard error were compared across the four statistical methods through computer simulations. In their simulation, they only considered balanced designs with same number of clusters in each of two arms and the same number of individuals in each cluster. They concluded that the performance of the full likelihood and the penalized quasi-likelihood methods were superior to GEE for analysis of clustered binary observations. In a subsequent paper [51], they compared MELR performance for binary outcomes between CRTs with equal cluster size and CRTs with unequal cluster size, using uniform distributions to generate unequal cluster size. They concluded that the performance of MELR was very similar, regardless of inequality in cluster size. However, they did not compare MELR with other analysis methods for CRTs with unequal cluster sizes.

Ukoumunne et al. [52] used simulation to compare the accuracy of estimation and confidence interval coverage of several methods for analyzing binary outcomes from CRTs. The emphasis of their study was on estimation rather than hypothesis testing. They considered GEE and also used ordinary least square regression to analyze cluster-level summary measures. They also considered a modified GEE approach, in which Wald tests of significance and confidence intervals were based on quantiles from a t-distribution rather than quantiles from the standard normal distribution. They assumed equal cluster size and equal number of clusters in the two arms. In general, GEE was better than the other methods. In a subsequent paper [53], they compared GEE and cluster level t-tests for binary outcomes when the outcome measures were RD, RR and OR. Since distributions of RR and OR are skewed, they used log transformations for

these measures. They found that the cluster level t-test often had large bias when log RR and log OR were used.

Pacheco et al. [54] compared methods for the analysis of CRTs with count data. They considered cluster-level t test, GEE and MELR, which they called generalized linear mixed models. For the GEE method, they used both the sandwich estimator and model-based estimator of the variance-covariance matrix. For MELR, they used both maximum-likelihood-based methods and Bayesian methods for parameter estimation. Unlike other authors, they assumed unequal cluster size. Cluster sizes were generated from normal distributions, with fractional cluster sizes rounded to the closest integer and the number of individuals per cluster truncated to a minimum of 8. They concluded that MELR performs better in general.

Some work on the comparative performance of different analysis methods in CRTs with binary outcomes and varying cluster size is reported by Austin [55]. Austin compared the cluster level t-test, Wilcoxon rank sum test, permutation test, adjusted chi-square test, MLER (which he called logistic-normal random effects model) and GEE method. In his simulation study, cluster size followed a Poisson distribution with mean of 7 or 39 and the ICC was assumed constant across arms. He concluded that the GEE method performs better in general. As far as we know, his work is the first to compare different methods to analyze binary outcomes in CRTs with varying cluster sizes. However his work has some limitations and in particular, reflects only limited imbalance of cluster sizes. In Austin's work, the CV of cluster size is only 0.38 and 0.16. In practice, the variance of cluster size may be larger than the mean of cluster size, but the Poisson distribution does not allow for such overdispersion. In addition, the assumption of a constant ICC across arms can be questionable, especially when proportions are very different in the two arms. In fact for binary outcomes, the ICC is a function of outcome prevalence. Hence, the ICC and the outcome prevalence are intrinsically related. With different outcome prevalences in two arms, a constant ICC in both arms may not hold. A discussion can be found in Crespi et al. [56].

The previous studies on analysis methods for binary outcomes in CRTs often overlook

imbalance of cluster size, and especially severe imbalance has not been considered. In addition, all previous studies have assumed the same ICC in the two arms, but in practice, this assumption is often violated. To address these issues, we conducted a simulation study to address two questions: (1) How does imbalance of cluster size affect the comparative performance of the different analysis methods, especially for severe imbalance of cluster sizes? (2) How does unequal ICC in two arms affect the performance of the methods if a constant ICC is assumed in the analysis?

Many analysis methods are available for CRTs with binary data. Many are described in Donner and Klar [1] and Hayes and Moulton [2]. We will consider four methods: the two-sample cluster-level t-test, the adjusted chi-square approach, GEE and MELR. Research such as Crespi et al. [57] has found the most commonly used analysis methods in cancer prevention and screening CRT are mixed models and GEE. Although the adjusted chi-square approach is less commonly used, it is an individual-level analysis method and unlike GEE and MELR, it is not a regression method. The cluster level t-test is a simple and commonly used cluster-level analysis method, and therefore we also include it. We first review these analysis methods. Next, we describe a method of indirectly generating correlated binary data, and modify this method to generate CRTs data with varying cluster size and different ICCs. Using this method to simulate data, we compare the four methods in terms of power and type 1 error rate under a broad range of scenarios.

6.2 Commonly used analysis methods

In this section, we review the four CRTs data analysis methods compared in our simulation study. We define some key notations here:

π_h: the success probability in the hth arm.

π_{hi}: the success probability in the ith cluster in the hth arm.

k_h: the number of clusters in the hth arm.

n_{hi}: the number of individuals in the ith cluster in the hth arm.

6.2.1 Cluster-level t-test

The first method we consider is the cluster-level t-test. This test is a two independent sample t-test on the cluster-level proportions which are regarded as independently distributed in each arm.

The null hypothesis is H_0 $\pi_1 = \pi_2$. The test statistic for a cluster-level t-test is

$$t = \frac{\bar{\pi}_1 - \bar{\pi}_2}{S\sqrt{\frac{1}{k_1} + \frac{1}{k_2}}}, \tag{6.1}$$

where $\bar{\pi}_h = \frac{\sum_{i=1}^{k_h} \hat{\pi}_{hi}}{k_h}$ is the mean proportion for treatment hth arm, $h = 1, 2$, $S^2 = \frac{\sum_{h,i}(\hat{\pi}_{hi} - \bar{\pi}_h)^2}{k_1 + k_2 - 2}$ is the pooled estimate of error variance, and k_h is the number of clusters in the hth arm.

Note that this statistic does not require calculation of the ICC. The observed proportion in each cluster is calculated and these are regarded as independent observations. Hence this method avoids estimating the ICC. Under the null hypothesis of no difference of proportions between the two arms, the statistic follows a t-distribution with $k_1 + k_2 - 2$ degrees of freedom [1].

