Statistical Issues in Platform Trials with a Shared Control Group

Jessica Overbey

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Public Health
in the Department of Biostatistics at the Mailman School of Public Health

COLUMBIA UNIVERSITY

2020

© 2019
Jessica Overbey
All rights reserved

ABSTRACT

Statistical Issues in Platform Trials with a Shared Control Group

Jessica Overbey

Platform trials evaluating multiple treatment arms against a shared control are an efficient alternative to multiple two-arm trials. Motivated by a randomized clinical trial of the effectiveness of two neuroprotection devices during aortic valve surgery against a control, this dissertation addresses two open questions in the optimal design of these trials. First, to explore whether multiplicity adjustments are necessary in a platform design, simulation studies evaluating the operating characteristics of platform designs relative to independent two-arm trials were conducted. Under the global null hypothesis, relative to a set of two-arm trials, we found that platform trials have slightly lower familywise error; however, conditional error rates for an experimental treatment being declared effective given another was declared effective are above the nominal alpha-level. Adjusting for multiplicity reduces familywise error, but has little impact on conditional error. These studies show that multiplicity adjustments are unnecessary in platform trials of unrelated treatments. Second, to determine the optimal approach for comparing delayed entry arms to the shared control, five methods for incorporating historical controls into two-arm trials were applied to the analyses of simulated open platform trials and compared to pooling all controls. We found that when response rates are constant, pooling yields the lowest error and most precise, unbiased estimates. However, if drift occurs, pooling results in type I error inflation or deflation depending on the direction of drift, as well as biased estimates. Although superior to naive pooling, none of the alternatives explored guarantee error control or unbiased estimates in the presence of drift. Thus, only concurrent controls should be used as comparators in the primary analysis of confirmatory studies. Finally, these findings were applied to assess the design and analysis of the neuroprotection trial.

TABLE OF CONTENTS

CHAPTER 1 Introduction and Motivating Example **1**

1.1 Call for More Efficient Trial Designs 1

1.2 Motivating Example 2

CHAPTER 2 Platform Trials with Shared Controls **6**

2.1 Introduction 6

2.2 Closed Platform Designs 8

2.2.1 Single-Stage Platform Design 8

2.2.2 Multi-arm, Multi-stage (MAMS) Platform Design 9

2.2.3 Sequential Approach 11

2.2.4 Bayesian Approaches 13

2.3 Open Platform Designs 14

2.4 Discussion 18

CHAPTER 3 Is Adjustment for Multiple Testing Needed in a Platform Trial? **21**

3.1 Introduction 21

3.2 Methods 23

3.2.1 Design Frameworks 23

3.2.2 Outcome Assessment 25

3.2.3 Type I Error and Multiple Testing 25

3.2.4 Early Stopping 25

3.2.5 Simulation Scenarios 26

3.2.6 Operating Characteristics 26

3.3 Simulation Results 27

3.3.1 Type I Error Rates under the Null Scenario ($\pi_0 = \pi_1 = \pi_2 = 0.5$) 27

3.3.2 Type I Error Rates and Power under Alternative Scenarios 29

3.3.3 Sample Size 30

3.3.4 Treatment Failure Rates 31

3.4 Discussion 38

CHAPTER 4 Integrating Non-concurrent Controls in Analyses of Open Platform Trials. 42

4.1 Introduction 42

4.2 Methods 47

4.2.1 General Trial Framework 47

4.2.2 Outcome Assessment 48

4.2.3 Concurrent Controls Only 48

4.2.4 All Controls (Pooled) 49

4.2.5 Test-then-Pool 49

4.2.6 Power Prior (Static) 50

4.2.7 Power Prior (Dynamic) 50

4.2.8 Multisource Exchangeability Models (MEMs) 52

4.2.9 Trial Design 53

4.2.10 Scenarios 53

4.2.11 Simulation Procedures and Operating Characteristics 54

4.3 Results 56

4.4 Discussion 66

CHAPTER 5 Applications to the Neuroprotection Trial 71

5.1 Introduction 71

5.2 Multiple Testing in the Neuroprotection Trial 74

5.3 Incorporating Non-concurrent Controls in the CardioGard Efficacy Analysis 75

5.4 Conclusions 85

CHAPTER 6 Conclusion and Future Directions 86

LIST OF TABLES

Table 3.1 Simulation Results under the Null Scenario 32

Table 3.2 Simulation Results under an alternative scenario where $\pi_0 = \pi_1 = 0.5$ and $\pi_2 = 0.35$ ... 33

Table 3.3 Simulation Results for varying levels of π_2 34

Table 3.4 Average treatment failure rates and total sample size under an alternative scenario where $\pi_0 = \pi_1 = 0.5$ and $\pi_2 = 0.35$ 35

Table 3.5 Average treatment failure rates and total sample size for varying levels of π_2 – No multiple comparison adjustment 36

Table 3.6 Average treatment failure rates and total sample size for varying levels of π_2 – Bonferroni adjustment 37

Table 4.1 Probability of Experimental Treatment being Declared Effective 62

Table 4.2 Mean and Standard Deviations of π_0 Estimates 63

Table 4.3 Mean and Standard Deviations of Relative Risk (RR= π_2/π_0) Estimates..................... 64

Table 4.4 Bias of Mean Relative Risk (RR= π_2/π_0) Estimates 65

Table 5.1 Baseline and Operative Characteristics 79

Table 5.2 Observed Primary Outcomes in the CardioGard Arm and in CardioGard's Non-Concurrent and Concurrent Controls for Varying Launch Scenarios 83

Table 5.3 Estimates for Probability of Failure in Control Arm for Varying Launch Scenarios and Analytical Approaches 83

Table 5.4 Estimates for the Relative Risk of Failure in CardioGard versus Control for Varying Launch Scenarios and Analytical Approaches 84

LIST OF FIGURES

Figure 1.1 Enrollment Timelines for the Neuroprotection Trial.. 4

Figure 3.1 Simulated Trial Frameworks 27

Figure 4.1 Timeline of Enrollment in Simulated Open Platform Trials 53

Figure 5.1 Original Trial Design Schematic 73

Figure 5.2 Cumulative Number of Sites Open to Enrollment by Quarter 77

Figure 5.3 Observed Treatment Failure Rates in the Control Arm by Quarter 78

ACKNOWLEDGEMENTS

Words cannot adequately express my gratitude to my advisor Dr. Emilia Bagiella. Her mentorship, support, and guidance over the past several years made this work possible. I also want to thank my co-advisor Dr. Ying Kuen Cheung for his help and encouragement. Discussions and reading recommendations from him helped guide and improve this work immensely.

I would also like to thank the other members of my dissertation committee, Dr. Cheng-Shiun Leu, Dr. Jay P. Mohr, and the committee chair Dr. Zhezhen Jin for their insightful comments and encouragement.

I would like to express my gratitude to the Cardiothoracic Surgical Trials Network for allowing me to use the neuroprotection trial data and to my colleagues at Mount Sinai who have encouraged and supported me along the way.

Finally, I want to thank my parents, siblings, friends, and incredible husband Mark Zheng for their love, understanding, and support throughout this process.

CHAPTER 1

Introduction and Motivating Example

1.1 Call for More Efficient Trial Designs

In 2016, the Biotechnology Innovation Organization reported that the probability of a drug tested in a phase I study ultimately gaining FDA approval was 9.6%[1]. In this report thousands of drug development programs were evaluated from 2006 to 2015 and transitions within these programs from each phase, New Drug Application or Biologics License Application (NDA/BLA) filing, and ultimately FDA approval were analyzed. Of the 1,491 programs in phase III observed, only 58.1% transitioned to a NDA/BLA filing. The probability of FDA approval from phase III was estimated to be 49.6%[1]. While the failure of some phase III studies can be attributed to safety concerns or methodological issues, such as a poor design or an insensitive primary outcome, in many cases the treatment being tested is truly ineffective. Given the resources needed to conduct confirmatory studies, the low rate of success for these studies is striking. A typical confirmatory trial will compare a single experimental treatment to the current standard of care (control). At best, interim analyses will be pre-specified with contingencies to stop early for futility if there is sufficient evidence that the primary hypothesis will not be rejected or for efficacy if there is overwhelming evidence of superiority. In the past two decades there has been a call for more efficient trial designs that minimize the resources and the time required to determine whether an intervention is efficacious[2, 3].

1.2 Motivating Example

Peri-operative stroke is a major concern after cardiac surgery. The majority of peri-operative strokes are embolic, resulting from emboli, such as blood clots, air bubbles, fat deposits, and other debris, that break free in the body during surgery and travel to the brain causing a cerebral artery blockage[4]. A number of embolic protection devices aimed at capturing emboli during surgery, to prevent them from traveling to the brain, have been developed. Two such devices are Embol-X (Edwards Lifesciences, Irvine, CA), an intra-aortic filtration device, and CardioGard (Cardiogard, Or-Yehuda, Israel), a suction-based extraction device. As of 2015, both of these devices showed promising results in early phase studies and had demonstrated the ability to capture debris during surgery; however, the efficacy of these devices for reducing peri-operative stroke had not been established.

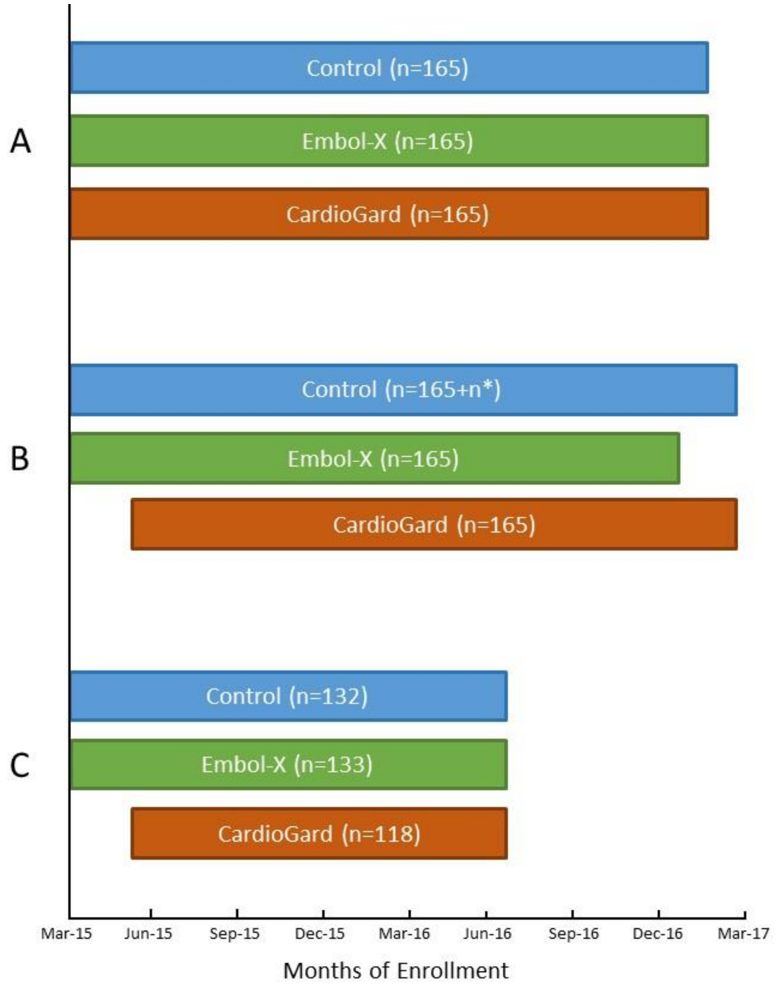
Typically, to establish efficacy, each company would sponsor a two-arm, confirmatory trial comparing their device to the standard of care. However, in collaboration with the Cardiothoracic Surgical Trials Network (CTSN), a National Institutes of Health (NIH) and Canadian Institutes of Health Research (CIHR) funded network aimed at improving cardiac surgery outcomes, Edwards and CardioGard agreed to a more efficient, single platform trial that simultaneously evaluated the efficacy of both devices. The primary outcome of the “Neuroprotection in Patients Undergoing Aortic Valve Replacement” study was a composite of death, clinically apparent stroke, or presence of post-operative emboli on diffusion-weighted MRI by post-operative day 7. At the end of the study, either or both of the devices could be declared effective relative to the shared control and no direct comparisons between the active arms were to be considered[5].

