Some New Advancements in Adaptive Randomization for Clinical Trials

By Guannan Zhai

B.S. in Financial Mathematics, June 2018, Northwest University
M.S. in Statistics, May 2021, University of Illinois Urbana-Champaign

A Dissertation submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

August 31, 2025

Dissertation directed by

Feifang Hu
Professor of Statistics

The Columbian College of Arts and Sciences of The George Washington University certifies that Guannan Zhai has passed the Final Examination for the degree of Doctor of Philosophy as of July 25 2025. This is the final and approved form of the dissertation.

Some New Advancements in Adaptive Randomization for Clinical Trials

Guannan Zhai

Dissertation Research Committee:

Feifang Hu, Professor of Statistics, Dissertation Director

Hua Liang, Professor of Statistics, Committee Member

Xiaoke Zhang, Associate Professor of Statistics, Committee Member

© Copyright 2025 by Guannan Zhai
All rights reserved

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Feifang Hu, who has been the most influential person in my academic journey since arriving in the United States. From taking his class in my first year to becoming his student, I have always felt incredibly fortunate to work under his guidance. Professor Hu gave me the freedom to explore my research interests while patiently guiding me through each step of the process.

I still remember my first summer as his research assistant, working on what seemed like a simple R package. This project helped me gradually enter the field and build confidence. Over the following years, under his mentorship, I transitioned from collaborative projects to independent research. Though the journey was challenging at times, Professor Hu made it feel well-structured and smooth. He provided numerous opportunities for growth, including research positions, internships, and speaking engagements.

Outside of academics, Professor Hu is also very inspiring. He's great at golf and skiing, and he maintains a youthful, optimistic, and easygoing attitude. His positive outlook has influenced me deeply, teaching me that happiness can be actively cultivated. He also fosters a strong sense of community in the Hu Group through various activities, making it feel like home.

In addition to Professor Hu, I would also like to sincerely thank my thesis advisory and examination committee members: Professor Zhaohai Li, Professor Hua Liang, Professor Qing Pan, Professor Xiaoke Zhang, and Professor Zhengling Qi. I am grateful for their careful review and constructive feedback. Special thanks to Professor Hua Liang and Professor Xiaoke Zhang for their valuable suggestions during my proposal stage. I also

benefited greatly from participating in Professor Xiaoke Zhang's reinforcement learning reading group.

My gratitude extends to Professor Tapan K. Nayak and Professor Hosam M. Mahmoud for their excellent instruction in foundational courses STAT 8263 and STAT 8257, which provided a strong foundation for my later research. I also want to give special thanks to my internship mentor and coauthor, Hongjian Zhu, for his valuable guidance and continuous support during my internship and our collaboration.

I want to thank my classmates and friends who made this journey memorable: Yang Liu, for helping me learn to write clear and compelling papers; Zexin Ren and Zixuan Zhao, for insightful discussions; Norah Alkhnefr, for being a wonderful collaborator; Fengyu Zhao, Jialu Wang, Jiaqian Yu, and Renjie Luo, for making conference trips unforgettable; Yawen Wu, for her friendship; and my cohort members Jilei Lin, Yipeng Wei, and Vikki Sui, for their constant companionship and support. Also, special thanks to my dear friend Geyue Sun. Finding someone who truly understands you during a Ph.D. program is rare and wonderful. Her companionship and emotional support meant everything to me. Thanks also to my friends Chong Dang and Ruixin Zheng for always being there whenever I needed help. I cherish all our time and conversations together in DC.

I would like to give the deep and special thanks to my parents, Yanli Yang and Yibing Zhai, for their unconditional support and respect. Their trust gave me the courage to express myself freely, pursue my dreams fearlessly, and grow into an independent person. My father, an incredibly determined and optimistic person, has inspired me with his discipline—quitting smoking and embracing exercise six years ago despite a busy career. He continues to maintain a healthy, balanced life filled with fishing, running, swimming,

and socializing. My mother, Yanli, is independent and passionate about life. We share a love for painting, though she has real talent while I simply enjoy appreciating her art. She is my best friend; we always enjoy each other's company, and I always feel completely accepted around her. From both of them, I have learned the value of living life richly from within. I'm also grateful to my two cute cats, Tangdou Zhai and Naigao Zhai, who joined my life in 2018, providing countless moments of joy and comfort.

I also want to thank my boyfriend, Rentian Yao, whom I met at UIUC. He helped deepen my understanding of statistical research and is always there during difficult moments, patiently listening and providing support.

I'm truly grateful for everyone who made this journey memorable and fulfilling.

Abstract of Dissertation

Some New Advancements in Adaptive Randomization for Clinical Trials

Clinical trial methodology continues to evolve as researchers strive to enhance efficiency, ethics, flexibility, and accuracy in response to the growing complexity and stringent regulatory requirements of modern clinical research. Despite these important methodological advances, significant practical challenges persist. Among the most pressing concerns are the development of trial designs that can adapt effectively to real-world conditions, inference methods capable of accommodating diverse data types, and statistical approaches that align closely with evolving regulatory standards. These ongoing issues highlight a gap between theoretical developments and practical application in clinical research. This dissertation addresses these critical gaps by identifying key limitations in current clinical trial methods and proposing practical, theoretically grounded solutions. The contributions presented here are intended not as incremental improvements but as meaningful advances capable of transforming clinical trial practice to better serve both researchers and patients.

In Chapter 3, we focus on Response-adaptive Randomization (RAR). Traditional RAR requires updates after every individual patient response, which is often impractical in real-world clinical trials. To address this limitation, we introduce new group-based RAR procedures. These methods update randomization periodically, either in groups or at fixed intervals like weekly or biweekly. Our theoretical analyses show these group-based methods maintain the strengths of traditional doubly adaptive biased coin designs (Hu and Zhang (2004)) and effectively handle delayed or missing responses. We illustrate the practicality and effectiveness of this approach through a real clinical trial example.

Chapter 4 explores statistical inference challenges in covariate-adaptive randomization (CAR) for multi-arm trials with general outcomes. Most existing inference approaches address continuous outcomes, leaving a notable gap for common trial endpoints such as binary or count data. We assess the performance of the standard Wald test under model misspecification, highlighting problems like inflated or overly conservative Type I error rates due to omitted covariates and correlation among tests from shared control groups. We propose adjusted test statistics that effectively control Type I error and may improve statistical power. This research represents the first comprehensive study on inference for multi-arm CAR trials with GLM endpoints.

In Chapter 5, we address seamless phase II/III clinical trial designs, highly relevant in oncology due to their potential to accelerate drug development. Current seamless methods predominantly target continuous endpoints, which do not align with oncology trials often relying on time-to-event primary outcomes and binary surrogate endpoints during interim stages. We propose a covariate-adaptive seamless phase II/III trial design specifically for trials using binary surrogate endpoints for interim analyses and time-to-event outcomes for final assessments. The accompanying inference method we introduce enhances statistical power, remains robust against model misspecification, and reliably controls Type I error. Simulation studies validate the practical effectiveness and regulatory suitability of our approach, emphasizing its usefulness for oncology trials.

Overall, these methodological improvements contribute significantly to clinical trial practice, addressing key limitations in current approaches. The proposed designs and inference methods enhance the practicality, adaptability, and accuracy of clinical trials, paving the way for more efficient drug development, improved regulatory alignment, and

better patient outcomes across various therapeutic fields.

Table of Contents

Acknowledgments . . . . . **iv**

Abstract of Dissertation . . . . . **vii**

List of Tables . . . . . **xiv**

1 Introduction **1**

1.1 Background 1

1.2 Overview of Adaptive Randomization 2

1.2.1 Response-adaptive Randomization 2

1.2.2 Covariate-adaptive Randomization 3

1.3 Practical Gap 3

1.4 Research Goal & Contributions 5

2 Literature Review **7**

2.1 Traditional Randomization: Complete Randomization (CR) 7

2.2 Response-adaptive Randomization (RAR) 8

2.3 Covariate-adaptive Randomization (CAR) 11

2.4 Statistical Inference Under CAR 15

3 Group Response-Adaptive Randomization with Delayed and Missing Response . . . **18**

3.1 Group DBCDs and their properties 18

3.1.1 Framework 18

3.1.2 Asymptotic properties 21

3.2 Group DBCD with delayed and missing responses 24

3.2.1 Framework 25

3.2.2 Asymptotic Results 28

3.3 Numerical studies and application 31

3.3.1 Some numerical studies . . . 31
3.3.2 Re-designing the HIV transmission trial . . . 34
3.4 Discussion . . . 37
3.5 Appendix . . . 38
3.5.1 Conditions . . . 38
3.5.2 Lemmas . . . 38
3.5.3 Proof of the Main Theroems . . . 45

4 Valid Test for Multi-arm Trials with GLM endpoints under CAR . . . 50
4.1 Introduction . . . 50
4.2 Test under Generalized Linear Model . . . 52
4.2.1 Framework and Notation . . . 52
4.2.2 Working Model and Test . . . 53
4.2.3 Control of Family-wise Type I error Rate under Multiplicity . . . 55
4.3 Main Results . . . 57
4.3.1 Main Theorem . . . 58
4.3.2 Test Size under Some Commonly Used Models . . . 60
4.3.3 Additional Results with non-canonical link . . . 62
4.4 Constructing Valid Test Statistics . . . 64
4.5 Numerical Studies . . . 67
4.6 Analysis of Metastatic Breast Cancer Trial . . . 71
4.7 Discussion . . . 74
4.8 Appendix . . . 76
4.8.1 Ancillary Results . . . 76
4.8.2 Proof of the Main Theorems . . . 79
4.8.3 Additional Working Variance Analysis in GLMs . . . 86

5 Robust Test in Seamless Phase II/III Design under CAR . . . 90

5.1 Background 90
5.2 Notation 92
5.3 Framework of Design 94
5.4 Framework of Proposed Inference 96
5.4.1 Inference in Dose Selection Stage 96
5.4.2 Inference in Confirmatory Stage 98
5.5 Numerical Studies 99
5.6 Discussion 104

6 Conclusion & Future Work 105

Bibliography 106

List of Tables

3.1 Power and allocation proportion for binary cases with 5000 replications (RSIHR target) 32

3.2 Delayed and Missing Response: Power and allocation proportion for binary cases with 5000 replications (RISHR target) 33

3.3 Power and allocation proportion for continuous cases with 5000 replications (Zhang and Rosenberger's target) 34

3.4 Delayed and Missing Response: Power and allocation proportion for continuous cases with 5000 replications (Zhang and Rosenberger's target) 34

3.5 The Re-designed HIV Transmission Trial with Original Sample Size $(n = 478)$ 35

4.1 Summary of Type I Error Behavior Across GLMs 63

4.2 Simulated Type I error for Logistic Regression 68

4.3 Simulated Type I error for Poisson Regression 68

4.4 Simulated Type I error for Exponential Model 68

4.5 Simulated power for logistic regression with N = 300 70

4.6 Simulated power for Logistic regression with N = 300 70

4.7 Simulated power for Poisson regression with N = 300 70

4.8 Simulated power for exponential model with $N = 300$ 71

4.9 Re-designed Metastatic Breast Cancer Trial (N = 1000) 74

5.1 Empirical power and No. futility for Stratifies Per-mutated block randomization under Wald test and adjusted test 100

5.2 Empirical power comparison of 4 designs based on stratified permuted block randomization with discrete covariates 102

5.3 Empirical power comparison of 4 designs based on Hu & Hu covairates adaptive randomization with discrete covariates 102

5.4 Empirical power comparison of 4 designs based on stratified permuted block randomization with continuous covariates 103

5.5 Empirical power comparison of 4 designs based on Hu & Hu covairates adaptive randomization with continuous covariates 103

CHAPTER 1

Introduction

1.1 Background

The randomized controlled trial (RCT) is widely recognized as the gold standard for evaluating medical interventions, because it effectively reduces bias and establishes clear causal relationships. Traditional RCTs use fixed designs with predetermined sample sizes and complete randomization (CR), with analysis performed only after participants complete their follow-up. Although historically effective, these methods often lack flexibility. Such rigidity may unnecessarily expose participants to inferior treatments and delay the identification of superior therapies. Additionally, fixed designs often fail to adequately account for patient heterogeneity, potentially compromising trial validity.

Adaptive randomization addresses these limitations through various strategies, including response-adaptive, covariate-adaptive, and covariate-response-adaptive randomization procedures. These approaches dynamically adjust the allocation probabilities based on accumulated treatment outcomes, patient characteristics, or a combination of both. Such flexibility allows trials to respond in real-time to accumulating evidence, thereby enhancing efficiency and improving statistical power.

Recent high-impact clinical trials demonstrate significant ethical, operational, and statistical advantages of adaptive randomization. For instance, the REMAP-CAP trial [2], conducted during the COVID-19 pandemic, employed response-adaptive randomization (RAR) to efficiently identify beneficial treatments, rapidly informing clinical care and resource allocation. Additionally, Ciolino et al. [10] conducted a systematic review of 298 RCTs published in leading medical journals from 2009 to 2014, reporting that 237 (80%) employed covariate-adaptive randomization (CAR), with stratified block randomization accounting for 69% and other CAR methods for 11%. These approaches improved baseline covariate balance, reduced confounding, and increased the accuracy

of trial results.

Thus, adaptive randomization is a meaningful advancement in clinical trial methodology. Combined with rigorous statistical inference, it optimizes patient care, resource allocation, and scientific discovery.

1.2 Overview of Adaptive Randomization

Having outlined the limitations of traditional RCT designs, we now introduce adaptive randomization methods designed to overcome these issues. Traditional RCTs utilize CR, assigning participants to treatment arms with fixed, pre-specified probabilities (typically a 1:1 ratio), regardless of accumulating data. Adaptive randomization relaxes this constraint by incorporating accumulating responses or patient characteristics to dynamically update allocation rules. Below, we briefly describe two distinct classes of adaptive randomization procedures, each particularly advantageous under different scenarios.

1.2.1 Response-adaptive Randomization

One key adaptive randomization method addressing the limitations of CR is RAR, which was developed to allocate participants to treatment that appears superior as trial data accumulate. Broadly, existing RAR methods fall into two categories: randomized urn-based models and sequential estimation procedures [19]. Urn-based methods intuitively assign participants to better-performing arms, but typically fail formal optimality criteria for statistical power or patient benefit. To address this gap, a newer sequential estimation method first defines optimal allocation targets by solving optimization problems that balance efficiency and ethical considerations. Then, it sequentially adjusts allocations, guiding actual treatment proportions toward these optimal targets as the trial progresses.

However, a significant practical challenge remains: traditional RAR procedures require sequential outcome observation before subsequent allocations. Consequently, delays in outcome reporting

significantly extend trial durations. This critical limitation is currently a primary focus in RAR research.

1.2.2 Covariate-adaptive Randomization

Another important adaptive randomization method is CAR, which emphasizes balancing prognostic baseline variables (e.g., age, disease stage, biomarkers) rather than observed outcomes. After each participant enrollment, CAR algorithms evaluate current covariate imbalances across treatment arms, adjusting assignment probabilities accordingly. Commonly used CAR techniques include stratified permuted block randomization, Pocock and Simon's minimization, stratified biased-coin designs, and advanced frameworks like the Hu and Hu (2012) approach. Each approach possesses unique strengths and limitations, detailed further in Section 2.

Significant research has concurrently targeted improving statistical inference methods under CAR. Standard statistical tests, relying on pooled variance estimates, often yield overly conservative results due to induced correlation between treatment assignments and covariates. Robust variance estimators, permutation tests, and bootstrap methods have been developed to correct this issue, maintaining nominal type I error rates and enhancing statistical power by improving covariate balance and reducing residual variance.

Despite these advancements, statistical inference remains challenging in complex trials involving multiple treatment arms or diverse endpoints, such as binary or time-to-event outcomes. Ensuring inference robustness and feasibility in these settings remains an ongoing research priority.

1.3 Practical Gap

As highlighted in the preceding sections, methodological advances in both RAR and CAR have significantly improved clinical trial designs. Nonetheless, critical practical gaps remain unresolved, limiting their widespread adoption and effectiveness in realistic clinical settings.

For RAR, the most substantial practical challenge is its inherent reliance on sequential en-

rollment and outcome observation. Hence, RARs require timely availability of patient outcomes before subsequent patients can be randomized. This sequential dependency substantially prolongs trial duration, particularly when outcomes are delayed due to lengthy follow-ups or logistical complexities in multi-center trials. The extended duration increases operational burdens, and poses substantial challenges to timely clinical decision-making, limiting broader practical adoption.

Similarly, despite substantial methodological progress, CAR faces distinct but related practical challenges when trials grow more complex. Specifically, in multi-arm trials, dose-ranging studies, and seamless trials, existing inference methods frequently encounter substantial limitations. These limitations arise primarily because traditional statistical tests, designed initially for straightforward two-arm trials with simple endpoints, become overly conservative or inaccurate when used in more complex settings. In particular, when outcomes include generalized linear model (GLM)-based endpoints, count data, binary events, or long-term survival outcomes, current methods often inadequately adjust for correlations induced by adaptive randomization. As trials increasingly integrate complex endpoints, the need for reliable statistical inference methods becomes more pressing, both practically and from a regulatory perspective.

Moreover, from an implementation standpoint, complex randomization algorithms require additional transparency and ease of use. Clinical teams and regulatory agencies demand clear documentation and interpretability of the allocation process. Consequently, bridging these practical gaps, particularly the reduction of trial duration in RAR and the development of robust and adaptable inference methods for CAR in complex scenarios, remains critical. Addressing these issues through innovations such as group-based randomization strategies and robust inference methods for complex endpoints, as explored in this dissertation, would significantly advance adaptive randomization practices and their practical applicability.

1.4 Research Goal & Contributions

The primary aim of this dissertation is to improve adaptive randomization methods by addressing key practical issues that limit their broader use in clinical trials. Previous studies established foundational approaches, yet existing methods often face practical challenges, including delayed responses, missing data, complex covariates, generalized outcomes, and complex trial structure. This dissertation focuses on bridging these gaps by providing novel approaches that enhance both statistical rigor and practical usability.

The main contributions distinguishing this research from prior work include (1) developing a novel class of group-based response-adaptive randomization that accommodates group enrollment and effectively handles delayed and missing responses; (2) proposing robust inference techniques for multi-arm clinical trials using covariate-adaptive randomization with generalized linear model endpoints; and (3) establishing a comprehensive inferential framework specifically for seamless phase II/III trials involving multiple endpoints.

In Chapter 3, we develop a new group response-adaptive randomization (group RAR) procedure capable of simultaneously addressing delayed responses and missing data under the missing-at-random (MAR) mechanism. Unlike existing sequential RAR methods, the proposed approach allows randomization by patient groups rather than individual patients, significantly reducing trial duration and logistical complexity. We establish its theoretical properties and validate practical performance through extensive simulations.

In Chapter 4, we propose robust inference procedures for multi-arm clinical trials utilizing CAR. The primary contribution of this project is extending statistical inference beyond standard continuous outcomes to endpoints modeled via GLMs. We systematically investigate how CAR affects parameter estimation and hypothesis testing under GLMs, establish theoretical properties such as asymptotic variance, and develop adjusted inference methods to maintain nominal type I error rates and potentially increase statistical power. Numerical studies and clinical trial examples demonstrate the superiority and robustness of our proposed methods compared to traditional

inference approaches.

In Chapter 5, we propose a statistical inference framework for seamless phase II/III clinical trials using CAR designs. Specifically, we focus on trials involving multiple endpoints, combining short-term binary outcomes and long-term time-to-event (survival) endpoints. Our approach addresses the practical challenges associated with adaptive allocation by explicitly accounting for dependencies among covariates, endpoints, and treatment assignments. We outline a practical inference framework emphasizing robust testing strategies designed to maintain type I error control and ensure sufficient statistical power in complex trial scenarios. Extensive simulation studies and real-world examples validate our proposed methods, providing robust evidence and practical guidance essential for clinical teams and regulators to confidently adopt seamless adaptive trial designs.

The rest of the dissertation is organized as follows. To systematically address these practical gaps and position our methodological innovations effectively, Chapter 2 first provides an extensive literature review detailing foundational concepts, existing randomization methods, and inference strategies under CAR. Chapters 3 to 5 introduce our proposed methodologies to address the practical gaps identified earlier, specifically: group randomization for delayed responses (Chapter 3), robust inference in multi-arm GLM-based trials (Chapter 4), and seamless phase II/III trial designs (Chapter 5). Each chapter presents theoretical developments, simulation studies, and illustrative examples. Finally, Chapter 6 summarizes the findings, discusses their practical implications, and highlights potential areas for future research.

CHAPTER 2

Literature Review

This literature review provides comprehensive theoretical background for the three research projects presented in this dissertation. The review begins with traditional complete randomization (Section 2.1) as the foundational framework, then progresses through response-adaptive randomization methods (Section 2.2) and covariate-adaptive randomization approaches (Section 2.3). Finally, Section 2.4 specifically addresses statistical inference under CAR, a topic particularly relevant to the projects presented in Chapters 4 and 5.

Each section traces the historical development of these methods and identifies their current limitations and practical challenges. The organization mirrors the logical progression from basic concepts to advanced methodologies, ultimately connecting to the three specific research gaps addressed in this dissertation: group-based RAR for delayed and missing responses (Chapter 3), robust inference in multi-arm trials with GLM-based endpoints under CAR (Chapter 4), and robust inference methods for seamless phase II/III designs with multiple endpoints (Chapter 5). This comprehensive foundation enables readers to understand both the theoretical motivation and practical necessity for the proposed methodological innovations.

2.1 Traditional Randomization: Complete Randomization (CR)

We first introduce CR, the simplest and widely used allocation method in RCTs, as a foundational concept. CR independently allocates participants with fixed probabilities, typically equal across treatment arms (e.g., a probability of 0.5 for two-arm trials). Because each assignment is independent, CR essentially corresponds to repeated Bernoulli trials with known probabilities of success

[44]. Formally, let T_{ik} indicate the treatment assignment for the i th patient in treatment arm k :

$$T_{ik} = \begin{cases} 1 & \text{if patient } i \text{ is assigned to arm } k \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

Under CR, the assignment probabilities satisfy:

$$P\left(T_{ik} = 1\right) = p_k, \quad \text{where} \quad p_k = \frac{1}{K}$$

with K representing the total number of treatment arms. The key advantages of CR are its simplicity and strong theoretical guarantees of unbiasedness. The independence of assignments prevents selection bias, while the law of large numbers ensures that, in sufficiently large samples, treatment groups will on average be balanced with respect to both observed and unobserved covariates [28, 44]. Regulators are highly familiar with this design, and most classical statistical tests assume an underlying CR mechanism.

However, the same independence that guarantees unpredictability also means CR does not actively control covariate balance in finite samples. Chance imbalance on important prognostic factors can inflate variance and reduce statistical power, especially when total sample size is modest [53, 28]. In multi-arm studies, CR further allows substantial variation in arm-wise sample sizes, again hurting efficiency. Finally, pure CR offers no intrinsic protection against exposing many early participants to a potentially inferior treatment, motivating the development of adaptive schemes.

2.2 Response-adaptive Randomization (RAR)

Given these limitations of CR, particularly ethical concerns about exposing patients to inferior treatments, RAR was developed. RAR dynamically adjust the probability of assigning patients to different treatment arms based on accumulating outcome data. Over time, this method allocates more participants to treatments showing better performance, directly addressing ethical concerns

[44]. As we mentioned in section 1.2.1, RAR approaches include intuitive urn-based methods and optimal designs that explicitly balance ethical considerations and statistical efficiency.

Among RAR methods, urn-based procedures have been extensively studied due to their intuitive appeal and straightforward implementation. A classic example is the RPW design [56], an extension of Zelen's original play-the-winner scheme [59], incorporating probabilistic treatment assignments. RPW adaptively assigns treatments in a two-arm trial using a simple urn model. Initially, the urn contains an equal number of balls representing each treatment. To assign a treatment, a ball is randomly drawn and then replaced. If a patient responds successfully to the assigned treatment, an additional ball of the same type is added, increasing its probability of future selection. Conversely, if the response is unsuccessful, a ball of the opposite treatment type is added, balancing probabilities. Over time, this approach favors assigning more patients to the more effective treatment.

However, despite their intuitive simplicity, urn-based designs, including RPW, face notable limitations. RPW specifically exhibits instability, as early random fluctuations significantly influence subsequent allocations, potentially causing extreme imbalances. A real-world example is the ECMO trial conducted in 1985 [3]. The trial resulted in a severe imbalance: 11 infants assigned to the ECMO group all survived, while only one infant assigned to the control group died. This severe imbalance prompted ethical concerns and methodological critiques. Such sensitivity can lead to misleading assessments of treatment performance and compromise statistical power [45].

To overcome these weaknesses, several variants of urn-based designs have been proposed. For instance, Tamura et al. [52] conducted a notable trial in depression using an RPW design with a "burn-in" period, initially fixing allocations to ensure a minimum number of control patients before switching to adaptive mode. Ivanova [26] introduced the drop-the-loser urn design to address instability by systematically reducing the likelihood of assigning poorly performing treatments. Another significant improvement is the generalized Pólya urn (GPU) model, incorporating adjustable parameters for better control over allocation proportions, thereby enhancing stability and balance. These procedural advancements mitigate some of RPW's vulnerabilities. Nevertheless, these extensions were not specifically designed to optimize any statistical criterion; their primary goal was placing

more patients on superior treatments. Unfortunately, many remain slow to converge and produce less powerful treatment comparison tests. A comprehensive historical perspective on this topic is discussed in Rosenberger and Lachin (2002) [43].

Considering these limitations, researchers in the early 2000s began developing optimization-based adaptive methods, which explicitly balance ethical concerns and statistical efficiency within a rigorous mathematical framework. Hu and Zhang (2004) developed the doubly biased coin design (DBCD), a general adaptive randomization method that targets pre-specified optimal allocation ratios based on a chosen optimization criterion, such as maximizing statistical power, minimizing variance, or enhancing ethical considerations by preferentially assigning patients to superior treatments [21]. Formally, let $\rho = (\rho_1, \rho_2, \ldots, \rho_K)$ represent the target allocation proportions for K treatment arms, and let $N_{i,k}$ denote the number of patients assigned to arm k after enrolling i patients. The allocation probability for the $(i+1)$th patient being assigned to treatment k under DBCD is given by:

$$P(T_{i+1} = k) = \frac{\hat{\rho}_k(\hat{\rho}_k/(N_{i,k}/i))^\alpha}{\sum_{j=1}^{K} \hat{\rho}_j(\hat{\rho}_j/(N_{i,j}/i))^\alpha}, \tag{2.2}$$

where $\alpha \geq 0$ is a tuning parameter controlling the randomness and sensitivity of the allocation. Larger α values yield more deterministic allocations aligned with the targeted proportions, while smaller values introduce greater randomness.

Efficient response-adaptive designs (ERADE) further refine allocation rules and have been shown to minimize the asymptotic variance of the estimated treatment effect, while maintaining ethically motivated allocation constraints. ERADE can be viewed as an advancement of the DBCD, achieving optimal precision for a targeted allocation ratio [24].

Building upon the optimization-based frameworks, recent approaches further integrated response-guided and non-myopic perspectives. Response-guided designs adjust treatment allocations based on interim data, enhancing ethical allocation and potentially reducing the required sample size. Zhu and Hu [66, 67] provided rigorous proofs confirming convergence and proper type I error control for sequential tests in DBCD and urn models. Non-myopic methods, inspired by multi-armed bandit theory, consider the full trial horizon, rather than just the next patient allocation. For

instance, the forward-looking Gittins index approach proposed by Villar et al. (2015) optimizes overall patient outcomes throughout the trial duration, preserving robust statistical performance. These innovations have significantly strengthened the ethical and statistical foundations of RARs.

Nevertheless, several methodology-specific practical challenges persist in RAR implementations. Most existing RAR methods require observing previous participants' outcomes before allocating the subsequent patient. In trials involving endpoints with long follow-up periods, such as survival or chronic disease progression, this sequential dependence can stall enrollment and substantially prolong trial duration, as real-time updating becomes impractical until outcomes become available. Hu et. al. (2008) [23] examined the impact of delayed responses on DBCDs, finding these designs to be relatively robust and insensitive to delays under commonly satisfied conditions. Despite this theoretical robustness, practical operational complexities remain significant. Implementing RAR in real-time imposes considerable demands on data capture, centralized coordination, and frequent updating of randomization procedures. These operational complexities significantly strain resources and increase logistical burdens on trial teams. Addressing these issues necessitates the development of smoothing or stabilization methods for allocation updates. Finally, advanced adaptive procedures, particularly those utilizing Bayesian or multi-armed bandit frameworks, may face regulatory scrutiny due to perceived complexity and lack of transparency. Thus, clearly specified and thoroughly documented adaptation rules are critical for ensuring regulatory acceptance and reproducibility. These detailed algorithm-specific difficulties directly motivate the novel group-enrollment-based RAR procedure proposed in Chapter 3, explicitly designed to handle delayed and missing responses encountered in clinical trials.

2.3 Covariate-adaptive Randomization (CAR)

While RAR addresses ethical concerns through dynamic allocation based on treatment outcomes, CAR methods specifically target balancing patient characteristics across treatment groups. Imbalances in important prognostic covariates, especially in smaller trials, can reduce statistical power

and complicate interpretation. CAR dynamically adjusts assignment probabilities using available baseline covariate information, thus effectively ensuring balance. Due to their simplicity and effectiveness, CAR approaches have been widely adopted since the 1970s and remain common in clinical trials today.

The simplest approach to balance covariates is stratified permuted block randomization (SPB) [60]. In stratified randomization, patients are grouped into strata defined by R key covariates, forming $S = \prod_{r=1}^{R} m_r$ strata, where the rth covariate has m_r discrete levels. Within each stratum defined by the covariate profile $(m_1^*, m_2^*, \cdots, m_R^*)$, where $m_r^* \in \{1, \cdots, m_r\}$ for each covariate r $(1 \leq r \leq R)$, permuted blocks of predetermined sizes ensure nearly equal allocation of patients to treatment arms. Mathematically, the within-stratum imbalance for a specific treatment arm k can be represented as:

$$D_k\left(m_1^*, m_2^*, \cdots, m_R^*\right) = n_k\left(m_1^*, m_2^*, \cdots, m_R^*\right) - \bar{N}\left(m_1^*, m_2^*, \cdots, m_R^*\right), \tag{2.3}$$

where $n_k\left(m_1^*, m_2^*, \cdots, m_R^*\right)$ denotes the number of patients allocated to treatment arm k within that stratum, and

$$\bar{N}\left(m_1^*, m_2^*, \cdots, m_R^*\right) = \frac{1}{K} \sum_{k'=1}^{K} n_{k'}\left(m_1^*, m_2^*, \cdots, m_R^*\right)$$

is the average number of patients across all treatment arms within that stratum.

SPB guarantee balance on the stratifying factors within each block of a fixed size, and thus overall balance if the number of patients per stratum is large. However, when there are many stratification factors or levels, the number of possible strata grows combinatorially and many strata may have few patients. In such cases, stratified blocking may fail to achieve good overall balance, and some strata will by chance be unfilled or uneven. This is a known limitation: with a large number of strata, the method does not scale well and can leave substantial imbalance in combined covariate distributions.

To overcome the drawbacks of SPB, minimization was introduced independently by Taves (1974) and further refined by Pocock and Simon (1975). Minimization is a flexible randomization

approach designed explicitly to control marginal covariate imbalances across treatment arms. Unlike SPB, minimization dynamically allocates new patients by assessing marginal imbalance scores for each covariate separately.

Mathematically, the marginal imbalance for treatment arm k concerning the rth covariate at level m_r^* is expressed as:

$$D_k\left(r; m_r^*\right) = n_k\left(r; m_r^*\right) - \bar{N}\left(r; m_r^*\right) \tag{2.4}$$

where $n_k(r; m_r^*)$ is the count of patients with the rth covariate at level m_r^* allocated to treatment k, and

$$\bar{N}\left(r; m_r^*\right) = \frac{1}{K}\sum_{k'=1}^{K} n_{k'}\left(r; m_r^*\right) \tag{2.5}$$

is the marginal average across treatments.

In minimization, the next patient is typically allocated to the treatment that minimizes a weighted sum of these marginal imbalances. Pocock and Simon emphasized adding a probabilistic element to this allocation (commonly assigning the minimizing treatment with a high probability, such as 0.8 or 0.9) to mitigate predictability and selection bias, maintaining balance while retaining randomness.

As trials began to incorporate more covariates (e.g. genetic biomarkers, multi-center indicators, etc.), the limitations of previous CAR methods motivated new designs. One key development was the formulation of CAR designs with formal probability models, allowing researchers to prove asymptotic properties. In an influential paper, Hu and Hu (2012) proposed a novel covariate-adaptive randomization procedure (HuHu CAR) that can control multiple types of imbalances simultaneously. They pointed out two critical shortcomings of earlier methods: (1) Stratified block randomization fails when there are a large number of strata (as discussed), and (2) traditional minimization can still lead to large within-stratum imbalances meaning that even if marginal totals are balanced, finer subgroup balances might not be. HuHu CAR design tackles these issues via assigning patients sequentially by minimizing a weighted imbalance score, combining overall imbalance (D_k), marginal imbalance ($D_k(r; m_r^*)$), and within-stratum imbalance ($D_k(m_1^*, \ldots, m_R^*)$).

The overall imbalance D_k is defined as:

$$D_k = n_k - \bar{N}$$

Marginal imbalance and within-stratum imbalance are calculated similarly, as shown above. Formally, the imbalance score for assigning the ith patient to treatment arm k is computed as:

$$\mathrm{Imb}_i^{(k)} = w_o \left\{ \sum_{k'=1}^{K} \left(D_{i,k}^{(k')} \right)^2 \right\} + \sum_{r=1}^{R} w_{m,r} \left\{ \sum_{k'=1}^{K} \left(D_{i,k}^{(k')} \left(r, m_r^* \right) \right)^2 \right\} + w_s \left\{ \sum_{k'=1}^{K} \left(D_{i,k}^{(k')} \left(m_1^\star, \ldots, m_R^* \right) \right)^2 \right\} \tag{2.6}$$

where ω_o, $\omega_{m,r}$, and ω_s are non-negative weights satisfying $\omega_o + \sum_{r=1}^{R} \omega_{m,r} + \omega_s = 1$. This formulation provides robust covariate balancing across all dimensions of interest, enhanced by rigorous theoretical guarantees of convergence and optimal allocation properties. For detailed procedures, see Hu and Hu (2012) for two-arm trials (Annals of Statistics, 40(3): 1794–1815) or Hu, Ye, and Zhang (2023) on multi-arm trials (Science China Mathematics, 66(1): 163–190).