According to the standard t-test assumptions, this method assumes that the cluster-specific proportions are normally distributed with common variance. The variance of $\hat{\pi}_{hi}$ is $\frac{\pi_h(1-\pi_h)(1+(n_{hi}-1)\rho_h)}{n_{hi}}$. Hence when there is considerable variation in cluster size, this assumption is commonly violated. However, simulation studies [58, 59] have shown that the test is robust to violations of the underlying assumptions, especially when the numbers of clusters in the two treatment arms are equal.

The cluster-level t-test ignores any variation in cluster size. An alternative is the weighted t-test, in which weights are given to clusters. Recall in Section 5.2, we described various weight schemes and used *minimal variance weight* in our sample size calculation. That weight scheme is given in equation (5.4) and we use this weight scheme here. Let

b_{hi} be minimal variance weights. The test statistic for the weighted t-test is

$$t_{weight} = \frac{\bar{\pi}_1 - \bar{\pi}_2}{\sqrt{(\frac{S_1^2(k_1-1)+S_2^2(k_2-1)}{k_1+k_2-2})(\frac{1}{k_1}+\frac{1}{k_2})}}, \tag{6.2}$$

where

$\bar{\pi}_h = \frac{\sum_{i=1}^{k_h} b_{hi}\hat{\pi}_{hi}}{k_h}$ is the mean proportion for treatment arm h, $h = 1, 2$, and

$S_h^2 = \frac{\sum_{i=1}^{k_h} b_{hi}(\pi_{hi}-\bar{\pi}_h)^2}{1-\sum_{i=1}^{k_h} b_{hi}^2}$ is the estimate of the sample variance in the hth arm, $h = 1, 2$.

6.2.2 Adjusted chi-Square test

The second method we consider is the adjusted chi-square test. This test was developed by Donner and Donald [60]. The idea behind this method is to adjust the usual Pearson chi-square statistic using by an estimate of the ICC. The test statistic is given by

$$\chi_A^2 = \sum_{h=1}^{2} \frac{\sum_{i=1}^{k_h} n_{hi}(\hat{\pi_h} - \hat{\pi})^2}{C_h\hat{\pi}(1-\hat{\pi})} \tag{6.3}$$

where $\hat{\pi_h}$ is the success rate in the hth arm, $\hat{\pi}$ is the success rate across both arms and n_{hi} is the cluster size in the hth arm and the ith cluster, $C_h = \frac{\sum_{i=1}^{k_h} n_{hi}[1+(n_{hi}-1)\hat{\rho}]}{\sum_{i=1}^{k_h} n_{hi}}$ and $\hat{\rho}$ is an estimate of ρ.

Recall that if cluster size is m, then the design effect is defined as $1+(m-1)\rho$, which measures the inflation in variance of $\hat{\pi_h}$ that can be attributed to clustering [1]. The factor C_h can be regarded as the estimated design effect in the hth arm. Donner and Donald [60] assumed the population design effects are the same in both treatment arms and argued that the assumption is guaranteed to hold for experimental comparisons due to randomization. The statistic χ_A^2 approximately follows a chi-square distribution with 1 degree of freedom when H_0 holds.

An estimate of the ICC is required for this method. There are several different methods available to estimate the ICC, for example, see Ridout et al. [21] and Wu et al.[61]. Donner and Donald use the ANOVA method to estimate ρ and we follow this method. Since in practice the ICCs may be different in the two arms, we also consider a modification of this method to estimate and use arm-specific estimates $\hat{\rho}_1$ and $\hat{\rho}_2$ in

C_1 and C_2. When $\rho_1 = 0$ and $\rho_2 = 0$, the adjusted chi-square test reduces to a standard Pearson chi-square test.

6.2.3 Mixed logistic regression model

The third method we consider is the mixed logistic regression model (MELR), which is an individual-level analysis method. Many textbooks have discussions about MELR, for example, see Fitzmaurice [62] and McCulloch [63]. The MELR model can be expressed as:

$$log\left(\frac{\pi_{hij}}{1-\pi_{hij}}\right) = logit(\pi_{hij}) = \alpha + \beta_{cs}T_{hi} + u_{hi}, \tag{6.4}$$

where $\pi_{hij} = Pr(x_{hij} = 1|u_{hi}, T_{hi})$ is the probability of success for the jth individual in the ith cluster in the hth arm conditional on u_{hi}, T_h, $u_{hi} \sim N(0, \sigma_u^2)$ is a cluster-level random effect for the ith cluster in the hth arm, T_{hi} is a dummy variable with $T_{hi} = 1$ for the intervention arm and $T_{hi} = 0$ for the control arm.

The random effects u_{ij} account for between-cluster variation and are assumed to follow a normal distribution with mean 0 and variance σ_u^2. The coefficient β_{cs} is the log odds of success in the intervention arm compared to the control arm. The subscript cs in β_{cs} indicates that this is a cluster-specific effect. Note that $logit(\pi_{hij}|T_{hi} = 1)$ - $logit(\pi_{hij}|T_{hi} = 0) = \beta_{cs}$. Hence the term $exp(\beta_{cs})$ can be interpreted as the cluster-specific odds ratio for the effect of intervention. The treatment effect can be tested by testing the hypothesis $H_0 : \beta_{cs} = 0$ using a Wald test. MELR belongs to the class of generalized linear mixed models (GLMM), for which there are several methods of estimation, including penalized quasilikelihood, Laplace approximation, Gauss-Hermite quadrature and Markov chain Monte Carlo [63].

Individual-level covariates can be added into the MELR model, which is an advantage of MELR compared to a non-regression method such as the adjusted chi-square test. Hayes [2] indicates that MELR does not provide reliable results when there are fewer than about 15 clusters per arm.

In the MELR model, the ICC is defined on the logistic scale. On this scale, cluster and

individual effects are assumed additive and the within-cluster individual-level variance $\frac{\Pi^2}{3}$ does not depend on within-cluster prevalence. In the MELR model, the ICC can be expressed as:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + (\Pi^2/3)}. \tag{6.5}$$

In equation (6.4), the random effects u_{ij} follow the same distribution in both arms, which implies that ICCs are the same in the two arms. When ICCs are different in the two arms, Omar [64] used a model to allow different random effect variances. We can modify the MELR model to allow different random effect variances; we call this model MELR2. This allows ICC to be different in the two arms. The modified model is:

$$log\left(\frac{\pi_{hij}}{1-\pi_{hij}}\right) = \alpha + \beta_{cs}T_{hi} + u_{1i}(1-T_{hi}) + u_{2i}T_{hi}, \tag{6.6}$$

where $u_{1i} \sim N(0, \sigma_{u1}^2)$ and $u_{2i} \sim N(0, \sigma_{u2}^2)$.