Patients were to be randomized with equal allocation to one of the two embolic protection devices or to a standard aortic cannula (control). Assuming a composite event rate of 50% in the control group, a sample size of 165 patients in each group yielded ~90% power to detect a 35% reduction in risk for each device compared to control using a 0.05 level chi-square test. A single interim analysis, based on group-sequential monitoring using efficacy boundaries specified by the Lan-DeMets approach with an O'Brien-Fleming type spending function, was prespecified. In addition, at the interim analysis a consideration for dropping either arm or halting the trial for futility was prespecified if the conditional power under the original alternative hypotheses for either was below 20%. No adjustment was made to the type I error rate in the sample size calculations as each comparison of device versus control was viewed as separate[5].

All three arms were intended to start enrolling at the same time. However, due to an unexpected delay in 510(k) approval by the FDA, the CardioGard device was not available at trial launch. Rather than delay the start of the trial, 1:1 randomization began into the Embol-X and control groups, with a plan to introduce the CardioGard device when it was available and change the randomization ratio to 1:1:1. When the Embol-X arm reached 165 enrolled it would close and 1:1 randomization into the control and CardioGard arms would continue until the CardioGard group reached 165.

The trial launched as planned with the CardioGard arm opening approximately 2 months into the study. At the interim analysis, after 132 patients had been randomized to the control, 133 to Embol-X, and 118 to CardioGard, the Data Safety and Monitoring Board (DSMB) recommended halting enrollment for futility of both experimental devices. In all analyses, experimental arms were compared only to concurrently randomized controls. Twelve patients were randomized to the control prior to the initiation of the CardioGard arm. Figure 1.1 shows the expected timelines of enrollment under the original and revised design as well as the actual timeline of enrollment.

Figure 1.1 Enrollment Timelines for the Neuroprotection Trial. A depicts enrollment under the original platform design where all arms were to open simultaneously, B depicts enrollment under the revised design that allowed for CardioGard to enter the trial after launch (n* is the number of controls randomized after the Embol-X arm closes), and C depicts the actual trial enrollment.



At the time the trial was launched, few methodological papers existed on how to design a multi-arm trial where treatment arms open in a staggered manner. It was not well established whether comparisons of each treatment to the shared control needed to be adjusted for multiple testing. Further, there was little published on analyzing platform trial data where not all arms were enrolled concurrently. Approaches that compare treatments to only concurrent controls, as was done in the neuroprotection trial, avoid any biases due to drift in the response rate, but lose efficiency by not using all available information. Platform trials are becoming more common across a broad range of therapeutic areas to address the growing costs of the "gold-standard" randomized controlled clinical trial and the call for more efficient designs. As experimental treatments are frequently at different stages in development, these issues are not unique to the neuroprotection trial. It is imperative that these methodological issues be examined carefully to optimize efficiency and ensure validity of trial results.

The work presented here is motivated by the neuroprotection trial. In Chapter 2, we introduce platform trials with a shared control group and review available designs. Chapter 3 assesses the operating characteristics of simulated open and closed platform trials with and without multiplicity adjustments and evaluates the need for multiple testing corrections under these frameworks. Next, Chapter 4 reviews analytical methods developed to incorporate historical control data into analyses of two-arm clinical trials and applies these methods to the analysis of an open platform trial. The efficiency and bias of these approaches are explored under varying scenarios of drift in response rates via a simulation study. In Chapter 5 we apply the findings of Chapters 3 and 4 to the neuroprotection trial and discuss whether the methods implemented in the trial were appropriate or whether alternatives may have been more efficient. Finally, in Chapter 6 we summarize findings and discuss future directions.

CHAPTER 2

Platform Trials with Shared Controls

2.1 Introduction

When several experimental treatments show signals for efficacy in early phase studies, a platform trial can offer several efficiencies relative to independent two-arm trials of each treatment. Under a single master protocol, each experimental treatment is evaluated simultaneously relative to a shared control group. Considering the number of patients, the time needed to develop protocols and set up infrastructure, and the costs that would be needed for multiple, independent two-arm trials, the resources required for a single platform trial are substantially lower. Additionally, multi-arm trials may have faster accrual than traditional two-arm trials, as patients may be more willing to enroll given they are more likely to be assigned to an experimental treatment[6]. Further, a single trial of multiple treatments reduces the competition for enrollment between trials within the target population. Despite these efficiencies, relatively few platform trials have been conducted. Prominent examples include the I-SPY2 and STAMPEDE trials in breast and prostate cancer respectively[7, 8]. Although most frequently conducted in cancer patients, platform trials have also been implemented in pneumonia, tuberculosis, and Alzheimer's disease[9].

Many designs that compare multiple treatments to a shared control have been developed. These designs can be categorized as controlled selection trials or controlled screening trials. Controlled selection trials combine the selection element of phase II studies with the confirmatory component of phase III studies. The goal of these trials is twofold: first, to determine whether any of the experimental treatments are superior to the control and second, to determine which of the experimental treatments shows the greatest superiority over the control. Several approaches to controlled selection exist including a single-stage approach, in which a selection is made after all patients have completed the trial[10], and multi-stage approaches, in which a selection is made at the end of a first stage and subsequent patients are randomized to only the selected treatment and control in ensuing stages[11, 12]. The latter approach will have a smaller sample size than the single-stage; however, selection is based off of a relatively small number of patients. If two or more experimental treatments show similar efficacy at the first interim analysis, it may be worthwhile to continue enrollment in multiple experimental arms to gain more information about the treatments before making a selection. To address this limitation, alternative multi-stage designs have been developed including an approach that allows a pre-specified number of treatments to continue at each stage[13], and others that allow any number to continue through each stage[14-17]. Regardless of which approach is chosen, at the end of a controlled selection trial, either a single experimental treatment will be selected or none will be deemed superior to the control. In contrast, a controlled screening trial can declare any number of experimental treatments effective relative to the shared control.

In the neuroprotection trial, the goal was to determine if either device was effective relative to the standard of care. This is not a selection design, but rather a screening design, in the sense that any number of treatments can be declared effective and treatments that are not effective will be screened out. Several controlled screening platform designs are available. Many were developed as closed trials, meaning all arms open to enrollment at the same time and no additional arms are added during the trial. Under this framework, single-stage platform designs, group-sequential multi-stage, multi-arm designs (MAMS)[18], a fully sequential approach[17], and Bayesian alternatives[9, 19] are available. More recently, open platform designs that allow for arms to be added during the course of the trial have been proposed. Saville et al.'s approach allows new arms to be added only when other arms are closed after an interim analysis[9], Elm et al. allow only a single arm to be added[20], while Hobbs et al. and Ventz et al. allow rolling entry of any number of arms throughout the course of the study[19, 21]. In what follows, the available closed and open controlled screening platform designs are reviewed. For open designs, approaches to analysis with concurrent versus all controls are also discussed.

2.2 Closed Platform Designs

2.2.1 Single-Stage Platform Design

In a single-stage platform design, patients are randomized to one of K active treatment groups or to a control group. After all patients are enrolled and primary endpoint data collection is complete, a single analysis is conducted in which a family of K null hypotheses of each active treatment versus control (H_{0j} for $j=1,\dots,K$) is tested. H_{0j} is the null hypothesis that arm j has an equivalent response to the control. If any H_{0j} are rejected, the corresponding treatments are declared effective. For a trial with a binary outcome, several design parameters are considered including θ_0, the response rate of the control group, and θ_j, the response rates of each experimental treatment group. An experimental treatment with a response rate of $\theta_0 + \delta_1$ would be considered sufficiently effective warranting either additional studies or a definitive declaration of efficacy. Under this design, a type I error would occur if a truly ineffective experimental treatment is declared effective and a type II error would occur if a truly effective treatment was declared ineffective.

Sample size for a single-stage platform design can be computed using standard sample size formulas. If the familywise type I error rate of the *K* null hypotheses is to be controlled, a Bonferroni correction can be applied. In a less conservative approach, Dunnett's multiple comparison procedure for comparing multiple treatments versus a control can also be applied[22]. Unlike the Bonferroni correction, which splits the type I error for each hypothesis test by the number of comparisons, Dunnett's procedure takes the fact that each comparison shares the same control into account and considers each hypothesis test conditionally independent given the control group. As a result, confidence intervals around the group differences calculated using Dunnett's procedure are narrower compared to those calculated under a Bonferroni correction.

Compared to *K* two-arm trials of each experimental treatment versus control, a closed screening platform design with *K* active treatment arms and one shared control arm will have a smaller total sample size. If multiple testing is not accounted for, given the same design parameters and equal allocation between groups, a series of *K* two-arm trials will have a total sample size of $2Kn$ whereas a single platform trial will have a total sample size of only $(K+1)n$. If multiple testing is accounted for, the sample size savings will be less striking, but the resource savings of the shared infrastructure will remain.

2.2.2 Multi-arm, Multi-stage (MAMS) Platform Design

Single-stage trials are straight-forward to design and conduct; however, more efficient designs allow treatment arms to be dropped as evidence accumulates that they are not effective and allow for a study to be terminated if no experimental treatments appear promising. Follmann et al. proposed a group sequential monitoring method for multi-arm trials based on pairwise test statistics performed at interim looks of the data using alpha spending functions of information time[18, 23]. At the start of the trial, the maximum sample size per arm is fixed. Information time ranges from 0 to 1 and is defined as the number of patients evaluated from the current set of arms divided by the number planned for the current set of arms. For the family of hypotheses H_{0j}, $Z_j(t)$ is the test statistic for testing H_{0j} at information time t. In the case where the trial aims to compare all active arms to a shared control, $Z_j(t)$ can be derived using Dunnett's multiple comparison procedure. At an interim analysis, the information time t is assumed to be the same across all comparisons. Blocked randomization is recommended to ensure that the number randomized over the maximum number planned in each arm is nearly equivalent at each interim analysis. The criteria for strong type I error control can be met under a variety of monitoring procedures including Pocock's and O'Brien and Fleming's. Calculating these monitoring boundaries is computationally intensive and computation time increases with the number of arms. Follman et al. recommend simulations to derive the boundaries. Alternatively, if multiple testing is not considered, each comparison of an experimental treatment versus the shared control can be viewed separately and the sample size and group-sequential boundary calculations for a two-arm trial can be applied.

Holding the design parameters, type I error rate, and power constant, the maximum sample size of a MAMS trial will be larger than that of a single-stage trial. However, the expected sample size will be lower as the design allows for arms to be dropped for futility. When used to the purposes of controlled screening, Follman et al.'s design does not allow for early stopping for efficacy. Efficacy stopping boundaries could be implemented; however, the ethical implications of early stopping for efficacy are complex. If only one arm passes the efficacy boundary and the trial is stopped, potentially effective treatments in other arms would be discarded. Since the goal of these trials is to determine if any of the treatments are effective, rather than identify a single effective treatment, efficacy stopping recommendation rules should be pre-specified over various scenarios prior to any interim looks at the data.

Compared to multiple two-arm trials, a MAMS design is more efficient. However, the maximum sample sizes for these trials, especially when multiple testing is accounted for, can be prohibitive. Although the expected sample size is smaller than a comparable single-stage platform design, the maximum sample size is larger and the actual sample size is uncertain as researchers cannot predict if and when futility boundaries will be crossed. In addition, when multiple testing is accounted for, designing a MAMS trial is challenging. Choosing the number of stages and calculating sample size and critical values from futility boundaries requires a great deal of computation time.

2.2.3 Sequential Approach

As an alternative to the MAMS design, Cheung proposed easy to compute sequential selection boundaries that are as flexible as group sequential boundaries in that interim analyses can be specified in any number and at any time[17]. Cheung's design was initially developed for controlled selection where a single effective treatment is selected at the end of the study. The boundaries were developed by extending Levin and Robbin's sequential elimination procedure to handle the constraint that experimental treatments not sufficiently superior to the control should not be selected[24].