Modern CAR methods have evolved from ad hoc algorithms into well-developed designs supported by rigorous theoretical frameworks, including guarantees of statistical stability like Markov chain properties and sufficient randomness to prevent predictability and selection bias. CAR is now commonly integrated into practical tools such as web-based randomization platforms and statistical software packages like carat [39], facilitating broad adoption in clinical trials. Recent literature surveys confirm widespread use of CAR techniques, particularly minimization and stratification, to maintain covariate balance. This is especially beneficial in trials with smaller sample sizes or multiple influential prognostic factors, substantially enhancing statistical power and credibility of trial outcomes.

2.4 Statistical Inference Under CAR

As CAR methods have become common in clinical trials, researchers have paid increasing attention to statistical inference after CAR. Early studies by Birkett [6] and Forsythe [13] showed through simulations that standard tests, such as the two-sample t-test or regression models without covariate adjustment, are often overly conservative. Shao et al. [49] later provided theoretical evidence that CAR methods introduce correlation among treatment assignments, changing the distribution of unadjusted test statistics. This correlation reduces variability between groups, leading to Type I error rates lower than the nominal level. Further theoretical studies [34, 36, 30] confirmed that incorporating covariates used during randomization into regression models can restore correct inference and improve statistical efficiency.

Current statistical inference methods under CAR are generally classified into two categories: model-free and model-based approaches. The model-free approaches typically use t-tests with adjusted standard errors or permutation tests conditioned on strata, providing asymptotically valid inference across diverse CAR designs. For instance, Bugni et al. [8] proposed t-tests with adjusted standard errors and developed regression-based methods that include strata indicators as covariates. However, these methods can become impractical when the number of strata is large. Additionally, permutation tests typically have high computational complexity. More recently, model-based inference methods have emerged as particularly attractive alternatives, as they remain valid even if the underlying statistical model is misspecified. Ma et al. (2015) [34] studied two-arm trials under CAR, focusing on scenarios where randomization covariates are partially or completely omitted in analysis. They showed that traditional model-based inference becomes conservative and proposed adjusted variance estimators to correct this issue. Following this, Li et al. (2021) [31] explored the behavior of Wald tests under GLMs in similar two-arm trial settings, providing modified variance estimators to maintain valid inference when covariates are omitted. Ye and Shao (2020) [57] further derived the asymptotic properties of partial likelihood score tests under potential model misspecification, offering a variance estimation adjustment to ensure robustness across a broad

class of CAR designs in two-arm trials.

Recent methodological extensions further consider more complex clinical scenarios, such as multi-arm trials, diverse endpoint types, dose-escalation studies, and seamless designs. For instance, Bugni et al. (2019) [9] generalized previous theoretical results to multiple-treatment scenarios, developing adjusted variance estimators that ensure valid inference under CAR. Additionally, Ma et al. (2020) [36] considered CAR in seamless phase II/III trials, proposing adjusted inference procedures that maintain correct Type I error control and efficiency despite the complexities introduced by interim adaptations and dose escalation. These recent developments broaden the applicability and reliability of CAR inference methods in increasingly complex trial designs.

Despite these advances, significant methodological gaps remain, particularly for multi-arm trials and seamless phase II/III trials involving non-continuous endpoints. A crucial challenge is statistical inference for multi-arm trials with endpoints modeled by GLMs, such as binary responses or count data, commonly encountered in clinical research. Although methods for two-arm GLM trials under CAR are well-established [48, 31], extending them to multi-arm contexts remains essential yet challenging. Multi-arm trials typically use a shared control group, introducing dependencies that violate assumptions of traditional inference methods. Additionally, analyzing CAR data with GLM endpoints often involves omitting covariates from working models, potentially invalidating standard inference approaches.

Similarly, existing research on seamless designs commonly assumes the same endpoint for dose selection and final analysis across trial phases [5, 32, 38], a notable limitation for oncology trials where using survival endpoints for both stages significantly prolongs trial duration [41].

To directly address these methodological gaps, this dissertation undertakes two targeted research projects: (1) Investigating asymptotic properties of standard Wald tests in multi-arm GLM trials under CAR, specifically handling cases of misspecified working models due to omitted covariates. Based on these theoretical findings, we propose adjusted tests to control Type I error effectively and potentially enhance statistical power. (2) Proposing a novel seamless phase II/III design integrating CAR, using binary surrogate endpoints for interim dose selection and survival endpoints

for final analysis, accompanied by corresponding robust inference procedures. These approaches substantially enhance the practicality and efficiency of clinical trials.

CHAPTER 3

Group Response-Adaptive Randomization with Delayed and Missing Response

As mentioned in the previous two sections, there is extensive literature on RAR procedures and their underlying theories. However, their practical application is hindered by the requirements of RAR, which require patients to enroll sequentially and update the allocation probability immediately after the results for each patient become available. In this section, we introduce our proposed group DBCD procedure, which allows for batch enrollment of patients and periodic updates of allocation probabilities, addressing the practical constraints associated with RAR.

3.1 Group DBCDs and their properties

3.1.1 Framework

In this section, we propose a new family of RAR procedures in which the patients are enrolled or observed in sequential groups. We first describe the framework of the new adaptive design and then discuss its asymptotic properties.

Consider a clinical trial with K treatment arms, where the total n participants enter the trial in groups sequentially, and their responses to the treatments are immediate. The sizes of the groups are assumed to be χ_1, χ_2, $\cdots$, which are random integers with $\chi_j \geq 1$. Each patient is assigned to one of the K treatments. We assume that the patient responses with each treatment are independent and identically distributed with the probability distribution indexed by a parameter $\boldsymbol{\theta}_k \in \mathscr{R}^d$, and denote $\boldsymbol{\xi}_{m,k}$ to be the response of the mth patient on treatment k. As usual, we let $\boldsymbol{X}_m = (X_{m,1}, \ldots, X_{m,K})$ be the allocation vector of the mth patient (i.e., $X_{m,k} = 1$ if the mth patient is assigned to treatment k, and 0 otherwise), and let $N_{n,k} = \sum_{j=1}^{n} X_{j,k}$ be the number of

patients assigned to treatment k, $k \in \{1, \cdots, K\}$.

In clinical trials, we aims to adjust the allocation proportion $N_{n,k}/n$ to converge towards the optimal allocation target by adapting the allocation probability for each group. The desired targets are functions of $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, which are derived as a solution to a formal optimization problem. For instance, the RSIHR targets, proposed by Rosenberger et al,[45] are derived to minimize the number of failures while maintaining a fixed power using Wald Z test statistics. In this paper, we implement Hu and Zhang's $g(\cdot)$ function [21] to push the allocation proportions toward the targets, and we outline the proposed Group DBCD procedure as follows.

(a) (Initial Step). To start, we assign the first one or two groups of patients to treatments via restricted randomization, ensuring a balance in treatment assignments. And let $\widehat{\boldsymbol{\Theta}}_0$ be an initial estimator of $\boldsymbol{\Theta}$.

(b) (Estimation Step) Assume that j groups of patients have been assigned to the trial, and their responses are observed. Let $\tau_j = \sum_{i=1}^{j} \chi_i$ be the number of patients assigned. When the $(j+1)$th group of patients is ready for allocation, the observed data include responses and assignments of previous τ_j patients, and the group sizes χ_1,..., χ_{j+1}. Let $\widehat{\boldsymbol{\Theta}}_{\tau_j}$ be the estimator of $\boldsymbol{\Theta}$ derived from these τ_j responses.

(c) (Allocation Step) We assign each patient in the $(j+1)$th group to treatment k with the probability

$$Pr_{j+1}^{G} = g_k\left(\frac{\boldsymbol{N}_{\tau_j}}{\tau_j}, \boldsymbol{\rho}\big(\widehat{\boldsymbol{\Theta}}_{\tau_j}\big)\right), \quad k = 1, \ldots, K.$$

That is

$$\begin{aligned}
&\Pr\left(\boldsymbol{X}_{m,k} = 1 \big| \chi_1, \ldots, \chi_{j+1}; \boldsymbol{\xi}_{i,l}, \boldsymbol{X}_i, i = 1, \ldots, \tau_j, l = 1, \ldots, K\right) \\
&\quad = g_k\left(\frac{\boldsymbol{N}_{\tau_j}}{\tau_j}, \boldsymbol{\rho}\big(\widehat{\boldsymbol{\Theta}}_{\tau_j}\big)\right), \quad k = 1, \ldots, K, \quad m = \tau_j + 1, \ldots, \tau_{j+1}.
\end{aligned}$$

Here $\boldsymbol{N}_{\tau_j}/\tau_j$ is the sample allocation proportion after the jth group assignments, and ρ is the desired

proportion which is a function dependent on the true parameters of the response distribution. The allocation function $\boldsymbol{g} = (g_1, \ldots, g_K)$ *in the procedure is for forcing the allocation proportion* $\boldsymbol{N}_{\tau_j}/\tau_j$ *close to the desired proportion* $\boldsymbol{\rho}$*. In Hu and Zhang (2004),[21] they had proposed a family of allocation functions to achieve this goal.*

$$g_k(x, y) = \frac{y_k \left(\frac{y_k}{x_k}\right)^{\gamma}}{\sum_{j=1}^{K} y_j \left(\frac{y_j}{x_j}\right)^{\gamma}}, \quad k = 1, \ldots, K$$

where $\gamma \geq 0$*.*

The specific form of $g_k(\cdot)$ *applied to this step is*

$$g_k\left(\frac{\boldsymbol{N}_{\tau_j}}{\tau_j}, \boldsymbol{\rho}\big(\widehat{\boldsymbol{\Theta}}_{\tau_j}\big)\right) = \frac{\boldsymbol{\rho}\big(\widehat{\boldsymbol{\Theta}}_{\tau_j,k}\big) \left(\frac{\boldsymbol{\rho}\left(\widehat{\boldsymbol{\Theta}}_{\tau_j,k}\right)}{\boldsymbol{N}_{\tau_j,k}/\tau_{j,k}}\right)^{\gamma}}{\sum_{i=1}^{K} \boldsymbol{\rho}\big(\widehat{\boldsymbol{\Theta}}_{\tau_j,i}\big) \left(\frac{\boldsymbol{\rho}\left(\widehat{\boldsymbol{\Theta}}_{\tau_j,i}\right)}{\boldsymbol{N}_{\tau_j,i}/\tau_{j,i}}\right)^{\gamma}}, \quad k = 1, \ldots, K.$$

(d) (Repeated Step) The estimation step and the allocation step are then repeated continuously until the end of the clinical trial.

When $\chi_j \equiv 1$, that is, only 1 people in each group are considered, the proposed design reduces to a special case of Hu and Zhang's DBCD. Through this paper, for simplicity and without loss of generality, we assume that the estimated $\boldsymbol{\theta}_k$ is the expectation of the observed responses on treatment k. To estimate it, we use the modified sample mean

$$\widehat{\boldsymbol{\theta}}_{m,k} = \frac{\sum_{i=1}^{m} X_{i,k}\boldsymbol{\xi}_{i,k} + \boldsymbol{\theta}_{0,k}}{N_{m,k} + 1}, \quad k = 1, \ldots, K \quad \text{and} \quad \widehat{\boldsymbol{\Theta}}_m = (\widehat{\boldsymbol{\theta}}_{m,1}, \ldots, \widehat{\boldsymbol{\theta}}_{m,K}).$$

When the $(j+1)$th group of patients arrives, the observed estimator is $\widehat{\boldsymbol{\Theta}}_{\tau_j}$. Here $\boldsymbol{\xi}_{i,k}$ is the response of the ith patient on treatment k. Write $\boldsymbol{\xi}_i = (\boldsymbol{\xi}_{i,1} \ldots, \boldsymbol{\xi}_{i,K})$ and assume that $\boldsymbol{\xi}_i$, $i \in \{1, 2, \cdots\}$, are independent identically distributed random vectors. $\boldsymbol{\Theta}_0 = (\boldsymbol{\theta}_{0,1}, \boldsymbol{\theta}_{0,2}, \cdots, \boldsymbol{\theta}_{0,K})$ is an initial value that may be based on historical data.

Regarding the group sizes χ_j, we make the following assumption.

Assumption 3.1. For some $\psi > 0$,

$$\chi_j = o\Big(\tau_j^{1/2-\psi}\Big) \text{ almost surely, where } \tau_j = \sum_{i=1}^{j} \chi_i. \tag{3.1}$$

This assumption means that, comparing with the sample size, the group size is not that large so that at least few adaptions during the trial. It is satisfied if the $(2+\epsilon_0)$ moments $E(\chi_j^{2+\epsilon_0})$ are bounded. In fact, for $0 < \psi < 1/(2(2+\epsilon_0))$,

$$\sum_j \Pr(\chi_j \geq \epsilon j^{1/2-\psi}) \leq \sum_j c j^{-(1/2-\psi)(2+\epsilon_0)} < \infty, \ \forall \epsilon > 0,$$

which implies that

$$\chi_j = o(j^{1/2-\psi}) = o(\tau_j^{1/2-\psi}) \text{ almost surely}$$

by the Borel-Cantelli lemma and noticing $\tau_j \geq j$.

3.1.2 Asymptotic properties

In this section, the asymptotic properties of the proposed Group DBCD in Section 3.1.1 are established. The following two theorems and one corollary are the extension of the properties of Hu and Zhang's DBCD within the context of group enrollment, so the three groups of conditions used for Hu and Zhang's DBCD are also applied here. In addition, Assumption 3.1 is introduced for the Group DBCD's asymptotic theorems. All conditions and assumptions in Section 3.1.2 are widely satisfied. For simplicity, we present the three conditions from Hu and Zhang's DBCD in the supplementary material.

Theorem 3.1 (Strong Consistency)**.** *Suppose that Condition 3.3.1 is satisfied,* $E(\|\boldsymbol{\xi}_{1,k}\|) < \infty$*,* $k = 1, \ldots, K$*, and that* $\boldsymbol{\rho}(\cdot) : \mathcal{R}^{d\times K} \to (0,1)^K$ *is a continuous function with* $\boldsymbol{\rho}(\boldsymbol{\Theta}) = \boldsymbol{v}$*. If the*

group size χ_j *satisfies* $\chi_j / \sum_{i=1}^{j} \chi_i \to 0$ *almost surely as* $j \to \infty$*, then* $\boldsymbol{N}_n/n \to \boldsymbol{v}$ *almost surely, and* $\widehat{\boldsymbol{\Theta}}_n \to \boldsymbol{\Theta}$ *almost surely.*

For the asymptotic normality, we use the Gaussian approximation method.[64, 62, 65] We need to introduce the following notation. We write

$$\boldsymbol{H} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\Big|_{(\boldsymbol{v},\boldsymbol{v})} =: \Big(\frac{\partial g_j(\boldsymbol{x},\boldsymbol{y})}{\partial x_i}\Big|_{(\boldsymbol{v},\boldsymbol{v})}\Big)_{i,j=1}^{K} \text{ and } \boldsymbol{E} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{y}}\Big|_{(\boldsymbol{v},\boldsymbol{v})},$$

where $\boldsymbol{x} = (x_1, \ldots, x_K)$ and $\boldsymbol{y} = (y_1, \ldots, y_K)$. It can easily be seen that $\boldsymbol{H}\mathbf{1}' = \boldsymbol{E}\mathbf{1}' = \mathbf{0}'$ and $\lambda_1 = 0$ is an eigenvalue of both $\boldsymbol{H}$ and $\boldsymbol{E}$. We let $\lambda_2, \ldots, \lambda_K$ be the other $K-1$ eigenvalues of $\boldsymbol{H}$, and let $\lambda = \max\{Re(\lambda_2), \ldots, Re(\lambda_K)\}$. Denote $\boldsymbol{V}_k = var(\boldsymbol{\xi}_{1,k})$, $k = 1, \ldots, K$, $\boldsymbol{V} = diag(\boldsymbol{V}_1/v_1, \ldots, \boldsymbol{V}_K/v_K)$, $\boldsymbol{\Sigma}_1 = diag(\boldsymbol{v}) - \boldsymbol{v}'\boldsymbol{v}$,

$$\boldsymbol{\Sigma}_3 = \Big(\frac{\partial \boldsymbol{\rho}}{\partial \boldsymbol{\Theta}}\Big)' \boldsymbol{V} \Big(\frac{\partial \boldsymbol{\rho}}{\partial \boldsymbol{\Theta}}\Big) = \sum_{k=1}^{K} \frac{1}{v_k} \Big(\frac{\partial \boldsymbol{\rho}}{\partial \boldsymbol{\theta}_k}\Big)' \quad \boldsymbol{\Sigma}_2 = \boldsymbol{E}'\boldsymbol{\Sigma}_3\boldsymbol{E}.$$

Theorem 3.2 (Asymptotic Normality)**.** *Suppose that* $\boldsymbol{N}_n/n \to \boldsymbol{v}$ *almost surely and Conditions 1, 2, 3.2 and Assumption 3.1 are satisfied. If* $\lambda < 1/2$*, then*

$$\frac{\boldsymbol{N}_n}{n} - \boldsymbol{v} = O\Big(\sqrt{\frac{\log\log n}{n}}\Big) \text{ and } \widehat{\boldsymbol{\Theta}}_n - \boldsymbol{\Theta} = O\Big(\sqrt{\frac{\log\log n}{n}}\Big) \text{ almost surely} \tag{3.2}$$

and there are two independent standard multi-dimensional Brownian motions $\boldsymbol{W}_t \in \mathscr{R}^K$ *and* $\boldsymbol{B}_t \in \mathscr{R}^{d\times K}$ *such that*

$$\frac{\boldsymbol{N}_n}{n} - \boldsymbol{v} = \frac{\boldsymbol{G}_n}{n} + o(n^{-1/2-\kappa}) \text{ and } \widehat{\boldsymbol{\Theta}}_n - \boldsymbol{\Theta} = \frac{\boldsymbol{B}_n}{n}\boldsymbol{V}^{1/2} + o(n^{-1/2-\kappa}) \text{ almost surely} \tag{3.3}$$

for some $\kappa > 0$*, where the Gaussian process* $\boldsymbol{G}_t$ *is the solution of the equation*

$$d\,\boldsymbol{G}_t = (d\,\boldsymbol{W}_t)\boldsymbol{\Sigma}_1^{1/2} + \frac{\boldsymbol{B}_t}{t}\boldsymbol{V}^{1/2}\frac{\partial \boldsymbol{\rho}}{\partial \boldsymbol{\Theta}}\boldsymbol{E}\,dt + \frac{\boldsymbol{G}_t}{t}\boldsymbol{H}\,dt, \;\; \boldsymbol{G}_0 = \mathbf{0} \tag{3.4}$$

and

$$\boldsymbol{G}_t = \int_0^t (d\,\boldsymbol{W}_x)\boldsymbol{\Sigma}_1^{1/2}\big(\frac{t}{x}\big)^{\boldsymbol{H}} + \int_0^t \frac{\boldsymbol{B}_x}{x}\boldsymbol{\Sigma}_2^{1/2}\big(\frac{t}{x}\big)^{\boldsymbol{H}}\,dx.$$

In particular, as $n \to \infty$*,*

$$n^{-1/2}(\boldsymbol{N}_{[nt]} - nt\boldsymbol{v}) \xrightarrow{\mathcal{D}} \boldsymbol{G}_t \;\; \textit{and} \;\; tn^{1/2}(\widehat{\boldsymbol{\Theta}}_{[nt]} - \boldsymbol{\Theta}) \xrightarrow{\mathcal{D}} \boldsymbol{B}_t\boldsymbol{V}^{1/2} \tag{3.5}$$

in the space $\mathscr{D}_{[0,1]}$ *with the Skorohod topology, and*

$$n^{1/2}(\boldsymbol{N}_n/n - \boldsymbol{v}) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{\Lambda}), \quad n^{1/2}(\widehat{\boldsymbol{\Theta}}_n - \boldsymbol{\Theta}) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{V}) \tag{3.6}$$

where

$$\boldsymbol{\Lambda} = var(\boldsymbol{G}_1) = \int_0^1 \big(\frac{1}{x}\big)^{\boldsymbol{H}'}\boldsymbol{\Sigma}_1\big(\frac{1}{x}\big)^{\boldsymbol{H}}dx + \int_0^1 dx\left[\int_x^1 \frac{1}{y}\big(\frac{1}{y}\big)^{\boldsymbol{H}}dy\right]'\boldsymbol{\Sigma}_2\left[\int_x^1 \frac{1}{y}\big(\frac{1}{y}\big)^{\boldsymbol{H}}dy\right].$$

We now apply the above theorem to the important class of DBCD treatment allocation functions[21] that is given in Hu & Zhang (2004). Their proposed treatment allocation function is

$$g_k(\boldsymbol{x}, \boldsymbol{y}) = \frac{y_k\big(\frac{y_k}{x_k}\big)^{\gamma}}{\sum_{j=1}^K y_j\big(\frac{y_j}{x_j}\big)^{\gamma}}, \quad k = 1, \ldots, K, \tag{3.7}$$

where $\gamma \geq 0$ is a constant that controls the degree of randomness of the procedure, from most random when $\gamma = 0$ to deterministic when $\gamma = \infty$.

Corollary 3.1. *Let the allocation function* $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y})$ *be defined as in (3.7). Suppose that Conditions 1 and 2 and Assumption 3.1 are satisfied. Then, (3.2), (3.3), (3.5) and (3.6) hold with*

$$\boldsymbol{G}_t = t^{-\gamma}\int_0^t x^{\gamma}(d\boldsymbol{W}_x)\boldsymbol{\Sigma}_1^{1/2} + (1+\gamma)t^{-\gamma}\int_0^t x^{\gamma-1}\boldsymbol{B}_x\boldsymbol{V}^{1/2}\frac{\partial\boldsymbol{\rho}}{\partial\boldsymbol{\Theta}}dx, \tag{3.8}$$

$$\boldsymbol{\Lambda} = \frac{1}{1+2\gamma}\boldsymbol{\Sigma}_1 + \frac{2(1+\gamma)}{1+2\gamma}\boldsymbol{\Sigma}_3. \tag{3.9}$$

Further, if the group sizes $\{\chi_j\}$ *are independent random variables with* $\sup_j E(\chi_j^{2+\epsilon_0}) < \infty$ *for*

some $\epsilon_0 > 0$, then Assumption 3.1 is satisfied and

$$\tau_j^{1/2}(\boldsymbol{N}_{\tau_j}/\tau_j - \boldsymbol{v}) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{\Lambda}), \quad \tau_j^{1/2}(\widehat{\boldsymbol{\Theta}}_{\tau_j} - \boldsymbol{\Theta}) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{V}), \quad \textit{as } \ j \to \infty. \tag{3.10}$$

Note that in Corollary 3.1, both τ_j and n in equation (3.6) *pertain to the number of patients assigned to the trial.*

Remark 3.1. (1) The theoretical results presented in Section 3.1.2 is an extension of the asymptotic theory in Hu and Zhang's DBCD. Theorems 3.1 and 3.2 state the strong consistency and asymptotic normality under the Conditions 1-3 and Assumption 3.1. Assumption 3.1 pertains to the group size within the context of group enrollment and is widely satisfied. For instance, if each group size within a trial remains around a constant, then Assumption 3.1 is fulfilled. (2) Corollary 3.1 is derived from Theorem 3.2. Theorem 3.2 is a general theoretical result that applies to any $g(\cdot)$ function that satisfies the requisite conditions. On the other hand, Corollary 3.1 demonstrates specific theoretical findings when Hu and Zhang's allocation function is used in the Group DBCD methods. (3) The theoretical results discussed in Section 3.1.2 are valid for various types of target allocations that satisfy Condition 3.2.

3.2 Group DBCD with delayed and missing responses

Although we consider the immediate responses above, in reality, the response to many diseases cannot be immediately available. For example, in some chronic diseases, the manifestation of treatment effects is usually gradual, and it may take days, weeks, or even months before we can collect patients' responses. Consequently, clinical trials should consider the possibility of delayed responses or long-term effects. Moreover, missing responses due to participant dropout or protocol violations present another challenge that should be taken into account. In this section, we assess the performance of the proposed Group DBCD in another practical scenarios where responses may be delayed or missing at random.

3.2.1 Framework

We use the same setting as subsection 2.1. Besides, we introduce additional notations subsequently. Let t_m be the entry time of patient m, $m \in \{1, 2, \cdots, n\}$. Patients from the same group j will be recorded as entering the trial simultaneously, meaning they are considered to have a shared entry time in the records. This entry time for group j is denoted by t'_j, where $j \in \mathbb{Z}$. We can infer that $\{t'_i\}_{i=1}^{j}$ is an increasing arithmetic sequence (weekly or monthly) or an increasing sequence of random variables. Specifically, $t'_j = t_{\tau_{j-1}+1} = \cdots = t_{\tau_j} < t'_{j+1}$, where τ_j is the index of the last patient in group j. The response time to treatment k of the m-th patient with response $\boldsymbol{\xi}_{m,k}$ is a random variable denoted by $\eta_m(k)$, $k \in \{1, .., K\}$. Alternatively, it can also be described as the time when this response becomes available to the clinic. Let $M_k(m, l) = I(t_m + \eta_m(k) \in [t_{m+l}, t_{m+l+1}))$, which is an indicator function that takes the value 1 if the response of patient m on treatment k, $\boldsymbol{\xi}_{m,k}$, is observed between the assignment of the treatments of the $(m + l)$th patient and the $(m + l + 1)$th patient. Moreover, let $I_{mis,k}(m)$ be an indicator function that takes the value 1 if the response of patient m on treatment k, $\boldsymbol{\xi}_{m,k}$, has been identified as missing data. Write $\boldsymbol{I}_{mis}(m) = (I_{mis,1}, \ldots, I_{mis,K})$ and assume that $\{\boldsymbol{I}_{mis}(m); m = 1, 2, \ldots\}$ is a sequence of independent random vectors, and that $I_{mis,k}(m)$ is independent of $\boldsymbol{\xi}_{m,k}$, $k = 1, \ldots, K$. $I_{mis,k}(m)$ follows the Bernoulli distribution with the missing rate α_k ($0 \leq \alpha_k < 1$) as the parameter. The total number of responses in treatment k before the assignment of the $(j + 1)$-th group is $\sum_{m=1}^{\tau_j} \sum_{l=0}^{\tau_j - m} M_k(m, l)(1 - I_{mis,k}(m))X_{m,k}$, and the summation of theses responses is $\sum_{m=1}^{\tau_j} \sum_{l=0}^{\tau_j - m} M_k(m, l)(1 - I_{mis,k}(m))X_{m,k}\boldsymbol{\xi}_{m,k}$.

The Group DBCD procedure with the missing and delayed response is defined in the following:

(a) (Initial Step). To start, we assign the first one or two groups of patients to treatment by restricted randomization and let $\widehat{\boldsymbol{\Theta}}_0^{dm}$ be an initial estimator of $\boldsymbol{\Theta}$.

(b) (Estimation Step). Assume that j groups of patients have been assigned in the trial. There is a positive probability that the total number of responses observed up to this time is less than $\tau_j = \sum_{i=1}^{j} \chi_i$ because of delayed and potential missing responses. We let $T_{\tau_j,k}$ be the

number of the observed responses on treatment k up to the time when group $(j+1)$ is ready for treatment assignment. And let the corresponding sum of the responses be $S_{\tau_j,k}$. Besides, noticed that $\boldsymbol{\Theta}_0 = (\boldsymbol{\theta}_{0,1}, \boldsymbol{\theta}_{0,2}, \cdots, \boldsymbol{\theta}_{0,K})$ is an initial value which may based on historical data.

Thus, the estimated $\widehat{\boldsymbol{\Theta}}_{\tau_j}^{dm}$ is

$$\widehat{\boldsymbol{\Theta}}_{\tau_j}^{dm} = \frac{S_{\tau_j,k} + \boldsymbol{\theta}_{0,k}}{T_{\tau_j,k} + 1},$$

where,

$$T_{\tau_j,k} = \sum_{m=1}^{\tau_j} \sum_{i=0}^{\tau_j - m} M_k(m,i)(1 - I_{mis,k}(m))X_{m,k},$$
$$S_{\tau_j,k} = \sum_{m=1}^{\tau_j} \sum_{i=0}^{\tau_j - m} M_k(m,i)(1 - I_{mis,k}(m))X_{m,k}\boldsymbol{\xi}_{m,k}.$$

(c) (Allocation Step). Assign each patients in the $(j+1)$th group to treatment k with the probability

$$Pr_{j+1}^{G} = g_k\left(\frac{\boldsymbol{N}_{\tau_j}}{\tau_j}, \boldsymbol{\rho}\left(\widehat{\boldsymbol{\Theta}}_{\tau_j}^{dm}\right)\right), \quad k = 1, 2, \cdots, K.$$

Here $\boldsymbol{N}_{\tau_j}/\tau_j$ represents the sample allocation proportion after the jth group assignments, and $\boldsymbol{\rho}$ is the target allocation proportion. It's typically a function of response distributions' parameter $\boldsymbol{\theta}$. These target allocation proportion is derived through the specific optimization objectives, for example, Neyman allocation is from the optimization which maximizing power under controlled sample size. Hu and Zhang (2004) introduced the allocation function $\boldsymbol{g} = (g_1, g_2, \cdots, g_K)$ for DBCD,[21] which determines the probability of the next patient being assigned to each group based on the current sample allocation proportions and the estimated target proportion. The fundamental purpose of this allocation function is to force the sample allocation proportion to the target allocation proportion. The formula is shown

below,

$$g_k(x,y) = \frac{y_k \left(\frac{y_k}{x_k}\right)^{\gamma}}{\sum_{j=1}^{K} y_j \left(\frac{y_j}{x_j}\right)^{\gamma}}, \quad k = 1, \ldots, k,$$

where $\gamma \geq 0$.

(d) (Repeated Step) The estimation step, and allocation step are then repeated continuously until the end of the clinical trial.

Remark 3.2. In the setup above, the group size depends on the number of individuals recruited at each fixed time interval, signifying that patients recruited within each interval (weekly or monthly) formed a group and entered the trial together. Hence, the number of participants recruited within each time interval is a random variable, and if its distribution with proper bounded finite second moments, it does not impact the theoretical outcomes or the finite samples results. In some of clinical trials, there are still some measured in terms of 'information time'. This means that a group joins the trial after recruiting a specific number of individuals. From this perspective, the number of patients per episode is treated as fixed, but the time intervals for each group recruitment become random variables. These two scenarios are fundamentally equivalent. The proposed Group DBCD is applicable to both situations, and the theoretical results remain valid in both cases.

Remark 3.3. Missing responses are a common issue in real clinical trials. The existing literature presents many strategies for handling them, yet there is very little discussion of this in the context of RAR. Most trials that address missing data tend to neglect them or impute them as failures if the response are binary. Alternatively, Kim et al. [27] advocated the use of linear regression to predict missing responses in RAR, using explanatory variables. Biswas et al.[7] suggested imputing missing data at each stage by utilizing the data available at that time. If covariates for the patients are available, the method of Kim et al.[27] is undoubtedly a good approach; however, most RAR studies do not collect covariate profiles of patients. In this paper, our procedure simply disregards missing responses. However, it is important to note that our procedure is equally applicable to

scenarios where missing responses are not ignored but are instead imputed or interpolated in the case of missing at random. More complex scenarios, for example, data not missing at random but depend on treatment or outcome, still remain unexplored, but this is the case for adaptive trials in general except for some simple setting.[42]

3.2.2 Asymptotic Results

Because in the same group, $M_k(m+1,l)$, $\tau_j < m+1 \leq \tau_{j+1}$, may be dependent, instead of assuming that $\{M_k(m,l); m = 1,2,\ldots\}$ is a sequence of i.i.d. random variables as in Hu et al.[23], we assume that for each fixed k and l, when $\tau_j < m+1 \leq \tau_{j+1}$, $\boldsymbol{\xi}_{m+1,k}$, $I_{mis,k}(m+1)$, $M_k(m+1,l)$ are independent of the history sigma field $\mathscr{G}_{\tau_j}$ of previous patients when group $(j+1)$ is ready for treatment assignment, which is generated by $\tau_1,\ldots,\tau_j$, $\boldsymbol{\xi}_{i,k}$, $I_{mis,k}(i)$, $M_k(i,s)$, $s = 0,\ldots,\tau_j - i$, $i = 1,\ldots,\tau_j$, $k = 1,\ldots,K$. Note that $\sum_l M_k(m,l) = 1$ when $I_{mis,k}(m) = 0$, which means that, when the delayed response will eventually be observed when it is not missed, even if they are not immediately available. And, $\sum_l M_k(m,l) = 0$ when $I_{mis,k}(m) = 1$. Define $\mu_{l,k}^{(m)} = \mathrm{E}\{M_k(m,l)|I_{mis,k}(m) = 0\}$. Thus, $\sum_i \mu_{i,k}^{(m)} = 1$.

Assumption 3.2. For some $\phi > 0$ and $C > 0$,

$$\sum_{i=l}^{\infty} \mu_{i,k}^{(m)} = \Pr(\eta_m(k) > t_{m+l} - t_m | I_{mis,k}(m) = 0) \leq Cl^{-\phi}, \quad k = 1,\cdots,K. \tag{3.11}$$

The Probability in Assumption 3.2 illustrates the probability that a patient on treatment k responds after at least another l patients have been assigned to treatments. Assumption 3.2 implies that the delay time cannot be too large compared to the enrollment period of patients.

Let

$$\widehat{\boldsymbol{\Theta}}_n^{dm} = \frac{S_{n,k} + \boldsymbol{\theta}_{0,k}}{T_{n,k} + 1},$$

where

$$T_{n,k} = \sum_{m=1}^{n} \sum_{i=0}^{n-m} M_k(m,i)(1 - I_{mis,k}(m))X_{m,k},$$
$$S_{n,k} = \sum_{m=1}^{n} \sum_{i=0}^{n-m} M_k(m,i)(1 - I_{mis,k}(m))X_{m,k}\boldsymbol{\xi}_{m,k}$$

are the number and the total sum of the responses on treatment k that are observed up to the time patient $(n+1)$ is ready for treatment allocation, respectively.