6.2.4 GEE method

The fourth method we consider is the GEE approach developed by Liang and Zeger [65]. The GEE model we consider is:

$$log\left(\frac{\pi_{hij}}{1-\pi_{hij}}\right) = \alpha + \beta_{pa}T_{hi},, \tag{6.7}$$

where $\pi_{hij} = Pr(x_{hij} = 1|T_{hi})$ and correlation within clusters is accounted for by using unknown correlation matrix $R(a)$. As for MELR, $T_{hi} = 0$ in the control arm and $T_{hi} = 1$ in the intervention arm.

In GEE, the working correlation matrix $R(a)$ can have different structures [65]. We assume a common correlation model for CRT data. Subjects from different clusters are independent and the correlation between pairs of subjects in the same cluster is identical. Therefore, an exchangeable working correlation matrix with 1 in the diagonal and a constant a elsewhere is assumed. It can be expressed as:

$$R(a) = \begin{pmatrix} 1 & a & a & ... \\ a & 1 & a & ... \\ a & a & 1 & ... \end{pmatrix}.$$

Note that if a constant ICC is accounted for in all clusters, the constant a in $R(a)$ is across all clusters. If the ICC in each arm is different, we can use modified working correlation structures in our GEE model. In this case, we still assume an exchangeable working correlation matrix. However, the working correlation matrix in each arm has its own constant in off-diagonal positions. In the control arm, the working correlation matrix is:

$$R(a_0) = \begin{pmatrix} 1 & a_0 & a_0 & ... \\ a_0 & 1 & a_0 & ... \\ a_0 & a_0 & 1 & ... \end{pmatrix}.$$

In the intervention arm, the working correlation matrix is $R(a1)$:

$$R(a_1) = \begin{pmatrix} 1 & a_1 & a_1 & ... \\ a_1 & 1 & a_1 & ... \\ a_1 & a_1 & 1 & ... \end{pmatrix}.$$

More discussions can be found in Crespi [66].

In this GEE model, β_{pa} denotes the intervention effect on the log odds scale. The subscript pa in β_{pa} indicates the population average effect. The treatment effect is tested by testing $H_0 : \beta_{pa} = 0$. A robust variance estimator and Wald test are used. Like MELR, GEE can accommodate cluster-level and individual-level covariates. Interpretation of β_{pa} is a little different from β_{cs}. Here $exp(\beta_{pa})$ is the population-averaged odds ratio for the effect of intervention, and $exp(\beta_{cs})$ is the odds ratio for the effect of innervation for a specific cluster. Like MELR, GEE also requires a large numbers of clusters per arm to obtain reliable results and at least 15 per arm is recommended [2].

6.3 Simulation method

We use Emrich and Piedmonte's method [67] to generate correlated binary data, which is an indirect method to generate correlated binary data from a multivariate normal distribution. Within each cluster, the generated binary outcomes have the given marginal expectation and pairwise correlation.

Suppose the cluster size is m. In general, the steps to simulate an m-dimensional vector $\mathbf{X}$ with binary elements $X_1, ..., X_m$ with $E(X_j) = \pi_j$ and $Corr(X_j, X_k) = \rho_{jk}$, $j \neq k$ are as follows:

(1): Let Φ denote the CDF for a standard bivariate normal random variable with correlation coefficient δ_{jk} and let $w(\pi)$ denote the πth quantile of the standard normal distribution. Solve the following equation for δ_{jk}:

$$\Phi[w(\pi_j), w(\pi_k), \delta_{jk}] = \rho_{jk}[\pi_j(1-\pi_j)\pi_k(1-\pi_k)]^{1/2} + \pi_j\pi_k \tag{6.8}$$

(2): Simulate an m-dimensional multivariate normal random variable $\mathbf{W} = (W_1, ..., W_J)^T$ with mean $\mathbf{0}$ and correlation matrix $\mathbf{\Sigma}$=(δ_{jk}).

(3): Generate the vector $\mathbf{X}$ with components $X_j = I[W_j \leq w(\pi_j)]$ for $j = 1, ..., m$. Note $I[W_j \leq w(\pi_j)]$ is an indicator function which transforms continuous components in the vector $\mathbf{X}$ to binary components.

It can be shown that under this setting, $E(X_j) = \pi_j$ and $Corr(X_j, X_k) = \rho_{jk}$. There is a very brief proof in Emrich and Piedmonte's paper [67]. We give a more detailed proof here.

Proof:

The expectation of the jth component of the vector $\mathbf{X}$ is $E(X_j) = E\{I[W_j \leq w(\pi_j)]\}$ $= \Pr[W_j \leq w(\pi_j)] = \pi_j$ since $w(\pi)$ denotes the πth quantile of the standard normal distribution and W_j is the jth component of a multivariate normal random variable.

The covariance of the jth and kth components of the vector $\mathbf{X}$ is:

$Cov(X_j, X_k)$=$E(X_jX_k)$− $E(X_j)E(X_k)$=$Pr(X_j = 1, X_k = 1)$− $\pi_j\pi_k$=$Pr[W_j \leq w(\pi_j), W_k \leq$

$w(\pi_k)]- \pi_j\pi_k=\Phi[w(\pi_j), w(\pi_k), \delta_{jk}] - \pi_j\pi_k$.

From equation (6.8) in step (1), we have $Cov(X_j, X_k)=\rho_{jk}[\pi_j(1-\pi_j)\pi_k(1-\pi_k)]^{1/2}$. Hence we have $Corr(X_j, X_k)=\frac{Cov(X_j,X_k)}{\sqrt{Var(X_j)Var(X_k)}}=\frac{\rho_{jk}[\pi_j(1-\pi_j)\pi_k(1-\pi_k)]^{1/2}}{[\pi_j(1-\pi_j)\pi_k(1-\pi_k)]^{1/2}}=\rho_{jk}$.