In a study with a binary outcome, each response (x_{ij} = 0 or 1 for the *i*th patient in the *j*th treatment arm) is transformed to $Y_{ij} =1- a_j + a_j x_{ij}$. At an interim analysis, S_{jn}, the partial sum of the first n transformed responses in arm *j*, are computed and compared between the arms. For all *j*, a_j is prespecified with a_0 for the control group and $a_1 = a_2 = \ldots = a_k$ for the active arms. The a_0 and a_1 are asymmetrizing parameters set such that the control arm is favored in the comparisons of the partial sums. At an interim analysis, arm k is closed if $S_{kn} \leq max[S_{jn}] - d$ where *d* is a prespecified design parameter greater than 0. Additional design parameters include $\theta_{0,}$ the response rate of the control, and $\theta_{[1]}, \theta_{[2]}, \ldots, \theta_{[K],}$ the ranked response rates of each experimental treatment such that $\theta_{[1]} \leq \theta_{[2]} \leq \ldots \leq \theta_{[K]}$ where $\theta_{[K]}$ indicates the best treatment. An experimental treatment with a response rate of $\theta_0 + \delta_0$ would be considered ineffective while an experimental treatment with a response rate of $\theta_0 + \delta_1$ would be considered sufficiently effective. Any *j* experimental treatment with response rate $\theta_j \geq \theta_0 + \delta_1$ is acceptable for selection. Under the null hypothesis that none of *K* experimental treatments are effective, a type I error would be committed if any experimental treatment was selected. Under the alternative hypothesis that there are *m* effective treatments, a type II error would occur if no experimental treatment was selected.

Using Cheung's selection boundaries, two designs can be applied: the ELIM and the ELIM$_0$. With the ELIM design, the trial continues until either the maximum sample size is enrolled or until only one arm remains. In the ELIM$_0$ design the trial can be stopped as soon as the control arm is dropped. If the control arm is dropped the arm with the highest response rate is selected. By incorporating an additional early stopping rule, ELIM$_0$ will have a lower sample size on average compared to ELIM; however, the probability of selecting a suboptimal treatment is slightly increased.

Although developed for selection purposes, Cheung's sequential boundaries can be extended for screening. As with the MAMs design, sample sizes for trials designed using Cheung's sequential boundaries will, on average, be smaller than the single-stage design. Although Cheung's approach offers easier to compute monitoring boundaries, comparisons of the operating characteristics of Cheung's approach versus the MAMs approach have not been explored.

2.2.4 Bayesian Approaches

Bayesian alternatives are available for each of the closed screening platform designs discussed above. Rather than evaluating efficacy with frequentist hypothesis testing approaches, efficacy decisions are based on Bayesian posterior probabilities. In addition, futility monitoring is done using either Bayesian posterior predictive probabilities or Bayesian posterior probabilities in lieu of alpha spending functions[9, 19]. For a Bayesian closed platform design with a fixed randomization ratio across all open arms, a binary primary outcome, and prespecified futility rules, many of the design parameters needed for the frequentist approaches discussed above remain the same, including the minimum improvement in the response rate from θ_0 deemed sufficiently effective and clinically meaningful (δ_1), the maximum sample size (N_{max}), and the frequency of interim analyses. Additional parameters include the prior distributions of the response probabilities of each treatment. The response rate of each arm has an assumed prior distribution $\theta_j \sim Beta(\alpha, \beta)$. Criteria for stopping the trial for futility (F) are also predefined. In Saville et. al's approach, at each interim analysis, if the Bayesian posterior probability that a treatment j is superior to control is less than F (i.e. $Pr\ (\theta_j > \theta_0 + \delta_1 | data) < F$) arm j will be dropped. If the futility boundary is not crossed by any of the active treatment arms, the trial proceeds with all arms open[9]. In contrast to Saville et al.'s approach, Hobbs et al. bases futility monitoring on the Bayesian posterior predictive probability which, unlike the posterior probability, which only takes observed interim data into account, also accounts for the uncertainty of future responses[19]. Regardless of the futility monitoring approach used, at the end of the study, a type I error occurs if any ineffective treatment is declared superior to control. Each treatment has its own pre-specified type I error rate that is equal across the active arms. Type II error is defined as the trial not identifying truly effective treatments.

The Bayesian approaches are viable alternatives to the approaches discussed above. They can also be more efficient as the designs are easily extended to include flexible features such as outcome adaptive randomization, more frequent interim monitoring, and adaptive sample size estimation. Bayesian methods are being used with increasing frequency[25]. However, despite their flexible features, Bayesian methods are not widely accepted in the Phase III realm. As such, the more popular group sequential approaches will likely have an easier path through regulatory agencies.

2.3 Open Platform Designs

The closed platform designs discussed above assumed that all arms would start enrolling at the same time and that no additional treatment arms would be added during the course of the study. However, it is common for experimental treatments to be at different stages in development. As in the case of the neuroprotection trial, delays in approval could mean one treatment arm is not ready to start enrolling at trial launch. Additionally, if an unexpected new treatment becomes available, it would be opportune to add it to an already established platform trial.

Methods for open platform designs are not well developed. Saville et.al proposed a Bayesian open selection platform design with the goal of finding a single effective treatment to move forward for future studies or approval. Under the design, a prespecified number of treatments are open at the start of the trial and interim analyses are prespecified across fixed time or enrollment units. At interim analyses, arms can be dropped for futility and replaced with new treatments. The trial stops when an experimental treatment's Bayesian posterior probability that it is superior to control exceeds a pre-specified efficacy boundary[9]. Saville et al. suggest that their design can be extended to a screening paradigm but do not specify the design characteristics.

For active arms that are introduced during the study rather than included from the beginning, the experimental treatment can be compared to concurrent controls only or to all controls enrolled both prior to the opening of the experimental arm and concurrently. In Saville et al., simulation studies compared an open selection platform design that used concurrent controls only versus an open platform design that used all controls. They found that the use of all controls was more efficient than the use of concurrent controls. On average, fewer patients needed to be randomized prior to selecting a treatment in studies that used all controls compared to concurrent controls. However, the simulation studies kept the response rate of all treatments constant overtime and did not consider any drift in the response rate. In an applied example of Saville et al.'s design, Berry et al. consider drift[26]. This example was designed during the recent Ebola epidemic when the need for an effective treatment to reach patients was paramount to both increase infected patients' survival and decrease the rate of incident cases. Berry et al. proposed an open selection platform design where multiple treatments could be evaluated and an optimal treatment selected relative to a standard of care control arm. The primary endpoint of the study was death by 14 days after randomization. To account for drift in mortality rates, all analyses of the primary endpoint were to include month as a covariate to treatment assignment. The study was approved but never launched due to the decline of the Ebola epidemic in 2015. The design was criticized in an editorial by Brittain and Proschan where they pointed out that including month as a covariate in the model could be inadequate if the model is misspecified. Changes in the mortality rate could occur more frequently than monthly or there could be a treatment by time interaction in which a treatment helps only the recently infected and the proportion of recently infected changes over time[27].

In an alternate Bayesian approach, Hobb's et al. developed an open screening platform design where any number of treatments can be declared effective. Under this design, new treatments can be introduced to a trial at any time. Patients are randomized equally across all concurrently active arms and enrollment into each experimental arm continues to a prespecified maximum unless a futility boundary is crossed. Enrollment to the control arm continues as long as an experimental arm remains open. In order to control the familywise error rate and prespecify futility thresholds, Hobbs et al. require that the maximum number of arms be pre-specified[19]. To avoid any biases in analyses of active arms versus the control due to drift, Hobbs et al.'s design recommends that experimental treatments be compared to concurrent controls only. This approach requires extensive simulation studies to optimize the design. Furthermore, the need to prespecify the number of additional arms could be problematic in the event that an unanticipated new treatment becomes available.

In a frequentist approach, Elm et al. describe an open platform design where a single arm is added during the course of the trial and the sample size of the control arm is fixed regardless of when the new arm is added[20]. For example, a trial could open with two arms, an experimental arm and the control arm. Both arms have a maximum sample size of *n* and are randomized with equal allocation during a first stage. When a new arm is introduced, a second stage begins and the total sample size of the study increases from 2*n* to 3*n*. The randomization allocation ratio changes so that more patients are randomized to the new arm and enrollment into each arm finishes around the same time. To account for the different stages in the trial, Elm et al. recommend analyzing the primary endpoint data using a linear model with a fixed or random effect for stage or using an adaptive combination rule. Introducing a new arm if enrollment in the opening arms exceeds 50% showed a substantial decrease in power for the comparison of the new arm versus the control in simulation studies. As such, this design is likely not suitable if the timing of the additional arm is uncertain. Furthermore, it was developed for the addition of a single treatment arm and likely would not be appropriate if several new arms may be introduced.

To simplify the design of an open screening platform, Ventz et al. proposed a rolling-arms design[21]. Under this design, the trial opens with two arms, an experimental arm and the control. Sample size for the two arms, *n* for each, is based on standard sample size calculations with group sequential monitoring. As new treatments become available, new arms are added to the trial. Under this design, multiplicity is ignored. Assuming each experimental arm shares the same null and design alternative hypotheses relative to the shared control, the maximum sample size for each experimental arm is *n*. However, as experimental arms are added, the control arm's sample size increases so that *n* patients are concurrently randomized to the control arm while each active arm is open. Throughout the trial the randomization allocation is equal across all open arms. Active arms will stop randomizing either when *n* patients are randomized or if futility boundaries are crossed at an interim analysis. All interim and final analyses of an active treatment are done by comparing the active arm to controls that were randomized concurrently.

As an alternative, Ventz et al. explored the use of all controls with simulation studies. These studies assumed designs with one control arm and 5 experimental arms with staggered entry. The primary endpoint was time to disease progression. In the control arm, they assumed varying linear increases in the progression free survival of the population every 10 months. In the experimental arms, they assumed the same linear scaling factors so that the treatment effect remained constant. Type I error rates of rolling arm designs that compared experimental treatments to concurrent controls only were unaffected by the time trends. However, approaches that considered all controls produced biased hazard ratio estimates and inflated type I error rates. The authors did not explore analysis plans that would adjust for the drift and instead recommended that the use of all controls be avoided.

Open platform trials are still in early stages of development and implementation. There is no consensus on whether adjustments for multiplicity are needed for these designs. Methods that do not adjust for multiple testing are the most flexible as they allow any number of treatments to be introduced. The majority of open platform designs discussed in this chapter recommend the use of concurrent controls in the analysis of experimental treatments. However, using all information available may yield more efficient studies. Further exploration of various methods of analysis that incorporate all controls is needed.

2.4 Discussion

There are several efficiencies of platform trials relative to multiple two-arm trials including the need for fewer patients, faster accrual rates, and lower infrastructure costs. Despite these efficiencies, competing sponsors may be hesitant to enter their experimental treatments into the same study in order to avoid direct comparisons with other potentially efficacious experimental treatments. Although sample size, cost, and time are decreased under a single platform trial compared to multiple two-arm trials, for a sponsor with a single therapy, testing their treatment alone in a single two-arm trial may be more efficient. Similarly, in rare disease targets, a platform trial may prove prohibitive as not enough patients are available to test multiple experimental treatments simultaneously. Another inhibitor of platform trials for sponsors is the concern that another experimental treatment being tested will be found efficacious before their arm completes enrollment. This is not an issue in closed platform trials if efficacy monitoring is not implemented, but in open designs some treatments will reach final analysis while others are still enrolling. If a treatment is found effective it could change standard of care and cause enrollment into the platform trial to slow down or be permanently halted. However, a similar situation could also occur in the two-arm setting where a competing two-arm trial releases results prior to the completion of an ongoing trial in the same target population.

Despite these issues, platform trials have been successfully implemented. For many sponsors, the efficiencies may outweigh the concerns. In cases where the efficacy of multiple experimental treatments needs to be evaluated, screening platform trials offer an efficient alternative to multiple two-arm studies. However, more methodological research is needed to optimize these designs.