Theorem 3.3 (Strong Consistency)**.** *Suppose that Condition 3.3.1 and Assumption 3.2 are satisfied,* $E(\|\boldsymbol{\xi}_{1,k}\|) < \infty$, $k = 1, \ldots, K$, *and that* $\boldsymbol{\rho}(\cdot) : \mathcal{R}^{d\times K} \to (0,1)^K$ *is a continuous function with* $\boldsymbol{\rho}(\boldsymbol{\Theta}) = \boldsymbol{v}$. *If the group size* χ_j *satisfies* $\chi_j / \sum_{i=1}^{j} \chi_i \to 0$ *almost surely as* $j \to \infty$, *then* $\boldsymbol{N}_n/n \to \boldsymbol{v}$ *almost surely, and* $\widehat{\boldsymbol{\Theta}}_n^{dm} \to \boldsymbol{\Theta}$ *almost surely.*

Theorem 3.4 (Rates of Consistency)**.** *Suppose that* $\boldsymbol{N}_n/n \to \boldsymbol{v}$ *almost surely under the delayed and missing circumstance, that Conditions 1, 2, 3.2 and Assumptions 3.1 and 3.2 are satisfied, and that* $\lambda < 1$. *Let* $\varphi' = (2+\varepsilon_0) / (2 + \varepsilon_0 + (1+\varepsilon_0)\phi)$. *Then*

$$\widehat{\boldsymbol{\Theta}}_n^{dm} - \boldsymbol{\Theta} = O\left(\sqrt{\frac{\log\log n}{n}}\right) + o\left(n^{\varphi'-1}(\log n)^2\right) \quad \textit{almost surely} \tag{3.12}$$

$$\textit{and } n^{-\kappa}(\boldsymbol{N}_n - n\boldsymbol{v}) \to 0 \ \textit{ almost surely for } \kappa > \max\{1/2, \lambda, \varphi'\}. \tag{3.13}$$

Furthermore, if $\lambda < 1/2$ *and* $\phi > (2+\varepsilon_0)/(1+\varepsilon_0)$, *then*

$$\boldsymbol{N}_n - n\boldsymbol{v} = O(\sqrt{n\log\log n}) \ \textit{ almost surely}. \tag{3.14}$$

To discuss the asymptotic normality, we need the following notation. Recall that $\boldsymbol{V}_k = var(\boldsymbol{\xi}_{1,k})$, $\boldsymbol{\Sigma}_1 = diag(\boldsymbol{v}) - \boldsymbol{v}'\boldsymbol{v}$, $\boldsymbol{H}$ and $\boldsymbol{E}$ are defined in Section 3.1.2. Denote $\boldsymbol{V}^{mis} =$

$diag(\frac{\boldsymbol{V}_1}{(1-\alpha_1)v_1}, \ldots, \frac{\boldsymbol{V}_K}{(1-\alpha_K)v_K})$,

$$\boldsymbol{\Sigma}_3^{mis} = \Big(\frac{\partial \boldsymbol{\rho}}{\partial \boldsymbol{\Theta}}\Big)' \boldsymbol{V}^{mis} \Big(\frac{\partial \boldsymbol{\rho}}{\partial \boldsymbol{\Theta}}\Big) = \sum_{k=1}^{K} \frac{1}{(1-\alpha_k)v_k} \Big(\frac{\partial \boldsymbol{\rho}}{\partial \boldsymbol{\theta}_k}\Big)' \boldsymbol{V}_k \Big(\frac{\partial \boldsymbol{\rho}}{\partial \boldsymbol{\theta}_k}\Big), \quad \boldsymbol{\Sigma}_2^{mis} = \boldsymbol{E}' \boldsymbol{\Sigma}_3^{mis} \boldsymbol{E}.$$

Theorem 3.5 (Asymptotic normality)**.** *Suppose that $\boldsymbol{N}_n/n \to \boldsymbol{v}$ almost surely under the delayed and missing circumstance, and Conditions 1, 2, 3.2 and Assumptions 3.1 and 3.2 are satisfied. If $\lambda < 1/2$ and $\phi > (2+\epsilon_0)/(1+\epsilon_0)$, then there are two independent standard multi-dimensional Brownian motions $\boldsymbol{W}_t \in \mathscr{R}^K$ and $\boldsymbol{B}_t \in \mathscr{R}^{d\times K}$ such that*

$$\frac{\boldsymbol{N}_n}{n} - \boldsymbol{v} = \frac{\boldsymbol{G}_n^{mis}}{n} + o(n^{-1/2-\kappa}) \text{ and } \widehat{\boldsymbol{\Theta}}_n^{dm} - \boldsymbol{\Theta} = \frac{\boldsymbol{B}_n}{n} (\boldsymbol{V}^{mis})^{1/2} + o(n^{-1/2-\kappa}) \text{ almost surely} \quad (3.15)$$

for some $\kappa > 0$. In particular,

$$\sqrt{n}\,(\boldsymbol{N}_n/n - \boldsymbol{v}) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{\Lambda}^{mis}), \quad \sqrt{n}\left(\widehat{\boldsymbol{\Theta}}_n^{dm} - \boldsymbol{\Theta}\right) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{V}^{mis}),$$

$$\sqrt{(1-\alpha_k)v_k n}\left(\widehat{\boldsymbol{\theta}}_{n,k}^{dm} - \boldsymbol{\theta}_k\right) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{V}_k),\ k = 1, \ldots, K.$$

Here $\boldsymbol{G}_t^{mis}$ and $\boldsymbol{\Lambda}^{mis}$ are defined the same as $\boldsymbol{G}(t)$ and $\boldsymbol{\Lambda}$ in Theorem 3.2, respectively, with $\boldsymbol{V}$ being replaced by $\boldsymbol{V}^{mis}$.

We investigate the asymptotic properties of group DBCD with delayed and missing responses. Despite these issues, our results show that the estimation of $\widehat{\boldsymbol{\Theta}}^{dm}$ and allocation proportion $\boldsymbol{N}_n/n$ will still converge to the true parameter $\boldsymbol{\Theta}$ and target allocation value $\boldsymbol{v}$, respectively, under widely satisfied conditions. However, we proved that delayed responses can affect the convergence rate of the estimation of $\boldsymbol{\Theta}$, as shown by a comparison of Theorem 3.3-3.4 and Theorem 3.2. Additionally, we find that the asymptotic variance of allocation proportion and estimation of $\boldsymbol{\Theta}$ depends on the missing rate $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_K)$, resulting in a slight increase in the asymptotic variance when randomly missing responses are present.

3.3 Numerical studies and application

3.3.1 Some numerical studies

In this section, we conduct numerical studies to compare the performance of the Group DBCD method with the Hu and Zhang's DBCD [21]. Our goal is to demonstrate that the Group DBCD and its modified version not only achieve practicality but also preserve the advantages of the Hu and Zhang's DBCD. The Hu and Zhang's DBCD algorithms offer the advantage of allocating more patients to superior treatments and allow for the dynamic adjustment of the allocation rate to any specified optimal targets. We evaluate the mean and standard deviation of allocation proportions, along with power and failure rates, as our primary comparison metrics.

In our numerical studies, we compared the Group DBCD with the DBCD proposed by Hu and Zhang[21], and the modified version of Group DBCD with the Hu and Zhang's DBCD with delayed response proposed by Hu et al.[23], respectively. We initially assigned 20 patients to each treatment via restricted randomization, and then allocates the rest of patients (180 for binary cases and 100 for continuous cases) according to each design. We have the following setting for the simulations:

- **Group Size:** Assume that the group size is random and follows a Poisson distribution with mean λ.
- **Delayed Time:** Assume the response lag time for treatment follows an exponential distribution with mean $1/\lambda_k$ for $k \in \{1, 2, \cdots, K\}$.
- **Missing Rate:** Assume the missing rates follow a Bernoulli distribution with mean α_k for $k \in \{1, 2, \cdots, K\}$.

To simplify our simulation and ensure comparability of results, we assumed that the missing rate parameters are identical across all treatment groups, which means $\alpha = \alpha_1 = \cdots = \alpha_K$. This setup in our numerical study aligns with the assumptions of the theory. To investigate the influence of group size, we conducted simulations for the Group DBCD and Group DBCD with missing and

delayed responses, using $\lambda = 5$ and $\lambda = 10$, respectively. We also varied parameter combinations to explore the impact of response lag time and the proportion of missing data.

The Case of Binary Response

We utilized the parameter settings described in the paper of Hu & Rosenberger [18] and present our findings. This analysis evaluates the performance of the Group DBCD method and its modified version, which handles delayed and missing responses, targeting both RSIHR and Neyman optimal goals. The RSIHR target seeks to reduce the number of failures while maintaining fixed power, whereas Neyman target aims at maximizing the power. To save space in the paper, we present the results for the RSIHR target as follows, while the results and findings for the Neyman target have been moved to the Supplementary material.

Table 3.1: Power and allocation proportion for binary cases with 5000 replications (RSIHR target)

		Hu and Zhang's DBCD			**Group DBCD** ($\lambda = 5$)			**Group DBCD** ($\lambda = 10$)		
(p_A, p_B)	ρ	| Power	$\mathbb{E}(\frac{n_A}{n})$(SD)	$\mathbb{E}$(No.Fail)(SD)	| Power	$\mathbb{E}(\frac{n_A}{n})$(SD)	$\mathbb{E}$(No.Fail)(SD)	| Power	$\mathbb{E}(\frac{n_A}{n})$(SD)	$\mathbb{E}$(No.Fail)(SD)
(0.8, 0.8)	0.50	0.051	0.500 (0.019)	0.201 (0.029)	0.047	0.500 (0.019)	0.200 (0.029)	0.048	0.500 (0.019)	0.200 (0.028)
(0.8, 0.7)	0.52	0.369	0.517 (0.020)	0.248 (0.030)	0.373	0.517 (0.020)	0.248 (0.030)	0.366	0.517 (0.020)	0.248 (0.030)
(0.7, 0.5)	0.54	0.822	0.542 (0.023)	0.392 (0.034)	0.824	0.542 (0.024)	0.391 (0.033)	0.814	0.542 (0.023)	0.391 (0.034)
(0.7, 0.3)	0.60	1.000	0.606 (0.030)	0.457 (0.031)	1.000	0.605 (0.029)	0.458 (0.032)	1.000	0.606 (0.031)	0.457 (0.031)
(0.5, 0.5)	0.50	0.046	0.500 (0.025)	0.500 (0.035)	0.052	0.500 (0.026)	0.501 (0.036)	0.053	0.500 (0.026)	0.500 (0.036)
(0.5, 0.2)	0.61	0.995	0.616 (0.039)	0.615 (0.033)	0.997	0.615 (0.039)	0.615 (0.033)	0.993	0.614 (0.039)	0.615 (0.034)
(0.2, 0.2)	0.50	0.058	0.500 (0.046)	0.800 (0.029)	0.058	0.499 (0.046)	0.800 (0.029)	0.064	0.501 (0.047)	0.801 (0.028)

In Table 3.1-3.2, we investigate the performance of Group DBCD and modified version of Group DBCD. Across a range of values for p_A and p_B, we found that the Group DBCD method, even when modified to handle delayed and missing responses, successfully met the target allocations while maintaining a lower number of failures. Our results indicate that lower missing rates lead to better performance, because a large missing rate leads to an increase in the standard deviation of both of the mean of the allocation proportion and failure rate. We also found, through simulations, that when Assumption 3.2 is satisfied — stating that delayed times are not that large compared to the whole trial duration — it has a minimal impact on the results.

The Case of Continuous Response

In this section, we explore the Group DBCD's performance when the responses are continuous type. Similarly, the goal is to demonstrate the proposed Group DBCD and its modified version

Table 3.2: Delayed and Missing Response: Power and allocation proportion for binary cases with 5000 replications (RISHR target)

		Hu and Zhang's DBCD			**Group DBCD** ($\lambda = 5$)			**Group DBCD** ($\lambda = 10$)		
(α, p_A, p_B)	ρ	| Power	$\mathbb{E}(\frac{n_A}{n})$(SD)	$\mathbb{E}$(No.Fail)(SD)	| Power	$\mathbb{E}(\frac{n_A}{n})$(SD)	$\mathbb{E}$(No.Fail)(SD)	| Power	$\mathbb{E}(\frac{n_A}{n})$(SD)	$\mathbb{E}$(No.Fail)(SD)
(0.0, 0.8, 0.8)	0.50	0.047	0.499 (0.020)	0.200 (0.028)	0.054	0.500 (0.019)	0.200 (0.029)	0.044	0.500 (0.019)	0.200 (0.028)
(0.0, 0.8, 0.7)	0.52	0.362	0.516 (0.021)	0.248 (0.030)	0.369	0.517 (0.020)	0.248 (0.030)	0.372	0.517 (0.020)	0.248 (0.031)
(0.0, 0.8, 0.6)	0.54	0.881	0.535 (0.023)	0.293 (0.031)	0.870	0.536 (0.021)	0.293 (0.031)	0.876	0.536 (0.021)	0.293 (0.031)
(0.0, 0.7, 0.5)	0.54	0.819	0.541 (0.026)	0.392 (0.033)	0.834	0.542 (0.023)	0.391 (0.034)	0.832	0.543 (0.024)	0.392 (0.032)
(0.0, 0.7, 0.3)	0.60	1.000	0.606 (0.036)	0.457 (0.033)	1.000	0.606 (0.030)	0.458 (0.031)	1.000	0.606 (0.030)	0.457 (0.032)
(0.0, 0.5, 0.5)	0.50	0.055	0.498 (0.030)	0.501 (0.036)	0.053	0.500 (0.026)	0.499 (0.035)	0.056	0.500 (0.026)	0.500 (0.036)
(0.0, 0.5, 0.2)	0.61	0.996	0.617 (0.049)	0.615 (0.034)	0.996	0.615 (0.038)	0.615 (0.033)	0.995	0.615 (0.039)	0.616 (0.033)
(0.0, 0.2, 0.2)	0.50	0.056	0.498 (0.055)	0.800 (0.028)	0.053	0.500 (0.046)	0.800 (0.029)	0.053	0.501 (0.047)	0.800 (0.029)
(0.2, 0.8, 0.8)	0.50	0.050	0.499 (0.027)	0.200 (0.032)	0.049	0.500 (0.027)	0.200 (0.032)	0.044	0.500 (0.027)	0.199 (0.032)
(0.2, 0.8, 0.7)	0.52	0.307	0.516 (0.029)	0.248 (0.034)	0.313	0.517 (0.027)	0.248 (0.033)	0.298	0.517 (0.028)	0.248 (0.033)
(0.2, 0.8, 0.6)	0.54	0.803	0.535 (0.030)	0.292 (0.035)	0.785	0.536 (0.028)	0.293 (0.036)	0.789	0.536 (0.028)	0.293 (0.034)
(0.2, 0.7, 0.5)	0.54	0.746	0.542 (0.033)	0.392 (0.038)	0.731	0.542 (0.031)	0.391 (0.037)	0.729	0.542 (0.031)	0.392 (0.038)
(0.2, 0.7, 0.3)	0.60	1.000	0.606 (0.043)	0.458 (0.038)	0.999	0.606 (0.037)	0.457 (0.036)	0.999	0.606 (0.037)	0.458 (0.036)
(0.2, 0.5, 0.5)	0.50	0.051	0.498 (0.036)	0.501 (0.040)	0.052	0.499 (0.033)	0.500 (0.039)	0.050	0.500 (0.032)	0.499 (0.040)
(0.2, 0.5, 0.2)	0.61	0.980	0.615 (0.053)	0.616 (0.039)	0.981	0.614 (0.047)	0.616 (0.037)	0.986	0.615 (0.047)	0.616 (0.037)
(0.2, 0.2, 0.2)	0.50	0.050	0.497 (0.064)	0.800 (0.031)	0.057	0.500 (0.054)	0.800 (0.031)	0.054	0.500 (0.056)	0.800 (0.032)
(0.3, 0.8, 0.8)	0.50	0.046	0.500 (0.031)	0.200 (0.034)	0.051	0.500 (0.031)	0.200 (0.034)	0.047	0.499 (0.031)	0.200 (0.034)
(0.3, 0.8, 0.7)	0.52	0.276	0.517 (0.032)	0.248 (0.036)	0.262	0.516 (0.032)	0.249 (0.037)	0.279	0.517 (0.032)	0.249 (0.036)
(0.3, 0.8, 0.6)	0.54	0.729	0.535 (0.034)	0.292 (0.038)	0.735	0.536 (0.033)	0.293 (0.037)	0.737	0.536 (0.033)	0.292 (0.037)
(0.3, 0.7, 0.5)	0.54	0.684	0.542 (0.037)	0.391 (0.041)	0.686	0.543 (0.034)	0.392 (0.040)	0.665	0.542 (0.035)	0.392 (0.040)
(0.3, 0.7, 0.3)	0.60	0.999	0.607 (0.045)	0.457 (0.041)	0.998	0.607 (0.043)	0.457 (0.038)	0.999	0.607 (0.042)	0.458 (0.038)
(0.3, 0.5, 0.5)	0.50	0.053	0.499 (0.040)	0.500 (0.043)	0.052	0.500 (0.038)	0.500 (0.043)	0.052	0.499 (0.037)	0.502 (0.042)
(0.3, 0.5, 0.2)	0.61	0.965	0.616 (0.057)	0.615 (0.041)	0.970	0.616 (0.052)	0.615 (0.040)	0.968	0.616 (0.052)	0.616 (0.040)
(0.3, 0.2, 0.2)	0.50	0.060	0.497 (0.068)	0.800 (0.034)	0.058	0.500 (0.061)	0.800 (0.034)	0.052	0.501 (0.060)	0.800 (0.034)

(with delayed and missing responses) perform as good as the Hu and Zhang's DBCD while enhance the practicality. We applied all procedures with a sample size of 120 and the optimal target as the Neyman allocation or Zhang and Rosenberger's allocation, following the parameter settings described in Zhang & Rosenberger (2006) [63]. Zhang and Rosenberger's target minimize the total expected response from all patients when a small response is desirable. The findings with this target are presented in Table 3.3-3.4, where we examine various combinations of parameters $(\mu_A, \sigma_A^2, \mu_B, \sigma_B^2)$. The simulation results indicate a slight decrease in power under the case with missing and delayed responses, which is expected as missing responses result in a smaller effective sample size. However, this effect is not a significant issue, and under moderate group size, the power of the trial is negligibly affected. Furthermore, the allocation proportions of the proposed procedures and its modified version handling delayed and missing responses consistently converge to the target in studies with lower standard deviation. Additionally, instead of reporting the failure rate as in the numerical study of the binary case, the mean of observed responses is reported when the response is continuous. It is evident that the mean of observed responses is similar for both Group DBCD and Hu and Zhang's DBCD, and they are smaller because of Zhang and

Rosenberger's target is used as the optimal allocation in the trial. These findings demonstrate the robustness and reliability of the proposed procedures in practical scenarios with moderate delay times, missing rates, and group sizes. Besides, the type I error rate is well-controlled. The results with the Neyman target are included in the supplementary material, from which similar results are found.

Table 3.3: Power and allocation proportion for continuous cases with 5000 replications (Zhang and Rosenberger's target)

		Hu and Zhang's DBCD			**Group DBCD** ($\lambda = 5$)			**Group DBCD** ($\lambda = 10$)		
$(\mu_A, \sigma_A, \mu_B, \sigma_B)$	ρ	| Power	$\mathbb{E}(\frac{n_A}{n})$(SD)	$\mathbb{E}$(Obs.Rsp)(SD)	| Power	$\mathbb{E}(\frac{n_A}{n})$(SD)	$\mathbb{E}$(Obs.Rsp)(SD)	| Power	$\mathbb{E}(\frac{n_A}{n})$(SD)	$\mathbb{E}$(Obs.Rsp)(SD)
(13, 16, 15, 6.25)	0.63	0.919	0.635 (0.042)	13.729 (0.336)	0.915	0.635 (0.043)	13.728 (0.335)	0.917	0.634 (0.043)	13.723 (0.337)
(13, 6.25, 15, 16)	0.40	0.911	0.399 (0.042)	14.200 (0.321)	0.909	0.400 (0.043)	14.191 (0.325)	0.919	0.400 (0.044)	14.206 (0.321)
(15, 16, 17, 6.25)	0.63	0.914	0.633 (0.042)	15.736 (0.333)	0.916	0.633 (0.043)	15.733 (0.338)	0.922	0.633 (0.044)	15.732 (0.333)
(15, 6.25, 17, 16)	0.40	0.916	0.398 (0.043)	16.202 (0.325)	0.916	0.397 (0.043)	16.205 (0.322)	0.917	0.399 (0.045)	16.203 (0.323)

Table 3.4: Delayed and Missing Response: Power and allocation proportion for continuous cases with 5000 replications (Zhang and Rosenberger's target)

		Hu and Zhang's DBCD			**Group DBCD** ($\lambda = 5$)			**Group DBCD** ($\lambda = 10$)		
$(\alpha, \mu_A, \sigma_A, \mu_B, \sigma_B)$	ρ	| Power	$\mathbb{E}(\frac{n_A}{n})$(SD)	$\mathbb{E}$(Obs.Rsp)(SD)	| Power	$\mathbb{E}(\frac{n_A}{n})$(SD)	$\mathbb{E}$(Obs.Rsp)(SD)	| Power	$\mathbb{E}(\frac{n_A}{n})$(SD)	$\mathbb{E}$(Obs.Rsp)(SD)
(0,13, 16, 15, 6.25)	0.63	0.923	0.633 (0.041)	13.734 (0.332)	0.913	0.633 (0.043)	13.733 (0.338)	0.919	0.631 (0.044)	13.746 (0.331)
(0,13, 6.25, 15, 16)	0.40	0.915	0.401 (0.043)	14.203 (0.326)	0.920	0.401 (0.044)	14.200 (0.322)	0.919	0.400 (0.046)	14.208 (0.326)
(0,15, 16, 17, 6.25)	0.63	0.926	0.631 (0.042)	15.730 (0.328)	0.917	0.631 (0.043)	15.740 (0.331)	0.920	0.631 (0.045)	15.741 (0.341)
(0,15, 6.25, 17, 16)	0.40	0.921	0.399 (0.042)	16.207 (0.324)	0.915	0.399 (0.043)	16.209 (0.325)	0.921	0.400 (0.043)	16.204 (0.327)
(0.2,13, 16, 15, 6.25)	0.63	0.844	0.634 (0.053)	13.735 (0.383)	0.845	0.633 (0.054)	13.739 (0.375)	0.859	0.635 (0.054)	13.729 (0.379)
(0.2,13, 6.25, 15, 16)	0.40	0.850	0.400 (0.054)	14.210 (0.367)	0.842	0.400 (0.055)	14.192 (0.372)	0.842	0.399 (0.057)	14.204 (0.368)
(0.2,15, 16, 17, 6.25)	0.63	0.861	0.633 (0.053)	15.732 (0.381)	0.847	0.633 (0.054)	15.732 (0.378)	0.858	0.632 (0.056)	15.726 (0.386)
(0.2,15, 6.25, 17, 16)	0.40	0.831	0.399 (0.053)	16.203 (0.367)	0.845	0.398 (0.054)	16.197 (0.367)	0.855	0.398 (0.056)	16.209 (0.371)
(0.3,13, 16, 15, 6.25)	0.63	0.797	0.634 (0.060)	13.731 (0.405)	0.801	0.635 (0.061)	13.731 (0.406)	0.795	0.635 (0.061)	13.724 (0.404)
(0.3,13, 6.25, 15, 16)	0.40	0.793	0.399 (0.061)	14.202 (0.400)	0.797	0.399 (0.062)	14.203 (0.396)	0.795	0.400 (0.064)	14.200 (0.398)
(0.3,15, 16, 17, 6.25)	0.63	0.796	0.634 (0.061)	15.729 (0.414)	0.788	0.632 (0.062)	15.740 (0.412)	0.796	0.631 (0.062)	15.741 (0.418)
(0.3,15, 6.25, 17, 16)	0.40	0.802	0.397 (0.060)	16.210 (0.399)	0.784	0.397 (0.061)	16.206 (0.402)	0.797	0.398 (0.063)	16.196 (0.398)

3.3.2 Re-designing the HIV transmission trial

Connor et al.[11] reported a clinical trial to evaluate the drug AZT (Zidovudine treatment) of reducing the risk of maternal-infant HIV transmission. The trial enrolled 477 HIV-infected pregnant women between April 1991 and December 1993, with 239 women in the treatment group and 238 in the placebo group. The trial employed a 50-50 restricted randomization scheme in the Zidovudine treatment group and placebo group, and was randomized, double-blind, and placebo-controlled. At the end of the trial, 8.3% of infants in the treatment group and 25.5% in the placebo group were HIV positive.

Table 3.5: The Re-designed HIV Transmission Trial with Original Sample Size ($n = 478$)

Method	Update Cycle	ρ	Effective n	Power	Patients in AZT	Total Infected Infants
CR	-	0.50	478	0.998	239.2 (11.6)	80.8 (8.2)
RR	-	0.50	478	0.998	239.0 (0.00)	80.8 (7.9)
RPW	-	0.75	478	0.993	330.3 (54.1)	64.9 (11.5)
DBCD	-	0.75	478	0.997	359.1 (23.1)	60.3 (8.4)
DBCD^{DM}	-	0.75	416	0.991	312.6 (24.3)	52.4 (7.9)
Group DBCD	Biweekly	0.75	478	0.997	359.4 (22.7)	60.4 (8.7)
	Monthly	0.75	478	0.999	358.8 (23.6)	60.4 (8.5)
Group DBCD^{DM}	Biweekly	0.75	416	0.998	309.6 (24.1)	52.6 (8.0)
	Monthly	0.75	416	0.992	309.6 (23.2)	52.8 (7.9)

^{DM} specifies the procedures deal with the delayed and missing responses.

To provide a more realistic comparison, we re-design this trial using our proposed Group DBCD method (with $\gamma = 2$) and its modified version for handling delayed and missing responses. In our re-design study, we utilize true success rates, specifically $p_A = 0.917$ for the treatment group (AZT) and $p_B = 0.745$ for the placebo group, as reported in the original paper. It is important to note that these true success rates were employed only for generating simulated responses, and the assumption throughout the trial was that they were unknown. The trial enrolled HIV-infected pregnant women between 14 and 34 weeks of gestation who were not receiving antiretroviral therapy during their current pregnancy. The recruitment period for the study lasted approximately 31 months (142 weeks). The comparison of efficacy between treatment groups was based on the percentage of infants who were infected. For the redesign, we assumed a random delay time following a uniform distribution between 12 and 32 weeks, and a missing rate of $\alpha_A = \alpha_B = 0.13$, as reported in the original paper. We considered two scenarios: (i) updating the group monthly with $\lambda = 15.4$, and (ii) updating the group biweekly with $\lambda = 6.6$.

Same as the previous numerical studies, we use restricted randomization to assign the first 20 patients to treatment to obtain initial parameter estimates. And then we adopted different designs for assigning the remaining 458 patients. Besides, we additionally compare them with complete randomization (CR), restricted randomization (RR) and randomized play-the-winner rule (RPW). The results are displayed in Table 3.5.

Based on the results of our redesign, we achieve higher power with any method if we keep

the sample size same as the original trial due to sufficient data. The proposed design and DBCD showed similar results in terms of allocation proportion, the total infected infants and their standard deviation, as well as power. Furthermore, DBCD and Group DBCD had fewer patients assigned to the inferior treatment compared to CR and RR. Since we chose urn allocation as our target allocation, we compared the redesign under the RPW with it. Our results showed that RPW had a slower convergence to the target allocation and a higher standard deviation than the proposed design. All three methods, DBCD, Group DBCD, and RPW rule had fewer failures than CR and RR, with the number decreasing from around 80 to around 60. However, we did not compare the total number of infected infants under the missing response scenario with the scenario without missing response because the effective dataset was smaller when missing responses were present, making the number of failures incomparable to the redesign without missing data.

We also examined the re-designed trial under a reduced sample size. The power of the Group DBCD with delayed and missing responses slightly increased compared to the Hu and Zhang's DBCD with delayed and missing responses. This increase was reasonable because group adaptive allocation reduced the probability of allocation shift, resulting in slightly more patients being assigned to the inferior group but still maintaining a lower failure rate.

Upon conducting a comprehensive evaluation of the proposed design and five other commonly used trial designs, we conclude that (1) the Group DBCD algorithm, along with its modified version for delayed and missing responses, outperforms CR, RR and the RPW rule in terms of efficiency and ethics, and (2) the proposed designs yield results similar to those of DBCD ($\gamma = 2$). This demonstrates that the new design not only inherits the favorable properties and advantages of DBCD but also broadens the algorithm's practical applicability. Additionally, group entry has negligible effects on the effectiveness and ethics of designs. In summary, the Group DBCD procedure, along with its modified version, effectively reduces the number of patients in the inferior group, offers expected benefits, and can be applied to a wide range of clinical trials to satisfy diverse practical needs.

3.4 Discussion

In practical clinical trial scenarios, participants are often recruited in batches at regular intervals, which can be challenging to accommodate in the designs that do not take this practical situation into account. To address this issue, we propose a family of group response adaptive randomization procedures, Group DBCD, based on the DBCDs[21] developed by Hu & Zhang. One key advantage of our new proposed procedure is that the patients enter the trial by group periodically, such as weekly and monthly. Hence, the proposed procedure, Group DBCD, enhance the practicality while preserve the strengths of Hu and Zhang's DBCD.[21] We also propose the modified version of Group DBCD which address the issue of delayed or missing responses making the procedure more accommodate to the common practical issues. Moreover, we derive the asymptotic theory of the proposed designs which demonstrates that our Group DBCD, along with its modified version for handling delayed and missing data, exhibit similar asymptotic properties to Hu and Zhang's DBCDs [21] under a broad range of conditions. Numerical studies are conducted to complement our theoretical results. All results and findings indicate that the proposed Group DBCD and its modified version, which accommodate delayed and missing responses, are capable of enhancing practicality while preserving all the advantages of the Hu and Zhang's DBCD.

In the existing literature on RAR, updates are made on a per-subject basis. Even though the term 'group' appears in RAR literature, its primary focus has been to conduct tests and adjust sample re-estimation based on information from a group of individuals. However, no studies have delved into the procedure of group enrollment itself. Hence, the methodology proposed in this paper represents a distinct and important contribution in the domain of RAR literature. Furthermore, the 'group' concept presented in this paper may be applicable to other RAR designs. For instance, there should be no difficulty to consider the group ERADEs[24] proposed by Hu, Zhang & He . Urn models, the most extensively studied RAR procedures, originating with Wei in 1978 [55] and further developed by Zhang et al. in 2011 [65] among others, also offer potential for a group-based extension of their work. We leave these as future research projects.

3.5 Appendix

3.5.1 Conditions

To establish the asymptotic properties, the following three groups of conditions are required. They are also used for the Hu and Zhang's DBCD [21].

Condition 3.1: For the response sequence $\{\boldsymbol{\xi}_n = (\boldsymbol{\xi}_{n,1}, \cdots, \boldsymbol{\xi}_{n,K})\}$, assume that $E\|\boldsymbol{\xi}_{1,k}\|^{2+\epsilon_0} < \infty$ for some $\epsilon_0 > 0$, $k = 1, \ldots, K$.

Condition 3.2: The proportion function $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_d) \to \boldsymbol{\rho}(\boldsymbol{z}) : \mathscr{R}^{d\times K} \to (0,1)^K$ is a continuous function and is twice differentiable at $\boldsymbol{\Theta}$ with $\boldsymbol{\rho}(\boldsymbol{\Theta}) = \boldsymbol{v}$.

Condition 3.3: The function of the allocation rule $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) = (g_1(\boldsymbol{x}, \boldsymbol{y}), \ldots, g_K(\boldsymbol{x}, \boldsymbol{y}))$ satisfies $\boldsymbol{g}(\boldsymbol{v}, \boldsymbol{v}) = \boldsymbol{v}$ and the following conditions.

(3.3.1). There exists a constant $0 \le \lambda_0 < 1$, such that for each $k = 1, \ldots, K$,

$$\frac{g_k(\boldsymbol{x}, \boldsymbol{y}) - y_k}{x_k - y_k} \le \lambda_0 \text{ for all } \boldsymbol{x}, \boldsymbol{y} \text{ with } x_k > y_k, \ \boldsymbol{x}\boldsymbol{1}' = \boldsymbol{y}\boldsymbol{1}' = 1.$$

(3.3.2). There exists $\delta > 0$, for which the function $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y})$ satisfies

$$\begin{aligned}\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) =& \boldsymbol{g}(\boldsymbol{v}, \boldsymbol{v}) + \sum_{k=1}^{K}(x_k - v_k)\frac{\partial \boldsymbol{g}}{\partial x_k}\Big|_{(\boldsymbol{v},\boldsymbol{v})} + \sum_{k=1}^{K}(y_k - v_k)\frac{\partial \boldsymbol{g}}{\partial y_k}\Big|_{(\boldsymbol{v},\boldsymbol{v})} \\ &+ o(\|\boldsymbol{x} - \boldsymbol{v}\|^{1+\delta}) + o(\|\boldsymbol{y} - \boldsymbol{v}\|^{1+\delta}) \text{ as } (\boldsymbol{x}, \boldsymbol{y}) \to (\boldsymbol{v}, \boldsymbol{v}).\end{aligned}$$

3.5.2 Lemmas

For the doubly adaptive biased coin design without delayed and missing responses, certain values such as $I_{mis,k}(m) \equiv 0$ and $M_{k,l} \equiv 0$ for $l \ge 1$ and all m and k. Therefore, $\alpha_k = 0$, and Assumption 3.2 is satisfied for any $\phi > 0$. Hence, it is sufficient to prove Theorems 3.3-3.5. The following lemmas are essential in supporting the proofs of the three theorems.

Before we begin, we first introduce some settings used in the lemmas. For the $(m+1)$th patient,

we define k_m, $\widetilde{\boldsymbol{q}}_m = (\widetilde{q}_{m,1}, \ldots, \widetilde{q}_{m,K})$, $\widetilde{\boldsymbol{\Theta}}_m^{dm} = \big(\widetilde{\boldsymbol{\theta}}_{m,1}^{dm}, \ldots, \widetilde{\boldsymbol{\theta}}_{m,K}^{dm}\big)$ as

$$k_m = \tau_j, \;\; \widetilde{\boldsymbol{q}}_m = \boldsymbol{N}_{\tau_j}/\tau_j, \;\; \widetilde{\boldsymbol{\Theta}}_m^{dm} = \widehat{\boldsymbol{\Theta}}_{\tau_j}^{dm} \quad \text{for } \tau_j < m+1 \le \tau_{j+1}.$$

Then

$$\Pr(X_{m+1,k} = 1 | \mathscr{F}_m) = g_k\big(\widetilde{\boldsymbol{q}}_m, \boldsymbol{\rho}(\widetilde{\boldsymbol{\Theta}}_m^{dm})\big),$$

where $\mathscr{F}_m = \sigma(\mathscr{G}_{k_m}; \boldsymbol{X}_i, \boldsymbol{\xi}_i, \boldsymbol{I}_{mis}(i), k_i : i = 1, \ldots, m)$. Further, let $\widehat{\boldsymbol{\Theta}}_m^{mis} = (\widehat{\boldsymbol{\theta}}_{m,1}^{mis}, \ldots, \widehat{\boldsymbol{\theta}}_{n,K}^{mis})$ and

$$\widehat{\boldsymbol{\theta}}_{m,k}^{mis} = \frac{\sum_{i=1}^m X_{i,k}\boldsymbol{\xi}_{i,k}(1 - I_{mis,k}(i)) + \boldsymbol{\theta}_{0,k}}{\sum_{i=1}^m X_{i,k}(1 - I_{mis,k}(i)) + 1}, \quad k = 1, \ldots, K. \tag{3.16}$$

Based on this setting, we presented the lemmas.