We implement this method in R. Since we use the common correlation model, we have $E(X_j) = \pi \ \forall \ j$ and $Corr(X_j, X_k) = \rho \ \forall \ j, k$. We write an R function to calculate the CDF of a bivariate normal distribution. Then we specify π and ρ and use the R function *uniroot* to solve equation (6.8) to obtain δ. After we obtain δ, the correlation matrix $\mathbf{\Sigma}$ is obtained, which has 1 on the diagonals and δ on the off-diagonals. Then we use the R function *rmvnorm* in R package *mvtnorm* to generate multivariate normal distribution of a m-dimensional random vector with mean $\mathbf{0}$ and correlation matrix $\mathbf{\Sigma}$. The dimension m of this vector is equal to the size of the corresponding cluster. Finally we write an indicator function to transform the continuous elements in that vector to be binary elements as in step (3). The transformed vector is a cluster of size m with binary outcomes with the desired mean and correlation structure. We repeat this procedure p times to obtain p clusters.

6.4 Simulation specification

The purpose of our simulation study is to compare different methods to analyze CRT data with varying cluster size. We set up the simulation study so that the mean of cluster size is fixed and CV of cluster size varies. Austin [55] who assumed varying cluster size following a Poisson distribution. However, the CV of cluster size in his simulation was low (CV=0.38 or 0.16). The Poisson distribution also has limitations; we can not vary the mean of cluster size and CV independently. The negative binomial distribution is more flexible. However, since one parameter of the negative binomial distribution must be an integer, this approach would not allow us to find appropriate parameters to achieve a pre-specified mean and CV. Heo et al. [51] used uniform distributions to generate unequal cluster size, but it can be shown that the CV of any uniform distribution is less than 0.58. To see this, let a uniform distribution have parameters a and b. The CV of

this uniform distribution is $\frac{\sqrt{(b-a)^2/12}}{(a+b)/2}=\frac{b-a}{b+a}\frac{1}{3} \leq \frac{1}{3} = 0.577$. Pacheco et al. [54] considered varying cluster size by generating normal distributions. However, when using normal distributions, if the CV is large, we can expect that many cluster would have negative cluster size. Hence none of these distributions are appropriate for our simulation study.

In our simulation, we let cluster size follow a gamma distribution. The reasons we choose a gamma distribution are as follows. First, CV of cluster size can be easily controlled by the shape and scale parameters of the gamma distribution. Therefore we can easily fix the mean cluster size and vary CV to pre-specified values. Second, a gamma distribution can be skewed and in practice, cluster size distributions are often skewed. Since the gamma distribution is a continuous distribution, but cluster size must take integer values, we round fractional cluster sizes value to the nearest integer. In addition, cluster size cannot be 0, hence we add 1 to all numbers obtained from the gamma distribution. Using this method, the CV of the simulated data is not exactly the same as the CV we want but the difference is negligible.

A factorial design is used in our simulation study. The factors include the number of clusters in the two arms, the CV of cluster size, the success probabilities in the two arms and the ICCs in the two arms. With respect to the number of clusters, we consider both equal and unequal numbers of clusters in the two arms. For equal number of clusters, we let $k_1 = 30, k_2 = 30$. For unequal number of clusters, which may occur when optimal allocation is used, we let $k_1 = 40, k_2 = 20$ and $k_1 = 20, k_2 = 40$. The mean of cluster size is set to be 20. CV of cluster size is set to be 0, 0.25, 0.5, 0.75, 1 and 1.5. The case of CV=0 corresponds to constant cluster size. When cluster size varies, we assume it to follow a gamma distribution. As CV increases, the imbalance of cluster size becomes more severe. The success rate π_1 in the intervention arm is set to be 0.3, 0.4 and 0.5. The success rate π_2 in the control arm is set to 0.3. When $\pi_1 = \pi_2$, we assess the type 1 error. When $\pi_1 \neq \pi_2$, we assess the statistical power. With respect to the ICCs, we consider both equal ICCs and unequal ICCs in two arms. For the former, we set $\rho_1 = \rho_2 = 0.1$. For the latter, we let $\rho_1 = 0.05$, $\rho_1 = 0.2$ and $\rho_=0.3$ while keeping $\rho_2 = 0.1$. Hence we

have 216 different scenarios.

For each scenario, 2000 data sets were generated. The MELR model is fitted in R using the *lme4* package and the GEE model is fitted in R using the *geepack* package. For the GEE method, we use the "robust error" estimator ("sandwich variance" estimator). We write R functions to calculate the cluster-level t-test and the adjusted chi-square test statistics. The weighted t-test and the adusted chi-square test need estimated ICCs values. We write an R function to estimate ICCs values using the ANOVA method. More detailed discussion of estimating ICC can be found in Ridout et al.[21].

The performance of methods is evaluated based on type I error rate and statistical power. Type I error rate is computed as the proportion of p values less than 0.05 under a null hypothesis of no intervention effect ($\pi_1 = \pi_2$, $\beta_{cs} = 1$, β_{pa}=1). Statistical power is computed as the proportion of p values less than 0.05 under an alternative hypothesis of intervention effect ($\pi_1 \neq \pi_2$, $\beta_{cs} \neq 1$, $\beta_{pa} \neq 1$).

6.5 Simulation results

In order to save space, we show selected simulation results. Recall in adjusted chi-square test shown in equation (6.3), the factor C_h is calculated using the common $\hat{\rho}$ or the arm-specific $\hat{\rho}_h$. In our simulations we found that the results for the two versions of the adjusted chi-square test are exactly the same. Therefore, we present results of only the common ICC adjusted chi-square test.

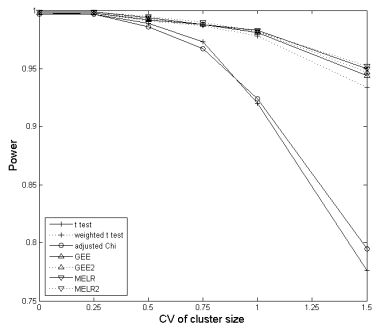
In Figures 6.1 and 6.2, π_2 is fixed as 0.3, and π_1 is equal to 0.5 and 0.4 respectively, and the statistical power of different methods is compared. In each figure, we include different sub-figures for different combinations of ρ_1 and ρ_2. More specifically, we fix ρ_2 as 0.1 and assign 0.05, 0.1, 0.2, 0.3 to ρ_1 sequentially.