Across both closed and open designs, the decision to correct or not correct the type I error rate for multiplicity is an issue. Although multiple tests are evaluated, many argue that multiplicity adjustments are not needed as the familywise error rate of a platform study unadjusted for multiple comparisons will be equivalent to that of multiple two-arm trials[28, 29]. The neuroprotection trial followed this approach. However, since each comparison within the multi-arm trial shares the same control group, the comparisons are not independent and the test statistics will be positively correlated. Simulation studies are needed to explore this issue in the open platform setting. Ventz et al. demonstrated that their rolling arms design had nearly identical familywise type I error rates and power relative to multiple two-arm trials; however, they did not report the conditional probabilities of a type I error given another had been made[21]. Correcting for multiple comparisons in a closed selection trial is straightforward; but in an open platform trial where the number and timing of arm entry is not fixed, type I error control is a challenge.

Methods for closed screening platform trials are well established. However, as with the neuroprotection trial, there is an increasing need for platform designs that allow for experimental arms to be added during the study. In open designs, no consensus exists on whether analyses of experimental treatments should incorporate concurrent controls or all available controls. Studies exploring how much efficiency is lost by not using all controls and studies exploring analytical approaches that use all controls under various drift scenarios are needed. This thesis seeks to address these two gaps identified above. In subsequent chapters, the need for multiple testing corrections in platform trials is assessed and approaches for incorporating all controls into the analyses of open platform trials are explored.

CHAPTER 3

Is Adjustment for Multiple Testing Needed in a Platform Trial?

3.1 Introduction

In the context of a clinical trial, a type I error results in promoting a drug or an intervention to a next step when, in fact, the intervention is not efficacious. Depending on the phase of the study, this can mean a truly ineffective treatment is moved forward for further study or potentially even for Food and Drug Administration (FDA) approval. In this case resources that could be devoted to other promising treatments are wasted, and more consequentially, an ineffective treatment could be translated to patient care. In a two-arm trial comparing a single experimental treatment to a control, the type I error probability is protected by pre-specifying a stringent threshold (significance level) for testing and rejecting the null hypothesis that the experimental treatment and control have an equivalent effect. The significance threshold should be sufficiently low such that if the null is rejected, stakeholders feel confident that the observed result very likely represents a true difference. However, there is a tradeoff between error protection and sample size; the lower the significance level, the higher the sample size required to test the hypothesis. Significance levels of 0.05 and 0.025 for two-sided and one-sided hypotheses, respectively, are standard for phase III, two-arm trials.

In a platform trial, protection of type I error is more complicated, as multiple null hypotheses, one for each experimental treatment being evaluated, are tested. Therefore, multiple type I errors can be made. It has not been well established whether tests of each experimental treatment's efficacy relative to the control need to be adjusted for multiple testing in a platform trial. In a systematic review of multi-arm trials published in 2012 across four major medical journals (British Medical Journal, The Lancet, New England Journal of Medicine, and PLoS Medicine), 9 out of 20 (45%) exploratory multi-arm trials and 21 out of 39 (54%) confirmatory multi-arm trials did not apply adjustments for multiple testing[28]. Although multiple tests are conducted under a platform design with a shared control group, a common view is to treat each comparison of an active group versus control as a separate trial[29]. In this sense, the familywise type I error in the multi-arm trial would be equivalent to the overall type I error incurred by multiple two-arm trials.

The familywise error rate (FWER) is the probability of at least one type I error occurring across multiple hypothesis tests. Simulation studies comparing type I error rates between platform studies and multiple two-arm trials with the same number of subjects in each arm demonstrated that FWER is similar between the two approaches[19, 21]. FWER is a marginal probability that does not consider the occurrence of multiple type I errors across the hypotheses, but rather the probability of any single error occurring. In the setting of multiple, two-arm trials this is not an issue as tests of each experimental treatment versus control are independent and the probability of multiple type I errors occurring simultaneously is minimal, and generally not a concern. However, under a platform design, since each experimental treatment shares the same control group, test statistics for each hypothesis test are positively correlated. As such, the conditional probability of a type I error for the comparison of one experimental treatment versus the control given a type I error has been made on another comparison will be higher than the nominal significance level[30]. This means that the occurrence of one type I error in a platform design increases the likelihood of a type I error for the other comparisons so that if one ineffective treatment has been declared effective, there is a heightened probability that another ineffective treatment will be declared effective as well.

Several approaches, both frequentist and Bayesian, have been developed for evaluating multiple treatments relative to a shared control[9, 18-22]. Given that the number of experimental treatments in a platform trial is fixed, FWER can be controlled under any of these designs. Previous simulation studies exploring the operating characteristics of platform studies have reported FWER, but have not reported the conditional probability of a type I error given another has been made[19, 21].

In this simulation study, conditional type I error rates are explored under platform designs with two experimental treatment groups and a shared control group with and without adjustments for multiple testing. The study aims to evaluate the impact of adjusting versus not adjusting for multiple testing on FWER and conditional type I error rates in a platform trial relative to multiple, two-arm trials.

3.2 Methods

3.2.1 Design Frameworks

Two platform trial frameworks were explored, a closed design where all three arms enroll simultaneously and an open design where the first experimental treatment group and the control group begin enrolling before the second experimental treatment is introduced. For comparison purposes, independent two-arm trials were also simulated (Figure 3.1). An open framework was explored in addition to the closed as open designs that allow for new treatments to be added to an existing, ongoing trial are being increasingly used[9, 19, 21]. Three arms were chosen for simplicity of interpretation, but results can be extrapolated to additional armed studies. Under the closed platform framework, enrollment into all three treatment arms opens simultaneously with up to 200 patients enrolled in each group for a maximum trial sample size of 600. Under the open platform framework, trial enrollment opens with only the control arm and the first experimental treatment. After 100 patients are randomized with equal allocation to the two groups, the second experimental arm is opened to enrollment and the randomization ratio changes to 1:1:1. Once the first experimental treatment arm reaches a maximum sample size of 200, it is closed. Enrollment continues to the control and second experimental treatment arms until the second experimental treatment arm reaches a maximum sample size of 200. Enrollment into the control arm is extended to a maximum sample size of 250 so that 200 controls are concurrently randomized to each of the experimental treatment arms giving a maximum trial sample size of 650. Under the independent two-arm, trial framework, rather than a single trial being conducted, two independent trials of each experimental treatment versus independent control groups are simulated. Each two-arm trial has a maximum sample size of 400 and patients are randomized to the experimental treatment or control with equal allocation so that a set of two trials under the independent, two-arm framework has an overall maximum sample size of 800.

3.2.2 Outcome Assessment

A binary outcome of treatment success or failure was considered. For each treatment group *k,* where *k*=0 denotes the control group and *k*=1 or 2 denotes the experimental groups, x_k denotes the number of treatment failures in *n* patients and $x_k \sim Binomial(n, \pi_k)$. For each experimental treatment, the one-sided null hypothesis that the failure rate of the experimental treatment is equivalent or worse than the control (H_0: $\pi_k \geq \pi_0$) was tested using a Z-statistic, $Z_{obs(k)}$, derived from a Chi-square test. If $Z_{obs(k)} \leq Z_{critical}$ then treatment k was declared effective relative to control.

3.2.3 Type I Error and Multiple Testing

A type I error occurs if, under the null hypothesis, an experimental treatment is declared effective relative to control. Initial simulations included no adjustment for multiple testing and derived $Z_{critical}$ so that each comparison of experimental treatment versus control was evaluated using a one-sided, 0.025 level test. Additional simulations were conducted using a Bonferroni adjustment to account for multiple testing so that each comparison of experimental treatment versus control was evaluated using a one-sided, 0.0125 level test.

3.2.4 Early Stopping

Simulations were conducted under each framework with and without early stopping for efficacy and futility. Trials with 0, 1, 2 or 3 equally spaced interim analyses were evaluated. When 3 interim analyses were pre-specified, analyses occurred after 50, 100, and 150 patient responses were observed in each treatment group. Randomization was blocked to ensure equal treatment allocation at interim analysis points. Efficacy boundaries were determined using

O'Brien-Fleming spending functions. Futility was defined as the conditional power under the current trend being less than 10%. At an interim analysis, if both treatments crossed a stopping boundary, all three arms closed. However, if only one treatment crossed a stopping boundary, only that treatment arm closed while the control arm and the other experimental arm stayed open.

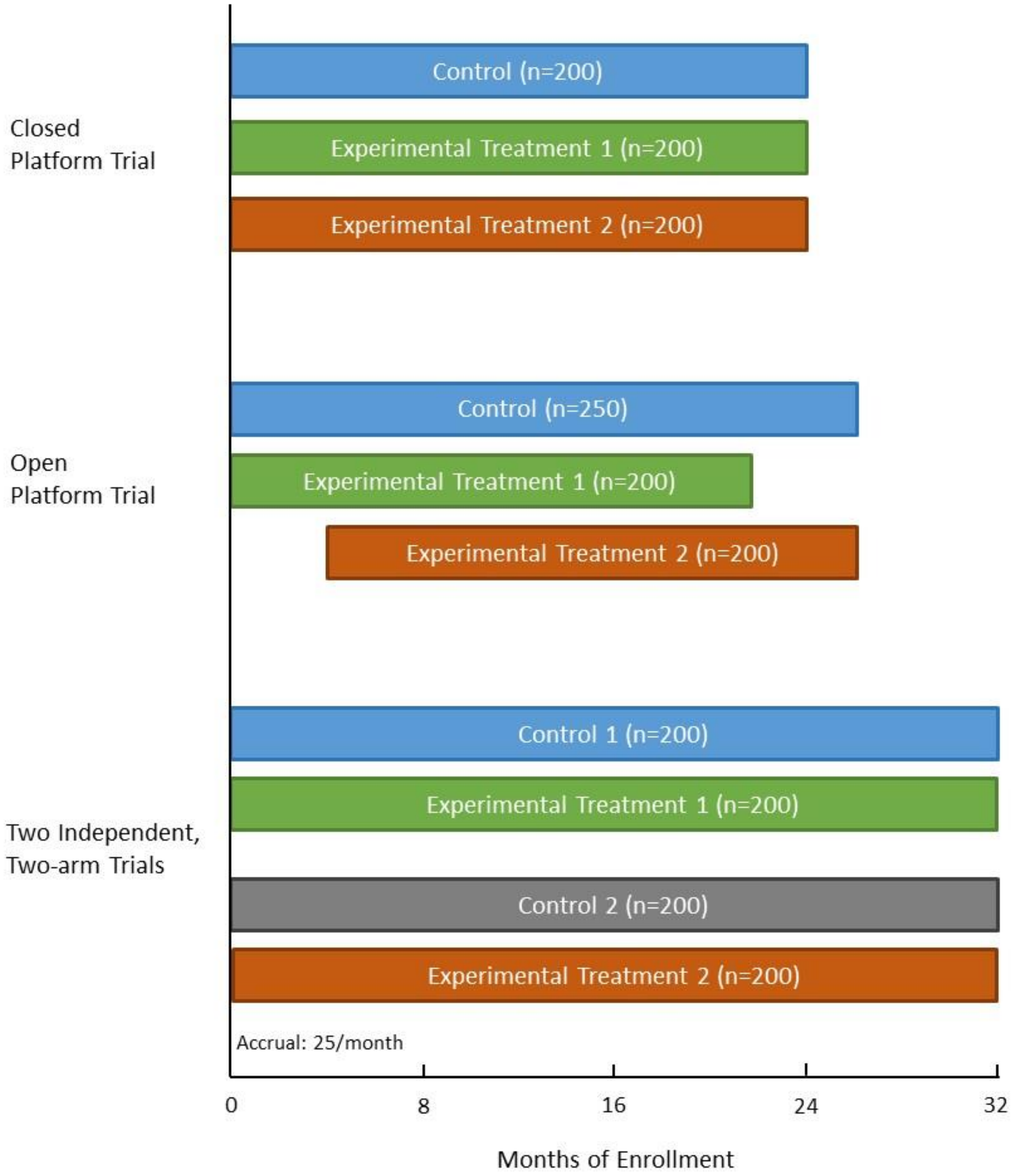
3.2.5 Simulation Scenarios

The failure rate of both the control group and the first experimental group across all scenarios was set to $\pi_0 = \pi_1 = 0.5$. The failure rate of the second experimental group (π_2) varied from 0.3 to 0.7. For each framework, number of interim analyses, and set of failure rates, 100,000 iterations were simulated. All simulations we conducted using R v3.5.0[31]. For each trial simulated, uniform random variates were used to simulate patient response. If the uniform variate of a subject was smaller than the "true" failure rate of the subject's assigned group, the subject was deemed to have failed. Blocked randomization was done using the blockrand package[32].