Lemma 3.1. *If $\chi_j/\tau_j \to 0$ almost surely (a.s.) as $j \to \infty$, then*

$$\widetilde{\boldsymbol{q}}_m - \boldsymbol{N}_m/m \to 0 \;\; \textit{a.s.} \tag{3.17}$$

If Assumption 3.1 is satisfied, then

$$\widetilde{\boldsymbol{q}}_m - \boldsymbol{N}_m/m = o\big(m^{-1/2-\psi}\big) \;\; \textit{a.s.} \tag{3.18}$$

Proof. Note for $\tau_j < m + 1 \le \tau_{j+1}$,

$$\left|\widetilde{\boldsymbol{q}}_m - \frac{\boldsymbol{N}_m}{m}\right| = \left|\frac{\boldsymbol{N}_{\tau_j}}{\tau_j} - \frac{\boldsymbol{N}_m}{m}\right| = \left|\frac{\boldsymbol{N}_{\tau_j}}{\tau_j}\frac{m - \tau_j}{m} - \frac{N_m - N_{\tau_j}}{m}\right| \le \frac{m - \tau_j}{\tau_j} \le \frac{\chi_{j+1}}{\tau_j}.$$

(3.17) and (3.18) follow immediately. ■

Lemma 3.2. *Let $\{\eta_k, k = 1, 2, \ldots\}$ be a sequence of independent and identically distributed random variables with zero means and let $\{\varepsilon_k, k = 1, 2, \ldots\}$ be another sequence of random variables that takes only two values, 0 or 1. Denote $\zeta_n = \sum_{i=1}^n \varepsilon_i$. Suppose that for each n, η_n is independent of*

$\{\eta_1, \ldots, \eta_{n-1}, \varepsilon_1, \ldots, \varepsilon_n\}$. *Then, in the event that* $\{\zeta_n \to \infty\}$,

$$\frac{\sum_{i=1}^{n} \varepsilon_i \eta_i}{\zeta_n} \to 0 \ \textit{a.s., if} \ \mathrm{E}\,|\eta_1| < \infty,$$

and

$$\limsup_{n\to\infty} \frac{|\sum_{i=1}^{n} \varepsilon_i \eta_i|}{\sqrt{2\zeta_n \log\log \zeta_n}} = \left(\mathrm{E}\eta_1^2\right)^{1/2} \ \textit{a.s., if} \ \mathrm{E}\eta_1^2 < \infty.$$

Proof. The proof is similar to Lemma A.4 in Hu and Zhang (2004) [21]. ■

Lemma 3.3. *For the doubly adaptive biased coin design with delayed and missing responses, if* $E(\|\boldsymbol{\xi}_{1,k}\|) < \infty$, $k = 1, \ldots, K$, *and Assumption 3.2 is satisfied, then the event* $\{N_{n,k} \to \infty\}$ *implies almost surely that* $T_{n,k}/N_{n,k} \to 1 - \alpha$, $\widetilde{\boldsymbol{\theta}}_{n,k}^{dm} \to \boldsymbol{\theta}_k$, *and* $\widehat{\boldsymbol{\theta}}_{n,k}^{dm} \to \boldsymbol{\theta}_k$.

Proof. Recall that

$$T_{n,k} = \sum_{m=1}^{n} \sum_{i=0}^{n-m} M_k(m,i)(1 - I_{mis,k}(m))X_{m,k},$$
$$S_{n,k} = \sum_{m=1}^{n} \sum_{i=0}^{n-m} M_k(m,i)(1 - I_{mis,k}(m))X_{m,k}\boldsymbol{\xi}_{m,k}$$

are respectively, the number and the total sum of the responses on treatment k that are observed up to the time patient $(n+1)$ is ready for treatment allocation, that $I_{mis,k}(m)$ is an indicator function that takes 1 when the response of patient m is missing, and that $M_k(m,j)$ is an indicator function that only represents the period $[(m+j),(m+j+1))$ in which we obtain the response of patient m. When $I_{mis,k}(m) = 0$, the patient's results will be available after treatment, even if there is a delay. Hence, we have

$$\sum_{j=0}^{\infty} M_k(m,j) = 1 \text{ if } I_{mis,k}(m) = 0 \text{ and } \sum_{j=0}^{\infty} M_k(m,j) = 0 \text{ if } I_{mis,k}(m) = 1. \tag{3.19}$$

Then

$$S_{n,k} = \sum_{i=1}^{n} X_{i,k} \left(1 - I_{mis,k}(i)\right) \boldsymbol{\xi}_{i,k} - \sum_{i=1}^{n} \sum_{j=n-i+1}^{\infty} X_{i,k} M_k(i,j) \left(1 - I_{mis,k}(i)\right) \boldsymbol{\xi}_{i,k}$$

$$=: J_{n,k} - R_{n,k}. \tag{3.20}$$

Let $I_k(i,p) = \sum_{j=p+1}^{\infty} M_k(i,j)$. Choose $l = l_n = [(1+N_{n,k})^{1/2}]$. Similar to (17) of Hu et al. (2008) [23], we have

$$\begin{aligned}
\|R_{n,k}\| &= \left\| \Big(\sum_{i=n-l+1}^{n} + \sum_{i=1}^{n-l} \Big) X_{i,k} I_k(i, n-i) \left(1 - I_{mis,k}(i)\right) \boldsymbol{\xi}_{i,k} \right\| \\
&\leq \sum_{i=n-l+1}^{n} X_{i,k} \|\boldsymbol{\xi}_{i,k}\| + \sum_{i=1}^{n} X_{i,k} I_k(i,l) \|\boldsymbol{\xi}_{i,k}\| \\
&\leq l_n \mathrm{E}[\|\boldsymbol{\xi}_{1,k}\|] + \sum_{i=n-l_n+1}^{n} X_{i,k} (\|\boldsymbol{\xi}_{i,k}\| - \mathrm{E}[\|\boldsymbol{\xi}_{i,k}\|]) + \sum_{m=1}^{n} X_{m,k} I_k(m, l_m) \|\boldsymbol{\xi}_{m,k}\| \\
&\leq l_n \mathrm{E}[\|\boldsymbol{\xi}_{1,k}\|] + 2 \max_{m \leq n} \Big| \sum_{i=1}^{m} X_{i,k} (\|\boldsymbol{\xi}_{i,k}\| - \mathrm{E}[\|\boldsymbol{\xi}_{i,k}\|]) \Big| \\
&\quad + \sum_{m=1}^{n} X_{m,k} \|\boldsymbol{\xi}_{m,k}\| I\{\|\boldsymbol{\xi}_{m,k}\| \geq C\} + C \sum_{m=1}^{n} X_{m,k} I_k(m, l_m).
\end{aligned} \tag{3.21}$$

Note that, when $\tau_j < m+1 \leq \tau_{j+1}$, $\boldsymbol{\xi}_{m+1,k}$, $I_{mis,k}(m+1)$, $M_k(m+1,l)$ are independent of the history sigma field $\mathscr{G}_{\tau_j} = \sigma(\tau_1, \ldots, \tau_j, \boldsymbol{\xi}_{i,k}, I_{mis,k}(i), M_k(i,s), s = 0, \ldots, \tau_j - i, i = 1, \ldots, \tau_j, k = 1, \ldots, K)$ when group $(j+1)$ is ready for treatment assignment, and, the results of the assignments $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{m+1}$ are generated according to $\mathscr{G}_{\tau_j}$. Hence

$$\boldsymbol{\xi}_{m+1,k}, I_{mis,k}(m+1), M_k(m+1,l) \text{ are independent of } \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{m+1}. \tag{3.22}$$

It follows that

$$\begin{aligned}
&\mathrm{E}[M_k(m,j) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_m] = \mathrm{E}[M_k(m,j)] \\
=&\mathrm{E}[M_k(m,j) | I_{mis,k}(m) = 1] \alpha_k + \mathrm{E}[M_k(m,j) | I_{mis,k}(m) = 0](1 - \alpha_k) = \mu_{k,j}^{(m)} (1 - \alpha_k).
\end{aligned}$$

Note $l_m = [(N_{m,k}+1)^{1/2}] \in \sigma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)$. It follows that

$$\begin{aligned}
&\mathrm{E}\left[I_k(m, l_m)\middle|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m\right] = \sum_{j=l_m+1}^{\infty} \mathrm{E}\left[M(m,j)\middle|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m\right] \\
&= \sum_{j=l_m+1}^{\infty} \mu_{k,j}^{(m)}(1-\alpha_k) \leq C(1+l_m)^{-\phi} \leq C(1+N_{m,k})^{-\phi/2}.
\end{aligned}$$

Then

$$\begin{aligned}
&\mathrm{E}\left[\sum_{m=1}^{\infty} \frac{X_{m,k} I_k(m, l_m)}{1+N_{m,k}}\right] = \sum_{m=1}^{\infty} \mathrm{E}\left[\frac{X_{m,k}\mathrm{E}\left[I_k(m, l_m)\middle|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m\right]}{1+N_{m,k}}\right] \\
\leq & C\sum_{m=1}^{\infty} \mathrm{E}\left[\frac{X_{m,k}}{(1+N_{m,k})^{1+\phi/2}}\right] = C\mathrm{E}\left[\sum_{m=1}^{\infty} \int_{N_{m-1,k}}^{N_{m,k}} \frac{dx}{(1+N_{m,k})^{1+\phi/2}}\right] \\
\leq & C\mathrm{E}\left[\sum_{m=1}^{\infty} \int_{N_{m-1,k}}^{N_{m,k}} \frac{dx}{(1+x)^{1+\phi/2}}\right] \leq C\int_0^{\infty} \frac{dx}{(1+x)^{1+\phi/2}} < \infty,
\end{aligned}$$

which implies

$$\sum_{m=1}^{\infty} \frac{X_{m,k} I_k(m, l_m)}{1+N_{m,k}} < \infty \ \text{ a.s.}$$

And thus,

$$\frac{\sum_{m=1}^{n} X_{m,k} I_k(m, l_m)}{1+N_{n,k}} \to 0 \ \text{ a.s. on } \{N_{n,k} \to \infty\}.$$

By (3.22) and that $\{\boldsymbol{\xi}_{i,k}; i = 1, 2 \ldots\}$ is a sequence of i.i.d. random vector, $\boldsymbol{\xi}_{m,k}$ is independent of $\{X_{1,k}, \ldots, X_{m,k}, \boldsymbol{\xi}_{1,k}, \ldots, \boldsymbol{\xi}_{m-1,k}\}$. For the second and the third terms of (3.21), by Lemma 3.2 we have

$$\frac{\sum_{i=1}^{m} X_{i,k}(\|\boldsymbol{\xi}_{i,k}\| - \mathrm{E}[\|\boldsymbol{\xi}_{i,k}\|])}{N_{n,k}} \to 0 \ \text{ a.s. on } \{N_{n,k} \to \infty\},$$

$$\begin{aligned}
\frac{\sum_{i=1}^{m} X_{i,k}\|\boldsymbol{\xi}_{i,k}\| I\{\|\boldsymbol{\xi}_{i,k}\| \geq C\}}{N_{n,k}} \to & \mathrm{E}\left[\|\boldsymbol{\xi}_{1,k}\| I\{\|\boldsymbol{\xi}_{1,k}\| \geq C\}\right] \ \text{ a.s. on } \{N_{n,k} \to \infty\} \\
\to & 0 \text{ as } \ C \to \infty.
\end{aligned}$$

It follows that

$$\frac{R_{n,k}}{N_{n,k}} \to 0 \text{ a.s. on } \{N_{n,k} \to \infty\}. \tag{3.23}$$

For $J_{n,k}$ in (3.20), by applying Lemma 3.2 again we have

$$\frac{J_{n,k}}{N_{n,k}} \to \mathrm{E}[(1 - I_{mis,k}(1))\, \boldsymbol{\xi}_{1,k}] = (1-\alpha_k)\boldsymbol{\theta}_k \text{ a.s. on } \{N_{n,k} \to \infty\}. \tag{3.24}$$

Combing (3.20), (3.23) and (3.24) yields

$$\frac{S_{n,k}}{N_{n,k}} \to (1-\alpha_k)\boldsymbol{\theta}_k \text{ a.s. on } \{N_{n,k} \to \infty\}. \tag{3.25}$$

Note that $T_{n,k}$ is a special case of $S_{n,k}$ with $\boldsymbol{\xi}_{m,k} \equiv 1$. We have

$$\frac{T_{n,k}}{N_{n,k}} \to (1-\alpha_k) \text{ a.s. on } \{N_{n,k} \to \infty\}. \tag{3.26}$$

From (3.25) and (3.26), it follows that

$$\widehat{\boldsymbol{\theta}}_n^{dm} = \frac{S_{n,k} + \boldsymbol{\theta}_{0,k}}{T_{n,k}+1} \to \boldsymbol{\theta}_k \text{ and } \widehat{\boldsymbol{\theta}}_{\tau_j}^{dm} \to \boldsymbol{\theta}_k \text{ a.s. on } \{N_{n,k} \to \infty\}.$$

Hence $\widetilde{\boldsymbol{\theta}}_{n,k}^{dm} \to \boldsymbol{\theta}_k$ a.s. on $\{N_{n,k} \to \infty\}$. The proof is completed. ■

Lemma 3.4. *For the doubly adaptive biased coin design with delayed and missing responses, suppose that $\boldsymbol{N}_n/n \to \boldsymbol{v}$ a.s. If Condition 3.1, and Assumptions 3.1 and 3.2 are satisfied, then for $\varphi' = (2+\epsilon_0)/(2+\epsilon_0+(1+\epsilon_0)\phi)$,*

$$\widehat{\boldsymbol{\Theta}}_n^{dm} - \widehat{\boldsymbol{\Theta}}_n^{mis} = o\left(n^{\varphi'-1}(\log n)^2\right) + o\left(n^{-(1+\varepsilon_0)/(2+\varepsilon_0)}\right) \quad a.s. \tag{3.27}$$

$$\widetilde{\boldsymbol{\Theta}}_n^{dm} - \widehat{\boldsymbol{\Theta}}_n^{mis} = o\left(n^{\varphi'-1}(\log n)^2\right) + o\left(n^{-1/2-\delta}\right) \quad a.s. \tag{3.28}$$

for some $0 < \delta \leq \min\{\psi, \epsilon_0/(2(2+\epsilon_0))\}$. *Also,*

$$\widehat{\boldsymbol{\Theta}}_n^{mis} - \boldsymbol{\Theta} = O\left(\sqrt{(\log\log n)/n}\right) \text{ a.s.} \tag{3.29}$$

Proof. Write $N_{n,k}^{mis} = \sum_{i=1}^{n} X_{i,k}(1 - I_{mis,k}(i))$. Then by Lemma 3.2,

$$\frac{N_{n,k}^{mis}}{n} = \frac{N_{n,k}^{mis}}{N_{n,k}} \cdot \frac{N_{n,k}}{n} \to (1-\alpha_k)v_k \text{ a.s.}$$

and

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{n,k}^{mis} - \boldsymbol{\theta}_{n,k} =& \frac{\sum_{i=1}^{n} X_{i,k}(1 - I_{mis,k}(i))(\boldsymbol{\xi}_{i,k} - \boldsymbol{\theta}_k) + \boldsymbol{\theta}_{0,k} - \boldsymbol{\theta}_k}{N_{n,k}^{mis} + 1} \\ =& O\left(\sqrt{(\log\log N_{n,k}^{mis})/N_{n,k}^{mis}}\right) = O\left(\sqrt{(\log\log n)/n}\right) \text{ a.s.}\end{aligned}$$

(3.29) is proved.

For (3.27), recall (3.20). Note

$$\mathrm{E}[I_k(i,l)] = \sum_{j=l+1}^{\infty} \mathrm{E}[M_k(i,j)] = \sum_{j=l+1}^{\infty} \mu_{k,j}^{(i)}(1-\alpha_k) \leq C(l+1)^{-\phi}.$$

By choosing $l = l_n = [n^{\varphi'}]$ with $\varphi' = (2+\varepsilon_0) \,/\, (2+\varepsilon_0+(1+\varepsilon_0)\,\phi)$ and noticing

$$\|R_{n,k}\| \leq l_n \mathrm{E}[\|\boldsymbol{\xi}_{1,k}\|] + \sum_{i=n-l_n+1}^{n} (\|\boldsymbol{\xi}_{i,k}\| - \mathrm{E}[\|\boldsymbol{\xi}_{i,k}\|]) + \sum_{m=1}^{n} I_k(m, l_m)\|\boldsymbol{\xi}_{m,k}\|,$$

similar to (18) of Hu et al. (2008)[23], we have

$$R_{n,k} = o\big(n^{\varphi'}(\log n)^2\big) + o\big(n^{\frac{1}{2+\epsilon_0}}\big) \; a.s.$$

Hence

$$S_{n,k} = \sum_{i=1}^{n} X_{i,k}\,(1 - I_{mis,k}(i))\,\boldsymbol{\xi}_{i,k} + o\big(n^{\varphi'}(\log n)^2\big) + o\big(n^{\frac{1}{2+\epsilon_0}}\big) \text{ a.s.}$$

Similarly,

$$T_{n,k} = N_{n,k}^{mis} + o\big(n^{\varphi'}(\log n)^2\big) + o\big(n^{\frac{1}{2+\epsilon_0}}\big) = N_{n,k}^{mis}\Big(1 + o\big(n^{\varphi'-1}(\log n)^2\big) + o\big(n^{-\frac{1+\epsilon_0}{2+\epsilon_0}}\big)\Big) \text{ a.s.}$$

It follows that

$$\widehat{\boldsymbol{\theta}}_{n,k}^{dm} = \frac{\sum_{i=1}^{n} X_{i,k}\left(1 - I_{mis,k}(i)\right)\boldsymbol{\xi}_{i,k}}{N_{n,k}^{mis}} + o\big(n^{\varphi'-1}(\log n)^2\big) + o\big(n^{-\frac{1+\epsilon_0}{2+\epsilon_0}}\big) \text{ a.s. .}$$

(3.27) is proved. For $\tau_j < n+1 \leq \tau_{j+1}$, we have $\tau_j \sim n$,

$$\widehat{\boldsymbol{\Theta}}_{\tau_j}^{dm} - \widehat{\boldsymbol{\Theta}}_{\tau_j}^{mis} = o\left(n^{\varphi'-1}(\log n)^2\right) + o\left(n^{-(1+\varepsilon_0)/(2+\varepsilon_0)}\right) \quad \text{a.s.}$$

Also, for $\tau_j < n+1 \leq \tau_{j+1}$, we have $n - \tau_j \leq \chi_{j+1} = o(n^{1/2-\psi})$ almost surely. So

$$\begin{aligned}
&\|\widehat{\boldsymbol{\theta}}_{\tau_j,k}^{mis} - \widehat{\boldsymbol{\theta}}_{n,k}^{mis}\| \\
=&\left\| \frac{\sum_{i=\tau_j+1}^{n} X_{i,k}\boldsymbol{\xi}_{i,k}(1 - I_{mis}(i))}{N_{n,k}^{mis}+1} - \widehat{\boldsymbol{\theta}}_{\tau_j,k}\frac{\sum_{i=\tau_j+1}^{n} X_{i,k}(1 - I_{mis}(i))}{N_{n,k}^{mis}+1} \right\| \\
\leq&\frac{\sum_{i=\tau_j+1}^{n}\|\boldsymbol{\xi}_{i,k}\|}{N_{n,k}^{mis}+1} + \|\widehat{\boldsymbol{\theta}}_{\tau_j,k}\|\frac{n-\tau_j}{N_{n,k}^{mis}+1} \leq c\frac{\sum_{i=n-o(n^{1/2-\psi})}^{n}\|\boldsymbol{\xi}_{i,k}\|}{n} + c\frac{n-\tau_j}{n} \\
=&c\frac{\sum_{i=n-o(n^{1/2-\psi})}^{n}(\|\boldsymbol{\xi}_{i,k}\| - E(\|\boldsymbol{\xi}_{i,k}\|))}{n} + o(n^{-1/2-\psi}) = o(n^{-1/2-\delta}) \text{ a.s.}
\end{aligned}$$

for some $\delta > 0$. (3.28) is proved. ■

3.5.3 Proof of the Main Theroems

This section contains all the proofs of the results presented in Chapter 3. As mentioned in Section 1 this appendix, Theorems 3.1-3.2 apply in the case where $I_{mis,k}(m) \equiv 0$ and $M_{k,l} \equiv 0$ for $l \geqslant 1$ and all m and k. Therefore, we will focus on proving Theorems 3.3-3.5 in the following.

of Theorem 3.3. Define $\widetilde{\boldsymbol{\rho}}_m = \boldsymbol{\rho}(\widetilde{\boldsymbol{\Theta}}_m^{dm})$, $p_{m+1,k} = g_k(\widetilde{\boldsymbol{q}}_m, \widetilde{\boldsymbol{\rho}}_m)$. As we proved in Lemma 3.3,

$$\widehat{\boldsymbol{\theta}}_{m,k}^{dm} \to \boldsymbol{\theta}_k \text{ and } \widetilde{\boldsymbol{\theta}}_{m,k}^{dm} \to \boldsymbol{\theta}_k \quad \text{a.s. in the event that } \{N_{m,k} \to \infty\}. \tag{3.30}$$

On the other hand, if $\{\sup_m N_{m,k} < \infty\}$, then there is m_0 such that $X_{m,k} = 0$ for $m > m_0$, and so

$$\begin{aligned} S_{m,k} =& \sum_{i=1}^{m}\sum_{j=0}^{m-i} M_k(i,j)(1-I_{mis,k}(i))X_{i,k}\boldsymbol{\xi}_{i,k} = \sum_{i=1}^{m_0}\sum_{j=0}^{m-i} M_k(i,j)(1-I_{mis,k}(i))X_{i,k}\boldsymbol{\xi}_{i,k} \\ \to& \sum_{i=1}^{m_0}\sum_{j=0}^{\infty} M_k(i,j)(1-I_{mis,k}(i))X_{i,k}\boldsymbol{\xi}_{i,k} = \sum_{i=1}^{m_0}(1-I_{mis,k}(i))X_{i,k}\boldsymbol{\xi}_{i,k} \end{aligned}$$

and

$$T_{m,k} = \sum_{i=1}^{m_0}\sum_{j=0}^{m-i} M_k(i,j)(1-I_{mis,k}(i))X_{i,k} \to \sum_{i=1}^{m_0}(1-I_{mis,k}(i))X_{i,k}.$$

Hence, $\widehat{\boldsymbol{\theta}}_{m,k}^{dm}$ and $\widetilde{\boldsymbol{\theta}}_{m,k}^{dm}$ converge to $\widehat{\boldsymbol{\theta}}_{m_0,k}^{mis}$. In either case, $\widetilde{\boldsymbol{\theta}}_{m,k}^{dm}$ has a limit $\boldsymbol{\theta}_k^*$ in the parameter space, $k = 1, \ldots, K$. By the continuity of $\boldsymbol{\rho}(\cdot)$, $\widetilde{\boldsymbol{\rho}}_{m-1} \to \boldsymbol{\rho}(\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_K^*) := \boldsymbol{v}^* \in (0,1)^{\otimes K}$ almost surely. Denote $M_{n,k} = \sum_{m=1}^{n}(X_{m,k} - E[X_{m,k}|\mathcal{F}_{m-1}])$. Then, $\max_{m\le n}|M_{m,k}|/n \to 0$ almost surely by the law of large numbers of martingales.

Note $\widetilde{\boldsymbol{q}}_m - \boldsymbol{N}_m/m \to 0$ almost surely by Lemma 3.1. For any given $\epsilon > 0$, we can choose a m_0 such that $\|\boldsymbol{N}_m/m - \widetilde{\boldsymbol{q}}_m\| < \epsilon$ for $m \ge m_0$. Let $\nu_n = \max\{Km_0 + 1 \le m \le n : \widetilde{q}_{m-1,k} \le \widetilde{\rho}_{m-1,k}\}$ and $\max\{\emptyset\} = Km_0$. Note the allocation probability $p_{m,k} \le \widetilde{\rho}_{m-1,k} + \lambda_0(\widetilde{q}_{m-1,k} - \widetilde{\rho}_{m-1,k})$ if $\widetilde{q}_{m-1,k} > \widetilde{\rho}_{m-1,k}$ by Condition 3.1. It follows that

$$\begin{aligned} N_{n,k} =& N_{\nu_n,k} + \sum_{m=\nu_n+1}^{n}(X_{m,k} - E[X_{m,k}|\mathcal{F}_{m-1}]) + \sum_{m=\nu_n+1}^{n} p_{m,k} \\ \le& 1 + N_{\nu_n-1,k} + M_{n,k} - M_{\nu_n,k} + \sum_{m=\nu_n}^{n-1}\Big(\widetilde{\rho}_{m,k} + \lambda_0(\widetilde{q}_{m,k} - \widetilde{\rho}_{m,k})\Big). \end{aligned}$$

If $\nu_n = Km_0$, then $N_{\nu_n-1,k} \le Km_0$. If $\nu_n \ge Km_0 + 1$, then

$$\frac{N_{\nu_n-1,k}}{\nu_n - 1} \le \widetilde{q}_{\nu_n-1,k} + \epsilon \le \widetilde{\rho}_{\nu_n-1,k} + \epsilon.$$

It follows that

$$\begin{aligned} N_{n,k} \leq & 1 + Km_0 + 2\max_{m\leq n}|M_{m,k}| + (\nu_n - 1)\epsilon \\ & + (\nu_n - 1)\widetilde{\rho}_{\nu_n-1,k} + \sum_{m=\nu_n}^{n-1}\widetilde{\rho}_{m,k} + \lambda_0\sum_{m=\nu_n}^{n-1}(\widetilde{q}_{m,k} - \widetilde{\rho}_{m,k}). \end{aligned}$$

It then follows that

$$\begin{aligned} \limsup_{n\to\infty} N_{n,k}/n \leq & \epsilon + v_k^* + \lambda_0\big(\limsup_{m\to\infty}\widetilde{q}_{m,k} - v_k^*\big) \\ = & \epsilon + v_k^* + \lambda_0\big(\limsup_{m\to\infty} N_{m,k}/m - v_k^*\big) \text{ a.s.} \end{aligned}$$

Note $\lambda_0 < 1$. We conclude that $\limsup_{n\to\infty} N_{n,k}/n \leq v_k^* + \epsilon$ almost surely. Hence, $\limsup_{n\to\infty} N_{n,k}/n \leq v_k^*$ almost surely, $k = 1, \ldots, K$. From the fact that $\sum_{k=1}^n N_{n,k}/n = \sum_{k=1}^K v_k^* = 1$, we conclude that $N_{n,k}/n \to v_k^* \in (0,1)$ almost surely, which implies that $N_{n,k} \to \infty$ almost surely, $k = 1, \ldots, K$. Hence, the limits θ_k^* and θ_k ($\boldsymbol{v}^*$ and $\boldsymbol{v}$) must be identical by (3.30). The proof is now complete. ■

of Theorem 3.4. (3.12) follows from (3.27) and (3.29) immediately. By noticing (3.18) and (3.28)-(3.29), we have

$$\begin{aligned} & \Delta(\boldsymbol{N}_m - m\boldsymbol{v}) = \boldsymbol{X}_m - \boldsymbol{v} = \Delta\boldsymbol{M}_m + \boldsymbol{g}\big(\widetilde{\boldsymbol{q}}_{m-1}, \boldsymbol{\rho}(\widetilde{\boldsymbol{\Theta}}_{m-1}^{dm})\big) - \boldsymbol{g}(\boldsymbol{v}, \boldsymbol{\rho}(\boldsymbol{\Theta})) \\ = & \Delta\boldsymbol{M}_m + (\widetilde{\boldsymbol{q}}_{m-1} - \boldsymbol{v})\boldsymbol{H} + (\widetilde{\boldsymbol{\Theta}}_{m-1}^{dm} - \boldsymbol{\Theta})\frac{\partial\boldsymbol{\rho}}{\partial\boldsymbol{\Theta}}\boldsymbol{E} + O(\|\widetilde{\boldsymbol{q}}_{m-1} - \boldsymbol{v}\|^2) + O(\|\widetilde{\boldsymbol{\Theta}}_{m-1}^{dm} - \boldsymbol{\Theta}\|^2) \\ = & \Delta\boldsymbol{M}_m + \Big(\frac{\boldsymbol{N}_{m-1}}{m-1} - \boldsymbol{v}\Big)\boldsymbol{H} + (\widehat{\boldsymbol{\Theta}}_{m-1}^{mis} - \boldsymbol{\Theta})\frac{\partial\boldsymbol{\rho}}{\partial\boldsymbol{\Theta}}\boldsymbol{E} \\ & + O\Big(\Big\|\frac{\boldsymbol{N}_{m-1}}{m-1} - \boldsymbol{v}\Big\|^2\Big) + o\left(m^{\varphi'-1}(\log m)^2\right) + o(m^{-1/2-\delta}) \text{ a.s.} \qquad (3.31) \\ = & \Big(\frac{\boldsymbol{N}_{m-1}}{m-1} - \boldsymbol{v}\Big)\boldsymbol{H} + o\Big(\Big\|\frac{\boldsymbol{N}_{m-1}}{m-1} - \boldsymbol{v}\Big\|\Big) + o\left(m^{\varphi'-1}(\log m)^2\right) + O\left(\sqrt{\frac{\log\log m}{m}}\right) \text{ a.s.} \end{aligned}$$

(3.32)

Here and in the sequel, $\Delta\boldsymbol{u}_m = \boldsymbol{u}_m - \boldsymbol{u}_{m-1}$ for a sequence $\boldsymbol{u}_m$. By (3.32) and Lemma A.2 of Hu and Zhang (2004) [21], the proofs of (3.13) and (3.14) follow closely to those given in Theorem

4.2 of Hu and Zhang (2004) [21]. ■

of Theorem 3.5. By Lemma 3.2 and (3.14) we have

$$\begin{aligned}\frac{N_{n,k}^{mis}}{n}-(1-\alpha_k)v_k=&\left(\frac{N_{n,k}^{mis}}{N_{n,k}}-(1-\alpha_k)\right)\frac{N_{n,k}}{n}+\left(\frac{N_{n,k}}{n}-v_k\right)(1-\alpha_k)\\ =&O\left(\sqrt{\frac{\log\log n}{n}}\right)\ \text{a.s.},\end{aligned}$$

where $N_{n,k}^{mis}=\sum_{i=1}^{n}X_{i,k}(1-I_{mis,k}(i))$. By noticing (3.16), similar to (A.10) of Hu and Zhang (2004) [21] we have

$$\widehat{\boldsymbol{\Theta}}_n^{mis}-\boldsymbol{\Theta}=\frac{\boldsymbol{Q}_n^{mis}}{n}+O\Big(\frac{\log\log n}{n}\Big)\ \text{a.s.},$$

where $\boldsymbol{Q}_n^{mis}=(\boldsymbol{Q}_{n,1}^{mis},\ldots,\boldsymbol{Q}_{n,K}^{mis})$ is a martingale with $\Delta\boldsymbol{Q}_{n,k}^{mis}=X_{n,k}(1-I_{mis,k}(n))(\boldsymbol{\xi}_{n,k}-\boldsymbol{\theta}_k)/(v_k(1-\alpha_k))$. Following the lines at page 297 of Hu and Zhang (2004) [21] we can derive that $cov\{\Delta\boldsymbol{M}_n,\Delta\boldsymbol{Q}_n^{mis}|\mathscr{F}_{n-1}\}=0$,

$$var[\Delta\boldsymbol{M}_n|\mathscr{F}_{n-1}]=\boldsymbol{\Sigma}_1+O\Big(\Big\|\widetilde{\boldsymbol{q}}_{n-1}-\boldsymbol{v}\Big\|+\big\|\widetilde{\boldsymbol{\Theta}}_{n-1}^{dm}-\boldsymbol{\Theta}\big\|\Big)=\boldsymbol{\Sigma}_1+O\Big(\sqrt{\frac{\log\log n}{n}}\Big)\ \text{a.s.},$$

$$\begin{aligned}var[\Delta\boldsymbol{Q}_n^{mis}|\mathscr{F}_{n-1}]=&\boldsymbol{V}^{mis}+O\Big(\Big\|\widetilde{\boldsymbol{q}}_{n-1}-\boldsymbol{v}\Big\|+\big\|\widetilde{\boldsymbol{\Theta}}_{n-1}^{dm}-\boldsymbol{\Theta}\big\|\Big)\\ =&\boldsymbol{V}^{mis}+O\Big(\sqrt{\frac{\log\log n}{n}}\Big)\ \text{a.s.}\end{aligned}$$

On the other hand, it is easily seen that $\|\Delta\boldsymbol{M}_n\|\le 1$, $\mathrm{E}\|\Delta\boldsymbol{Q}_n\|^{2+\epsilon_0}\le C$. By Corollary 1.1 of Zhang (2004) [61], we can define two independent standard multi-dimensional Brownian motions $\boldsymbol{W}_t\in\mathscr{R}^K$ and $\boldsymbol{B}_t\in\mathscr{R}^{d\times K}$ such that for some $0<\kappa<\min\{1/2-\lambda,\delta\}$,

$$\boldsymbol{M}_n=\boldsymbol{W}_n\boldsymbol{\Sigma}_1^{1/2}+o(n^{1/2-\kappa})\ \text{ and }\ \boldsymbol{Q}_n=\boldsymbol{B}_n(\boldsymbol{V}^{mis})^{1/2}+o(n^{1/2-\kappa})\ \text{ a.s.}$$

Notice that $\varphi'=(2+\varepsilon_0)\,/\,(2+\varepsilon_0+(1+\varepsilon_0)\,\phi)<1/2$. We assume that $0<\kappa<\frac{1}{2}-\varphi'$. It

follows that $\widehat{\boldsymbol{\Theta}}_n^{mis} - \boldsymbol{\Theta} = \frac{\boldsymbol{B}_n}{n}(\boldsymbol{V}^{mis})^{1/2} + o(n^{-1/2-\kappa})$ almost surely and, by (3.31), (3.14) and (3.4),

$$\begin{aligned}
\boldsymbol{N}_n - n\boldsymbol{v} =& \boldsymbol{M}_n + \sum_{m=1}^{n-1}\Big(\frac{\boldsymbol{N}_m}{m} - \boldsymbol{v}\Big)\boldsymbol{H} + \sum_{m=0}^{n-1}(\widehat{\boldsymbol{\Theta}}_m^{mis} - \boldsymbol{\Theta})\frac{\partial \boldsymbol{\rho}}{\partial \boldsymbol{\Theta}}\boldsymbol{E} + o(n^{1/2-\kappa}) \\
=& \boldsymbol{W}_n\boldsymbol{\Sigma}_1^{1/2} + \sum_{m=1}^{n-1}\frac{\boldsymbol{B}_m}{m}(\boldsymbol{V}^{mis})^{1/2}\frac{\partial \boldsymbol{\rho}}{\partial \boldsymbol{\Theta}}\boldsymbol{E} + \sum_{m=1}^{n-1}\Big(\frac{\boldsymbol{N}_m}{m} - \boldsymbol{v}\Big)\boldsymbol{H} + o(n^{1/2-\kappa}) \\
=& \boldsymbol{W}_n\boldsymbol{\Sigma}_1^{1/2} + \int_0^n \frac{\boldsymbol{B}_t}{t}(\boldsymbol{V}^{mis})^{1/2}\frac{\partial \boldsymbol{\rho}}{\partial \boldsymbol{\Theta}}\boldsymbol{E}dt + \sum_{m=1}^{n-1}\Big(\frac{\boldsymbol{N}_m}{m} - \boldsymbol{v}\Big)\boldsymbol{H} + o(n^{1/2-\kappa}) \\
=& \boldsymbol{G}_n^{mis} - \int_0^n \frac{\boldsymbol{G}_t^{mis}}{t}\boldsymbol{H}dt + \sum_{m=1}^{n-1}\Big(\frac{\boldsymbol{N}_m}{m} - \boldsymbol{v}\Big)\boldsymbol{H} + o(n^{1/2-\kappa}) \\
=& \boldsymbol{G}_n^{mis} + \sum_{m=1}^{n-1}\frac{\boldsymbol{N}_m - m\boldsymbol{v} - \boldsymbol{G}_m^{mis}}{m}\boldsymbol{H} + o(n^{1/2-\kappa}) \ \text{ a.s.}
\end{aligned}$$

It follows that $\boldsymbol{N}_n - n\boldsymbol{v} - \boldsymbol{G}_n^{mis} = o(n^{1/2-\kappa})$ almost surely by Lemma A.2 of Hu and Zhang (2004) [21]. Further, by (3.27)

$$\widehat{\boldsymbol{\Theta}}_n^{dm} - \boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}_n^{mis} - \boldsymbol{\Theta} + o(n^{-1/2-\kappa}) = \frac{\boldsymbol{B}_n}{n}(\boldsymbol{V}^{mis})^{1/2} + o(n^{-1/2-\kappa}) \ \text{ a.s.}$$

The proof of (3.15) is completed. For deriving the variability $\boldsymbol{\Lambda}^{mis} = var(\boldsymbol{G}_1^{mis})$, we refer to Hu and Zhang (2004) [21] and Zhang (2007) [64]. ■

CHAPTER 4

Valid Test for Multi-arm Trials with Generalized Linear Models under Covariate-adaptive Randomization

4.1 Introduction

Clinical trials are widely utilized in medical research to evaluate treatment efficacy and establish causal relationships between interventions and health outcomes. Traditional clinical trials typically employed simple two-arm comparisons with clearly defined continuous endpoints. However, these conventional designs have become inadequate for the complexity of modern medical interventions and increased demand for trial efficiency and adaptability. Consequently, modern clinical trials have evolved towards more sophisticated designs. For example, dose-finding studies, seamless trials, and shared control designs are widely employed, often simultaneously comparing multiple treatments. Despite their wide application, statistical inference methods remain predominantly focused on two-arm trials with continuous outcomes, leaving methods for general outcomes in multi-arm settings underdeveloped.