In Figure 6.1, we see that with CV of cluster size increasing, which means cluster size becomes more unbalanced, the power of all methods decreases. But the rates of decrease are different for different methods. The cluster-level t-test and the adjusted

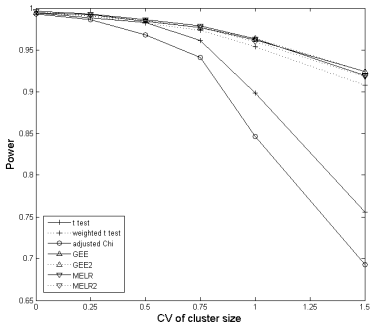
chi-square test are more sensitive to increasing CV of cluster size than other methods. For example, in Figure 6.1(a) in which $\rho_1 = 0.05$ and $\rho_2 = 0.1$, when cluster size is a constant (CV=0) the power is close to 1 for all methods. When CV of cluster size increases to 1, the power of t-test and the power of adjusted chi-square test decrease to 0.92. In contrast, all other methods have power around 0.98 when CV=1. When CV of cluster size increases to 1.5, the power of t-test decreases to 0.78 and the power of adjusted chi-square test decreases to 0.80, much less than power for constant cluster size. Note that although weighted t-test is a cluster-level analysis method, it has relatively high power.

In Figure 6.1(b) in which $\rho_1 = \rho_2$, we see that GEE and GEE2 have almost the same power, as do MELR and MELR2. In Figure 6.1(a), 6.1(c) and 6.1(d), ρ_1 and ρ_2 are not the same, and we use these figures to check whether the power of GEE2 is different from that of GEE and the power of MELR2 is different from that of MELR when $\rho_1 \neq \rho_2$. In Figures 6.1(a) and 6.1(c), we see that GEE and GEE2 almost have the same power, even when the CV is large. In Figure 6.1(d), in which the difference of ρ_1 and ρ_2 and the absolute value of ρ_1 and ρ_2 are relatively large, GEE2 has slightly higher power than GEE when CV of cluster size is as large as 1.5. However, the power differences for GEE and GEE2 are negligible when CV is less than 1. For MELR and MELR2, the situation is different. In Figure 6.1(a), we see that MELR and MELR2 have almost the same power. However, in Figures 6.1(c) and 6.1(d), MELR2 has less power than MELR2 for different CV values, including for constant cluster size.

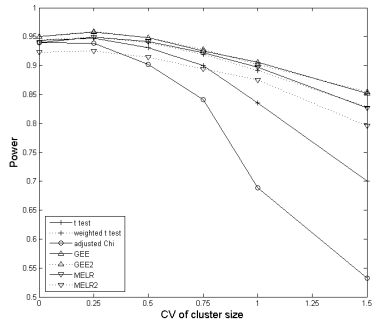
Figure 6.2 also compares the power of different methods. In this figure, the success rate π_1 reduced to 0.4, so the true difference of π_1 and π_2 is only 0.1. We see that the power for all methods are smaller than those in Figure 6.1. However, the pattern of the power change with CV increasing is similar in those two figures. In Figure 6.2(b), when $\rho_1 = \rho_2$, we see that GEE and GEE2 have almost the same power. In Figures 6.2(a), GEE and GEE2 also have the similar power. In Figures 6.2(c) and 6.2(d), when the difference between ρ_1 and ρ_2 and the absolute values of ρ_1 and ρ_2 are large, GEE2 has a slightly higher power than GEE. In Figure 6.2(a), we see that MELR2 has higher power



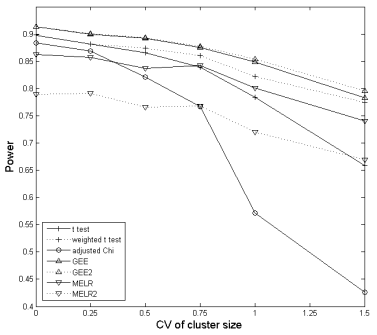
(a) $\rho_1 = 0.05, \rho_2 = 0.1$



(b) $\rho_1 = 0.1, \rho_2 = 0.1$



(c) $\rho_1 = 0.2, \rho_2 = 0.1$

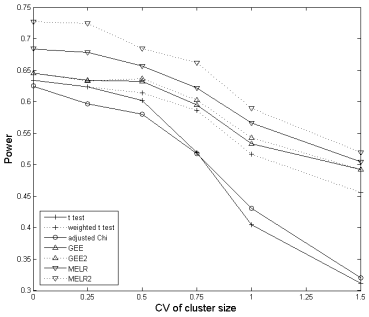


(d) $\rho_1 = 0.3, \rho_2 = 0.1$

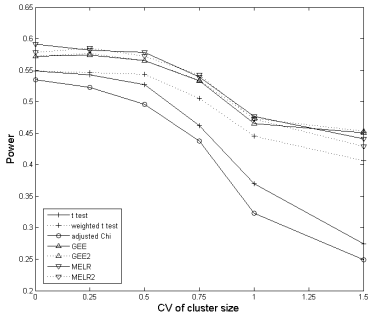
Figure 6.1: Comparison of power for different analysis methods when $\pi_1 = 0.5, \pi_2 = 0.3, \bar{m} = 20$ and $k_1 = k_2 = 30$.

than MELR, but in Figures 6.2(b) and Figure 6.2(d), MELR2 has smaller power than MELR. In fact, in those two figures, MELR2 even has smaller power than t-test and adjusted chi square test.