3.2.6 Operating Characteristics

Several operating characteristics were evaluated. The primary measure of interest was the number of trials that resulted in either or both experimental treatments being declared effective. Under the null scenario ($\pi_0 = \pi_1 = \pi_2 = 0.5$), in addition to FWER, conditional type I errors rates (i.e. the probability of a type I error for one treatment group given the other was declared effective) were also explored. In addition, the total sample size and the failure rate across the entire trial or trials for the independent, two-arm trial framework were evaluated.

Figure 3.1 Simulated Trial Frameworks



3.3 Simulation Results

3.3.1 Type I Error Rates under the Null Scenario ($\pi_0 = \pi_1 = \pi_2 = 0.5$)

Table 3.1 shows the simulation results for each design framework and number of interim analyses under the null scenario where $\pi_0 = \pi_1 = \pi_2 = 0.5$. The probability of a type I error for each individual comparison of an experimental treatment versus control was equivalent across the three frameworks. The family wise error rate (FWER), the probability of at least one type I error occurring, was also comparable across the three frameworks. As expected, in the independent, two-arm trial framework with no interim analyses, the FWER was approximately 0.05 and 0.025 for the unadjusted and Bonferroni adjusted scenarios respectively. Compared to the two-arm trial framework with no interim analyses, the closed platform framework had a slightly lower FWER (0.046 and 0.022), as did the open platform framework (0.048 and 0.024).

Although FWER was comparable across the frameworks, under the null scenario, the conditional probability of either experimental treatment being declared effective given the other had been declared effective were substantially higher under the platform frameworks. In the independent two-arm framework, given no interim analyses or multiple comparison adjustment, the marginal probability of experimental treatment 1 being declared effective was 0.025 whereas the probability of experimental treatment 1 being declared effective given treatment 2 had been was slightly lower at 0.023. Under the closed platform framework, the marginal probability of experimental treatment 1 being declared effective was 0.026; however, the conditional probability of experimental treatment 1 being declared effective given experimental treatment 2 had been declared effective was approximately seven times higher at 0.184. A Bonferroni correction improved the conditional error rate to 0.145; however, this remained substantially higher compared to the independent two-arm framework. The open platform framework had lower conditional error rates compared to the closed platform with conditional probabilities of a type I error for experimental arm 1, assuming no interim analyses, being 0.121 and 0.089 for the unadjusted and Bonferroni adjusted scenarios respectively. However, these rates were still high relative to the unadjusted, independent, two-arm trial framework. Marginal and conditional error rates for experimental treatment 2 were equivalent to that of experimental treatment 1 across all three frameworks. The marginal probability of both treatments being declared effective simultaneously was at most 0.005 and 0.003 under the closed and open platform frameworks respectively and 0.001 for the independent two-arm framework. As the number of interim analyses increased, marginal and conditional type I error rates generally decreased (Table 3.1).

3.3.2 Type I Error Rates and Power under Alternative Scenarios

Table 3.2 shows simulation results for each design framework and number in interim analyses under an alternative scenario where $\pi_0 = \pi_1 = 0.5$ and $\pi_2 = 0.35$. When experimental treatment 2 is truly effective with a lower failure rate of 0.35, the probability of a type I error in which experimental treatment 1 is declared effective is equivalent across all three frameworks for all number of interim analyses. Conditional error rates of treatment 1 being declared effective given treatment 2 is declared effective remain higher in the closed and open frameworks compared to the independent two-arm trial framework; however, the inflation is much smaller than that observed under the null scenario. For scenarios with no interim analyses, the conditional type I error rate is 0.026 in the independent two-arm framework and increases by ~ 0.004 in the closed and open frameworks (0.030 and 0.029 respectively). A Bonferroni correction lowers the conditional type I error rate to 0.015 under both the closed and open frameworks.

Conditional type I error rates for other scenarios where $\pi_0 = \pi_1 = 0.5$ and π_2 varies from 0.3 to 0.7 are shown in Table 3.3. As experimental treatment 2 becomes more effective, the conditional error rate of experimental treatment 1 being declared effective given treatment 2 is declared effective decreases to the unconditional probability of experimental treatment 1 being declared effective. As experimental treatment 2 becomes worse than the control, the conditional type I error rate of experimental arm I also decreases to the unconditional type I error rate.

Across all three frameworks, when experimental treatment 2 is truly effective, power is equivalent. In the alternative scenario where $\pi_0 = \pi_1 = 0.5$ and $\pi_2 = 0.35$, with no Bonferroni correction or interim analyses, power is approximately 86% in all three frameworks and a Bonferroni correction reduced power to ~78% (Table 3.2). As π_2 increases and decreases, power also remains similar across the three frameworks (Table 3.3).

3.3.3 Sample Size

By design, the maximum sample sizes of the closed and open platform designs are 25% and 19% smaller than the maximum sample size of the independent, two-arm trial framework. When interim analyses are considered, under the null scenario ($\pi_0 = \pi_1 = \pi_2 = 0.5$), sample size for the closed and open platform frameworks remain smaller on average compared to the independent, two-arm trial framework (Table 3.1). For one, two and three interim analyses the closed design yields average sample sizes ~22%, ~21%, and ~20% smaller than the independent, two-arm trial framework. Comparing the open and closed platform framework under the null scenario, for zero, one, two, and three interim analyses, the open platform gives average sample sizes ~8%,~12%, ~15%, and ~17% larger than the closed platform framework. Table 3.4 shows additional simulation results for each design framework and number in interim analyses under an alternative scenario where $\pi_0 = \pi_1 = 0.5$ and $\pi_2 = 0.35$. Under this alternative scenario, average sample sizes remain higher in the independent, two-arm trial framework compared to the closed and open platform frameworks.

3.3.4 Treatment Failure Rates

Under an alternative scenario where $\pi_0 = \pi_1 = 0.5$ and $\pi_2 = 0.35$, the overall treatment failure rate of patients across each framework was examined (Table 3.4). In the closed platform framework, the failure rate across all trial participants was lower on average compared to two comparable independent, two-arm trials. For all number of interim analyses considered, the treatment failure rate in the closed platform framework was 2% less compared to the independent, two-arm trial framework. The open platform also had lower failure rates on average compared to the independent, two-arm trial framework; however failure rates were slightly higher than those observed under the closed framework.

In other alternative scenarios where $\pi_0 = \pi_1 = 0.5$ and π_2 varied from 0.3 to 0.7, similar patterns where overall treatment failure rates were lower for the platform frameworks compared to the independent, two-arm trial framework were observed among scenarios when treatment 2 was superior to control (π_2 <0.5). However, when experimental treatment 2 was worse than control, (π_2 >0.5), the overall failure rate was higher for platform trial subjects compared to subjects enrolled in independent, two-arm studies. Average treatment failure rates and total sample sizes across these scenarios are provided in Table 3.5 for analyses with no multiple comparison adjustment and Table 3.6 for Bonferroni adjusted analyses.

Table 3.1 Simulation Results under the Null Scenario

						Either Effective		E1 Effective		E2 Effective		Both Effective		P(E1 Effective|E2)	P(E2 Effective|E1)	Total N	
Trial Type	IA	MC Adjustment	π_0	π_1	π_2	n	p	n	p	n	p	n	p	p	p	Mean	SD
Closed Platform	0	None	0.5	0.5	0.5	4636	0.046	2573	0.026	2527	0.025	464	0.005	0.184	0.180	600.0	0.0
		Bonferroni	0.5	0.5	0.5	2235	0.022	1211	0.012	1197	0.012	173	0.002	0.145	0.143	600.0	0.0
	1	None	0.5	0.5	0.5	4250	0.043	2333	0.023	2301	0.023	384	0.004	0.167	0.165	377.3	111.2
		Bonferroni	0.5	0.5	0.5	2044	0.020	1102	0.011	1085	0.011	143	0.001	0.132	0.130	364.3	104.5
	2	None	0.5	0.5	0.5	3991	0.040	2170	0.022	2159	0.022	338	0.003	0.157	0.156	290.8	121.6
		Bonferroni	0.5	0.5	0.5	1867	0.019	993	0.010	1002	0.010	128	0.001	0.128	0.129	271.6	109.3
	3	None	0.5	0.5	0.5	3258	0.033	1780	0.018	1744	0.017	266	0.003	0.153	0.149	244.2	124.2
		Bonferroni	0.5	0.5	0.5	1890	0.019	1015	0.010	998	0.010	123	0.001	0.123	0.121	237.5	116.1
Open Platform	0	None	0.5	0.5	0.5	4810	0.048	2548	0.025	2574	0.026	312	0.003	0.121	0.122	650.0	0.0
		Bonferroni	0.5	0.5	0.5	2362	0.024	1209	0.012	1265	0.013	112	0.001	0.089	0.093	650.0	0.0
	1	None	0.5	0.5	0.5	4386	0.044	2332	0.023	2312	0.023	258	0.003	0.112	0.111	422.2	101.0
		Bonferroni	0.5	0.5	0.5	2160	0.022	1112	0.011	1138	0.011	90	0.001	0.079	0.081	409.6	94.5
	2	None	0.5	0.5	0.5	4107	0.041	2177	0.022	2157	0.022	227	0.002	0.105	0.104	333.8	108.2
		Bonferroni	0.5	0.5	0.5	1910	0.019	973	0.010	1009	0.010	72	0.001	0.071	0.074	315.8	96.9
	3	None	0.5	0.5	0.5	3407	0.034	1770	0.018	1807	0.018	170	0.002	0.094	0.096	285.8	109.4
		Bonferroni	0.5	0.5	0.5	1934	0.019	992	0.010	1016	0.010	74	0.001	0.073	0.075	278.8	101.8
Independent Two-Arm Trials	0	None	0.5	0.5	0.5	5086	0.051	2538	0.025	2609	0.026	61	0.001	0.023	0.024	800.0	0.0
		Bonferroni	0.5	0.5	0.5	2474	0.025	1234	0.012	1249	0.012	9	0.000	0.007	0.007	800.0	0.0
	1	None	0.5	0.5	0.5	4628	0.046	2298	0.023	2376	0.024	46	0.000	0.019	0.020	486.2	116.3
		Bonferroni	0.5	0.5	0.5	2293	0.023	1133	0.011	1166	0.012	6	0.000	0.005	0.005	471.0	108.1
	2	None	0.5	0.5	0.5	4317	0.043	2160	0.022	2199	0.022	42	0.000	0.019	0.019	368.2	125.4
		Bonferroni	0.5	0.5	0.5	2055	0.021	1034	0.010	1028	0.010	7	0.000	0.007	0.007	346.4	111.8
	3	None	0.5	0.5	0.5	3623	0.036	1810	0.018	1843	0.018	30	0.000	0.016	0.017	304.8	127.6
		Bonferroni	0.5	0.5	0.5	2062	0.021	1014	0.010	1055	0.011	7	0.000	0.007	0.007	297.5	119.1

Table 3.2 Simulation Results under an alternative scenario where $\pi_0 = \pi_1 = 0.5$ and $\pi_2 = 0.35$