Several studies have begun addressing inference issues in complex trial designs under CAR. CAR adaptively assigns patients to treatment groups based on previous patient allocations and their baseline covariates (e.g., age, gender, disease severity, pre-treatment). Its randomization framework and associated theoretical properties have been extensively studied, providing robust foundations for their use in clinical trials [40, 25, 22, 33, 20, 39, 35]. Consequently, CAR methods are now widely adopted in practice to ensure efficient and precise treatment comparisons.

Building upon CAR's extensive theoretical foundation, recent research has started addressing inference issues in multi-arm trials. For instance, Nugni et al. (2019)[9] expanded their earlier work on two-arm trials [8] to multi-arm trials with continuous outcomes, proposing robust methods

such as two-sample t-tests and regression adjustments under stratified randomization. Additionally, Ma et al. (2022)[38] investigated adjusted inference methods in a multi-arm setting as part of a seamless trial design. However, these studies remain limited to continuous outcomes. Although some researchers have explored other outcome types, these investigations have mostly focused on two-arm trials [37, 48, 58, 57, 31]. Thus, further research specifically addressing inference methodologies for multi-arm trials with general outcome types is necessary and timely.

Currently, a critical gap remains in statistical methodologies for multi-arm trials with non-continuous outcomes. A particularly important class of such non-continuous endpoints includes binary responses and count data, which are fundamental to clinical research and are usually analyzed within the GLM framework. The methodological gap stems from two fundamental challenges. First, unlike two-arm trials, multi-arm trials typically employ a shared control group, creating dependencies among treatment comparisons that invalidate traditional statistical approaches. Second, when analyzing data under CAR where the endpoints accommodate GLMs, the common practice of omitting covariates from the working model raises issues about the validity of standard inference method.

To address these challenges, we develop a comprehensive framework for valid statistical inference in multi-arm trials with GLM-accommodating endpoints, particularly focusing on scenarios where the working model omits covariates. Our work makes three main contributions: (1) theoretically investigating the asymptotic properties of the standard Wald test under CAR in multi-arm trials with shared controls; (2) developing valid statistical inference methods that properly account for both the shared control structure and scenarios where working models cannot include all randomization covariates; and (3) establishing theoretical guarantees for the proposed methods under GLM settings. Our framework provides the first theoretical foundation for statistical inference in this important yet under-explored setting, while maintaining practical applicability in real clinical trials.

This chapter is organized as follows. Section 4.2 introduces the GLM-based testing framework, including notation, working model, test statistic, and Type I error control. Section 4.3 presents

our main theoretical results, examining test size under commonly used models (logistic, linear, Poisson, and exponential regression) and exploring extensions to non-canonical link functions. Section 4.4 introduces an adjusted test that maintains valid Type I error control. Sections 4.5 and 4.6 evaluate our method through simulation studies and an application to a metastatic breast cancer trial, respectively. Section 4.7 concludes with a discussion of our findings.

4.2 Test under Generalized Linear Model

4.2.1 Framework and Notation

We consider a clinical trial with K treatment groups ($K \geq 1$) and one control group, where a total of N subjects are allocated to one of $(K+1)$ arms based on their covariate profiles via a multi-arm CAR procedure proposed by Hu et al. (2023)[20]. For each subject i, let T_{ik} be the treatment assignment indicator, where $T_{ik} = 1$ if subject i is assigned to arm k, and 0 otherwise. Here, $k = 0$ denotes the control group and $k \in \{1, \ldots, K\}$ denotes the treatment groups. Let $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iq})$ represent the covariate profile of subject i, and y_i be their observed response under the assigned treatment. The true model is assumed to be:

$$\mathbb{E}\left[Y_i \mid \boldsymbol{T}_i, \boldsymbol{Z}_i\right] = h\left(\delta_0 + \delta_1 T_{i1} + \delta_2 T_{i2} + \cdots + \delta_K T_{iK} + \boldsymbol{\beta}^T \boldsymbol{Z}_i\right) \tag{4.1}$$

where $\boldsymbol{T}_i$ is the assignment vector of subject i, δ_k is treatment effect of group k, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_q)$ are coefficients for prognostic factors $\boldsymbol{Z}_i$.

In the GLM framework considered in this paper, the function $h(\cdot)$ denotes the inverse of a monotonic link function, establishing the relationship between the linear predictor and the expected outcome. For example, $h(u) = u$ corresponds to linear regression models suitable for continuous outcomes, while $h(u) = \exp(u)/\{1+\exp(u)\}$ corresponds to logistic regression models commonly used for binary responses. Similarly, $h(u) = \exp(u)$ corresponds to Poisson regression models suitable for count data. Our theoretical results and proposed inference procedures hold under

general assumptions on the link function $h(\cdot)$, including canonical as well as certain non-canonical link functions. Specific assumptions and conditions on $h(\cdot)$ required for our theoretical results are detailed in later sections.

The framework we discuss in this article is based on the following assumptions.

Assumption 4.1. Over the whole article, we assume the covariates are independent and identically distributed for each patient.

Assumption 4.2. All covariates are centered with $\mathbb{E}(\boldsymbol{Z}) = \mathbf{0}$

For the Assumption 4.2, even though the true mean of covariate j is c ($c \neq 0$), we could centralized it by subtracting c for each value to construct a new variable.

In GLMs, the distribution of Y given the linear predictor $\delta_0 + \delta_1 T_{i1} + \delta_2 T_{i2} + \cdots + \delta_K T_{iK} + \boldsymbol{\beta}^T \boldsymbol{Z}$ is assumed to follow an exponential family distribution, which has the form

$$\exp\left\{\frac{Y\theta - b(\theta)}{\phi} + c(Y, \phi)\right\} \tag{4.2}$$

where ϕ is the scale parameter, θ is the canonical parameter of the distribution, and $b(\cdot)$ and $c(\cdot)$ are known functions.

4.2.2 Working Model and Test

Before officially starting a trial, it is required a detailed planning, including how the qualified patients to be randomized and the statistical methodology for post-randomization data analysis. The post-randomization data analysis usually involves the inference method and associate test statistic's establishment. A well-established statistical method can effectively control the Type I error rate. More specifically, consider a test statistic S constructed from the observed outcome data of patients, where the null hypothesis is rejected if and only if $|S| > z_{\alpha/2}$. Under H_0, a valid well-controlled test statistic maintains a Type I error rate close to nominal significance level α. When all strata factors are used in the post-randomization analysis, the test statistic is well-controlled.

However, the covariates used in CAR may be partially or fully omitted from the post randomization data analysis due to simplicity of a testing procedure and some practical reasons. Hence, it is necessary to study the inference properties of the standard tests when the working model with omitted covariates used. Consider the following working model:

$$\mathbb{E}\left[Y_i \mid \boldsymbol{T}_i, \boldsymbol{Z}_i\right] = h\left(\delta_0 + \delta_1 T_{i1} + \delta_2 T_{i2} + \cdots + \delta_K T_{iK}\right) \tag{4.3}$$

where the link function and conditional distribution of Y given linear predictor are same as in the true model.

For comparing each treatment with control based on the working model (4.3), we test the individual null hypotheses $H_{0,k} : \delta_k = 0$ for $k = 1, \ldots, K$ and obtain their corresponding p-values. Since multiple comparisons are conducted simultaneously, there is a risk of family-wise Type I error (FWER) inflation that could lead to false-positive findings. To control this risk and ensure the validity of our results, we implement appropriate multiplicity adjustment methods, which are detailed in Section 4.2.3. The multivariate test can be formed as follows:

$$H_{0k} : \delta_k = 0 \qquad k = 1, 2, \cdots, K \tag{4.4}$$

with the test statistics

$$S_k = \frac{\hat{\delta}_k}{\text{se}(\hat{\delta}_k)} \tag{4.5}$$

Remark 4.1. The commonly used test statistic for the hypothesis (4.4) is normalized to have a unit variance.

Suppose that N patients have been enrolled in the trial via CAR. For the ith patient, we observe $(y_i, T_{ik}, \boldsymbol{z}_i)$ as a realization of (Y, T, Z). The log-likelihood function of the observed responses can

be written as:

$$\log \prod_{i=1}^{N} p(y_i \mid \theta_i) = \sum \left\{ \frac{\theta_i y_i - b(\theta)}{\phi} + c\left(\phi, y_i\right) \right\} \tag{4.6}$$

The maximum likelihood estimator (MLE) takes the form:

$$\hat{\delta}_0 + \hat{\delta}_k = h^{-1}\left(\frac{\sum_{i=1}^{N} y_i T_{ik}}{n_k} \right)$$

From which we can derive:

$$\hat{\delta}_0 = h^{-1}\left(\frac{\sum_{i=1}^{N} y_i T_{i0}}{n_0} \right) \tag{4.7}$$

$$\hat{\delta}_k = h^{-1}\left(\frac{\sum_{i=1}^{N} y_i T_{ik}}{n_k} \right) - h^{-1}\left(\frac{\sum_{i=1}^{N} y_i T_{i0}}{n_0} \right) \tag{4.8}$$

where $h^{-1}(\cdot)$ is the link function and $n_k = \sum_{i=1}^{N} T_{ik}$ denotes the number of patients in group k. These estimators form the basis for our hypothesis testing framework, which we next introduce with proper multiplicity adjustment.

4.2.3 Control of Family-wise Type I error Rate under Multiplicity

Comparisons across different treatment arms and a control lead to potential inflation of the FWER as we mentioned previously. For example, in a dose-finding trial, multiple effective doses of a new drug are commonly compared with a single control, resulting in simultaneous testing of multiple null hypotheses, denoted as $H_{01}, H_{02}, \cdots, H_{0K}$. A critical issue here is controlling the FWER. Hence, multiplicity adjustment methods that ensure the probability of rejecting at least one true null hypothesis does not exceed the predetermined significance level α are essential [46, 16, 4]. Existing literature offers several methods, including Bonferroni, Holm, Simes, and Dunnett adjustments. More advanced approaches are reviewed by Alosh et al. (2014)[1].

In this article, we explore the use of the Simes correction method [50, 47] followed by statistical

testing, due to its simplicity, greater statistical power compared to Bonferroni and Holm methods, and flexibility in accommodating correlated hypotheses. While the Dunnett method is specifically designed for multiple comparisons against a single control group, it requires stronger assumptions such as normality and equal variance, limiting its applicability in more generalized or exploratory contexts. The Simes test has been widely used in trials with continuous outcomes, but its use for multi-arm trials with general outcomes is quite limited. Specifically, the validity of the Simes procedure relies on assumptions such as independence or the Positive Regression Dependency on Subset (PRDS) property, as detailed in Proposition 4.1. However, the detailed characterization of dependency structures under general outcomes remains less studied in the existing literature.

Specifically, the Simes method operates by comparing ordered p-values to adjusted significance thresholds. For K null hypotheses with p-values $p_1, p_2, \ldots, p_K$, let the ordered p-values be denoted as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(K)}$. The Simes procedure rejects the global null hypothesis if there exists any $i \in 1, \ldots, K$ such that:

$$p_{(i)} \leq \frac{i}{K}\alpha. \tag{4.9}$$

To justify the validity of the Simes procedure, we introduce the following proposition specifying the necessary conditions under which the FWER is strongly controlled.

Proposition 4.1 (Validity Conditions for Simes Method)**.** *The Simes procedure provides strong control of the FWER at level α, meaning:*

$$P\,(\textit{ reject at least one true } H_{0i}) \leq \alpha, \tag{4.10}$$

under the following conditions:

(i) The test statistics (or corresponding p-values) are independent; or

(ii) The test statistics exhibit certain positive dependence properties, specifically the Positive Regression Dependency on Subset (PRDS) property. Formally, PRDS means that for every

subset $I \subseteq 1, \ldots, K$ and every increasing function ϕ, the conditional expectation:

$$E\left[\phi\left(p_i, i \in I\right) \mid p_j \leq t\right] \tag{4.11}$$

is non-decreasing in t for each $j \notin I$.

Under these conditions, the Simes procedure controls the FWER at or below the nominal significance level α.

Proof sketch (see Sarkar and Chang (1997)[47]): Condition (i) of independence is classical, and the validity under independence was originally established by Simes (1986)[50]. Condition (ii), the PRDS property, which covers broader realistic scenarios, was proven to ensure the Simes method's validity in more general dependence structures by Sarkar and Chang (1997)[47].

With these conditions clearly stated, we proceed to present our primary theoretical results in the subsequent sections.

4.3 Main Results

Before presenting the main theorems, we first describe how to measure different types of covariate balance under multi-arm CAR procedures. These measurements are essential for establishing our theoretical results. In this article, we consider a family of general multi-arm CAR procedures proposed by Hu et al. (2023)[20]. This family encompasses many commonly used CAR procedures, including stratified permuted block randomization (STR-PB), Pocock and Simon's minimization (PS), stratified biased coin design (STR-BCD), and Hu and Hu's general CAR (HuHuCAR). Consider a clinical trial, each patient's covariate profile includes R covariates, where rth covariate has m_r levels. Hence, there are $M = \sum_{r=1}^{R} m_r$ margins, and $S = \prod_{r=1}^{R} m_r$ strata. We denote $(r; m_r^*)$ as a margin formed by patients whose rth covariate falls into m_r^*th ($1 \leq m_r^* \leq m_r$) level. Similarly, we denote $(m_1^*, m_2^*, \cdots, m_R^*)$ as a strata formed by patients has the covariate profile $Z_1 = m_1^*, \cdots, Z_R = m_R^*$, where $m_r^* \in \{1, \cdots, m_r\}$ for any covariate r ($1 \leq r \leq R$). Three

types of covariate imbalance, overall imbalance $\boldsymbol{D}_N$, marginal imbalance $\boldsymbol{D}_N(r; m_r^*)$, and stratum imbalance $\boldsymbol{D}_N(m_1^*, m_2^*, \cdots, m_R^*)$ are measured respectively.

Overall Imbalance: $\boldsymbol{D}_N = (D_0, D_2, \cdots, D_K)$, for each arm k ($0 \leq k \leq K$), we denote the overall imbalance for treatment k as $D_k = n_k - \bar{N}$, where $\bar{N} = \frac{1}{K}\sum_{k'=0}^{K} n_{k'}$ is average number of patients over treatments.

Marginal Imbalance: $\boldsymbol{D}_N(r; m_r^*) = (D_1(r; m_r^*), \cdots, D_K(r; m_r^*))$, for each arm k ($0 \leq k \leq K$), we denote the marginal imbalance for treatment k as $\boldsymbol{D}_k(r; m_r^*) = n_K(r; m_r^*) - \bar{N}(r; m_r^*)$, where $\bar{N}(r; m_r^*) = \frac{1}{K}\sum_{k'=0}^{K} n_{k'}(r; m_r^*)$ is the marginal average number of patients over the treatments.

Within-stratum Imbalance: $\boldsymbol{D}_N(m_1^*, \cdots, m_R^*) = (D_1(m_1^*, \cdots, m_R^*), \cdots, D_K(m_1^*, \cdots, m_R^*))$, for each arm k ($0 \leq k \leq K$), we denote the within-stratum imbalance for treatment k as $D_k(m_1^*, m_2^*, \cdots, m_R^*) = n_K(m_1^*, m_2^*, \cdots, m_R^*) - \bar{N}(m_1^*, m_2^*, \cdots, m_R^*)$, where $\bar{N}(m_1^*, m_2^*, \cdots, m_R^*) = \frac{1}{K}\sum_{k'=0}^{K} n_{k'}(m_1^*, m_2^*, \cdots, m_R^*)$ is the within-stratum average number of patients over the treatments.

Remark 4.2. The multi-arm CAR procedures we considered in this article could be divided into two categories in terms of the within-stratum imbalance. In stratified randomization (when $w_s > 0$), all imbalances including within-stratum, marginal, and overall are bounded in probability ($O_p(1)$) [20]. In Pocock and Simon's procedure (when $\mathrm{wo} = \mathrm{ws} = 0$), this method effectively controls both marginal and overall imbalances. However, at the stratum level, all within-stratum imbalances $D_k(m_1^*, m_2^*, \cdots, m_R^*)$ grow at a rate of $\sqrt{N}$ and follow an asymptotic normal distribution [22, 20].

4.3.1 Main Theorem

We now present the asymptotic distribution of the standard Wald test statistic under the working model (4.3).

Theorem 4.1. *Suppose that a covariate-adaptive randomization procedure satisfies condition that all within-stratum imbalances are bounded in probability; then under null hypothesis* $H_{0k} : \delta_k = 0$ $k \in \{1, \cdots, K\}$*, we have, as* $n \to \infty$*,*

$$(S_1, \cdots, S_K) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{\mathbb{E}[Var(Y \mid Z)]}{\phi h' h^{-1}(\mathbb{E}(Y))} \boldsymbol{V}\right)$$

where $\boldsymbol{V}$ *a K-dimensional square matrix with* $v_{ij} = 1$ *if* $i = j$ *and* $v_{ij} = 1/2$ *if* $i \neq j$*.*

Based on Theorem 4.1, if the within-stratum imbalances are bounded in probability, the size of the test for treatment effect depends on a comparison between $\mathbb{E}[\text{Var}(Y \mid Z)]$ and $\phi h' h((E)[Y])$, i.e., in asymptotic sense,

- The test is conservative if $\mathbb{E}[\text{Var}(Y \mid Z) < \phi h' h^{-1}(\mathbb{E}(Y))$.
- The test is valid if $\mathbb{E}[\text{Var}(Y \mid Z) = \phi h' h^{-1}(\mathbb{E}(Y))$
- The test is inflated $\mathbb{E}[\text{Var}(Y \mid Z) > \phi h' h^{-1}(\mathbb{E}(Y))$

Besides, the test statistics $S_1, S_2, \cdots, S_K$ are dependent as same control used.

We assume two key conditions which are also stated in [31]:

Condition 4.1: The overall imbalance is bounded in probability $D_N = O_p(1)$

Condition 4.2: Under the null hypothesis H_{0k}: $\delta_k = 0$, for any stratum $(m_1^*, ..., m_R^*)$, as $n \to \infty$,

$$\frac{\sum_{i=1}^{N} (T_k - T_0)\, h\left(\mu + \boldsymbol{\beta} \boldsymbol{z}_i\right)}{\sqrt{N}} \xrightarrow{D} N\left(0, \sigma_h^2\right)$$

These conditions are generally satisfied by CAR procedures that balance discrete covariates. For CAR designs where within-stratum imbalances are bounded in probability, the distribution in Condition 4.2 reduces to $O_p(1/\sqrt{N})$, implying $\sigma_h^2 = 0$. It has been proved that these conditions hold for PS by showing that within-stratum imbalances, when scaled by $1/\sqrt{n}$, follow an asymptotic joint normal distribution with positive variances. In the following, we present our theorem 4.2 [22, 20].

Theorem 4.2. *Suppose that a covariate-adaptive design satisfies Condition 4.1-4.2; then under the null hypothesis H_{0k}: $\delta_k = 0$, we have, as $n \to \infty$,*

$$(S_1, \cdots, S_K) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{\mathbb{E}[Var(Y \mid Z)] + (K+1)\sigma_h^2/2}{\phi h' h^{-1}(\mathbb{E}(Y))} \boldsymbol{V}^*\right)$$

where $\boldsymbol{V}^*$ a K-dimensional square matrix with $v_{ij} = 1$ if $i = j$ and $v_{ij} = 1/(2 + \sigma_h^2)$ if $i \neq j$.

Remark 4.3. The Simes test is widely recognized as one of the most common multiplicity adjustments for handling positively related or mutually independent test statistics. In this context, our Theorems 1-2 provide test statistics $S_1, \cdots, S_K$ that are postively correlated, when combined with the Simes test, may effectively control the overall familywise Type I error rate.

4.3.2 Test Size under Some Commonly Used Models

In Section 4.3.1, we established the asymptotic joint distribution of Wald test statistics $S_1, \cdots, S_K$ under a general framework of GLM (4.3) and CAR. Now we examine several important special cases of GLMs that are commonly encountered in practice. For each case, we derive the specific form of their asymptotic properties by applying the general theory to the corresponding canonical link functions. Specifically, we consider the logistic regression for binary outcomes, linear regression model for continuous outcome, Poisson regression for count data, and exponential model.

Logistic Regression

In clinical trials, logistic regression model has been widely used because of the common binary response. For example, it commonly uses as the surrogate endpoint in oncology trials, such as overall response rate (ORR). Under the logistic regression, the conditional distribution of Y given θ is a Bernoulli distribution with $\phi = 1$. When the CAR procedures with within-stratum imbalance are bounded in probability are implemented, we have

$$(S_1, S_2, \cdots, S_K) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{\mathbb{E}[\text{Var}(Y \mid Z)]}{\text{Var}(Y)} \boldsymbol{V}\right).$$

Because we have $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid Z)] + \mathrm{Var}(\mathbb{E}[Y \mid Z])$ and $\mathrm{Var}(\mathbb{E}[Y \mid Z]) > 0$, it follows that $\mathrm{Var}(Y) > \mathbb{E}[\mathrm{Var}(Y \mid Z)]$. Hence, the test is conservative.

Linear Regression

Linear regression is common in dealing with the continuous response, especially in may medical areas like immunology's drugs. Here, the dispersion parameter $\phi = \sigma^2$, which is usually unknown and could be estimated via the observations. Under the linear regression, $h(\cdot)$ is an identity function with $h'h^{-1}(\mathbb{E}(Y)) = 1$. Hence, under the CARs with within-strata imbalance bounded and σ^2 is known, we have

$$(S_1, S_2, \cdots, S_K) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{V})$$

However, the σ^2 is usually unknown and must be estimated based on the data. In practice, the model with omitted covariate tends to overestimate the variance of the treatment effect estimate [49, 34]. An inflated variance results in test statistics that are smaller in absolute value, leads to the conservative type I error.

Poisson Regression

As clinical trials frequently involve count outcomes, such as the number of adverse events or disease recurrences over a specified time period, Poisson regression serves as a fundamental tool for analyzing such data while accounting for their discrete, non-negative nature. The distribution of $Y|\theta$ is a Poisson distribution with $\phi = 1$. Hence, under the null hypotheses, we have $\phi h'h^{-1}(\mathbb{E}(Y)) = \mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(\mathrm{Var}(Y|X))$. It follows by

$$(S_1, S_2, \cdots, S_K) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{V})$$

A unique property of Poisson regression is that the variance equals the mean by definition. Therefore, even when the analysis model omits certain covariates, the marginal behavior of Y still

maintains the relationship: $\mathbb{E}[\mathrm{Var}(Y \mid Z)] = \phi h'(h^{-1}(\mathbb{E}(Y)))$. This equality ensures that the asymptotic variance of the treatment effect estimate is correctly specified, resulting in test statistics that follow a normal distribution with unit variance asymptotically. Consequently, the type I error rate remains well-controlled, neither inflated nor conservative.

Exponential Model

Exponential model also widely used in the post-randomization analysis in many clinical trials. For example, when clinical trials measure intervals between recurring events, such as hospital readmissions or disease relapses, exponential regression offers a straightforward approach for modeling these gaps while accommodating varying follow-up times and patient-specific risk factors. In the exponential model, the response Y given θ is exponentially distributed with $\phi = 1$. Under the null hypothesis and canonical link $h(\theta) = -1/\theta$, we have $\phi h' h^{-1}(\mathbb{E}(Y)) = (\mathbb{E}(Y))^2 = \{\mathbb{E}[\mathbb{E}(Y|X)]\}^2$. For $Y \mid Z$ following an exponential distribution with rate parameter $\lambda(X)$, the mean is $1/\lambda(X)$ and variance is $1/[\lambda(X)]^2$, thus $E[\mathrm{Var}(Y|X)] = E\{[E(Y|X)]^2\}$. It follows by $\phi h' h^{-1}(\mathbb{E}(Y)) \leq E[\mathrm{Var}(Y|X)]$.

$$(S_1, S_2, \cdots, S_K) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{\mathbb{E}[\mathrm{Var}(Y \mid Z)]}{\{\mathbb{E}[\mathbb{E}(Y|X)]\}^2}\boldsymbol{V}\right)$$

where $\mathbb{E}[\mathrm{Var}(Y|X)]/\{\mathbb{E}[\mathbb{E}(Y|X)]\}^2 \geq 1$.

Consequently, when using stratified randomization without fully accounting for stratification factors in the analysis, tests for treatment effects tend to be Inflated, leading to inflated Type I error rates.

4.3.3 Additional Results with non-canonical link

While canonical link functions in GLMs provide computational convenience, they may not always be the most appropriate choice for real data analysis. In many practical situations, non-canonical links can offer better model fit, more interpretable results sometimes. For example, in biomedical

Table 4.1: Summary of Type I Error Behavior Across GLMs

GLM Models	Link Function	Working Variance Term	Type I Error Behavior
Logistic Reg.	Logit	$\mathbb{E}(Y)(1-\mathbb{E}(Y))$	Conservative
Logistic Reg.	Probit	$\varphi\left(\Phi^{-1}(\mathbb{E}(Y))\right)$	Conservative
Logistic Reg.	Complementary Log-Log	$-\ln(1-\mathbb{E}(Y))(1-\mathbb{E}(Y))$	Conservative
Linear Reg.	Identity	σ^2 (overestimated)	Conservative
Poisson Reg.	Log	$\mathbb{E}(Y)$	Valid
Exponential	Inverse	$(\mathbb{E}(Y))^2$	Inflated
Negative Binomial Reg.	Log	$\phi E(Y)$	Inflated
Gamma Reg.	Log	$\phi E(Y)$	Typically Inflated
Ordinal Logistic Reg.	Cumulative Logit	$E\left(\pi_j\right)\left[1-E\left(\pi_j\right)\right]$ for threshold j	Conservative

studies, the probit link may be more suitable than the canonical logit link when the underlying process follows a normal tolerance distribution. Similarly, the complementary log-log link might be more appropriate when analyzing time-to-event data in discrete time. Additionally, some non-canonical links, such as the identity link for count data, can provide more direct interpretation of the effects. Therefore, extending our theoretical results to non-canonical links could broaden the practical applicability of our testing procedure while maintaining its statistical robustness. Hence, in this section, Theorems 4.1-4.2 which derived under canonical link functions are extended to the GLMs with non-canonical link functions.

In GLMs with general link functions, the relationship between the natural parameter θ and the linear predictor $\eta = \delta_0 + \delta_1 T_1 + \cdots + \delta_K T_K$ is characterized by $\theta = \gamma(\eta)$, where γ is some function. Unlike canonical links where γ is the identity function (i.e., $\theta = \eta$), non-canonical links involve a more general transformation that connects the linear predictor to the natural parameter of the exponential family.

Theorem 4.3. *Suppose that a covariate-adaptive randomization procedure satisfies condition that all within-stratum imbalances are bounded in probability; then under null hypothesis $H_{0k} : \delta_k = 0$, we have, as $n \to \infty$,*

$$(S_1, \cdots, S_K) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{\mathbb{E}[Var(Y \mid Z)]}{\phi h' h^{-1}(\mathbb{E}(Y))/\gamma' h^{-1}(\mathbb{E}(Y))} \boldsymbol{V}\right)$$

where $\boldsymbol{V}$ a K-dimensional square matrix with $v_{ij} = 1$ if $i = j$ and $v_{ij} = 1/2$ if $i \neq j$.

Theorem 4.4. *Suppose that a covariate-adaptive design satisfies Condition 4.1-4.2; then under the null hypothesis H_{0k}: $\delta_k = 0$, we have, as $n \to \infty$,*

$$(S_1, \cdots, S_K) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{\mathbb{E}[Var(Y \mid Z)] + (K+1)\sigma_h^2/2}{\phi h' h^{-1}(\mathbb{E}(Y))/\gamma' h^{-1}(\mathbb{E}(Y))} \boldsymbol{V}^*\right)$$

where $\boldsymbol{V}^*$ a K-dimensional square matrix with $v_{ij} = 1$ if $i = j$ and $v_{ij} = 1/(2 + \sigma_h^2)$ if $i \neq j$.

Remark 4.4. Let $\eta = \delta_0 + \delta_1 T_1 + \ldots + \delta_k T_k$ where exactly one of $T_1, \ldots, T_k$ equals 1 and others are 0. For binary endpoint, besides the canonical logit link $\log(\mu/(1-\mu)) = \delta_0 + \delta_k T_k$, the probit link $\Phi^{-1}(\mu) = \delta_0 + \delta_k T_k$ and complementary log-log link $\log(-\log(1-\mu)) = \delta_0 + \delta_k T_k$ are common alternatives. For Poisson regression modeling count data, although the canonical link is $\log(\mu) = \delta_0 + \delta_i$, the identity link $\mu = \delta_0 + \delta_k T_k$ and square root link $\sqrt{\mu} = \delta_0 + \delta_k T_k$ are sometimes preferred for direct interpretation. In gamma regression for positive continuous responses, while the inverse link $1/\mu = \delta_0 + \delta_k T_k$ is canonical, the $\log \operatorname{link} \log(\mu) = \delta_0 + \delta_k T_k$ is widely used as it ensures positive fitted values. Here μ represents the conditional mean of the response given $\mathrm{T}_k = 1$, and these non-canonical options expand the flexibility of GLMs while maintaining the theoretical properties we established under mild regularity conditions.

4.4 Constructing Valid Test Statistics

Theorem 4.1, along with the discussion in Section 4.3.1 for common GLMs, shows that under CAR, the standard Wald test statistics fail to converge to a standard multivariate normal distribution in the absence of a fully adjusted model. This deviation arises from two primary sources: (1) The implementation of CAR and the potential working model misspecification can lead to test statistics that fail to converge to the correct asymptotic distribution under the null hypothesis H_0. (2) The presence of multiple treatment arms and their associated subject assignment mechanisms induces dependence among test statistics, which subsequently affects multiple testing corrections. While existing methods can adjust for dependent test statistics or p-values, they typically yield conservative family-wise Type I error. To address these limitations, we propose the adjusted test

statistics that converges to $N(\mathbf{0}, \boldsymbol{I})$ under null hypotheses for the given randomization scheme, regardless of potential covariate omission. Such test would simultaneously achieve robustness to model misspecification and valid family-wise Type I error control.

The scenarios where the within-stratum imbalances are bounded in probability, such as STR-PB, STR-BCD, and HuHuCAR are consider first. It should be noted that in HuHuCAR [25], the within-stratum imbalances remain bounded when non-zero weights are assigned to stratum-level balance. Based on the proof of Theorem 4.1 in the Supplementary material, under the working model and null hypotheses $H_k : \delta_k = 0$ $(k = 1, \ldots, K)$, we can readily establish that:

$$\sqrt{N}\left(\frac{h'[h^{-1}\{\mathbb{E}(Y)\}]}{\sqrt{2(K+1)\mathbb{E}[\text{Var}(Y \mid Z)]}}\right)(\boldsymbol{V})^{-1/2}\hat{\boldsymbol{\delta}} \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{4.12}$$

Hence the newly constructed statistics are

$$\boldsymbol{S}_{adj} = \sqrt{N}\frac{h'\{h^{-1}(\bar{Y})\}}{\sqrt{2(K+1)\sigma^2_{\text{pooled}}}}(\boldsymbol{V})^{-1/2}\hat{\boldsymbol{\delta}} \tag{4.13}$$

where the pooled variance estimator is

$$\sigma^2_{\text{pooled}} = \sum_{s=1}^{S}\frac{n_s}{N}\frac{\sum_{i=1}^{N}(y_i - \hat{\mu}_s)^2\mathbf{1}\{z_i = s\}}{n_s - 1}. \tag{4.14}$$

Here $\bar{Y}$ is the overall sample mean, and S is the total number of strata. The indicator $I\{z_i = s\}$ takes value 1 if observation i is in stratum s, and 0 otherwise. The pooled variance estimator σ^2_{pooled} weights each stratum-specific variance by the relative stratum size n_s/N.