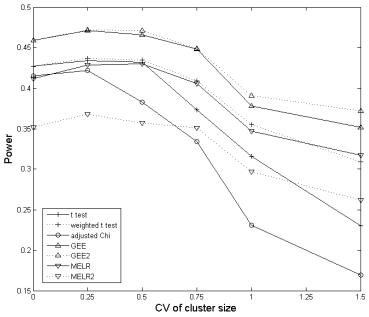
The results of comparison of type 1 error rates are shown in Figures 6.3. It seems that the type 1 error rate of cluster-level t-test is very close to the nominal value of 0.05, regardless of CV of cluster size; the same is true for the adjusted chi square test. The type 1 error rate of the weighted t-test is also close to the nominal value of 0.05 when CV is less than 1. When CV is as large as 1.5, the weighted t-test has slightly inflated type 1 error rate. GEE, GEE2, MELR and MELR2 all have inflated type 1 error rate. Unlike power decreasing with CV increasing, the change of type 1 error rate is not so obvious with CV increasing, unless the CV is larger than 1. In Figures 6.3(c) and 6.3(d),



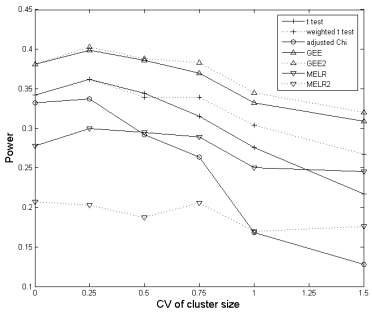
(a) $\rho_1 = 0.05, \rho_2 = 0.1$



(b) $\rho_1 = 0.1, \rho_2 = 0.1$

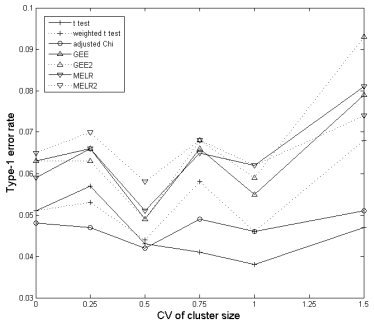


(c) $\rho_1 = 0.2, \rho_2 = 0.1$

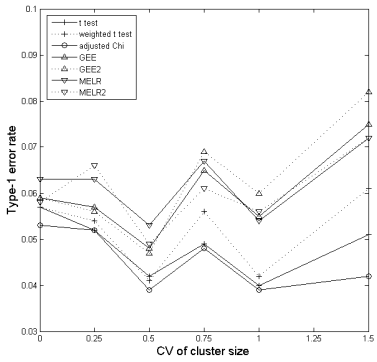


(d) $\rho_1 = 0.3, \rho_2 = 0.1$

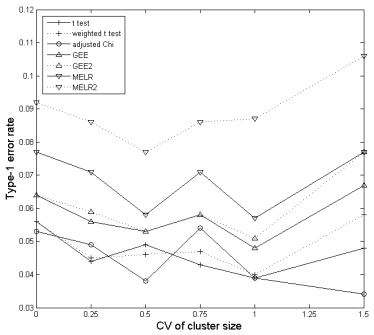
Figure 6.2: Comparison of power for different analysis methods when $\pi_1 = 0.4, \pi_2 = 0.3, \bar{m} = 20$ and $k_1 = k_2 = 30$.



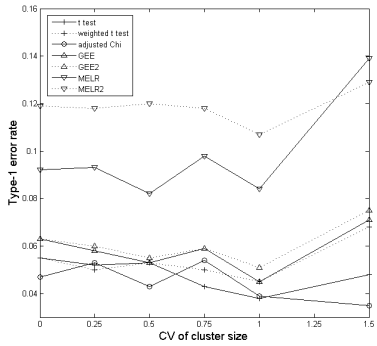
(a) $\rho_1 = 0.05, \rho_2 = 0.1$



(b) $\rho_1 = 0.1, \rho_2 = 0.1$



(c) $\rho_1 = 0.2, \rho_2 = 0.1$



(d) $\rho_1 = 0.3, \rho_2 = 0.1$

Figure 6.3: Comparison of type 1 error rate for different analysis methods when $\pi_1 = 0.3, \pi_2 = 0.3, \bar{m} = 20$ and $k_1 = k_2 = 30$.

we see that MELR2 has high type 1 error rate. MELR also has high type 1 error rate when CV is equal to 1.5.

6.6 Chapter summary and discussion

In this chapter, we reviewed several commonly used CRTs analysis methods: cluster-level t-test, weighted t-test, adjusted chi-square test, MELR method and GEE method. We showed how to modify the adjusted chi-square test, MELR and GEE to incorporate arm-specific ICCs. Then we reviewed Emrich and Piedmonte's method to generate correlated binary data. Based on a modification of this method, we simulated CRT data with varying cluster size and different ICCs in the two arms. We used simulated data

to compare the performance of different analysis methods.

In previous research, although it is known that power may decrease when the cluster is varying, few papers compare different analysis methods with pre-selected CV of cluster size. As far as we know, we are the first to select CV values and compare the performance of different analysis methods with increasing CV. We find that the power of all methods decreases when cluster sizes become more unbalanced. When CV is very large, the power of the cluster-level t-test is very low. In addition, although Donner and Klar [1] and Donner [68] provide the adjusted chi-square test to analyze CRT data, our simulation results indicate that such method may have poor performance when CV of cluster size is large and ICCs are large. The weighted t-test is relatively insensitive to increasing CV of cluster size. When the CV is less than 0.75, the weighted t-test has slightly lower power than GEE. Therefore, the weighted t-test may be an alternative method to GEE when CV of cluster size is not very large. GEE and MELR are relatively robust to increasing CV in terms of power. Ahn, Jung and Kang [69] provided a weighted version of the adjusted Chi square test. We have not checked its performance in our simulation, but we expect it may have better performance than the adjusted chi-square test we used.

All previous work on comparing different analysis methods has simulated CRT data with the same ICCs in the two arms. In our work, we simulated different ICCs in the two arms, which is more realistic. We find the results are identical for adjusted chi-square test with estimated common $\hat{\rho}$ and adjusted chi-square test with arm-specific estimated $\hat{\rho_1}$ and $\hat{\rho_2}$. According to Jung, Ann and Donner [70], this is due to the fact that the mean cluster sizes in the two arms are the same. We also modify GEE and MELR to consider different ρ_1 and ρ_2 in two arms (GEE2 and MELR2). We find that when $\rho_1 \neq \rho_2$, GEE2 has the same power or only slightly higher power than GEE2, depending on CV of cluster size and ρ_1 and ρ_2. Unless CV of cluster size is large and ρ_1 and ρ_2 are large, e.g., when CV is larger than 1 and ρ_1 and ρ_2 are larger than 0.1, GEE2 is not superior to GEE in terms of power. This is reasonable, since for GEE method, even though the working correlation matrix is misspecified, the estimation is robust. We find that in most instances, MELR2 actually has lower power than MELR when $\rho_1 \neq \rho_2$. When CV

is very large, MELR2 even has lower power than the t-test. We used R package *lme4* to fit MELR2 and these problems may be specific to this package.