						Either Effective		E1 Effective		E2 Effective		Both Effective		P(E1 Effective|E2)	P(E2 Effective|E1)
Trial Type	IA	MC Adjustment	π_0	π_1	π_2	n	p	n	p	n	p	n	p	p	p
Closed Platform	0	None	0.5	0.5	0.35	86064	0.861	2573	0.026	86054	0.861	2563	0.026	0.030	0.996
		Bonferroni	0.5	0.5	0.35	78416	0.784	1211	0.012	78409	0.784	1204	0.012	0.015	0.994
	1	None	0.5	0.5	0.35	82864	0.829	2333	0.023	82845	0.828	2314	0.023	0.028	0.992
		Bonferroni	0.5	0.5	0.35	75216	0.752	1102	0.011	75205	0.752	1091	0.011	0.015	0.990
	2	None	0.5	0.5	0.35	78696	0.787	2170	0.022	78648	0.786	2122	0.021	0.027	0.978
		Bonferroni	0.5	0.5	0.35	71282	0.713	993	0.010	71257	0.713	968	0.010	0.014	0.975
	3	None	0.5	0.5	0.35	74586	0.746	1780	0.018	74531	0.745	1725	0.017	0.023	0.969
		Bonferroni	0.5	0.5	0.35	68772	0.688	1015	0.010	68743	0.687	986	0.010	0.014	0.971
Open Platform	0	None	0.5	0.5	0.35	85960	0.860	2548	0.025	85906	0.859	2494	0.025	0.029	0.979
		Bonferroni	0.5	0.5	0.35	78532	0.785	1209	0.012	78493	0.785	1170	0.012	0.015	0.968
	1	None	0.5	0.5	0.35	82659	0.827	2332	0.023	82596	0.826	2269	0.023	0.027	0.973
		Bonferroni	0.5	0.5	0.35	75230	0.752	1112	0.011	75184	0.752	1066	0.011	0.014	0.959
	2	None	0.5	0.5	0.35	78652	0.787	2177	0.022	78550	0.786	2075	0.021	0.026	0.953
		Bonferroni	0.5	0.5	0.35	71399	0.714	973	0.010	71337	0.713	911	0.009	0.013	0.936
	3	None	0.5	0.5	0.35	74652	0.747	1770	0.018	74544	0.745	1662	0.017	0.022	0.939
		Bonferroni	0.5	0.5	0.35	68966	0.690	992	0.010	68897	0.689	923	0.009	0.013	0.930
Independent Two-Arm Trials	0	None	0.5	0.5	0.35	86141	0.861	2538	0.025	85798	0.858	2195	0.022	0.026	0.865
		Bonferroni	0.5	0.5	0.35	78628	0.786	1234	0.012	78363	0.784	969	0.010	0.012	0.785
	1	None	0.5	0.5	0.35	82980	0.830	2298	0.023	82583	0.826	1901	0.019	0.023	0.827
		Bonferroni	0.5	0.5	0.35	75454	0.755	1133	0.011	75168	0.752	847	0.008	0.011	0.748
	2	None	0.5	0.5	0.35	78929	0.789	2160	0.022	78465	0.785	1696	0.017	0.022	0.785
		Bonferroni	0.5	0.5	0.35	71522	0.715	1034	0.010	71231	0.712	743	0.007	0.010	0.719
	3	None	0.5	0.5	0.35	75012	0.750	1810	0.018	74524	0.745	1322	0.013	0.018	0.730
		Bonferroni	0.5	0.5	0.35	69088	0.691	1014	0.010	68765	0.688	691	0.007	0.010	0.681

Table 3.3 Simulation Results for varying levels of π_2

		MC				E1 Effective		E2 Effective		Both Effective		Either Effective		P(E1 Effective|E2)
Trial Type	IA	Adjustment	π_0	π_1	π_2	n	p	n	p	n	p	n	p	p
Closed Platform	0	None	0.5	0.5	0.3	2573	0.026	98425	0.984	2573	0.026	98425	0.984	0.026
		Bonferroni	0.5	0.5	0.3	1211	0.012	96981	0.970	1210	0.012	96982	0.970	0.012
		None	0.5	0.5	0.35	2573	0.026	86054	0.861	2563	0.026	86064	0.861	0.030
		Bonferroni	0.5	0.5	0.35	1211	0.012	78409	0.784	1204	0.012	78416	0.784	0.015
		None	0.5	0.5	0.4	2573	0.026	51890	0.519	2368	0.024	52095	0.521	0.046
		Bonferroni	0.5	0.5	0.4	1211	0.012	40136	0.401	1063	0.011	40284	0.403	0.026
		None	0.5	0.5	0.5	2573	0.026	2527	0.025	464	0.005	4636	0.046	0.184
		Bonferroni	0.5	0.5	0.5	1211	0.012	1197	0.012	173	0.002	2235	0.022	0.145
		None	0.5	0.5	0.6	2573	0.026	1	0.000	1	0.000	2573	0.026	1.000
		Bonferroni	0.5	0.5	0.6	1211	0.012	1	0.000	1	0.000	1211	0.012	1.000
		None	0.5	0.5	0.7	2573	0.026	0	0.000	0	0.000	2573	0.026	-
		Bonferroni	0.5	0.5	0.7	1211	0.012	0	0.000	0	0.000	1211	0.012	-
Open Platform	0	None	0.5	0.5	0.3	2548	0.025	98362	0.984	2548	0.025	98362	0.984	0.026
		Bonferroni	0.5	0.5	0.3	1209	0.012	96956	0.970	1208	0.012	96957	0.970	0.012
		None	0.5	0.5	0.35	2548	0.025	85906	0.859	2494	0.025	85960	0.860	0.029
		Bonferroni	0.5	0.5	0.35	1209	0.012	78493	0.785	1170	0.012	78532	0.785	0.015
		None	0.5	0.5	0.4	2548	0.025	52387	0.524	2157	0.022	52778	0.528	0.041
		Bonferroni	0.5	0.5	0.4	1209	0.012	40428	0.404	961	0.010	40676	0.407	0.024
		None	0.5	0.5	0.5	2548	0.025	2574	0.026	312	0.003	4810	0.048	0.121
		Bonferroni	0.5	0.5	0.5	1209	0.012	1265	0.013	112	0.001	2362	0.024	0.089
		None	0.5	0.5	0.6	2548	0.025	1	0.000	1	0.000	2548	0.025	1.000
		Bonferroni	0.5	0.5	0.6	1209	0.012	1	0.000	0	0.000	1210	0.012	0.000
		None	0.5	0.5	0.7	2548	0.025	0	0.000	0	0.000	2548	0.025	-
		Bonferroni	0.5	0.5	0.7	1209	0.012	0	0.000	0	0.000	1209	0.012	-
Independent Two-Arm Trials	0	None	0.5	0.5	0.3	2538	0.025	98411	0.984	2489	0.025	98460	0.985	0.025
		Bonferroni	0.5	0.5	0.3	1234	0.012	97078	0.971	1194	0.012	97118	0.971	0.012
		None	0.5	0.5	0.35	2538	0.025	85798	0.858	2195	0.022	86141	0.861	0.026
		Bonferroni	0.5	0.5	0.35	1234	0.012	78363	0.784	969	0.010	78628	0.786	0.012
		None	0.5	0.5	0.4	2538	0.025	51871	0.519	1339	0.013	53070	0.531	0.026
		Bonferroni	0.5	0.5	0.4	1234	0.012	40095	0.401	486	0.005	40843	0.408	0.012
		None	0.5	0.5	0.5	2538	0.025	2609	0.026	61	0.001	5086	0.051	0.023
		Bonferroni	0.5	0.5	0.5	1234	0.012	1249	0.012	9	0.000	2474	0.025	0.007
		None	0.5	0.5	0.6	2538	0.025	1	0.000	0	0.000	2539	0.025	0.000
		Bonferroni	0.5	0.5	0.6	1234	0.012	0	0.000	0	0.000	1234	0.012	-
		None	0.5	0.5	0.7	2538	0.025	0	0.000	0	0.000	2538	0.025	-
		Bonferroni	0.5	0.5	0.7	1234	0.012	0	0.000	0	0.000	1234	0.012	-

Table 3.4 Average treatment failure rates and total sample size under an alternative scenario where $\pi_0 = \pi_1 = 0.5$ and $\pi_2 = 0.35$

						Overall Failure Rate		Total N	
Trial Type	IA	MC Adjustment	π_0	π_1	π_2	Mean	SD	Mean	SD
Closed Platform	0	None	0.5	0.5	0.35	0.450	0.020	600.0	0.0
		Bonferroni	0.5	0.5	0.35	0.450	0.020	600.0	0.0
	1	None	0.5	0.5	0.35	0.446	0.025	472.0	91.4
		Bonferroni	0.5	0.5	0.35	0.444	0.024	479.3	86.4
	2	None	0.5	0.5	0.35	0.444	0.027	396.6	103.9
		Bonferroni	0.5	0.5	0.35	0.442	0.027	400.2	106.1
	3	None	0.5	0.5	0.35	0.443	0.030	351.0	114.3
		Bonferroni	0.5	0.5	0.35	0.442	0.029	362.6	115.0
Open Platform	0	None	0.5	0.5	0.35	0.454	0.019	650.0	0.0
		Bonferroni	0.5	0.5	0.35	0.454	0.019	650.0	0.0
	1	None	0.5	0.5	0.35	0.451	0.023	515.8	96.1
		Bonferroni	0.5	0.5	0.35	0.450	0.023	526.1	89.3
	2	None	0.5	0.5	0.35	0.450	0.025	441.7	100.9
		Bonferroni	0.5	0.5	0.35	0.449	0.025	447.4	103.6
	3	None	0.5	0.5	0.35	0.450	0.028	395.8	109.6
		Bonferroni	0.5	0.5	0.35	0.449	0.027	409.5	110.9
Independent Two-Arm Trials	0	None	0.5	0.5	0.35	0.462	0.017	800.0	0.0
		Bonferroni	0.5	0.5	0.35	0.462	0.017	800.0	0.0
	1	None	0.5	0.5	0.35	0.456	0.022	583.7	122.9
		Bonferroni	0.5	0.5	0.35	0.455	0.022	591.5	112.4
	2	None	0.5	0.5	0.35	0.454	0.025	481.7	126.5
		Bonferroni	0.5	0.5	0.35	0.452	0.025	482.5	123.7
	3	None	0.5	0.5	0.35	0.452	0.028	419.8	133.5
		Bonferroni	0.5	0.5	0.35	0.451	0.027	432.3	131.9

Table 3.5 Average treatment failure rates and total sample size for varying levels of π_2 – No multiple comparison adjustment