Similarly, for general CAR procedures under Conditions 1-2, the within-stratum-level imbalance may be not bounded. Under the working model and the null hypothesis $H_K : \delta_k = 0(k = 1, \ldots, K)$, we have:

$$\sqrt{N}\left(\frac{h'[h^{-1}\{\mathbb{E}(Y)\}]}{\sqrt{2(K+1)\left(\mathbb{E}[\text{Var}(Y \mid Z)] + (K+1)\sigma_h^2/2\right)}}\right)(\boldsymbol{V})^{-1/2}\hat{\boldsymbol{\delta}} \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{4.15}$$

It follows that the adjusted test statistics are

$$\boldsymbol{S}_{adj} = \sqrt{N}\frac{h'\{h^{-1}(\bar{Y})\}}{\sqrt{2(K+1)\sigma^2_{\text{pooled,adj}}}}(\boldsymbol{V}^*)^{-1/2}\hat{\boldsymbol{\delta}} \tag{4.16}$$

where the adjusted pooled variance estimator is

$$\sigma^2_{\text{pooled,adj}} = \sum_{s=1}^{S}\frac{n_s}{N}\frac{\sum_{i=1}^{N}(y_i-\hat{\mu}_s)^2\mathbf{1}\{z_i=s\}}{n_s-1} + \frac{(K+1)\hat{\sigma}^2_h}{2} \tag{4.17}$$

Compared with Equation (4.13), the only additional quantity that needs to be considered in Equation (4.16) is σ^2_h. This arises because the CAR procedures used here (e.g., Pocock & Simon's minimization) do not necessarily bound the within-stratum imbalance, thus introducing additional variance. Specifically, for the general K-treatment scenario, σ^2_h can be expressed as a weighted sum of the within-stratum imbalances across all treatment groups:

$$\sigma^2_h = \sum_{s=1}^{S}\sum_{k=1}^{K} D_k\left(s\right) h\left(\mu+\beta s\right), \tag{4.18}$$

In practice, Monte Carlo simulation is employed to numerically estimate σ^2_h. The simulation process involves repeatedly generating treatment assignments under the CAR method, keeping the observed covariates fixed. Specifically, the simulation follows these steps:

Step 1*: Repeatedly simulate the CAR algorithm (e.g., minimization) under the null hypothesis, generating B (e.g.* $B = 500$*) independent replicates of treatment assignments.*

Step 2*: For each replicate, calculate the within-stratum imbalance measures* $D_k(s)$ *for each treatment and stratum.*

Step 3*: Estimation of Variance: Compute variance contributions for each replicate by replacing* $h(\mu+\beta s)$ *with its sample-based estimate* $h(\hat{\mu}_j)$*, yielding replicate-specific variance estimates:*

$$\sigma^2_{h,b} = \sum_{s=1}^{S}\sum_{k=1}^{K} D_k^{(b)}\left(s\right) h\left(\hat{\mu}_j\right), \quad b = 1,\ldots,B.$$

***Step 4**: Finally, average across the B replicates to obtain a stable numerical estimate:*

$$\hat{\sigma}_h^2 = \frac{1}{B}\sum_{b=1}^{B}\sigma_{h,b^*}^2.$$

Remark 4.5. The proposed adjusted test ensures valid Type I error control by accounting for the variance-covariance structure inherent in the dependency among treatment groups. While this adjustment ensures convergence to a standard multivariate normal distribution and addresses conservative Type I error in the standard Wald test, it does not always improve power. This is due to potential variance inflation, which can reduce sensitivity to detect the effects, especially for weak treatment effects or low correlation between test statistics. However, the adjusted test is advantageous in small samples, imbalanced group sizes, or highly correlated test statistics, where it provides robust error control and reliability in complex study designs. Thus, its application should balance Type I error control against potential power trade-offs, depending on the trial context.

4.5 Numerical Studies

In this section, we present simulation studies that evaluate our proposed adjusted test statistics in (4.13) and (4.16). Our primary goal is to demonstrate that these tests maintain valid FWER while potentially improving power under various CAR designs. We evaluate both the family-wise and individual Type I error rates for the original and adjusted tests when testing each null hypothesis H_k ($k = 1, \ldots, K$). The simulations cover three commonly used GLMs: logistic regression, Poisson regression, and exponential model with canonical link functions. For all simulation scenarios, we consider 3-arm trials with sample sizes of $n = 300$ and $n = 600$ for Type I error assessment, and $n = 300$ for power analysis. Each scenario is replicated 5000 times.

In Tables 4.2-4.4, we present the family-wise and individual Type I errors for common-used GLMs under both the standard Wald test statistics and our proposed adjusted test statistics. The results demonstrate that the adjusted test statistics provide more efficient control of Type I errors.

For the treatment effect tests based on logistic regression, we set the parameters $\boldsymbol{\delta} = (-1, 0, 0)$

Table 4.2: Simulated Type I error for Logistic Regression

		Wald Test			**Adjust Test**		
Sample Size	Randomization	Family-wise	H_1	H_2	Family wise	H_1	H_2
N = 300	STR-PB	0.0190	0.0108	0.0114	0.0480	0.0270	0.0226
	PS	0.0140	0.0090	0.0101	0.0433	0.0259	0.0247
	HuHuCAR	0.0230	0.0116	0.0160	0.0534	0.0268	0.0298
N = 600	STR-PB	0.0184	0.0102	0.0118	0.0480	0.0262	0.0234
	PS	0.0192	0.0109	0.0112	0.0476	0.0254	0.0272
	HuHuCAR	0.0258	0.0154	0.0156	0.0540	0.0298	0.0330

Table 4.3: Simulated Type I error for Poisson Regression

Sample Size	Randomization	**Wald Test**			**Adjust Test**		
		Family-wise	H_1	H_2	Family wise	H_1	H_2
N = 300	STR-PB	0.0525	0.0300	0.0310	0.0555	0.0275	0.0315
	PS	0.0570	0.0313	0.0304	0.0521	0.0282	0.0297
	HuHuCAR	0.0442	0.0294	0.0236	0.0530	0.0296	0.0264
N = 600	STR-PB	0.0510	0.0295	0.0310	0.0540	0.0265	0.0320
	PS	0.0574	0.0304	0.0322	0.0529	0.0272	0.0299
	HuHuCAR	0.0440	0.0238	0.0270	0.0474	0.0226	0.0262

Table 4.4: Simulated Type I error for Exponential Model

Sample Size	Randomization	**Wald Test**			**Adjust Test**		
		Family-wise	H_1	H_2	Family wise	H_1	H_2
N = 300	STR-PB	0.1554	0.0956	0.0954	0.0538	0.0284	0.0298
	PS	0.1655	0.1087	0.1149	0.0522	0.0294	0.0303
	HuHuCAR	0.1794	0.1096	0.1170	0.0568	0.0338	0.0344
N = 600	STR-PB	0.1666	0.1068	0.0998	0.0550	0.0314	0.0268
	PS	0.1825	0.1077	0.1092	0.0544	0.0301	0.0289
	HuHuCAR	0.1730	0.1044	0.1078	0.0572	0.0296	0.0306

and $\beta = (1, 2)$ with fixed sample sizes of $N = 300$ and $N = 600$. The covariates used in simulations include two binary variables Z_1 and Z_2, each evenly distributed with probability 0.5. Under this logistic regression setup, while the standard Wald test statistics yield conservative but acceptable FWER, our adjusted test statistics demonstrate improved performance. Specifically, they maintain valid FWER and reduce conservativeness in individual Type I error rates for each test. Therefore, the adjusted tests show superior performance in controlling Type I errors.

For the treatment effect tests based on Poisson regression, we set $\boldsymbol{\delta} = (0.8, 0, 0)$ and $\beta = (0.25, 0.25)$ with sample size $N = 300$ and $N = 600$. The covariates Z_1, Z_2 are assumed to

follow the same distribution as in the logistic regression scenario. Under this setting, Table 4.3 shows that both the standard Wald test and our adjusted test maintain good control of Type I error rates. The FWER stays close to the nominal level of 0.05 , and the individual Type I errors are also well controlled. This indicates that the asymptotic approximation works well for Poisson regression even without adjustment, likely due to the natural mean-variance relationship in the Poisson distribution. Nevertheless, our adjusted test statistics provide comparable performance while maintaining theoretical consistency with our general framework.

For the exponential model scenario, we specify parameters $\boldsymbol{\delta} = (0.2, 0, 0)$ and $\boldsymbol{\beta} = (0.5, 1)$. The covariates Z_1 and Z_2 are also generated following the same distribution described earlier in the logistic regression setting. Empirical Type I errors from this setup are summarized in Table 4.4. It can be seen that the standard Wald test leads to substantial inflation in both family-wise and individual Type I error rates, which are not acceptable in practical trial designs. Conversely, the adjusted test controls the Type I errors well.

Having established the Type I error control properties of both standard Wald test and adjusted test, we now evaluate their power performance through simulation studies. We consider a 3-arm trial setting with one control group and two different dosages of a new drug. To create a realistic scenario for power analysis, we designate one treatment arm as having an optimal effect and the other as having a suboptimal effect compared to the control. This design allows us to assess how well our testing procedures can detect both strong and moderate treatment effects, which is particularly relevant in dose-finding studies where identifying the most effective dose level is crucial.

For power analysis, we consider trials with 300 subjects ($N = 300$) using the HuHuCAR, which balances covariates among arms at stratum, marginal, and overall levels. While we present results for this specific CAR method, similar power performance was observed with other CAR procedures, including STR-PB, stratified biased coin design, and Pocock and Simon minimization. In the following tables, δ_1 and δ_2 denote the treatment effects of the two treatment arms relative to control.

Tables 4.5 and 4.6 present power analyses under the logistic regression model under two distinct

Table 4.5: Simulated power for logistic regression with N = 300

δ_1	δ_2	Test (Trt1)	Test (Trt2)	Adjusted Test (Trt1)	Adjusted Test (Trt2)
0.2	0.15	0.0296	0.0238	0.0484	0.0384
0.4	0.15	0.0906	0.0288	0.1652	0.0332
0.6	0.15	0.2270	0.0354	0.3640	0.0386
0.8	0.15	0.4276	0.0392	0.6204	0.0476
1.0	0.15	0.6404	0.0404	0.8148	0.0646
1.2	0.15	0.8202	0.0422	0.9414	0.0918

Table 4.6: Simulated power for Logistic regression with N = 300

δ_1	δ_2	Test (Trt1)	Test (Trt2)	Adjusted Test (Trt1)	Adjusted Test (Trt2)
0.2	0.15	0.0296	0.0238	0.0484	0.0384
0.4	0.35	0.0998	0.0848	0.1230	0.0888
0.6	0.55	0.2596	0.2252	0.2552	0.1926
0.8	0.75	0.4932	0.4502	0.4270	0.3574
1.0	0.95	0.7092	0.6736	0.6082	0.5328
1.2	1.15	0.8790	0.8566	0.7646	0.7118

Table 4.7: Simulated power for Poisson regression with N = 300

δ_1	δ_2	Test (Trt1)	Test (Trt2)	Adjusted Test (Trt1)	Adjusted Test (Trt2)
0.1	0.03	0.0658	0.0338	0.0654	0.0306
0.2	0.06	0.1876	0.0518	0.1838	0.0362
0.3	0.09	0.3862	0.0802	0.3922	0.0416
0.4	0.12	0.6138	0.1176	0.6246	0.0520
0.5	0.15	0.8072	0.1644	0.8172	0.0608
0.6	0.18	0.9222	0.2164	0.9296	0.0664

scenarios. In Table 4.5, we fix δ_2 at 0.15 while varying δ_1 from 0.2 to 1.2. As δ_1 increases, both tests show monotonically increasing power for Treatment 1, with our adjusted test consistently achieving higher power (e.g., 0.8148 versus 0.6404 when $\delta_1 = 1.0$). Meanwhile, the power for Treatment 2 remains stable, reflecting its fixed effect size. Table 4.6 examines a different scenario where both treatment effects increase in parallel, with δ_2 maintained slightly below δ_1. In this case, we observe a distinct pattern: while our adjusted test exhibits superior power for smaller treatment effects, the standard Wald test demonstrates better power when both treatment effects become large.

In Table 4.7, we evaluate the power performance for Poisson regression where δ_2 is set to be

Table 4.8: Simulated power for exponential model with $N = 300$

δ_1	δ_2	Adjusted Test (Trt1)	Adjusted Test (Trt2)
0.05	0.03	0.0878	0.0472
0.10	0.06	0.2236	0.0692
0.15	0.09	0.3890	0.1000
0.20	0.12	0.5576	0.1416
0.25	0.15	0.7022	0.1854
0.30	0.18	0.8068	0.2270
0.35	0.21	0.8802	0.2752
0.40	0.24	0.9258	0.3210

proportional to δ_1 ($\delta_2 = 0.3\delta_1$). For Treatment 1 , both the standard Wald test and adjusted test provide comparable power as δ_1 increases from 0.1 to 0.6 , with the adjusted test showing slightly higher power (e.g., 0.6246 vs 0.6138 when $\delta_1 = 0.4$). For Treatment 2, where δ_2 takes smaller values, the standard Wald test demonstrates marginally higher power than the adjusted test, though both maintain relatively low power due to the smaller values of δ_2.

For the exponential model, since we have demonstrated that the standard Wald test with Simes correction fails to control the FWER, we focus only on examining the power of our adjusted test. Table 4.8 presents the power analysis under our adjusted test where δ_2 is proportional to δ_1 ($\delta_2 = 0.6\delta_1$). As δ_1 increases from 0.05 to 0.40 , the power for Treatment 1 increases substantially from 0.0878 to 0.9258 . The power for Treatment 2 , though consistently lower due to smaller values of δ_2, also shows a steady increase from 0.0472 to 0.3210 . These results show that our adjusted test maintains good power for detecting treatment differences while preserving valid Type I error control.

4.6 Analysis of Metastatic Breast Cancer Trial

Hamilton et al. (2022) [15] reported a clinical trial (nextMONARCH) evaluating the efficacy of Abemaciclib, a cyclin-dependent kinase 4 and 6 (CDK4/6) inhibitor for treating metastatic breast cancer. The trial, registered under ClinicalTrials.gov Identifier NCT02747004, was an open-label, randomized, controlled, phase 2 study conducted at 60 sites across 14 regions/countries from September

2016 to June 2017. A total of 234 eligible patients were enrolled and randomly assigned in a 1:1:1 ratio to one of three groups: abemaciclib (150 mg) plus tamoxifen (78 patients) (group A+T), abemaciclib (150 mg) alone (79 patients) (group A-150), or the reference treatment abemaciclib (200 mg) alone (77 patients) (group A-200). Randomization was stratified by the presence of liver metastases (Yes or No) and prior tamoxifen therapy (Yes or No) in the advanced/metastatic setting. The overall baseline involved patients with heavily pretreated hormone receptor-positive, HER2-negative metastatic breast cancer. The primary endpoint was investigator-assessed progression-free survival (PFS), while Objective Response Rate (ORR) served as a common binary endpoint to assess tumor response. Among all endpoints, ORR is defined as the proportion of patients with a complete or partial response. It is a key and common measure of treatment efficacy in cancer clinical trials, providing insight into the immediate tumor response to therapeutic intervention. In this trial, the ORR was 34.6% for the abemaciclib plus tamoxifen group, 24.1% for the abemaciclib 150 mg alone group, and 33.8% for the abemaciclib 200 mg alone group.

In this section, we conduct a post-randomization analysis using the newly proposed adjust test. Our redesigned study utilizes the true ORR of three treatments and the specified strata factors and their proportions to generate an increased number (N = 1000) of the individual patient data via logistic regression model. Since the original paper did not provide the odds ratios for the strata factors, we assumed reasonable odds ratios based on published real-world and retrospective studies that provide information on ORR in HR+/HER2- metastatic breast cancer (mBC) patients, specifically examining the impact of liver metastases and prior tamoxifen use.

For liver metastases, consistent evidence across major trials demonstrated its substantial impact on treatment response. In the MONARCH 3 trial [14], patients with liver metastases showed a baseline ORR of 21.1%. Similar findings were reported in MONALEESA-2 [17] with 22% ORR, and PALOMA-2 [12] with 25.3% ORR in liver metastasis subgroups. Based on these data, an odds ratio of 0.30 was selected for liver metastases, representing approximately a 70% reduction in response odds. For prior tamoxifen exposure, we broaden the scope to prior endocrine therapy exposure, as data specifically related to tamoxifen are limited. The impact of prior endocrine

therapy exposure was demonstrated in several CDK4/6 inhibitor trials. In MONARCH 2 [51], which evaluated abemaciclib plus fulvestrant in patients with HR+/HER2- advanced breast cancer, all patients had progressed on prior endocrine therapy, and prior tamoxifen use was a stratification factor due to its potential impact on treatment outcomes. Similarly, in PALOMA-3 [54], which included patients who had progressed on prior endocrine therapy, the response rates were analyzed according to the line and type of previous endocrine treatment, supporting its inclusion as a stratification factor. Based on these trials demonstrating the clinical relevance of prior endocrine therapy exposure, an odds ratio of 0.75 was selected as a conservative estimate for the effect of prior tamoxifen use in our model. Hence, the individual patients data is generated based on the following logistic regression:

$$\begin{aligned}\mathbb{P}(\text{Objective Response} = 1 \mid \boldsymbol{Z}) = \text{logit}^{-1}\Big(& -0.684 + 0.032 T_{\text{A+T}} - 0.462 T_{\text{A-150}} \\ & - 1.20\, \mathbb{I}(\text{liver metastases} = \text{Yes}) \\ & - 0.29\, \mathbb{I}(\text{prior tamoxifen usage} = \text{Yes})\Big)\end{aligned}$$

where $T_{\mathrm{A+T}}$ is a binary indicator vector of length N representing assignment to treatment group $\mathrm{A} + \mathrm{T}$. Specifically, each element of $T_{\mathrm{A+T}}$ is coded as 1 if the patient is assigned to group $\mathrm{A} + \mathrm{T}$ and 0 otherwise. Similarly, $T_{\mathrm{A-150}}$ is a binary indicator vector indicating assignment to treatment group A-150, where each element is 1 for patients in group A-150 and 0 otherwise. Besides, We denote the covariate vector Z here as including treatment indicators, liver metastasis status, and prior tamoxifen usage.

Based on the generated patient data, we are now using a newly proposed adjusted inference method, which employs two-sided tests, to redesign and analyze the trial. Specifically, we are comparing the statistical power of the new inference method with that of the standard Wald test to demonstrate the improved performance of the new approach in detecting treatment effects within the context of the trial. The results in Table 4.9 demonstrate the superior power performance of the adjusted test compared to the standard Wald test across all randomization schemes within

Table 4.9: Re-designed Metastatic Breast Cancer Trial (N = 1000)

	Wald Test		**Adjust Test**	
Randomization	A+T	A-150	A+T	A-150
STR-PB	0.0326	0.5438	0.1690	0.7652
PS	0.0332	0.5492	0.1712	0.7657
HuHu	0.0346	0.5590	0.1720	0.7660

the working model. In the A+T scenario, the Wald test exhibited relatively low power (ranging from 0.0326 to 0.0346), whereas the adjusted test showed a substantial improvement in power (0.1690 to 0.1720). This enhancement was even more pronounced in the A-150 scenario, where the adjusted test achieved significantly higher power (approximately 0.76) compared to the Wald test (approximately 0.54). These improvements were consistent across all three randomization methods (STR-PB, PS, and HuHuCAR), which means that the enhanced performance of the adjusted inference method is robust to the choice of randomization scheme.

4.7 Discussion

In this article, we proposed a family of robust tests for treatment effects in multi-arm trials within GLMs under CAR. Through the theoretical proof, extensive simulation studies and a real-world trial application, we demonstrate that the proposed inference method could maintain the control of the Type I error, and also, in some cases, provide improved statistical power.

From a theoretical perspective, we proved that the standard Wald test may not adequately control Type I error because the test statistics may not converge to the standard multivariate normal distribution under the working model and CAR. Our analysis revealed distinct performance patterns across commonly used models: conservative Type I error in logistic regression, valid control in Poisson regression, and inflation in exponential models. Notably, this represents the first theoretical investigation of asymptotic convergence for multi-arm trials under both CAR and GLMs, as previous research primarily focused on linear models. Our findings also demonstrated that the test statistics are correlated, leading us to propose adjusted test statistics that ensure convergence to the standard normal distribution.

Our simulation studies corroborated these theoretical findings. The proposed adjusted test maintained FWER control across all scenarios, a crucial requirement for post-randomization analysis. This control was particularly important for GLMs that showed inflated FWER under the standard Wald test. Power analysis revealed that our adjusted test provided enhanced performance in certain scenarios under the logistic regression model. For the exponential model, we focused solely on the adjusted test's power, as the Wald test's failure to maintain Type I error control made comparisons meaningless. The practical application to a metastatic breast cancer trial validated our theoretical findings and demonstrated potential power advantages, suggesting tangible benefits for real-world clinical trials.

While our method shows promising results, several limitations should be noted. In scenarios with moderate or strong treatment effects, the adjustment may reduce power compared to the traditional Wald test due to variance inflation. However, the proposed adjusted test excels when treatment effects are weak, where its ability to maintain nominal Type I error ensures more reliable performance. Despite these nuances, the adjusted test remains a valuable tool for rigorous statistical inference, particularly in settings with small sample sizes and correlated test statistics.

Several directions for future research emerge from this work. An immediate extension would be to expand the current framework beyond CAR designs to incorporate response-adaptive randomization. The existing framework focuses primarily on achieving covariate balance during randomization; however, with the growing emphasis on precision medicine, solely balancing covariates may no longer suffice. Future randomization schemes should also account for individual patient responses to improve trial efficiency and enhance outcome precision. The proposed framework is well-suited for such advancements and can be readily extended to covariate-adjusted response-adaptive randomization, enabling more dynamic and personalized trial designs. Additionally, while the adjusted test was developed within the GLM framework, similar principles could be applied to time-to-event endpoints using a Cox proportional hazards working model, expanding its applicability to survival analysis and other time-to-event scenarios. These extensions represent promising areas for future research and practical development.

4.8 Appendix

This section provides the theoretical proof for the main results presented in Chapter 4. It is organized into three sections. Section 4.8.1 presents ancillary results, including technical lemmas and intermediate theoretical results that are essential for establishing the main theorems. Section 4.8.2 contains detailed mathematical proofs of all main theorems stated in Section 4.3.1 of the paper. Section 4.8.3 examines the working variance term in asymptotic theory under different GLMs. Section 4.8.3 examines the working variance term in asymptotic theory under different GLMs.

4.8.1 Ancillary Results

Lemma 4.1. *Under the assumption the number of patients in each treatment group is approximately equal,* n_k/N *converges in probability to* $1/(K+1)$ *as* $N \to \infty$*, for any* $k = 0, 1, \cdots, K$.

Proof. It has been proved in Lemma 1 in Ma et al. (2022) [38]. ■

Lemma 4.2. *Assume that the overall imbalance is bounded in probability, and if the null hypothesis* $H_{0,k} : \alpha_k = 0$ *are true for all* $k = 1, \cdots, K$*, then as* $n \to \infty$*, we have*

$$\sqrt{\frac{2(K+1)}{N}} \left\{ \left[\sum_{i=1}^{n} (T_{i1} - T_{i0})(y_i - \nu_i), \cdots, \sum_{i=1}^{n} (T_{iK} - T_{i0})(y_i - \nu_i) \right], \left[\sum_{i=1}^{n} (T_{i1} - T_{i0})\nu_i, \cdots, \sum_{i=1}^{n} (T_{iK} - T_{i0})\nu_i \right] \right\} \xrightarrow{D} (\xi_1, \xi_2),$$

where ξ_1 *and* ξ_2 *are independent, and* $\xi_1 \sim \mathcal{N}\left(\mathbf{0}, \mathbb{E}[\mathrm{Var}(Y \mid \mathcal{Z})]\left\{\mathrm{diag}\{2\mathbf{1}_K\} + 2\mathbf{1}_K\mathbf{1}_K^{\mathrm{T}}\right\}\right)$ *and* $\xi_2 \sim N(\mathbf{0}, \mathrm{diag}\{2(K+1)\sigma_h^2\mathbf{1}_K\})$.

proof of Lemma 4.2: Let $\mathcal{Z} = \{Z_1, Z_2, \cdots, Z_N\}$ and $\mathcal{T} = \{T_1, \cdots, T_N\}$. Since $(Y_{i0}, Y_{i1}, \cdots, Y_{iN})$ are conditionally independent of $\mathcal{T}$ given $\mathcal{Z}$, we have, under the null hypothesis,

$$\mathbb{E}\left(\frac{K+1}{N} \sum_{i=1}^{N} \{T_{ik}(Y_{ik} - v_i) - T_{i0}(Y_{i0} - v_i)\} \mid \mathcal{Z} \right) = 0,$$

where $v_i = E\left(Y_i \mid Z_i\right)$, and $\mathbb{E}(Y_{ik}) = \mathbb{E}(Y_{i0})$ asymptotically under null hypothesis. Besides, the variance-covariance has been shown in the following,

$$
\begin{aligned}
&\text{Var}\left(\frac{K+1}{N}\sum_{i=1}^{n}(T_{ik}-T_{i0})(y_i-\nu_i) \mid \mathcal{Z}\right) \\
&= \left(\frac{K+1}{N}\right)^2 \sum_{i=1}^{N}\mathbb{E}\left((T_{ik}-T_{i0})^2(y_i-\nu_i)^2 \mid \mathcal{Z}\right) \\
&= \left(\frac{K+1}{N}\right)^2 \sum_{i=1}^{N}\mathbb{E}\left((T_{ik}+T_{i0}) \mid \mathcal{Z}\right)\text{Var}(y_i \mid \mathcal{Z}) \\
&= \left(\frac{K+1}{N}\right)^2 \frac{2}{K+1}\sum_{i=1}^{N}\text{Var}(y_i \mid \mathcal{Z}) \\
&= \frac{2(K+1)}{N^2}\sum_{i=1}^{N}\text{Var}(Y \mid \mathcal{Z}),
\end{aligned}
$$

and

$$
\begin{aligned}
&\text{Cov}\left\{\frac{K+1}{N}\sum_{i=1}^{N}(T_{i1}-T_{i0})(y_i-\nu_i), \frac{K+1}{N}\sum_{i=1}^{N}(T_{i2}-T_{i0})(y_i-\nu_i) \mid \mathcal{Z}\right\} \\
&= \left(\frac{K+1}{N}\right)^2 \sum_{i=1}^{N}\sum_{j=1}^{N}\text{Cov}\left\{(T_{i1}-T_{i0})(y_i-\nu_i), (T_{j2}-T_{j0})(y_j-\nu_j) \mid \mathcal{Z}\right\} \\
&= \left(\frac{K+1}{N}\right)^2 \left\{\sum_{i=1}^{N}\mathbb{E}\left((T_{i1}-T_{i0})(T_{i2}-T_{i0})(y_i-\nu_i)^2 \mid \mathcal{Z}\right)\right. \\
&\qquad\qquad \left. + \sum_{i\neq j}^{N}\mathbb{E}\left((T_{i1}-T_{i0})(T_{j2}-T_{j0})(y_i-\nu_i)(y_j-\nu_j) \mid \mathcal{Z}\right)\right\} \\
&= \left(\frac{K+1}{N}\right)^2 \sum_{i=1}^{N}\frac{1}{K+1}\text{Var}(y_i \mid \mathcal{Z}) \xrightarrow{P} \frac{K+1}{N^2}\sum_{i=1}^{N}\text{Var}(Y \mid \mathcal{Z}).
\end{aligned}
$$

By law of large number theory,

$$
\frac{1}{N}\sum_{i=1}^{N}\text{Var}(Y \mid \mathcal{Z}) \xrightarrow{P} \mathbb{E}[\text{Var}(Y \mid \mathcal{Z})]
$$

By central limit theorem,

$$\frac{K+1}{\sqrt{N}}\left[\sum_{i=1}^{n}\left(T_{i1}-T_{i0}\right)\left(y_i-\nu_i\right),\cdots,\sum_{i=1}^{n}\left(T_{iK}-T_{i0}\right)\left(y_i-\nu_i\right)\right]\xrightarrow{D}\mathcal{N}\left(\mathbf{0},\boldsymbol{M}\right),$$

where $\boldsymbol{M}=\mathbb{E}[\mathrm{Var}(Y\mid\mathcal{Z})]\left\{\mathrm{diag}\left\{(K+1)\mathbf{1}_K\right\}+(K+1)\mathbf{1}_K\mathbf{1}_K^{\mathrm{T}}\right\}$.

All covariate-adaptive randomization methods have one thing in common: the assignments are dependent on the covariates $Z_1, Z_2, ..., Z_n$. Given the $\boldsymbol{\mathcal{Z}}$, then the T_i and Y_{ik} are conditionally independent. Here we let $A_n=\sqrt{\frac{K+1}{N}}\left[\sum_{i=1}^{n}\left(T_{i1}-T_{i0}\right)\left(y_i-\nu_i\right),\cdots,\sum_{i=1}^{n}\left(T_{iK}-T_{i0}\right)\left(y_i-\nu_i\right)\right]$ and $B_n=\sqrt{\frac{K+1}{N}}\left[\sum_{i=1}^{n}\left(T_{i1}-T_{i0}\right)\nu_i,\cdots,\sum_{i=1}^{n}\left(T_{iK}-T_{i0}\right)\nu_i\right]$. Because ν_i only depends on the $\mathcal{T}$ and $\mathcal{Z}$, $\{B_n\mid(\mathcal{X},\mathcal{T})\}$ is constant. Hence, A_n and B_n are independent given $(\mathcal{T},\mathcal{Z})$. Then we have for any sets D_1 and D_2,

$$\begin{aligned}
&\mathbb{P}\{A_n\in D_1,B_n\in D_2\}\\
&=\mathbb{E}[\mathbb{P}\{A_n\in D_1,B_n\in D_2\mid\mathcal{Z},\mathcal{T}\}]\\
&=\mathbb{E}[\mathbb{P}\{A_n\in D_1\mid\mathcal{Z},\mathcal{T}\}\mathbb{P}\{B_n\in D_2\mid\mathcal{Z},\mathcal{T}\}]\\
&=\mathbb{E}[\mathbb{P}\{A_n\in D_1\mid\mathcal{Z},\mathcal{T}\}I\{B_n\in D_2\}]\\
&=\mathbb{E}\left[\left(P\left\{A_n\in D_1\mid\mathcal{Z},\mathcal{T}\right\}-P\left\{\xi_1\in D_1\right\}\right)I\left\{B_n\in D_2\right\}\right]+P\left(\xi_1\in D_1\right)P\left(B_n\in D_2\right)\\
&\xrightarrow{P}P\left(\xi_1\in D_1\right)P\left(\xi_2\in D_2\right)\qquad\text{(By dominated convergence theorem)}
\end{aligned}$$

So far, the lemma 4.2 has been proved. ■

Lemma 4.3. *Suppose that the within-probability is bounded in probability; then under the null hypothesis $H_0:\boldsymbol{\alpha}=\mathbf{0}$ in $\mathbb{R}^K$, we have, as $N\to\infty$, for each $k=0,1,\ldots,K$,*

$$\frac{1}{n_k}\sum_{i=1}^{n}T_{ik}y_i\xrightarrow{P}\mathbb{E}[Y],$$

where $n_k=\sum_{i=1}^{N}T_{ik}$ represents the number of observations in group k, and T_{ik} is the indicator

variable that equals 1 *if the* i*-th observation belongs to group* k *and 0 otherwise. The expectation* $\mathbb{E}[Y]$ *remains the same across all groups under the null hypothesis.*

Proof. Under the CAR with within-probability bounded, we have

$$\frac{\sum_{i=1}^{N} T_{ik} y_k}{n_k} = \frac{K}{N} \sum_{i=1}^{N} T_{ik} y_i + o_P\left(\frac{1}{N^{1/2}}\right), \quad k \in \{1, 2, \cdots, K\}.$$

Because $(Y_{i0}, Y_{i1}, \cdots, Y_{iN})$ are conditionally independent of $\mathcal{T}$ given $\mathcal{Z}$, we have

$$\mathbb{E}(T_{ik} Y_i \mid \mathcal{Z}) = \mathbb{E}(T_{ik} \mid \mathcal{Z}) \mathbb{E}(Y_i \mid \mathcal{Z}) = \frac{1}{K} \nu_i,$$

and further we have

$$\frac{1}{N} \sum_{i=1}^{N} \nu_i \xrightarrow{P} \mathbb{E}(Y_i).$$

By large number theory, for any ϵ, we have,

$$\lim_{N \to \infty} \mathbb{P}\left(\left| \frac{K}{N} \sum_{i=1}^{N} T_{ik} Y_i - \frac{1}{N} \sum_{i=1}^{N} \nu_i \right| > \epsilon \,\middle|\, \mathcal{Z} \right) = 0.$$

Then, we have

$$\frac{1}{n_K} T_{iK} y_i \xrightarrow{P} \mathbb{E}[Y]$$

■

4.8.2 Proof of the Main Theorems

Proof of Theorem 4.4: Based on the remark 1 in the main paper, the test statistics S_k are expected to follow a normal distribution with a mean of $\mathbf{0}$ and an identity variance matrix. However, when a CAR is implemented during the trial, the variance of $\hat{\alpha}$ decreases, subsequently influencing the

asymptotic distribution. Theorem 4.4 is designed to re-estimate this asymptotic distribution and establish a corresponding adjusted test statistic. The proofs for Theorems 4.1, 4.2, and 4.3 can be readily derived from the proof of theorem 4.4.

We first re-represent the $\hat{\alpha}$ as follows

$$\hat{\alpha}_k = \frac{K+1}{N} \frac{1}{g'(\mathbb{E}(Y))} \sum_{i=1}^{n} y_i (T_{ik} - T_{i0}) + o_p(N^{-1/2}) \tag{4.1}$$

where $g(\cdot) = h^{-1}(\cdot)$.

Denote $\psi_k = \sum_{i=1}^{N} T_{ik} y_i / n_k$, and $\boldsymbol{g}\Big(\psi_0, \psi_1, \cdots, \psi_K\Big) = \Big(g(\psi_0), g(\psi_1), \cdots, g(\psi_K)\Big)$. And the use the Taylor expansion to expand $\boldsymbol{g}(\cdot)$ at $\boldsymbol{g}\left(\mathbb{E}(Y)\right)$, we have

$$\begin{aligned}
&\boldsymbol{g}\Big(\psi_0, \psi_1, \cdots, \psi_K\Big) - \boldsymbol{g}\Big(\mathbb{E}(Y), \mathbb{E}(Y), \cdots, \mathbb{E}(Y)\Big) \\
=&\mathrm{J}_{\boldsymbol{g}}\left(\mathbb{E}(Y), \mathbb{E}(Y), \cdots, \mathbb{E}(Y)\right) \begin{pmatrix} \psi_0 - \mathbb{E}(Y) \\ \psi_1 - \mathbb{E}(Y) \\ \vdots \\ \psi_K - \mathbb{E}(Y) \end{pmatrix} \\
&+ \frac{1}{2}\Big(\psi_0 - \mathbb{E}(Y), \cdots, \psi_K - \mathbb{E}(Y)\Big)^T \nabla^2 \boldsymbol{g}(\boldsymbol{\zeta}) \Big(\psi_0 - \mathbb{E}(Y), \cdots, \psi_K - \mathbb{E}(Y)\Big)
\end{aligned}$$

where $\nabla^2 \boldsymbol{g}(\cdot)$ is a Hessian matrix of $\boldsymbol{g}(\cdot)$, and ζ_k lies between ψ_k and $\mathbb{E}(Y)$.