Unlike power decreasing with increasing CV of cluster size for all methods, the imbalance of cluster size does not seem to affect type 1 error rate in a consistant manner. All regression methods (GEE, GEE2, MELR and MELR2) showed inflated type 1 error rates. GEE has smaller type 1 error rate than MELR. MELR2 has very large type 1 error rate when $\rho_1 \neq \rho_2$ and ρ_1 and ρ_2 are relatively large.

Based on our simulation study, we would recommend using GEE. Austin [55] also concluded that GEE is slightly better than other methods. GEE2 can also be used, but it does not have obvious advantages over GEE unless CV of cluster size is large and ρ_1 and ρ_2 are large. We did not consider the case of small number of clusters in a CRT. Previous studies indicate MELR and GEE are not robust if the number of clusters is smaller than 15, e.g., see Donner and Klar [1]. When the number of cluster is very small and cluster size is close to a constant, the weighted t-test may be used.

References

[1] Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. New York, N.Y.: Oxford University Press; 2000.

[2] Hayes RJ, Moulton LH. Cluster Randomised Trials. Boca Ration, FL: CRC Press; 2009.

[3] Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. American Journal of Epidemiology. 1981;114:906–914.

[4] Klar N, Donner A. Current and future challenges in the design and analysis of cluster randomization trials. Statistics in Medicine. 2001;20:3729–3740.

[5] Cornfield J. Randomization by group: a formal analysis. American Journal of Epidemiology. 1978;108:100–102.

[6] Campbell M, Walters S. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. Wiley; 2014.

[7] Walter SD. Determination of significant relative risk and optimal sampling procedures in prospective and retrospective comparative studies of varous sizes. American Journal of Epidemiology. 1977;105:387–397.

[8] Meydrich EF. Cost considerations and sample size requirements in cohort and case-conrol studies. American Journal of Epidemiology. 1978;107:201–205.

[9] Morgenstern H, Winn DM. A method for determining the sampling ratio in epidemiologic studies. Statistics in Medicine. 1983;2:387–396.

[10] Yanagawa T, Blot Wj. Optimal sampling ratios for prospective studies. American Journal of Epidemiology. 1977;106:436–437.

[11] Lubin JH. Some efficency comments on group size in study design. Amderican Journal of Epidemiology. 1980;111:347–359.

[12] Brittain E, Schlesselman JJ. Optimal allocation for the comparison of proportions. Biometrics. 1982;38:1003–1009.

[13] Guittet L, Ravaud P, Giraudeau B. Planning a cluster randomized trial with unequal cluster sizes: practical issues involving continuous outcomes. BMC Medical Research Methodology. 2006;6:17.

[14] Ahn C, Hu F, Skinner CS, Ahn D. Effect of imbalance and intracluster correlation coefficient in cluster randomization trials with binary outcomes when the available number of cluster is fixed in advance. Contemporary Clinical Trials. 2009;30:317–320.

[15] Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Statistics in Medicine. 2001;20:377–390.

[16] Manatunga AK, Hudgens MG. Sample size estimation in cluster randomized studies with varying cluster size. Biometrical journal. 2001;43:75–86.

[17] Jung SH, Kang SH, Aln C. Sample size calculations for clustered binary data. Statistics in Medicine. 2001;20:1971–1982.

[18] Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. Statistics in Medicine. 1996;15:1069–1092.

[19] Dette H. On robust and efficient designs for risk estimation in epidemiological studies. Scandinavian Journal of Statistics. 2004;31:319–331.

[20] Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster-randomized trials: a review of definitions. International Statistical Review. 2009;77:378–394.

[21] Ridout MS, Demetrio CGB, Firth D. Estimating intraclass correlation with binary data. Biometrics. 1999;55:137–148.

[22] Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. International Journal of Epidemiology. 2006;35:1292–1300.

[23] Mishra SI, Bastani R, Crespi CM, Chang LC, Luce PH, Baquet CR. Results of a randomized trial to increase mammogram usage among Samoan women. Cancer Epidemiology Biomarkers and Prevention. 2007;16:2594–2604.

[24] Sistrom CL, Garvan CW, Grobbee DE. Potential misinterpretation of treatment effects due to use of odds ratios and logistic regression in randomized controlled tirals. PLoS ONE. 2011;6:e21248.

[25] Knol NJ, and DRG. Proportions, odds and risk. Radiology. 2004;230:12–19.

[26] Miettinen OS. Individual mathcing with multiple controls in the case of all-or-none response. Biometrics. 1969;25:339–355.

[27] Breukelen GJ, Candel MJ. Calculating sample sizes for cluster randomized trials: we can keep it simple and efficient. Journal of Clinical Epidemiology. 2012;65:1212–1218.

[28] Raudenbush SW. Statistical analysis and optimal design for cluster randomized trials. Psychological Methods. 1997;2:173–185.

[29] Dette H, Haines ML, Imhof LA. Maximin and Bayesian optimal designs for regression model. Statistica Sinica. 2007;17:463–380.

[30] Hade EM, Murray DM, Pennell ML, Rhoda D, Paskett ED, et al . Intraclass correlation estimates for cancer screening outcome: estimates and applications in the design of group-randomized cancer screening studies. Journal of the National Cancer Institute. 2010;40:97–103.

[31] Kerman J. Neutral noninformative and informative conjugate beta and gamma prior distributions. Electronic Journal of Statistics. 2001;5:1450–1470.

[32] Turner RM, Prevost AT, Thompson SG. Allowing for imprecision of the intra-class correlation coefficient in the design of cluster randomized trials. Statistics in Medicine. 2004;23:1195–1214.