					Closed Platform				Open Platform				Independent, Two-Arm Trials			
	MC				Overall Response Rate		Total N		Overall Response Rate		Total N		Overall Response Rate		Total N	
IA	Adjustment	p0	p1	p2	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	None	0.5	0.5	0.3	0.433	0.020	600.0	0.0	0.438	0.019	650.0	0.0	0.450	0.017	800.0	0.0
1	None	0.5	0.5	0.3	0.432	0.028	439.2	100.0	0.438	0.026	479.3	103.0	0.445	0.024	544.7	129.4
2	None	0.5	0.5	0.3	0.426	0.029	376.1	83.7	0.435	0.026	419.2	82.4	0.439	0.026	458.0	113.0
3	None	0.5	0.5	0.3	0.425	0.031	332.7	90.9	0.435	0.028	375.6	87.6	0.437	0.029	399.3	115.8
0	None	0.5	0.5	0.35	0.450	0.020	600.0	0.0	0.454	0.019	650.0	0.0	0.462	0.017	800.0	0.0
1	None	0.5	0.5	0.35	0.446	0.025	472.0	91.4	0.451	0.023	515.8	96.1	0.456	0.022	583.7	122.9
2	None	0.5	0.5	0.35	0.444	0.027	396.6	103.9	0.450	0.025	441.7	100.9	0.454	0.025	481.7	126.5
3	None	0.5	0.5	0.35	0.443	0.030	351.0	114.3	0.450	0.028	395.8	109.6	0.452	0.028	419.8	133.5
0	None	0.5	0.5	0.4	0.467	0.020	600.0	0.0	0.469	0.019	650.0	0.0	0.475	0.018	800.0	0.0
1	None	0.5	0.5	0.4	0.464	0.024	461.9	105.2	0.468	0.023	510.1	103.3	0.471	0.022	577.8	124.9
2	None	0.5	0.5	0.4	0.463	0.028	383.5	129.0	0.468	0.026	430.8	122.5	0.470	0.025	469.6	142.3
3	None	0.5	0.5	0.4	0.463	0.031	336.2	138.9	0.468	0.028	382.5	131.1	0.469	0.028	405.5	150.5
0	None	0.5	0.5	0.5	0.500	0.020	600.0	0.0	0.500	0.020	650.0	0.0	0.500	0.018	800.0	0.0
1	None	0.5	0.5	0.5	0.500	0.027	377.3	111.2	0.500	0.025	422.2	101.0	0.500	0.023	486.2	116.3
2	None	0.5	0.5	0.5	0.500	0.031	290.8	121.6	0.500	0.029	333.8	108.2	0.500	0.027	368.2	125.4
3	None	0.5	0.5	0.5	0.500	0.035	244.2	124.2	0.500	0.031	285.8	109.4	0.500	0.031	304.8	127.6
0	None	0.5	0.5	0.6	0.533	0.020	600.0	0.0	0.531	0.019	650.0	0.0	0.525	0.018	800.0	0.0
1	None	0.5	0.5	0.6	0.530	0.028	345.2	85.4	0.527	0.026	384.9	66.1	0.523	0.024	446.0	85.6
2	None	0.5	0.5	0.6	0.529	0.033	254.9	93.2	0.525	0.030	292.6	73.5	0.523	0.029	323.6	92.8
3	None	0.5	0.5	0.6	0.529	0.038	208.0	95.7	0.523	0.033	244.0	75.6	0.522	0.033	260.0	95.1
0	None	0.5	0.5	0.7	0.566	0.020	600.0	0.0	0.562	0.019	650.0	0.0	0.550	0.017	800.0	0.0
1	None	0.5	0.5	0.7	0.561	0.029	343.4	82.5	0.553	0.026	382.5	61.9	0.546	0.025	443.2	82.3
2	None	0.5	0.5	0.7	0.558	0.035	251.5	89.0	0.549	0.031	288.0	68.9	0.545	0.030	318.3	88.7
3	None	0.5	0.5	0.7	0.556	0.040	203.2	90.9	0.545	0.034	237.8	71.1	0.543	0.034	252.9	90.3

Table 3.6 Average treatment failure rates and total sample size for varying levels of π_2 – Bonferroni adjustment

					Closed Platform				Open Platform				Independent, Two-Arm Trials			
	MC				Overall Response Rate		Total N		Overall Response Rate		Total N		Overall Response Rate		Total N	
IA	Adjustment	p0	p1	p2	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	Bonferroni	0.5	0.5	0.3	0.433	0.020	600.0	0.0	0.438	0.019	650.0	0.0	0.450	0.017	800.0	0.0
1	Bonferroni	0.5	0.5	0.3	0.428	0.027	460.2	91.9	0.435	0.025	503.4	97.0	0.442	0.023	567.1	121.9
2	Bonferroni	0.5	0.5	0.3	0.424	0.028	384.1	82.2	0.433	0.025	430.0	81.6	0.437	0.025	465.1	106.8
3	Bonferroni	0.5	0.5	0.3	0.423	0.030	349.3	89.5	0.433	0.027	394.2	88.3	0.435	0.028	416.9	113.9
0	Bonferroni	0.5	0.5	0.35	0.450	0.020	600.0	0.0	0.454	0.019	650.0	0.0	0.462	0.017	800.0	0.0
1	Bonferroni	0.5	0.5	0.35	0.444	0.024	479.3	86.4	0.450	0.023	526.1	89.3	0.455	0.022	591.5	112.4
2	Bonferroni	0.5	0.5	0.35	0.442	0.027	400.2	106.1	0.449	0.025	447.4	103.6	0.452	0.025	482.5	123.7
3	Bonferroni	0.5	0.5	0.35	0.442	0.029	362.6	115.0	0.449	0.027	409.5	110.9	0.451	0.027	432.3	131.9
0	Bonferroni	0.5	0.5	0.4	0.467	0.020	600.0	0.0	0.469	0.019	650.0	0.0	0.475	0.018	800.0	0.0
1	Bonferroni	0.5	0.5	0.4	0.463	0.024	454.0	107.4	0.467	0.023	503.9	104.2	0.471	0.022	568.3	121.7
2	Bonferroni	0.5	0.5	0.4	0.463	0.028	372.3	130.7	0.468	0.026	420.8	124.9	0.470	0.026	454.5	140.3
3	Bonferroni	0.5	0.5	0.4	0.462	0.031	336.4	139.2	0.468	0.028	383.8	131.8	0.469	0.028	405.6	148.7
0	Bonferroni	0.5	0.5	0.5	0.500	0.020	600.0	0.0	0.500	0.020	650.0	0.0	0.500	0.018	800.0	0.0
1	Bonferroni	0.5	0.5	0.5	0.500	0.027	364.3	104.5	0.500	0.025	409.6	94.5	0.500	0.024	471.0	108.1
2	Bonferroni	0.5	0.5	0.5	0.500	0.032	271.6	109.3	0.500	0.029	315.8	96.9	0.500	0.028	346.4	111.8
3	Bonferroni	0.5	0.5	0.5	0.500	0.036	237.5	116.1	0.500	0.032	278.8	101.8	0.500	0.031	297.5	119.1
0	Bonferroni	0.5	0.5	0.6	0.533	0.020	600.0	0.0	0.531	0.019	650.0	0.0	0.525	0.018	800.0	0.0
1	Bonferroni	0.5	0.5	0.6	0.531	0.028	337.2	79.1	0.527	0.026	378.4	60.8	0.523	0.024	437.4	78.9
2	Bonferroni	0.5	0.5	0.6	0.530	0.034	243.0	82.9	0.525	0.030	282.8	64.5	0.523	0.029	311.3	82.5
3	Bonferroni	0.5	0.5	0.6	0.529	0.038	204.1	89.5	0.523	0.033	240.0	69.9	0.522	0.033	256.0	88.9
0	Bonferroni	0.5	0.5	0.7	0.566	0.020	600.0	0.0	0.562	0.019	650.0	0.0	0.550	0.017	800.0	0.0
1	Bonferroni	0.5	0.5	0.7	0.562	0.029	335.9	76.8	0.554	0.026	376.8	57.5	0.547	0.025	435.5	76.4
2	Bonferroni	0.5	0.5	0.7	0.560	0.035	240.5	79.2	0.550	0.031	279.3	60.3	0.546	0.030	307.3	79.1
3	Bonferroni	0.5	0.5	0.7	0.556	0.040	199.6	84.8	0.545	0.034	234.1	65.3	0.544	0.034	249.2	84.3

3.4 Discussion

Existing arguments that call multiple testing adjustment in platform trials into question are largely philosophical. Freidlin et al. state that when multiple experimental treatments are compared to a shared control for the purposes of efficiency alone and the results of one experimental treatment have no impact on the interpretation of another, multiple testing adjustments are not needed since the underlying clinical question of each comparison is independent.[29] Other researchers agree that a correction for multiple comparisons is needed only if the treatments being tested are related. Examples of related treatments would be different doses of the same drug. If the treatments are unrelated, each comparison of an experimental treatment versus control can be viewed as an independent trial[28, 29]. While these arguments have merit, the shared control group makes the comparisons of each experimental treatment versus the control decidedly dependent.

This simulation study sought to quantitatively evaluate the impact of adjusting and not adjusting for multiple testing on FWER and conditional type I error rates in platform trials. The simulation results demonstrate that, on average, platform trial designs yield smaller sample sizes, better trial-wide treatment response rates when at least one treatment is more effective than the control, and slightly lower FWER when compared to a set of comparable two-arm trials evaluating the same set of treatments. These results are consistent with previously reported simulation studies that explored the operating characteristics of various platform designs[9, 19, 21]. Compared to closed platform studies where all treatments open to enrollment simultaneously, open platform designs that allow for additional treatments to enter after initial study launch have higher average sample sizes and, when at least one treatment is effective, lower trial-wide treatment response rates due to the higher number of patients allocated to the control group.

However, when compared to a set of comparable two-arm trials, the open framework still demonstrates lower average sample sizes as well as higher treatment response rates when at least one treatment is more effective than the control.

Conditional type I error rates under the null scenario for one experimental treatment being declared effective given the other had been declared effective were substantially higher under the platform frameworks compared to the independent, two-arm trial framework. Under scenarios with no interim analyses, conditional error rates for the closed and open platform frameworks with no multiple comparison adjustment were approximately 8 and 5 times higher respectively, when compared to the independent, two-arm trial framework. When a Bonferroni correction was applied, the conditional error rates improved to be approximately 6 and 4 times higher for the closed and open respectively when compared to a set of two-arm trials with no multiple comparison adjustment. Under an alternative scenario where $\pi_0 = \pi_1 = 0.5$ and $\pi_2 = 0.35$, when no interim analyses are used, power for testing the null hypothesis that $\pi_2 \geq \pi_0$ is ~86% across all three frameworks when no Bonferroni correction is applied and decreases to ~78% when the correction is applied. Although the correction does protect FWER, it does little to the conditional type I error rate and results in a substantial loss of power. Further, under the null scenario the highest marginal probability observed for both treatments being declared effective, in the case of the closed platform with no interim analyses, was only 0.5%.

In the open platform framework where the second experimental treatment is introduced later in the trial, conditional type I error rates are ~35% lower compared to the closed framework due to the fact that each experimental treatment is compared only to concurrently randomized controls making test statistics less correlated than they are under the closed framework. Similarly, under the null scenario, as the number of interim analyses increase, the conditional error rates decrease due to arms exiting the study early for futility resulting in less overlap between the control patients used for each experimental versus control comparison. Under alternative scenarios where $\pi_0 = \pi_1 = 0.5$ and π_2 varies, as π_2 decreases, the conditional probability of experimental treatment 1 being declared effective given treatment 2 is declared effective decreases downward to the rate observed in the independent, two-arm scenario as treatment 2 is declared effective more often.

Given these observations, multiple comparison adjustments are likely unnecessary in platform trials of unrelated treatments. While a Bonferroni correction will control FWER, it does not substantially impact conditional type I error rates. Furthermore, if flexible features are implemented, such as allowing arms to enter at varying time points under an open framework or allowing arms to exit early for efficacy or futility, conditional type I error rates decrease. Since FWER is comparable between the unadjusted platform and equivalent two-arm designs, the substantial drop in power yielded by adjusting for multiple testing is likely not worth the modest improvement in conditional type I error. As such, the results of this simulation study support that multiple testing does not need to be adjusted for in a platform trial if the treatments being evaluated are unrelated. For related treatments, such as different doses of the same drug, multiple testing should be accounted for as the drug is getting multiple chances to be declared effective.

Platform trials offer an efficient alternative to conducting multiple two-arm studies. However, stakeholders should be aware that, as with all efficient design, tradeoffs exist. One of which is that conditional error rates are higher than the nominal significance level. This simulation study shows that conditional error rates are elevated regardless of whether analyses are adjusted for multiple testing. When developing a platform design, simulation studies should be conducted to ensure study stakeholders understand this as well as other potential tradeoffs relative to conducting independent two-arm studies.

CHAPTER 4

Integrating Non-concurrent Controls in Analyses of Open Platform Trials

4.1 Introduction

When launching a trial with multiple experimental treatment arms, logistic or regulatory issues can result in a bottleneck, delaying enrollment until all experimental treatments are ready to enroll simultaneously. Open platform designs circumvent this issue by allowing treatment arms to enter and exit the platform at varying time points. Various open designs have been discussed previously[9, 19-21, 26]. Under these designs, a control arm opens to enrollment at the beginning of the trial and remains open as long as any experimental treatment is enrolling. For experimental treatment arms introduced to the platform after the trial's initial launch, whether to incorporate all available control data in analyses or use only controls randomized after the introduction of the experimental arm is not standardized. Using all available controls can increase power and yield more precise estimates; however, drift in population parameters over time can yield biased estimates and impact type I error rates.