Then, we re-represent the second term and third term in the Taylor expansion.

Based on the large number theory we showed in the Lemma 4.3, we have

$$\begin{aligned}
\psi_k - \mathbb{E}(Y) &= \frac{\sum_{i=1}^{N} T_{ik} y_i}{n_k} - \mathbb{E}(Y) \\
&= \frac{K+1}{N} \sum_{i=1}^{N} y_i T_{ik} - \mathbb{E}(Y) + \left[\frac{N - (K+1) n_k}{N} \right] \frac{\sum_{i=1}^{N} T_{ik} y_i}{n_k} \\
&= \frac{K+1}{N} \sum_{i=1}^{N} y_i T_{ik} - \mathbb{E}(Y) + \left[\frac{D_0 + D_1 + \cdots + D_k - (K+1) D_k}{N} \right] \frac{\sum_{i=1}^{N} T_{ik} y_i}{n_k} \\
&= \frac{K+1}{N} \sum_{i=1}^{N} y_i T_{ik} - \mathbb{E}(Y) + o_p(N^{-1/2})
\end{aligned}$$

where $D_k = n_k - n_0$. Furthermore, we have the following results via Lemma 4.2 and central limit theorem:

$$\begin{aligned}
\sqrt{N} \left(\frac{T_{i0} y_i}{n_0} - \mathbb{E}(Y) \right) &= \frac{\sum_{i=1}^{N} \left[(K+1) T_{i0} - 1 \right] y_i}{\sqrt{N}} - \frac{\sum_{i=1}^{N} y_i - N \mathbb{E}(Y)}{\sqrt{N}} + o_p(1) \\
&= \frac{\sum_{i=1}^{N} \left[T_{i0} - T_{i1} \right] y_i}{\sqrt{N}} + \cdots + \frac{\sum_{i=1}^{N} \left[T_{i0} - T_{iK} \right] y_i}{\sqrt{N}} - \frac{\sum_{i=1}^{N} y_i - N \mathbb{E}(Y)}{\sqrt{N}} + o_p(1) \\
&= O_p(1)
\end{aligned}$$

We obtain the the following result by symmetry.

$$\sqrt{N} \left(\frac{T_{ik} y_i}{n_k} - \mathbb{E}(Y) \right) = O_p(1)$$

Plugging the representation of second and third term into the Taylor expansion, and together with the law of large number theorem from lemma 4.3, we have

$$\begin{aligned}
\boldsymbol{g}\left(\psi_0, \cdots, \psi_K\right) =& \boldsymbol{g}(\mathbb{E}(Y), \cdots, \mathbb{E}(Y)) + \\
& \mathbf{J}_{\boldsymbol{g}}(\mathbb{E}(Y), \cdots, \mathbb{E}(Y)) \begin{pmatrix} \frac{K+1}{N} \sum_{i=1}^{N} y_i T_{i0} - \mathbb{E}(Y) \\ \vdots \\ \frac{K+1}{N} \sum_{i=1}^{N} y_i T_{ik} - \mathbb{E}(Y) \end{pmatrix} + o_P\left(N^{-1/2}\right)
\end{aligned}$$

Together with the MLE estimator of α_k which shown in equation (2.7) - (2.8) in the main paper, we have

$$\left(\hat{\alpha}_1, \cdots, \hat{\alpha}_K\right) = \mathbf{J}_g(\mathbb{E}(Y), \cdots, \mathbb{E}(Y)) \begin{pmatrix} \frac{K+1}{N}\sum_{i=1}^{N} y_i(T_{i1} - T_{i0}) \\ \vdots \\ \frac{K+1}{N}\sum_{i=1}^{N} y_i(T_{iK} - T_{i0}) \end{pmatrix} + o_p\left(N^{-1/2}\right)$$

The individual estimation for $\hat{\alpha}_k$ is:

$$\hat{\alpha}_k = g'(\mathbb{E}(Y))\frac{K+1}{N}\sum_{i=1}^{N} y_i(T_{ik} - T_{i0}) + o_p(N^{-1/2})$$

So far, the result equation (4.1) has been proved.

In the following, we obtain the variance-covariance matrix. Start with obtaining the Fisher information matrix,

$$\frac{1}{n}\mathcal{I}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} h'(\eta) & \frac{1}{n}\sum_{i=1}^{n} T_{i1}h'(\eta) & \frac{1}{n}\sum_{i=1}^{n} T_{i2}h'(\eta) & \cdots & \frac{1}{n}\sum_{i=1}^{n} T_{iK}h'(\eta) \\ \frac{1}{n}\sum_{i=1}^{n} T_{i1}h'(\eta) & \frac{1}{n}\sum_{i=1}^{n} T_{i1}^2h'(\eta) & 0 & \cdots & 0 \\ \frac{1}{n}\sum_{i=1}^{n} T_{i2}h'(\eta) & 0 & \frac{1}{n}\sum_{i=1}^{n} T_{i2}^2h'(\eta) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{i=1}^{n} T_{i(K-1)}h'(\eta) & 0 & 0 & \cdots & 0 \\ \frac{1}{n}\sum_{i=1}^{n} T_{i(K)}h'(\eta) & 0 & 0 & \cdots & \frac{1}{n}\sum_{i=1}^{n} T_{iK}^2h'(\eta) \end{bmatrix}$$

Notice that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} h'\left(\hat{\alpha}_0+\hat{\alpha}_1 T_{i,1}+\cdots+\hat{\alpha}_K T_{i,K}\right)\gamma'\left(\hat{\alpha}_0+\hat{\alpha}_1 T_{i,1}+\cdots+\hat{\alpha}_K T_{i,K}\right)\\
&=\frac{n_0}{n}h'(\hat{\alpha}_0)\gamma'(\hat{\alpha}_0)+\frac{n_1}{n}h'(\hat{\alpha}_0+\hat{\alpha}_1)\gamma'(\hat{\alpha}_0+\hat{\alpha}_1)+\cdots+\frac{n_K}{n}h'(\hat{\alpha}_0+\hat{\alpha}_K)\gamma'(\hat{\alpha}_0+\hat{\alpha}_K)\\
&=\frac{n_0}{n}h'(\frac{\sum T_{i,0}y_i}{n_0})\gamma'(\frac{\sum T_{i,0}y_i}{n_0})+\frac{n_1}{n}h'(\frac{\sum T_{i,1}y_i}{n_1})\gamma'(\frac{\sum T_{i,1}y_i}{n_1})+\cdots+\frac{n_K}{n}h'(\frac{\sum T_{i,K}y_i}{n_K})\gamma'(\frac{\sum T_{i,K}y_i}{n_K})\\
&\xrightarrow{P} h'h^{-1}(\mathbb{E}[Y])\gamma'h^{-1}(\mathbb{E}[Y]) \quad \text{(By Lemma 4.3)}
\end{aligned}
$$

and, similarly,

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} T_{ik}h'\left(\hat{\alpha}_0+\hat{\alpha}_1 T_{i,1}+\cdots+\hat{\alpha}_K T_{i,K}\right)\gamma'\left(\hat{\alpha}_0+\hat{\alpha}_1 T_{i,1}+\cdots+\hat{\alpha}_K T_{i,K}\right)\\
&=\frac{n_k}{n}h'(\hat{\alpha}_0+\hat{\alpha}_k)\gamma'(\hat{\alpha}_0+\hat{\alpha}_k)\\
&\xrightarrow{P}\frac{1}{K+1}h'h^{-1}(\mathbb{E}[Y])\gamma'h^{-1}(\mathbb{E}[Y])
\end{aligned}
$$

Then, we have

$$
\frac{1}{n}\mathcal{I}(\hat{\boldsymbol{\theta}})\xrightarrow{P}\frac{h'h^{-1}(\mathbb{E}[Y])\gamma'h^{-1}(\mathbb{E}[Y])}{\phi}\begin{bmatrix}
1 & 1/(K+1) & 1/(K+1) & \cdots & 1/(K+1)\\
1/(K+1) & 1/(K+1) & 0 & \cdots & 0\\
1/(K+1) & 0 & 1/(K+1) & \cdots & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots\\
1/(K+1) & 0 & 0 & \cdots & 1/(K+1)
\end{bmatrix}
$$

It follows that,

$$n\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}}) \xrightarrow{P} \frac{\phi}{h'h^{-1}(\mathbb{E}[Y])\gamma' h^{-1}(\mathbb{E}[Y])} \begin{bmatrix} K+1 & -(K+1) & -(K+1) & \cdots & -(K+1) \\ -(K+1) & 2(K+1) & K+1 & \cdots & K+1 \\ -(K+1) & K+1 & 2(K+1) & \cdots & K+1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -(K+1) & K+1 & K+1 & \cdots & 2(K+1) \end{bmatrix}$$

Because the matrix

$$\boldsymbol{\Sigma}_I = \begin{bmatrix} 2(K+1) & K+1 & \cdots & K+1 \\ K+1 & 2(K+1) & \cdots & K+1 \\ \vdots & \vdots & \ddots & \vdots \\ K+1 & K+1 & \cdots & 2(K+1) \end{bmatrix}$$

is symmetric with positive and real eigenvalues are $\{\lambda_1 = (K+1)^2, \lambda_2 = \lambda_3 = \cdots = \lambda_K = (K+1)\}$, its semi-definite. Thus, the eigenvalue decomposition could be used to obtain the $\boldsymbol{\Sigma}_I^{1/2}$.

$\boldsymbol{\Sigma_I} = SDS^{-1}$

$$= \begin{bmatrix} -1 & -1 & \cdots & -1 & 1 \\ 0 & 0 & \cdots & 1 & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} K+1 & 0 & \cdots & 0 & 0 \\ 0 & K+1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & K+1 & 0 \\ 0 & 0 & \cdots & 0 & (K+1)^2 \end{bmatrix} \begin{bmatrix} -1/K & -1/K & \cdots & -1/K & 1-1/K \\ -1/K & -1/K & \cdots & 1-1/K & -1/K \\ \vdots & \vdots & & \vdots & \vdots \\ -1/K & 1-1/K & \cdots & -1/K & -1/K \\ 1/K & 1/K & \cdots & 1/K & 1/K \end{bmatrix}$$

Then, we have

$$\boldsymbol{\Sigma}_I^{1/2} = \begin{bmatrix} \frac{\sqrt{K+1}(K-1)}{K} + \frac{K+1}{K} & -\frac{\sqrt{K+1}}{K} + \frac{K+1}{K} & \cdots & -\frac{\sqrt{K+1}}{K} + \frac{K+1}{K} & -\frac{\sqrt{K+1}}{K} + \frac{K+1}{K} \\ -\frac{\sqrt{K+1}}{K} + \frac{K+1}{K} & \frac{\sqrt{K+1}(K-1)}{K} + \frac{K+1}{K} & \cdots & -\frac{\sqrt{K+1}}{K} + \frac{K+1}{K} & -\frac{\sqrt{K+1}}{K} + \frac{K+1}{K} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{\sqrt{K+1}}{K} + \frac{K+1}{K} & -\frac{\sqrt{K+1}}{K} + \frac{K+1}{K} & \cdots & \frac{\sqrt{K+1}(K-1)}{K} + \frac{K+1}{K} & -\frac{\sqrt{K+1}}{K} + \frac{K+1}{K} \\ -\frac{\sqrt{K+1}}{K} + \frac{K+1}{K} & -\frac{\sqrt{K+1}}{K} + \frac{K+1}{K} & \cdots & -\frac{\sqrt{K+1}}{K} + \frac{K+1}{K} & \frac{\sqrt{K+1}(K-1)}{K} + \frac{K+1}{K} \end{bmatrix}$$

As a result,

$$\sqrt{n}\widehat{\mathrm{se}}(\hat{\boldsymbol{\alpha}}) \xrightarrow{P} \left\{ \frac{\phi}{h'h^{-1}(\mathbb{E}[Y])\gamma' h^{-1}(\mathbb{E}[Y])} \right\}^{1/2} \boldsymbol{\Sigma}_I^{1/2} \tag{4.2}$$

From the above proof, we can derive the following asymptotic distribution of the test statistics $(S_1, \cdots, S_K)$.

According to Lemma 4.2, we can readily obtain the following:

$$\sqrt{\frac{2(K+1)}{N}} \sum_{i=1}^{n} \left(T_{i1} - T_{i0}\right)(y_i) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \mathbb{E}[\mathrm{Var}(Y \mid \mathcal{Z})] \left\{ \mathrm{diag}\left\{2\mathbf{1}_K + 2(K+1)\sigma_h^2\right\} + 2\mathbf{1}_K \mathbf{1}_K^{\mathrm{T}} \right\}\right) \tag{4.3}$$

Combined (4.3) and the representation of $\hat{\boldsymbol{\alpha}}$ shown in (4.1), the asymptotic normality of the

$(\hat{\alpha}_1, \cdots, \hat{\alpha}_K)$ is as follows.

$$
\begin{aligned}
(\hat{\alpha}_1, \cdots, \hat{\alpha}_K) &= \frac{K+1}{N} \frac{1}{h' h^{-1}(\mathbb{E}(Y))} \Big\{ \sum_{i=1}^{n} y_i \left(T_{i1} - T_{i0}\right), \cdots, \sum_{i=1}^{n} y_i \left(T_{iK} - T_{i0}\right) \Big\} \\
&= \sqrt{\frac{K+1}{2N}} \frac{1}{h' h^{-1}(\mathbb{E}(Y))} \sqrt{\frac{2(K+1)}{N}} \Big\{ \sum_{i=1}^{n} y_i \left(T_{i1} - T_{i0}\right), \cdots, \sum_{i=1}^{n} y_i \left(T_{iK} - T_{i0}\right) \Big\}
\end{aligned}
$$

Then, we have

$$
\sqrt{N}(\hat{\alpha}_1, \cdots, \hat{\alpha}_K) = \mathcal{N}\left(\mathbf{0}, \frac{(K+1)\mathbb{E}[\mathrm{Var}(Y \mid \mathcal{Z})]}{2(h' h^{-1}(\mathbb{E}(Y)))^2} \left\{ \mathrm{diag}\left\{ 2\mathbf{1}_K + 2(K+1)\sigma_h^2 \right\} + 2\mathbf{1}_K \mathbf{1}_K^{\mathrm{T}} \right\}\right). \tag{4.4}
$$

Based on continuous mapping theorem, the asymptotic distribution of $(S_1, \cdots, S_K)$ is right followed by (4.2) and (4.4):

$$
(S_1, \cdots, S_K) = \left(\frac{\hat{\alpha}_1}{\mathrm{se}(\hat{\alpha}_1)}, \cdots, \frac{\hat{\alpha}_K}{\mathrm{se}(\hat{\alpha}_K)} \right) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{\mathbb{E}[\mathrm{Var}(Y \mid Z)] + (K+1)\sigma_h^2/2}{\phi h' h^{-1}(\mathbb{E}(Y))/\gamma' h^{-1}(\mathbb{E}(Y))} \boldsymbol{V}\right)
$$

where $\boldsymbol{V}$ a K-dimensional square matrix with $v_{ij} = 1$ if $i = j$ and $v_{ij} = \frac{1}{2+(K+1)\sigma_h^2}$ if $i \neq j$. ■

4.8.3 Additional Working Variance Analysis in GLMs

In this section, we provide additional derivations for various models to support the results presented in Table 4.3.3. *Logistic Regression with Probit Link*

The probit model is defined as $\mu = \Phi(\eta)$, hence $h(\eta) = \Phi(\eta)$, where $\Phi(\cdot)$ is the standard normal cumulative density of function (CDF). According to the Theorem 4.1 in the main paper, the working variance term is as follows:

$$
\begin{aligned}
&h'(\eta) = \varphi(\eta), \text{ where } \varphi(\eta) = 1/\sqrt{2\pi}\exp(-\eta^2/2), \\
&\phi h'(h^{-1}(\mathbb{E}(Y))) = \varphi(\Phi^{-1}(\mu)), \text{ where } \phi = 1 \text{ for binary endpoints.}
\end{aligned}
$$

Logistic Regression with Complementary Log-Log Link

The complementary log-log model is defined as $\mu = 1 - \exp(-\exp(\eta))$, hence, the inverse link is $h(\eta) = 1 - \exp(-\exp(\eta))$. According to the Theorem 4.1 in the main paper, the working variance term is as follows:

$$h'(\eta) = \frac{d}{d\eta}\left[1 - \exp(-\exp(\eta))\right] = \exp(\eta)\exp\left(-\exp(\eta)\right),$$
$$h^{-1}(\mathbb{E}(Y)) = \ln\left(-\ln(1-\mu)\right),$$
$$\phi h'\left(h^{-1}(E(Y))\right) = -\ln(1-\mu)(1-\mu), \text{ where } \phi = 1.$$

Gamma Regression with Log Link

The Gamma Regression with Log Link is defined as $\mu = \exp(\eta)$, hence, the inverse link is $h(\eta) = \exp(\eta)$. According to the theorem 4.1, the working variance term could be written as:

$$h'(\eta) = \frac{d}{d\eta}(\exp(\eta)) = \exp(\eta)$$
$$h^{-1}(\mathbb{E}(Y)) = \log(\mathbb{E}(Y))$$
$$\phi h' h^{-1}(\mathbb{E}(Y)) = \phi \exp\left[\log(\mathbb{E}(Y)\right] = \phi\mathbb{E}(Y)$$

Ordinal Logistic Regression with Logit Link

For an ordinal response with J categories, define the cumulative probabilities for category j (with $j = 1, \ldots, J-1$) as $\pi_j = P(Y \leq j)$. The canonical link for the binary endpoint is the logit function. Its inverse is $h(\eta) = \exp(\eta)/(1+\exp(\eta))$. According to the theorem 4.1, the working variance term could be derived as follows:

$$h'(\eta) = \frac{d}{d\eta}\left[e^{\eta}/(1+e^{\eta})\right] = e^{\eta}/\left(1+e^{\eta}\right)^2 = h(\eta)(1-h(\eta))$$

For a given threshold j, if we denote the expected cumulative probability by $\mathbb{E}(\pi_j)$ with $E(\pi_j) = h\left(\eta_j^*\right)$, where η_j^* is the value such that $h\left(\eta_j^*\right) = E(\pi_j)$. Then, we have

$$h'\left(\eta_j^*\right) = E(\pi_j)\left[1 - E(\pi_j)\right]$$

Because for the binary endpoint, it's common to set $\phi = 1$, for the cumulative logit model at threshold j we have:

$$\phi h'\left(h^{-1}\left(E(\pi_j)\right)\right) = 1 \cdot E(\pi_j)\left[1 - E(\pi_j)\right].$$

Negative Binomial Regression with Log Link

In a Poisson model, the variance is equal to the mean. However, in many practical situations, the observed variance exceeds the mean—a phenomenon known as overdispersion. The negative binomial model is commonly used to handle such data, and its variance is typically expressed as $\mathrm{Var}(Y \mid X) = \mu + \alpha\mu^2$. Here, α quantifies the extra variability (overdispersion) beyond what the Poisson model would predict. If $Y \mid X$ follows the negative binomial distribution, we have the $\mathbb{E}[\mathrm{Var}(Y \mid X)] = \mathbb{E}\left[\mu + \alpha\mu^2\right] = \mathbb{E}[\mu] + \alpha\mathbb{E}\left[\mu^2\right]$. With the log link, we have $h(\eta) = e^{\eta}$ and $h'(\eta) = e^{\eta}$. It follows that the working variance term is $\phi h'\left(h^{-1}(\mathbb{E}[Y])\right) = \phi\mathbb{E}(Y)$. According to the Theorem 4.1, we could compare the$\mathbb{E}[\mathrm{Var}(Y \mid X)]$ and the working variance term to determine the test size:

$$\frac{\mathbb{E}\left[\mu + \alpha\mu^2\right]}{\phi\mathbb{E}(Y)} = \frac{\mathbb{E}[\mu]}{\phi\mathbb{E}(Y)} + \frac{\alpha\mathbb{E}\left[\mu^2\right]}{\phi\mathbb{E}(Y)}.$$

When $\mu = e^{\eta}$ varies across the population due to the inherent heterogeneity, one has

$$\mathbb{E}\left[e^{2\eta}\right] \geq \left(\mathbb{E}\left[e^{\eta}\right]\right)^2 = \left(\mathbb{E}[\mu]\right)^2$$

with strict inequality. Hence, $\mathbb{E}[\mu^2]$ typically exceeds $(\mathbb{E}[\mu])^2$. Besides, the dispersion term $\alpha\mu_2$ with $\alpha > 0$ adds an additional μ^2 term to the variance, any tendency for μ^2 to exceed μ is further amplified. In other words, the variance grows faster than linearly in μ. Consequently, underestimating the variance of the estimated effect and resulting in an anti-conservative Type I error rate.

CHAPTER 5

Robust Inference for Covariate-adaptive Seamless Phase II/III Trials

5.1 Background

Clinical trials play a pivotal role in medical research, ensuring patient safety, and facilitating the development of innovative therapies. Traditionally, clinical trials have been categorized into four distinct phases, with Phase II primarily focused on dose selection and safety evaluation, while Phase III is dedicated to the assessment of treatment efficacy. However, in recent years, a novel and highly efficient trial design, known as the 'Seamless Phase II/III trials', has garnered significant attention from researchers, clinicians, and the pharmaceutical industry.

The Seamless Phase II/III trial design differs from the conventional model. It integrates both dose selection and the confirmation of a selected dose's efficacy within a single trial, resulting in a streamlined overall trial duration.

In a Seamless Phase II/III trial, time typically spent transitioning between traditional Phase II and Phase III is saved. Moreover, it reduces the required sample size, extends patient monitoring during Phase II, enabling a more comprehensive evaluation of treatment outcomes. This design also allows for the simultaneous comparison of various dosage levels, thereby reducing the number of required trials. Consequently, this innovative approach has gained popularity within the pharmaceutical and clinical research sectors.

Of particular note, Seamless trials are often employed in oncology due to the longer duration required to assess drug effectiveness, expediting the development of cancer treatments and providing patients with faster access to potentially life-saving options. Up to 2016, over 40 'first-in-human' cancer trials had adopted the Seamless strategy [41].

More recently, Ma, Wang, and Zhu (2022) [38] emphasized the significance and necessity of implementing covariate-adaptive randomization in Seamless Phase II/III trials. This approach mitigates imbalances in confounding covariates across treatment groups, addressing potential biases in results and further promoting the practical application of Seamless Phase II/III trials in real-world clinical research.

However, the application of seamless Phase II/III under CAR procedures is still in its developmental stages. Several challenges remain to be addressed: (1) In terms of design, CAR is widely adopted to ensure accurate treatment comparisons; however, research on seamless phase II/III trials under CAR remains limited. Although seamless phase II/III trials are appealing, they are unlikely to be accepted by current regulatory standards without adequately balanced prognostic factors. (2) Regarding the inference, there is still a lack of a well-defined inference framework for the use of seamless trials with CAR tailored to binary surrogate endpoint and time-to-event primary endpoint.

So far, there have been numerous studies on tests under the CAR framework [49, 48, 34, 8, 9, 36, 57, 31]. However, research specifically addressing seamless trials with CAR is relatively limited at present. Ma, Wang, and Zhu (2022) [38] introduced the use of the adjusted t-test for inference in seamless trials with CAR. Their study, however, primarily focused on continuous responses, whereas the predominant application of seamless trials, particularly in oncology trials, often involves surrogate binary endpoints and time-to-event primary endpoints. An inference framework for this particular aspect is currently lacking. Therefore, within this chapter, we propose a seamless Phase II/III design with CAR specifically designed for oncology trials featuring surrogate binary endpoints and time-to-event primary endpoints. We also introduce an inference framework that remains robust to model misspecification while ensuring the preservation of valid type 1 error rates.

5.2 Notation

Consider a seamless phase II/III clinical trial with K candidate treatment groups and one control group. At the end of the first stage of this seamless trial, we select the optimal treatment based on safety and efficacy, along with the control, to proceed to the second stage. During the whole process, CAR is implemented to allocate subjects based on their prognosis profiles.

Let n denote the total number of subjects planned for the entire clinical trial, with n_A and n_B representing the numbers of subjects enrolled in the first and second stages, respectively. We define T_{ik} as the treatment indicator for subject i with respect to treatment k, where k ranges from 0 to K, and i ranges from 1 to n. T_{ik} takes the value 1 if subject i was assigned to treatment k and 0 otherwise. Let $\widetilde{\boldsymbol{Z}}_i$ be a vector containing all prognostic factors of a patient that may influence patient responses. $\boldsymbol{Z}_i$ is a subset of $\widetilde{\boldsymbol{Z}}_i$, representing the covariate vector in our inference model. Let Y_{ik} represent the potential binary surrogate outcome for patient i if assigned to treatment arm k, which follows a Bernoulli distribution with success probability p_{ik}. Similarly, let S_{ik} denote the potential primary time-to-event outcome for the i-th patient if assigned to treatment arm k.

In a seamless phase II/III trial setting, for the initial n_1 subjects enrolled in Stage 1, the potential surrogate outcomes are denoted by $Y_{i0}, Y_{i1}, \ldots, Y_{iK}$, corresponding to the $K+1$ arms, including the control and K experimental treatments. Likewise, their potential primary time-to-event outcomes are represented as $S_{i0}, S_{i1}, \ldots, S_{iK}$. For the subsequent n_2 subjects enrolled in Stage 2, only the control arm and the optimal treatment arm identified in Stage 1 are considered. Thus, their potential surrogate outcomes are Y_{i1}, Y_{i2}, and their potential primary time-to-event outcomes are S_{i1}, S_{i2}. The observed surrogate outcome Y_i for patient i is defined as the potential outcome corresponding to the treatment arm actually assigned. Similarly, the primary outcome S_i for patient i is the potential primary outcome corresponding to the assigned treatment arm. Additionally, for patient i, the censoring time is C_i, and the potential censoring times corresponding to treatment arm k are denoted by C_{ik}. Consequently, the observed failure time for patient i is $\tilde{S}_i = S_i \wedge C_i$, and δ_i is an indicator function equal to 1 if $\tilde{S}_i = S_i$, and 0 otherwise.

The logistic regression model is very popular to be used to analyze the relationship between a binary outcome variable (e.g., yes/no, success/failure) and other predictor variables.Throughout this chapter, we assume that the underlying true model for the surrogate binary response is the logistic regression model:

$$Y_i \sim \text{Bernoulli}(p_i) \tag{5.1}$$

$$p_i = \text{logit}^{-1}\left(\delta_0 + \delta_1 T_{i1} + \delta_2 T_{i2} + \cdots + \delta_K T_{iK} + \boldsymbol{\beta}'\widetilde{Z}_i\right) \tag{5.2}$$

where δ_0 and δ_k represent the intercept and treatment effects for treatment k, $k \in \{1, 2, \cdots, K\}$, respectively. Moreover, $\boldsymbol{\beta}'$ denotes the transpose of the vector of coefficients associated with the all prognostic factors $\widetilde{\boldsymbol{Z}}_i$.

As for the primary time-to-event response S_i, we assume the underlying Cox proportional hazard model for the actual survival time. It describes the hazard function as

$$\lambda_i(t) = \lambda_0(t)\exp\left(\theta_1 T_1 + \cdots + \theta_K T_K + \widetilde{\boldsymbol{\gamma}}'\widetilde{\boldsymbol{Z}}_i\right) \tag{5.3}$$

where $\lambda_i(t)$ represents the hazard rate for subject i at time t, $\lambda_0(t)$ is the baseline hazard function, and $\widetilde{\boldsymbol{\gamma}}$ is the vector of coefficients corresponding to the all prognostic factors $\widetilde{\boldsymbol{Z}}_i$. The survival time S_i for subject i follows an exponential distribution with the hazard rate $\lambda_i(t)$.
Throughout, we have the following assumptions:

Assumption 5.1 (Discrete or Discretizable Covariates)**.** All covariates are either discrete variables or continuous variables that can be discretized into finitely many levels.

Assumption 5.2 (IID Sampling)**.** The observations $(S_{i0}, C_{i0}, S_{i1}, C_{i1}, \ldots, S_{iK}, C_{iK}, \widetilde{\boldsymbol{Z}}_i)$ for $i = 1, 2, \ldots, n$ are independent and identically distributed samples from the joint distribution of $(S_0, C_0, S_1, C_1, \ldots, S_K, C_K, \widetilde{\boldsymbol{Z}})$.

Assumption 5.3 (Conditional Independence of Treatment)**.** Given the covariate vector $\widetilde{\boldsymbol{Z}}$, the

treatment assignment T_i is independent of the potential failure times and censoring times:

$$T_i \perp\!\!\!\perp (S_{i0}, C_{i0}, \ldots, S_{iK}, C_{iK}) \mid \widetilde{\boldsymbol{Z}}_i$$

Assumption 5.4 (Independent Censoring)**.** The failure time S_i and censoring time C_i are conditionally independent given the covariates:

$$S_i \perp\!\!\!\perp C_i \mid \widetilde{\boldsymbol{Z}}_i$$

Assumption 5.5 (Proportional Censoring Hazards)**.** For any $k_1, k_2 \in \{0, 1, \ldots, K\}$, the ratio

$$\frac{P(C_{k_1} \geq t \mid \widetilde{\boldsymbol{Z}})}{P(C_{k_2} \geq t \mid \widetilde{\boldsymbol{Z}})} = g_{k_1,k_2}(t)$$

is a function of time t only, where $g_{k_1,k_2}(\cdot)$ does not depend on $\widetilde{\boldsymbol{Z}}$.

Assumption 5.1 ensures that the CAR balances the treatment assignments across each arm. Assumption 5.2 posits the iid (independent and identically distributed) property for each patient. In Assumption 5.3, the potential failure time and censoring are assumed to be unaffected by the treatment; however, only the observed time can be influenced. Assumption 5.4 adheres to the non-informative censoring principle commonly seen in survival studies. Assumption 5.5 pertains to robustness. It guarantees that the later-defined partial likelihood score function has an asymptotic mean of 0 under H_0. Assumptions 5.3 through 5.5 are consistent with those presented in Ye and Shao (2020) [57].

5.3 Framework of Design

Given that the primary application of seamless phase II/III designs is in oncology trials, we consider here a CAR-based design tailored specifically to a binary surrogate endpoint and a time-to-event primary endpoint. The procedure is outlined below.

Stage 1: *(Dose Selection Stage)* The first n_A patients are sequentially allocated to one of the K treatment groups or the control group using the CAR method. At the end of the first stage, an interim analysis is conducted, followed by dose selection. Based on the available binary surrogate endpoint data, we determine whether at least one dose is statistically significantly different from the control. If not, the current trial is terminated. If there is, the subsequent dose selection continues. Guided by predefined criteria, we select the optimal dose, denoted as k^*, using the surrogate binary response. Typically, the optimal dose is selected as the treatment arm demonstrating the most statistically significant effect compared to the control, provided it also has an acceptable safety profile.

Stage 2: *(Confirmatory Stage)* The selected arm k^* in the first stage, along with the control arm, proceeds to this stage. An additional n_B patients are sequentially assigned to treatment k^* or the control group using CAR. At the end of the trial, time-to-event data from both stages are available for the final analysis, comparing the treatment effect of the selected dose with the control arm.

The multi-arm CAR procedures described in Hu et al. (2023) [20] are employed within this framework. This unified class of CAR procedures for multi-arm trials aims to minimize a weighted average of stratum-specific, marginal, and overall imbalances. Specifically, this general approach reduces to stratified permuted block randomization when the weights for marginal and overall imbalance are set to 0, and it becomes equivalent to Pocock and Simon's minimization when the weights for stratum-specific and overall imbalance are set to 0. A detailed definition of imbalance is provided in Section 4.3; alternatively, readers may refer directly to the original paper by Hu et al. (2023) [20]. As described previously, throughout the seamless phase II/III design, CAR is utilized to generate treatment assignments T_i for n_A patients in stage 1 and n_B patients in stage 2.

Throughout this chapter, we focus on randomization procedures satisfy one of the following conditions.

Condition 5.1: $n_{\tilde{\mathbf{z}}}^{-1/2} D_n(\tilde{\mathbf{z}}) = o_p(1)$ for every $\tilde{\mathbf{z}}$

Condition 5.2: $n_{\tilde{\mathbf{z}}}^{-1/2} D_n(\tilde{\mathbf{z}}) \to^{\mathrm{D}} N(0, \nu_D)$ for every $\tilde{\mathbf{z}}$ and a known ν_D satisfying $0 < \nu_D \leqslant \frac{1}{4}$, and $D_n(\tilde{\mathbf{z}})$ and $D_n(\tilde{\mathbf{z}}')$ are independent for all $\tilde{\mathbf{z}} \neq \tilde{\mathbf{z}}'$, where '$\to$ D ' denotes convergence in distribution as $n \to \infty$.

5.4 Framework of Proposed Inference

Many clinical trials prefer non-parametric inference methods due to their simplicity and robustness to model misspecification. However, under CAR, non-parametric approaches can lead to conservative type I errors, especially when continuous covariates are discretized [57]. To address these challenges, we propose a comprehensive, robust inference framework tailored specifically for the covariate-adaptive seamless Phase II/III trial described above. Our framework consists of two distinct inference stages:

Dose Selection Stage: A robust adjusted Wald test for the binary surrogate endpoint in the dose selection stage.

Confirmatory Stage: A robust adjusted score test for the primary time-to-event endpoint in the confirmatory stage.

This framework effectively mitigates type I error conservatism and ensures robust and reliable inference under CAR.

5.4.1 Inference in Dose Selection Stage

Developing universally robust statistics against all forms of covariate omission can be challenging and impractical [48]. Therefore, for simplicity and robustness, we intentionally exclude covariates used in CAR from the working model during the dose selection stage. Specifically, while the true binary surrogate response model is assumed as (5.1), we adopt the following simplified working

model:

$$p_i = \text{logit}^{-1}\left(\delta_0 + \delta_1 T_{i1} + \delta_2 T_{i2} + \cdots + \delta_K T_{iK}\right), \tag{5.4}$$

where δ_k represents the treatment effect for arm k.