[33] Turner RM, Thompson SG, Spiegelhalter DJ. Prior distributions for the intraclass correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. Clinical Trials. 2005;2:108–118.

[34] D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues the encounters of academic consultants in statistics. Statistics in Medicine. 2003;22:169–186.

[35] John C, A H. Handbook of Statistics in Clinical Oncology. 3rd ed. Boca Raton, FL: CRC Press; 2012.

[36] Hilton JF. Noninferiority trials design for odds ratio and risk difference. Statistics in Medicine. 2008;29:982–993.

[37] Dixon A, Clarkin C, Barrowman N, Correll R, Osmond MH, Plint AC. Reduction of radial-head subluxation in children by triage nurses in the emergency department: a cluster-randomized controlled trial. Canadian Medical Association Journal. 2014;186:E317–323.

[38] Bastani R, Glenn BA, Maxwell AE, Ganz PA, Mojica CM, Chang LC. Validation of self-reported colorectal cancer (CRC) screening in a study of ethnically diverse first-degree relatives of CRC cases. Cancer Epidemiology Biomarkers and Prevention. 2008;17:791–798.

[39] Glenn BA, Herrmann AK, Crespi CM, Mojica CM, Chang LC, Maxwell AE, et al. Changes in risk perceptions in relation to self-reported colorectal cancer screening among first-degree relatives of colorectal cancer cases enrolled in a randomized trial. Health Psychology. 2011;30:481491.

[40] Crespi CM, Wong WK, Wu S. A new dependence parameter approach to improve the design of cluster randomized trials with binary outcomes. Clinical Trials. 2011;8:687–698.

[41] Donner A, Klar K. Design and analysis of cluster randomization trials in health research. New York, NY: Oxford University Press Inc; 2000.

[42] Lee EW, Dubin N. Estimation and sample size considerations for clustered binary responses. Statistics in Medicine. 1994;13:1241–1252.

[43] Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data. Biometrics. 1992;48:577–585.

[44] Jung SN, Ahn C. Estimation of response probability in correlated binary data: a new approach. Drug Information Journal. 2000;34:599–604.

[45] Breukelen GJ, Candel MJ, Berger MP. Relative efficiency of unequal versus equal cluster sizes in clster randomoized and multicentre trials. Statistics in Medicine. 2007;26:2589–2603.

[46] Carter B. Cluster size varibility and imbalance in cluster randomized controlled trials. Statistics in Medicine. 2010;29:2984–2993.

[47] Kang SH, Ahn CW, Jung SH. Sample size calculations for dichotomous outcomes in cluster randomization trials with varying cluster size. Drug Information Journal. 2003;37:109–114.

[48] Breukelen GJ, Candel MJ, Berger MP. Relative efficiency of unequal cluster sizes for variance component estimation in clster randomoized and multicentre trials. Statical Methods in Medicine Research. 2008;17:439–458.

[49] Candel MJ, Breukelen GJ, L K, Berger MP. Optimality of unequal cluster sizes in multilevel studies with realistic sample sizes. Communications in Statistics Simulation and Computation. 2008;37:222–239.

[50] Heo M, Leon A. Comparison of statistical methods for analysis of clustered binary observations. Statistics in Medicine. 2005;24:911–923.

[51] Heo M, Leon A. Performance of a mixed effects logistic regression model for binary outcomes with unequal cluster size. Journal of Biopharmaceutical Statistics. 2005;15:513–526.

[52] Ukoumunne OC, Carlin JB, Gulliford MC. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. Statistics in Medicine. 2007;26:3415–3428.

[53] Ukoumunne OC, Forbes B, Carlin JB, Gulliford MC. Comparison of the risk difference, risk ratio and odds ratio scales for quantifying the unadjusted intervention effect in cluster randomized trials. Statistics in Medicine. 2008;27:5143–5155.

[54] Pacheco GD, Hattendorf J, Colford JM, Mäusezahl D, Smith T. Performance of analytical methods for overdispersed counts in cluster randomized trials: sample size, degree of clustering and imbalance. Statistics in Medicine. 2009;28:2989–3011.

[55] Austin PC. A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. Statistics in Medicine. 2007;26:35503565.

[56] Crespi CM, Wong WK, Wu S. A new dependence parameter approach to improve the design of cluster randomized trials with binary observations. Clinical Trials. 2011;8:687–698.

[57] Crespi CM, Maxwell AE, Wu S. Cluster randomized trials of cancer screening interventions: Are appropriate statistical methods being used? Contemporary Clinical Trials. 2011;32:477484.

[58] Donner A, Klar N. Statistical considerations in the design and analysis of community intervention trials. J Clin Epidemiol. 1996;49:435–439.

[59] Heeren T, D'Agostino R. Robustness of the two-independent sample t-test when applied to ordinal scaled data. Statistics in Medicine. 1987;6:79–90.

[60] Donner A, Donald A. The statistical analysis of multiple binary measurements. Journal of Clinical Epidemiology. 1988;41:899–905.

[61] Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemporary Clinical Trials. 2012;33:869–880.

[62] Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. Hoboken, NJ: John Wiley; 2004.

[63] McCulloch CE, Searle SR. Generalized, Linear and Mixed Models. New York, NY: John Wiley and Sons; 2001.

[64] Omar RZ, Thompson SG. Analysis of a cluster randomized trial with binary outcome data using a multilevel model. Statistics in Medicine. 2000;19:2675–2688.

[65] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13–22.

[66] Crespi CM, Wong WK, Mishra S. Using second-order generalized estimating equations to model heterogeneous intraclass correlation for cluster-randomized trials. Statistics in Medicine. 2009;28:814–827.

[67] Emrich LJ, Piedmonte MR. A method for generating high-dimensional multivariate binary variates. American Statistician. 1991;45(4):302–304.

[68] Donner A. Statistical methods in ophthalmology: an adjusted chi-squared approach. Biometrics. 1989;45:605–611.

[69] Ahn C, Jung SH, Kang SH. An evaluation of weighted chi-square statistics for clustered binary data. Drug Information Journal. 2003;37:91–99.

[70] Jung SH, Ahn C, Donner A. Evaluation of an adjusted chi-square statistic as applied to observational studies involving clustered binary data. Statistics in Medicine. 2001;20:2149–2161.