At the interim analysis, we test $H_{0,k} : \delta_k = 0$ versus $H_{a,k} : \delta_k \neq 0$. Due to variance reduction under CAR, the unadjusted Wald test statistic $\hat{\delta}_k / \operatorname{se}\left(\hat{\delta}_k\right)$ is conservative. As demonstrated in Chapter 4, under logistic regression and CAR (conditions in Hu et al. (2023)), we have the asymptotic normality result:

$$\left\{ \frac{\hat{\delta}_1}{\text{se}(\hat{\delta}_1)}, \frac{\hat{\delta}_2}{\text{se}(\hat{\delta}_2)}, \cdots, \frac{\hat{\delta}_K}{\text{se}(\hat{\delta}_K)} \right\} \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{\mathbb{E}[\text{Var}(Y \mid \boldsymbol{Z})] + (K+1)\sigma_h^2/2}{\text{Var}(Y)} \boldsymbol{V}^*\right), \tag{5.5}$$

where $\boldsymbol{V}^*$ a K-dimensional square matrix with $v_{ij} = 1$ if $i = j$ and $v_{ij} = 1/(2 + \sigma_h^2)$ if $i \neq j$. (Under Condition 5.1, $\sigma_h^2 = 0$.) Accordingly, we construct adjusted statistics to achieve valid inference under CAR as follows:

$$\boldsymbol{S}_{adj} = \sqrt{N} \left(\frac{\left(\bar{Y}_n\right)\left(1 - \bar{Y}_n\right)}{\sqrt{2(K+1)\left(\sum_{s=1}^{S} \frac{n_s}{n} \cdot \frac{\sum_{i=1}^{n} (y_i - \hat{\mu}_s)^2 I\{z_i = s\}}{n_s - 1} + (K+1)\hat{\sigma}_h^2/2\right)}} \right) (\boldsymbol{V}^*)^{-1/2} \hat{\boldsymbol{\delta}}, \tag{5.6}$$

with asymptotic distribution:

$$\boldsymbol{S}_{adj} \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \mathbf{I}_K\right). \tag{5.7}$$

The null hypothesis is rejected when $|S_{adj,k}| > z_{\alpha/2}$, where $z_{\alpha/2}$ denotes the upper $\alpha/2$-quantile of the standard normal distribution.

Since the inference involves simultaneously performing multiple hypothesis tests, controlling for multiplicity becomes essential. To address this issue, we adopt the Simes test to effectively control the FWER. At the end of the dose selection stage, the treatment arm associated with the smallest Simes-adjusted p-value is selected as the optimal dosage and proceeds to the confirmatory

stage. However, if none of the adjusted p-values are less than 0.05 (i.e., all Simes-adjusted p-values > 0.05), the trial is terminated early for futility, indicating insufficient evidence of efficacy among the candidate doses.

5.4.2 Inference in Confirmatory Stage

To evaluate the treatment effectiveness at the end of the trial, we perform a final analysis of the primary time-to-event endpoint using data from both stages. Commonly, log-rank tests (unstratified or stratified) are used in survival analyses. However, as shown by Ye and Shao (2020) [57], the conventional stratified log-rank test may still exhibit conservatism under CAR, especially when continuous covariates are discretized.

To address these limitations, we adopt the adjusted robust score test proposed by Ye and Shao (2020). This approach is based on a working hazard function incorporating treatment assignment and partially covariates:

$$\lambda(t) = \lambda_0(t) \exp\left(\theta_{k^*} T_{k^*} + \boldsymbol{\gamma}' \boldsymbol{Z}\right) \tag{5.8}$$

Here, $\mathbf{Z}$ denotes the covariate vector included in the working model, with dimension q, typically satisfying $q < p$. The true underlying model (5.3), however, includes the full covariate vector $\tilde{\mathbf{Z}}$ of dimension p. Parameters $\theta_{k^*} \in \mathbb{R}$ and $\gamma \in \mathbb{R}^q$ are estimates obtained from the working model. Under this framework, the robust score test statistic for testing the null hypothesis $(\theta = 0)$ is given by:

$$T_{\mathrm{RScore}} = n^{-1/2} U_\theta(0, \hat{\boldsymbol{\gamma}}) / \hat{B}_{\mathrm{RScore}}(0, \hat{\boldsymbol{\gamma}})^{1/2} \tag{5.9}$$

where $U_\theta(0, \hat{\boldsymbol{\gamma}})$ is the score function evaluated under the null hypothesis, and $\hat{B}_{\mathrm{RScore}}$ is the robust variance estimator derived from stratified sample variances and means of specific influence functions (for detailed formulations, see Ye and Shao, 2020 [57]).

The key advantage of this adjusted robust score test is its validity and robustness under CAR,

regardless of potential model misspecification. This ensures reliable statistical inference in our seamless phase II/III design, particularly when handling continuous covariates discretized during randomization.

Furthermore, since we have two-stage data at the end of the trial, it is essential to integrate the inference results obtained from both stages to perform an overall hypothesis test. To accomplish this, we adopt the combination test approach, specifically the weighted inverse normal method [29]. In this method, we first obtain the p-values $P_{1,I}$ and $P_{2,I}$ from each stage separately, using Ye and Shao's robust score test. Then, we combine these p-values into a single statistic:

$$C\left(P_{1,I}, P_{2,I}\right) = w_1 \Phi^{-1}\left(1 - P_{1,I}\right) + w_2 \Phi^{-1}\left(1 - P_{2,I}\right)$$

and reject the null hypothesis if the absolute value $|C\left(P_{1,I}, P_{2,I}\right)|$ exceeds the $1 - \alpha/2$ quantile of the standard normal distribution.

5.5 Numerical Studies

Part 1: Dose Selection based on Binary Response This subsection presents simulations to evaluate the performance of our adjusted test statistics for dose selection in the covariate-adaptive seamless Phase II/III design. We compare these adjusted statistics with the traditional Wald test in terms of type I error control, statistical power, and early stopping thresholds for futility under the working logistic model (5.4).

The true underlying model assumes a logistic relationship between the success probability p_i and covariates for each subject. Specifically, the binary response Y_i follows a Bernoulli distribution with parameter p_i :

$$p_i = \text{logit}\left(\delta_0 + \delta_1 I_{i1} + \delta_2 I_{i2} + \beta_1 Z_{i1} + \beta_2 Z_{i2}\right), \quad Y_i \sim \text{Bernoulli}\left(p_i\right)$$

We set $\beta_1 = 4$ and $\beta_2 = 2$. To assess type I error control, we assume no treatment effect, i.e., $\delta_1 =$

Table 5.1: Empirical power and No. futility for Stratifies Per-mutated block randomization under Wald test and adjusted test

parameters			power for test δ_1		power for test δ_2		futility for Tr 1		futility for Tr 2	
δ_0	δ_1	δ_2	wald test	adjust	wald test	adjust	wald test	adjust	wald test	adjust
-1.0	0.3	0.2	0.022	0.088	0.015	0.073	4352	3874	4492	4081
-1.0	0.3	0.4	0.016	0.082	0.024	0.105	4355	3863	4182	3626
-1.0	0.3	0.6	0.019	0.090	0.057	0.179	4326	3811	3653	2995
-1.0	0.3	0.8	0.020	0.089	0.110	0.283	4321	3784	3035	2374
-1.0	0.3	1.0	0.018	0.088	0.176	0.375	4349	3865	2457	1827
-1.0	0.3	1.2	0.019	0.083	0.273	0.500	4374	3874	1836	1304
-1.0	0.3	1.4	0.022	0.091	0.378	0.608	4351	3889	1346	877
-1.0	0.3	1.6	0.019	0.084	0.487	0.706	4380	3841	913	560
-1.0	0.3	1.8	0.018	0.085	0.589	0.788	4378	3897	618	360
-1.0	0.3	2.0	0.021	0.088	0.685	0.865	4329	3812	401	178
-1.0	0.3	2.2	0.022	0.097	0.771	0.912	4324	3841	328	99
-1.0	0.3	2.4	0.021	0.088	0.807	0.942	4357	3850	339	61
-1.0	0.3	2.6	0.019	0.090	0.819	0.968	4332	3834	420	33
-1.0	0.3	2.8	0.022	0.097	0.800	0.976	4334	3788	630	22
-1.0	0.3	3.0	0.019	0.089	0.786	0.984	4355	3836	840	16

$\delta_2 = 0$. To evaluate statistical power, we vary δ_2 from 0.2 to 3 in increments of 0.2.

As table above shown, it is evident that the adjusted Wald test perform better over the traditional Wald test in terms of type I error, power and number of early stopping when treatment effect exists. While both tests are valid under working model, the adjusted Wald test achieves this with a higher statistical power. On the other hand, the Wald test tends to be overly conservative with its type I error rate and, as a result, exhibits a relatively lower power. Furthermore, when considering the decision to stop early for futility before selecting the dose, it is apparent that the adjusted Wald test significantly reduces the likelihood of terminating the treatment when a treatment effect is truly exists. These results highlight the efficacy of the adjusted Wald test in the first stage analysis. Next, we investigate the performance of the whole inference framework.

Part 2: Testing for treatment effect based on the time-to-event data from two stages given the first stage results In this section, we evaluate the overall performance of our proposed inference framework, which integrates our adjusted Wald test for dose selection in the first stage with Shao's robust score test for the final survival analysis. To demonstrate the advantage of this combined

approach, we perform comparative analyses in the final stage using multiple established tests, including the standard log-rank test, stratified log-rank test, classical score test, and Shao's robust score test.

Specifically, we assess the advantages of our combined framework across different covariate scenarios and model assumptions. For discrete covariates (categorical or binary), such as gender or smoking status, we show that our approach remains robust in situations commonly encountered in clinical trials. For continuous covariates, including age, blood pressure, or biomarker levels, our method effectively handles the complexity and variability inherent in these data, resulting in more reliable inference.

Moreover, we evaluate our framework's performance under correct model specification as well as under scenarios involving covariate omission, a common form of model mis-specification in clinical practice. Our simulation results indicate that the proposed inference framework is particularly robust against such misspecification, maintaining accurate type I error control and statistical power when compared with existing alternatives. The specific simulation scenarios are detailed below.

> *Case 1* The true hazard model is $\lambda_0 \exp\left(\theta_1 T_1 + \theta_2 T_2 + \theta_3 T_3 + 1.5\,\text{Gender} - 0.75\,\text{Age}\right)$, $C \sim \text{Unif}(20, 40)$, both gender and age are binary with $P(\text{Gender} = \text{Female}) = 0.5$ and $P(\text{Age} = \text{“} < 60\text{"}) = 0.5$.
>
> *Case 2* The true hazard model is $\lambda_0 \exp\left(\theta_1 T_1 + \theta_2 T_2 + \theta_3 T_3 + 1.5 Z_1 - 0.75 Z_2\right)$, $C \sim \text{Unif}(20, 40)$, both Z_1 and Z_2 are continuous with $Z_1 \sim \mathcal{N}(1, 1)$ and $Z_2 \sim \mathcal{N}(0, 1)$.

For each case, we analyze the tests' performance under both correctly specified models (all relevant covariates included) and mis-specified models (with certain covariates omitted). Specifically, in Case 1, the mis-specified model omits the covariate "Age," while in Case 2, it omits covariate Z_2.

In our numerical study of case 1, we initially assumed that our model was correctly specified. Upon analysis, as shown in the Table 5.2, it is evident that the stratified log-rank test, score test, and robust score test all exhibit excellent performance, displaying similarities. However, when the

Table 5.2: Empirical power comparison of 4 designs based on stratified permuted block randomization with discrete covariates

True Parameters			Correct Model				Mis-specified Model			
θ_1	θ_2	θ_3	T_L	T_SL	T_S	T_RS	T_L	T_SL	T_S	T_RS
-0.10	-0.10	-0.10	0.0300	0.1740	0.1925	0.1985	0.0300	0.1740	0.0365	0.1230
-0.15	-0.10	-0.10	0.0555	0.2915	0.3130	0.3205	0.0555	0.2915	0.0650	0.1950
-0.20	-0.10	-0.10	0.1135	0.4155	0.4340	0.4390	0.1135	0.4155	0.1295	0.2990
-0.25	-0.10	-0.10	0.1990	0.5540	0.5805	0.5860	0.1990	0.5540	0.2250	0.4215
-0.30	-0.10	-0.10	0.2980	0.6635	0.6845	0.6870	0.2980	0.6635	0.3210	0.5465
-0.35	-0.10	-0.10	0.4700	0.7815	0.8015	0.8020	0.4700	0.7815	0.4880	0.6720
-0.40	-0.10	-0.10	0.5640	0.8340	0.8480	0.8485	0.5640	0.8340	0.5815	0.7570
-0.45	-0.10	-0.10	0.6730	0.8825	0.8910	0.8930	0.6730	0.8825	0.6955	0.8340
-0.50	-0.10	-0.10	0.7705	0.8900	0.8935	0.8935	0.7705	0.8900	0.7755	0.8640
-0.55	-0.10	-0.10	0.8330	0.9060	0.9100	0.9100	0.8330	0.9060	0.8405	0.8950

[1] Weights: Equal weights for each stratum
[2] Assign Probability: $(\text{Imb}_{(0)} \leq \cdots \leq \text{Imb}_{(3)}) \longleftrightarrow \{11/16, 2/16, 2/16, 1/16\}$

Table 5.3: Empirical power comparison of 4 designs based on Hu & Hu covairates adaptive randomization with discrete covariates

True Parameters			Correct Model				Mis-specified Model			
θ_1	θ_2	θ_3	T_L	T_SL	T_S	T_RS	T_L	T_SL	T_S	T_RS
-0.10	-0.10	-0.10	0.026	0.153	0.170	0.167	0.024	0.144	0.147	0.175
-0.15	-0.10	-0.10	0.072	0.246	0.270	0.269	0.056	0.233	0.257	0.279
-0.20	-0.10	-0.10	0.115	0.360	0.394	0.394	0.110	0.341	0.352	0.384
-0.25	-0.10	-0.10	0.196	0.485	0.502	0.500	0.214	0.504	0.504	0.548
-0.30	-0.10	-0.10	0.275	0.590	0.604	0.607	0.311	0.596	0.615	0.638
-0.35	-0.10	-0.10	0.391	0.645	0.657	0.658	0.411	0.655	0.667	0.681
-0.40	-0.10	-0.10	0.506	0.721	0.731	0.733	0.489	0.732	0.731	0.740
-0.45	-0.10	-0.10	0.573	0.738	0.749	0.753	0.568	0.735	0.732	0.741
-0.50	-0.10	-0.10	0.650	0.749	0.752	0.753	0.670	0.767	0.765	0.769
-0.55	-0.10	-0.10	0.732	0.782	0.782	0.782	0.708	0.767	0.768	0.769

[1] Weights: (Overall imbalance, Each marginal imbalance, Stratum imbalance) = {1/6, 1/3, 1/3, 1/6}
[2] Assign Probability: $(\text{Imb}_{(0)} \leq \cdots \leq \text{Imb}_{(3)}) \longleftrightarrow \{11/16, 2/16, 2/16, 1/16\}$

model is mis-specified, as revealed in the table 5.3, the proposed inference framework (Adjusted Wald test combined with the robust score test) outperforms the other methods. In contrast, both the log-rank test and score test exhibit suboptimal performance under these conditions.

In case 2, when dealing with continuous covariates, it's important to note that under the correctly specified model, both the score test and the robust score test exhibit superior performance, followed

Table 5.4: Empirical power comparison of 4 designs based on stratified permuted block randomization with continuous covariates

True Parameters			Correct Model				Misspecified Model			
θ_1	θ_2	θ_3	T_L	T_SL	T_S	T_RS	T_L	T_SL	T_S	T_RS
-0.10	-0.10	-0.10	0.0145	0.1080	0.1670	0.1725	0.0145	0.1080	0.0675	0.1295
-0.15	-0.10	-0.10	0.0185	0.1530	0.2700	0.2755	0.0185	0.1530	0.1105	0.1890
-0.20	-0.10	-0.10	0.0430	0.2370	0.3900	0.4005	0.0430	0.2370	0.1705	0.2700
-0.25	-0.10	-0.10	0.0710	0.3390	0.5285	0.5385	0.0710	0.3390	0.2600	0.3695
-0.30	-0.10	-0.10	0.1055	0.4295	0.6545	0.6600	0.1055	0.4295	0.3545	0.4815
-0.35	-0.10	-0.10	0.1595	0.5310	0.7525	0.7555	0.1595	0.5310	0.4510	0.5765
-0.40	-0.10	-0.10	0.2590	0.6440	0.8425	0.8440	0.2590	0.6440	0.5960	0.7025
-0.45	-0.10	-0.10	0.3055	0.6955	0.8870	0.8860	0.3055	0.6955	0.6595	0.7505
-0.50	-0.10	-0.10	0.4220	0.7875	0.9090	0.9100	0.4220	0.7875	0.7645	0.8320
-0.55	-0.10	-0.10	0.5105	0.8310	0.9185	0.9195	0.5105	0.8310	0.8195	0.8640

[1] Weights: Equal weights for each stratum
[2] Assign Probability: $(\text{Imb}_{(0)} \leq \cdots \leq \text{Imb}_{(3)}) \longleftrightarrow \{11/16, 2/16, 2/16, 1/16\}$

Table 5.5: Empirical power comparison of 4 designs based on Hu & Hu covairates adaptive randomization with continuous covariates

True Parameters			Correct Model				Misspecified Model			
θ_1	θ_2	θ_3	T_L	T_SL	T_S	T_RS	T_L	T_SL	T_S	T_RS
-0.10	-0.10	-0.10	0.019	0.080	0.146	0.149	0.013	0.087	0.061	0.116
-0.15	-0.10	-0.10	0.023	0.130	0.226	0.231	0.025	0.126	0.107	0.170
-0.20	-0.10	-0.10	0.047	0.199	0.338	0.339	0.057	0.205	0.181	0.238
-0.25	-0.10	-0.10	0.070	0.279	0.440	0.443	0.077	0.262	0.226	0.335
-0.30	-0.10	-0.10	0.094	0.368	0.547	0.553	0.121	0.383	0.340	0.448
-0.35	-0.10	-0.10	0.139	0.422	0.627	0.622	0.156	0.462	0.400	0.498
-0.40	-0.10	-0.10	0.213	0.508	0.690	0.692	0.225	0.532	0.494	0.600
-0.45	-0.10	-0.10	0.294	0.627	0.764	0.764	0.296	0.626	0.611	0.682
-0.50	-0.10	-0.10	0.355	0.658	0.781	0.782	0.332	0.651	0.617	0.675
-0.55	-0.10	-0.10	0.421	0.700	0.770	0.772	0.421	0.696	0.703	0.746

[1] Weights: (Overall imbalance, Each marginal imbalance, Stratum imbalance) = {1/6, 1/3, 1/3, 1/6}
[2] Assign Probability: $(\text{Imb}_{(0)} \leq \cdots \leq \text{Imb}_{(3)}) \longleftrightarrow \{11/16, 2/16, 2/16, 1/16\}$

by the stratified log-rank test. Conversely, under the mis-specified model, the robust score test remains the top performer, closely followed by the stratified log-rank test. This outcome diverges somewhat from the performance observed with discrete covariates, but it aligns with the underlying rationale. In this context, the stratified log-rank test requires discretization of covariates, restricting it to discrete information. In contrast, the robust score test can directly incorporate continuous covariates, thus preserving more information. Consequently, the stratified log-rank test loses

valuable information compared to the robust score test.

5.6 Discussion

In practical applications of seamless phase II/III trials, the majority of scenarios involve trials whose endpoints require longer time to become available, leading to extended trial periods and elevated costs. Trials with these characteristics are predominantly associated with oncology. Hence, the seamless phase II/III trial has thus garnered significant attention from pharmaceutical companies specializing in cancer trials. However, until now, the researches in seamless phase II/III trials and their inference methods have largely focused on continuous covariates. In a bid to enhance the applicability of the seamless phase II/III trial in real-world scenarios, we introduced a seamless phase II/III design tailored for binary surrogate endpoints and time-to-event primary endpoints. Along with this, we have proposed corresponding inference methods.

The newly proposed inference method has been demonstrated to be more robust against model misspecification while maintain a valid type I error. Particularly, when covariates are continuous, it significantly outperforms non-parametric analytical methods which are widely used currently. The proposed design's framework allows for trials to either terminate early or continue at the end of the first stage, where a set of treatments may be reduced due to either efficacy concerns or safety issues. This procedure boasts considerable flexibility concerning distributional assumptions, stopping rules, and selection criteria. Nevertheless, the current inference framework has its limitations. Specifically, the analytical method is best suited for CAR procedures with within-stratum imbalances bounded in probability. However, for scenarios where this condition isn't met, such as in Pocock & Simon minimization, the extra layer of variation induced by within-stratum imbalances must be estimated. We recognize these challenges and earmark them for future research endeavors.

CHAPTER 6

Conclusion & Future Work

In this dissertation, we addressed three critical methodological issues that advance clinical trial methodology, making it more aligned with current clinical practice.

In Chapter 3, we proposed a novel class of RAR procedures known as Group Doubly-Biased Coin Design (Group DBCD). This new design effectively maintains the desirable theoretical properties of Hu and Zhang's DBCD while accommodating group enrollment, missing at random (MAR), and delayed outcomes. To our knowledge, this is the first RAR procedure explicitly designed for group enrollment. We established rigorous theoretical foundations for Group DBCD, including unbiasedness and asymptotic normality. Furthermore, we demonstrated that delayed responses affect the convergence rate of treatment effect estimates, and that randomly missing responses can lead to an increase in the asymptotic variance of both allocation proportions and treatment effect estimation. This research significantly advances the practical applicability of RAR by addressing concerns around prolonged trial durations and providing essential theoretical guidance for real-world implementations.

In Chapter 4, we shifted focus from RAR to CAR. Unlike RAR, CAR methods are widely used in practice, strongly endorsed by regulatory agencies like the FDA for their capability to balance prognostic factors. However, the increasing complexity of clinical trials introduces new challenges for post-randomization inference under CAR. This chapter specifically investigated the statistical inference challenges in multi-arm trials with general outcomes under CAR, an important and under-explored area given the prevalence of multi-arm designs in clinical research. We thoroughly explored the behavior of the standard Wald test under model misspecification, identifying issues such as conservative or inflated Type I errors caused by omitted covariates and correlations among test statistics due to shared controls. To address these issues, we proposed adjusted test statistics that effectively control Type I error rates and potentially enhance statistical power. This represents the

first comprehensive study of statistical inference in multi-arm CAR trials accommodating general outcomes.

Chapter 5 introduced an innovative covariate-adaptive seamless phase II/III trial design, utilizing binary surrogate endpoints for interim dose selection and time-to-event primary endpoints for final analysis. Seamless designs are increasingly valued for their potential to streamline drug development. Nevertheless, existing inference methods are often conservative and exhibit diminished statistical power. The proposed design and associated model-based inference method specifically address these challenges, offering robust control of Type I error and enhanced power while demonstrating resilience to model misspecification. Simulation studies confirmed the practical effectiveness and efficiency of this new approach, particularly within oncology trials.

Despite the advancements provided in this dissertation, numerous critical problems remain open for further research. First, while the CAR methods discussed here focus primarily on balancing categorical or discretized covariates, recent work by Ma et al. (2024) [35] introduces approaches to balance continuous covariates on multiple moments. Developing comprehensive post-randomization inference methods that effectively utilize this advanced covariate balancing represents an essential future research direction. Additionally, the exploration of Covariate-Adaptive Response-Adaptive (CARA) designs in the context of precision medicine deserves further attention. This emerging class of designs integrates both ethical and efficiency considerations, tailored for personalized treatment strategies. Further theoretical and practical advancements in CARA designs would greatly enhance their applicability and acceptance in clinical research. These areas represent critical and promising directions, which we leave for subsequent research efforts.

Bibliography

[1] M. Alosh, F. Bretz, and M. Huque. Advanced multiplicity adjustment methods in clinical trials. *Statistics in medicine*, 33(4):693–713, 2014.

[2] D. C. Angus, S. Berry, R. J. Lewis, F. Al-Beidh, Y. Arabi, W. van Bentum-Puijk, Z. Bhimani, M. Bonten, K. Broglio, F. Brunkhorst, et al. The remap-cap (randomized embedded multifactorial adaptive platform for community-acquired pneumonia) study. rationale and design. *Annals of the American Thoracic Society*, 17(7):879–891, 2020.

[3] R. H. Bartlett, D. W. Roloff, R. G. Cornell, A. F. Andrews, P. W. Dillon, and J. B. Zwischenberger. Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics*, 76(4):479–487, 1985.

[4] P. Bauer. Multiple testing in clinical trials. *Statistics in medicine*, 10(6):871–890, 1991.

[5] P. Bauer and M. Kieser. Combining different phases in the development of medical treatments within a single trial. *Statistics in medicine*, 18(14):1833–1848, 1999.

[6] N. J. Birkett. Adaptive allocation in randomized controlled trials. *Controlled clinical trials*, 6(2):146–155, 1985.

[7] A. Biswas and J. Rao. Missing responses in adaptive allocation design. *Statistics & probability letters*, 70(1):59–70, 2004.

[8] F. A. Bugni, I. A. Canay, and A. M. Shaikh. Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, 113(524):1784–1796, 2018.

[9] F. A. Bugni, I. A. Canay, and A. M. Shaikh. Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, 10(4):1747–1785, 2019.

[10] J. D. Ciolino, H. L. Palac, A. Yang, M. Vaca, and H. M. Belli. Ideal vs. real: a systematic review on handling covariates in randomized controlled trials. *BMC medical research methodology*, 19:1–11, 2019.

[11] E. M. Connor, R. S. Sperling, R. Gelber, P. Kiselev, G. Scott, M. J. O'sullivan, R. VanDyke, M. Bey, W. Shearer, R. L. Jacobson, et al. Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. *New England Journal of Medicine*, 331(18):1173–1180, 1994.

[12] R. S. Finn, M. Martin, H. S. Rugo, S. Jones, S.-A. Im, K. Gelmon, N. Harbeck, O. N. Lipatov, J. M. Walshe, S. Moulder, et al. Palbociclib and letrozole in advanced breast cancer. *New England journal of medicine*, 375(20):1925–1936, 2016.

[13] A. B. Forsythe. Validity and power of tests when groups have been balanced for prognostic factors. *Computational Statistics & Data Analysis*, 5(3):193–200, 1987.

[14] M. P. Goetz, M. Toi, M. Campone, J. Sohn, S. Paluch-Shimon, J. Huober, I. H. Park, O. Trédan, S.-C. Chen, L. Manso, et al. Monarch 3: abemaciclib as initial therapy for advanced breast cancer. *Journal of Clinical Oncology*, 35(32):3638–3646, 2017.

[15] E. Hamilton, J. Cortes, O. Ozyilkan, S.-C. Chen, K. Petrakova, A. Manikhas, G. Jerusalem, R. Hegg, J. Huober, W. Zhang, et al. nextmonarch phase 2 randomized clinical trial: overall survival analysis of abemaciclib monotherapy or in combination with tamoxifen in patients with endocrine-refractory hr+, her2–metastatic breast cancer. *Breast Cancer Research and Treatment*, 195(1):55–64, 2022.

[16] Y. Hochberg. Multiple comparison procedures. *Wiley Series in Probability and Statistics*, 1987.

[17] G. N. Hortobagyi, S. M. Stemmer, H. A. Burris, Y.-S. Yap, G. S. Sonke, S. Paluch-Shimon, M. Campone, K. L. Blackwell, F. André, E. P. Winer, et al. Ribociclib as first-line therapy for

hr-positive, advanced breast cancer. *New England journal of medicine*, 375(18):1738–1748, 2016.

[18] F. Hu and W. F. Rosenberger. Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association*, 98(463):671–678, 2003.

[19] F. Hu and W. F. Rosenberger. *The theory of response-adaptive randomization in clinical trials*. John Wiley & Sons, 2006.

[20] F. Hu, X. Ye, and L.-X. Zhang. Multi-arm covariate-adaptive randomization. *Science China Mathematics*, 66(1):163–190, 2023.

[21] F. Hu and L.-X. Zhang. Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *The Annals of Statistics*, 32(1):268–301, 2004.

[22] F. Hu and L.-X. Zhang. On the theory of covariate-adaptive designs, 2020.

[23] F. Hu, L.-X. Zhang, S. H. Cheung, and W. S. Chan. Doubly adaptive biased coin designs with delayed responses. *Canadian Journal of Statistics*, 36(4):541–559, 2008.

[24] F. Hu, L.-X. Zhang, and X. He. Efficient randomized-adaptive designs. *The Annals of Statistics*, pages 2543–2560, 2009.

[25] Y. Hu and F. Hu. Asymptotic properties of covariate-adaptive randomization. *The Annals of Statistics*, 40(3):1794–1815, 2012.

[26] A. Ivanova. A play-the-winner-type urn design with reduced variability. *Metrika*, 58:1–13, 2003.

[27] M.-O. Kim, C. Liu, F. Hu, and J. J. Lee. Outcome-adaptive randomization for a delayed outcome with a short-term predictor: imputation-based designs. *Statistics in medicine*, 33(23):4029–4042, 2014.

[28] J. M. Lachin. Statistical properties of randomization in clinical trials. *Controlled clinical trials*, 9(4):289–311, 1988.

[29] W. Lehmacher and G. Wassmer. Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):1286–1290, 1999.

[30] X. Li, J. Zhou, and F. Hu. Testing hypotheses under adaptive randomization with continuous covariates in clinical trials. *Statistical methods in medical research*, 28(6):1609–1621, 2019.

[31] Y. Li, W. Ma, Y. Qin, and F. Hu. Testing for treatment effect in covariate-adaptive randomized trials with generalized linear models and omitted covariates. *Statistical Methods in Medical Research*, 30(9):2148–2164, 2021.

[32] Q. Liu, M. A. Proschan, and G. W. Pledger. A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association*, 97(460):1034–1041, 2002.

[33] Y. Liu and F. Hu. Balancing unobserved covariates with covariate-adaptive randomized experiments. *Journal of the American Statistical Association*, 117(538):875–886, 2022.

[34] W. Ma, F. Hu, and L. Zhang. Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association*, 110(510):669–680, 2015.

[35] W. Ma, P. Li, L.-X. Zhang, and F. Hu. A new and unified family of covariate adaptive randomization procedures and their properties. *Journal of the American Statistical Association*, 119(545):151–162, 2024.

[36] W. Ma, Y. Qin, Y. Li, and F. Hu. Statistical inference for covariate-adaptive randomization procedures. *Journal of the American Statistical Association*, 115(531):1488–1497, 2020.

[37] W. Ma, F. Tu, and H. Liu. Regression analysis for covariate-adaptive randomization: a robust and efficient inference perspective. *Statistics in Medicine*, 41(29):5645–5661, 2022.

[38] W. Ma, M. Wang, and H. Zhu. Seamless phase ii/iii clinical trials with covariate adaptive randomization. *Statistica Sinica*, 32(2), 2022.

[39] W. Ma, X. Ye, F. Tu, and F. Hu. carat: an r package for covariate-adaptive randomization in clinical trials. *Journal of Statistical Software*, 107:1–47, 2023.

[40] S. J. Pocock and R. Simon. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, pages 103–115, 1975.

[41] T. M. Prowell, M. R. Theoret, and R. Pazdur. Seamless oncology-drug development. *The New England journal of medicine*, 374(21):2001–2003, 2016.

[42] D. S. Robertson, K. M. Lee, B. C. López-Kolkovska, and S. S. Villar. Response-adaptive randomization in clinical trials: from myths to practical considerations. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 38(2):185, 2023.

[43] W. F. Rosenberger. Randomized urn models and sequential design. *Sequential Analysis*, 21(1-2):1–28, 2002.

[44] W. F. Rosenberger and J. M. Lachin. *Randomization in clinical trials: theory and practice*. John Wiley & Sons, 2015.

[45] W. F. Rosenberger, N. Stallard, A. Ivanova, C. N. Harper, and M. L. Ricks. Optimal adaptive designs for binary response trials. *Biometrics*, 57(3):909–913, 2001.

[46] T. A. Ryan. Multiple comparison in psychological research. *Psychological Bulletin*, 56(1):26, 1959.

[47] S. K. Sarkar and C.-K. Chang. The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, 92(440):1601–1608, 1997.

[48] J. Shao and X. Yu. Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics*, 69(4):960–969, 2013.

[49] J. Shao, X. Yu, and B. Zhong. A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*, 97(2):347–360, 2010.

[50] R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[51] G. W. Sledge Jr, M. Toi, P. Neven, J. Sohn, K. Inoue, X. Pivot, O. Burdaeva, M. Okera, N. Masuda, P. A. Kaufman, et al. Monarch 2: abemaciclib in combination with fulvestrant in women with hr+/her2- advanced breast cancer who had progressed while receiving endocrine therapy. *Journal of clinical oncology*, 35(25):2875–2884, 2017.

[52] R. N. Tamura, D. E. Faries, J. S. Andersen, and J. H. Heiligenstein. A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *Journal of the American Statistical Association*, 89(427):768–776, 1994.

[53] D. R. Taves. Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics*, 15(5):443–453, 1974.

[54] N. C. Turner, J. Ro, F. André, S. Loi, S. Verma, H. Iwata, N. Harbeck, S. Loibl, C. Huang Bartlett, K. Zhang, et al. Palbociclib in hormone-receptor–positive advanced breast cancer. *New England Journal of Medicine*, 373(3):209–219, 2015.

[55] L. Wei. An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association*, 73(363):559–563, 1978.

[56] L. Wei and S. Durham. The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, 73(364):840–843, 1978.

[57] T. Ye and J. Shao. Robust tests for treatment effect in survival analysis under covariate-adaptive randomization. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(5):1301–1323, 2020.

[58] Y. Yi, T. Ye, M. Yu, and J. Shao. Cox regression with survival-time-dependent missing covariate values. *Biometrics*, 76(2):460–471, 2020.

[59] M. Zelen. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64(325):131–146, 1969.

[60] M. Zelen. The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases*, 27(7-8):365–375, Sept. 1974.

[61] L.-x. Zhang. Strong approximations of martingale vectors and their applications in markov-chain adaptive designs. *Acta Mathematicae Applicatae Sinica, English Series*, 20(2):337–352, 2004.

[62] L.-X. Zhang. Central limit theorems of a recursive stochastic algorithm with applications to adaptive designs. *The Annals of Applied Probability*, 26:3630–3658, 2016.

[63] L.-X. Zhang, F. Hu, and S. H. Cheung. Asymptotic theorems of sequential estimation-adjusted urn models. *The Annals of Applied Probability*, 16:340–369, 2006.

[64] L.-X. Zhang, F. Hu, S. H. Cheung, and W. S. Chan. Asymptotic properties of covariate-adjusted response-adaptive designs. *The Annals of Statistics*, 35:1166–1182, 2007.

[65] L.-X. Zhang, F. Hu, S. H. Cheung, and W. S. Chan. Immigrated urn models—theoretical properties and applications. *The Annals of Statistics*, 39:643–671, 2011.

[66] H. Zhu and F. Hu. Sequential monitoring of response-adaptive randomized clinical trials. *The Annals of Statistics*, 38(4):2218 – 2241, 2010.

[67] H. Zhu and F. Hu. Interim analysis of clinical trials based on urn models. *Canadian Journal of Statistics*, 40(3):550–568, 2012.