To avoid bias and potential type I error inflation, the rolling arm platform design proposed by Ventz et al. recommends the use of only concurrently randomized controls in the analyses of each experimental arm[21]. In the ongoing STAMPEDE and the recent CTSN Neuroprotection platform trials, only concurrent controls are incorporated in analyses of experimental arms[5, 33]. Open platform design proposals by Hobbs et al. and Yuan et al. pool all controls in analyses of experimental arms under the stipulation that drift is not a concern.[19, 34] If drift is a concern, both propose the use of concurrent controls only or the use controls that were enrolled within a pre-specified time interval before the experimental arm was introduced. In contrast to a simple pooled approach, Berry et al.'s platform design proposal incorporates all controls and accounts for possible drift in outcome rates by using a model with month as a covariate[26]. A similar approach was implemented in the I-SPY2 trial[9]. However, this approach has been criticized since including month as a covariate could be inadequate if the model is misspecified.[27] Recently Kaizer et al. introduced an open platform design aimed at finding an effective Ebola treatment quickly that leverages multi-source exchangeability models (MEMs) to incorporate non-concurrent controls in analyses. Experimental treatments are assessed sequentially and, through an adaptive randomization procedure, the proportion of patients allocated to the current experimental treatment increases if outcome rates in past control data are similar to concurrent control data[35].

With the exception of Kaizer et al. and Berry et al.'s approaches, previously proposed open platform designs have operated under two extremes. Either they ignore parameter drift and pool all available controls in analyses of experimental treatments, or they protect against drift by using concurrent controls only. A more desirable option would operate in between these extremes such that the information incorporated from controls randomized prior to the initiation of an experimental treatment is dependent on observed drift or heterogeneity. Methods developed to incorporate historical control data in the analyses of two-arm clinical trials could be applied to address this gap.

Historical control groups are commonly used in single-arm phase II studies evaluating a new experimental therapy relative to a performance goal determined by the historical data. However, several approaches are available in which historical control data can supplement concurrent control data in two-arm studies[36]. Leveraging historical control data can yield lower sample sizes and allow for higher patient allocation to the experimental arm[37]. FDA guidance encourages the use of historical controls in medical device trials and serious conditions[38, 39]. Pediatric, oncology, and rare disease therapeutic areas have increasingly been using these approaches[40, 41].

As is the case in an open platform framework, in a two-arm trial incorporating historical information, if historical and concurrent controls are homogenous, supplementing the concurrent data with the historical controls can increase power and decrease type I error rates and mean square error[40, 42, 43]. However, if the two groups are not homogenous, estimates can be biased and type I error inflated. To protect against these issues, in a seminal paper, Pocock introduced six requirements for incorporating historical controls into the analysis of a randomized clinical trial[44]. Pocock's six requirements are as follows:

1. The historical control group received the same treatment as the randomized control group
2. The historical control group was collected recently as part of a study with the same eligibility criteria
3. The treatment in the historical control group is evaluated in the same way as the randomized control group

4. Patient characteristics of the historical control group are similar to those of patients enrolled in the randomized trial
5. The historical control group comes from a study that was conducted by the same group of clinical investigators
6. There is no reason to expect different outcomes for the historical and randomized controls (i.e. patient accrual and selection is similar between the historical and randomized studies)

In an open platform, controls randomized prior to the initiation of a new experimental arm can be viewed as pseudo-historical controls. As all subjects are enrolled under the same protocol, with the same eligibility criteria, and at the same clinical centers, it is easy to argue that these pseudo-historical controls are likely to meet Pocock's requirements. This makes the application of methods developed to supplement the analyses of two-arm trials with historical control data well suited to open platforms.

Beyond naively pooling historical and current data, alternative methods such as down weighting the historical data, either with pre-specified weights in a frequentist analysis or with power priors in a Bayesian approach, can reduce the influence of the historical data in analysis[44, 45]. However, these methods depend on subjective choices of weights or priors to determine the amount of information borrowed from the historical data. Dynamic methods that allow the data to determine the degree of borrowing seem better suited to protect against drift in population parameters. The simplest dynamic method is a test-then-pool approach in which the historical control data are pooled with the current control only if one fails to reject the null hypothesis that the two are equivalent[36]. In contrast to this all or nothing approach, other dynamic methods allow for partial borrowing through Bayesian hierarchical modeling. In an extension of the static power prior, the weight used in a power prior can be treated as a random variable dependent on homogenity[45, 46]. In an alternative approach, Hobb's commensurate prior method considers the underlying population parameter of the historical data to be distinct from that of the current control and uses a commensurability parameter based on the closeness of the two parameters to determine the degree of borrowing[47, 48]. Methods based on meta-analysis have also been developed where multiple historical studies can be used to estimate between-study variability; however in the case of one historical sample, an informative prior distribution for the between-study variance is needed[49, 50].

Recently, Kaizer et al. introduced MEMs for integrating data from previous trials into the analysis of a current trial. In this approach, all possible combinations of exchangeability between the current data and the historical datasets are considered and an extension of Bayesian model averaging is used to estimate the parameter of interest[51]. As referenced above, this method has been applied in a platform design aimed at finding an effective Ebola treatment[35]. In this design, experimental treatments are considered one at a time, sequentially and compared to a control. For experimental treatments considered after the first experimental treatment arm is closed, a pre-specified number of patients are randomized with equal allocation between the current experimental arm and control. If the control data of the current segment is similar to that of previous segments, adaptive randomization is used to allocate more patients to the current experimental arm. A MEM is used to estimate the parameter of interest in the control group as well as the effective supplemental sample size (ESSS) gained by using the past control. If current control data is homogenous with past segments, the parameter estimate will have higher precision and therefore higher ESSS. The ESSS is used to determine the randomization allocation ratio with a higher ESSS corresponding to higher allocation to the experimental group.

MEMs have not yet been explored in an open platform framework where multiple experimental treatments could be declared effective and randomization is fixed.

This study aims to assess the operating characteristics of open platform designs that use methods developed for integrating historical control data into two-arm trials to incorporate non-concurrent controls in analyses of experimental treatments. Five approaches are explored: two test-then-pool approaches, a static power prior, a dynamic power prior, and a MEM approach. In a 2016 simulation study, Dejardin et al. assessed the use of various methods of incorporating a historical control group in a two-arm non-inferiority trial with a binary outcome. In this study, a dynamic power prior, commensurate prior, and meta-analytic based robust mixture prior were compared. All three approaches demonstrated comparable operating characteristics[52]. As such, of the three, the dynamic power prior was selected for this study over the commensurate and robust mixture priors as it is simpler to implement than the commensurate prior and does not require a static weight and informative prior distribution for between-study variance like the meta-analytic based approach. The five selected methods are compared to the two extremes of using concurrent controls only versus naively pooling the data across a variety of potential parameter drift scenarios.

4.2 Methods

4.2.1 General Trial Framework

An open platform design with two experimental treatments and a control was considered. Three arms were chosen for simplicity of interpretation, but results can be extrapolated to additional armed studies. The trial opens to enrollment with the first experimental treatment and control only. The second experimental treatment is subsequently added to the platform after a number of patients have been enrolled.

4.2.2 Outcome Assessment

Efficacy is evaluated with a binary outcome of treatment success or failure. Under the trial design, the efficacy of two experimental treatments is evaluated; however, the comparison of experimental treatment 2 versus control is the primary interest of this simulation study. This comparison is assessed using the one-sided null hypothesis that the failure rate of the second experimental treatment group is equivalent to or worse than the control (H_0: $\pi_2 \geq \pi_0$). The null hypothesis is rejected if the posterior probability that $\pi_2 < \pi_0$ is greater than 0.975. Multiple approaches to estimating this posterior probability were explored and are detailed below.

4.2.3 Concurrent Controls Only

In the most conservative approach, the second experimental treatment is compared to only concurrently randomized controls using Bayesian inference. The probability model is defined as follows:

For each treatment group k, where k=0 denotes the control group and k=1 or 2 denotes the experimental groups, x_k is the number of treatment failures observed in n_k patients. x_k follows a *Binomial(*n_k*,* π_k*)* distribution where π_k is the probability of treatment failure. The prior distribution of π_k is assumed to follow a *Beta(*α_k*,* β_k*)* distribution with known hyperparameters α_k and β_k. After x_k failures are observed in n_k patients, the posterior distribution of π_k is as follows:

$$\Pr(\pi_k | x_k) \propto \Pr(x_k | \pi_k)\Pr(\pi_k) \propto \pi_k^{\alpha_k + x_k - 1}(1 - \pi_k)^{\beta_k + n_k - x_k - 1}$$

so that Pr($\pi_k|x_k$) ~ $Beta(\alpha_k+x_k, \beta_k+n_k-x_k)$.[53] A non-informative uniform prior on the probability of failure ($\pi_k \sim Beta(1, 1)$) was used for all groups.

4.2.4 All Controls (Pooled)

In the least conservative approach, the second experimental treatment is compared to the pooled sample of all controls (i.e. concurrent controls and controls randomized prior to the initiation of experimental treatment 2). The probability model used to estimate the posterior distributions of π_2 and π_0 is identical to that of the analysis of concurrent controls only. The only difference is that n_0 and x_0 will have larger values that account for the sample size and number of observed failures in the non-concurrent controls.

4.2.5 Test-then-Pool

In the test-then-pool approach, the inclusion of non-concurrent controls in the analysis of experimental group 2 is dependent on failing to reject the null hypothesis that the probability of treatment failure is equivalent in the non-concurrent and concurrent controls (i.e. H_0: $\pi_{0_{nc}} = \pi_{0_c}$). The same probability model defined above is used to estimate the posterior distributions of $\pi_{0_{nc}}$ and π_{0_c}. The null hypothesis is tested using a Monte Carlo approach to draw random samples from the posterior distributions of $\pi_{0_{nc}}$ and $\pi_{0_c}$and estimate the posterior probability that $\pi_{0_{nc}} \neq \pi_{0_c}$. If the posterior probability that $\pi_{0_{nc}} > \pi_{0_c}$or $\pi_{0_{nc}} < \pi_{0_c}$is greater than *S*, the null is rejected and only concurrent controls are included in analysis. Otherwise, if there is insufficient evidence that $\pi_{0_{nc}}$ is not equal to π_{0_c}all controls are incorporated in the analysis. Two cut-offs for *S* were considered, 0.975 (test-then-pool 1) and 0.95 (test-then-pool 2).

4.2.6 Power Prior (Static)

A fixed power prior essentially down weights non-concurrent control data using a scalar, power parameter $\Theta \subset [0,1]$. The general form of the power prior is as follows:

$$\Pr\left(\pi_0 \middle| x_{0_{nc}}, \Theta\right) \propto \mathrm{L}\left(\pi_0 \middle| x_{0_{nc}}\right)^{\Theta} \Pr(\pi_0) = \Pr\left(x_{0_{nc}} \middle| \pi_0\right)^{\Theta} \Pr(\pi_0)$$

Where π_0 again represents the probability of a treatment failure in the control group, $x_{0_{nc}}$ is the number of failures observed in the $n_{0_{nc}}$ non-concurrent controls, and $\Pr(\pi_0)$ is the initial prior on π_0 (π_0~*Beta(*α_0*,* β_0*)*) as defined above. Using these, the power prior is:

$$\Pr(\pi_0|x_{0nc}, \Theta) \propto \pi_0^{\Theta x_{0nc}+\alpha_0-1}(1-\pi_0)^{\Theta(n_{0nc}-x_{0nc})+\beta_0-1}$$

so that $\Pr(\pi_0|x_{0nc}, \Theta)$ ~ *Beta(*$\Theta x_{0_{nc}} + \alpha_0$, $\Theta(n_{0_{nc}} - x_{0_{nc}}$*)* +$\beta_0$*).*

After all control data is observed, let x_{0c} be the number of failures observed in the n_{0c} concurrent controls. The posterior distribution of π_0 is then:

$$\Pr(\pi_0|x_{0c}, x_{0nc}, \Theta) \propto \Pr(x_{0c}|\pi_0)\Pr(x_{0nc}|\pi_0)^{\Theta} \Pr(\pi_0)$$

$$\propto \pi_0^{x_{0c}+\Theta x_{0nc}+\alpha_0-1}(1-\pi_0)^{(n_{0c}-x_{0c})+\Theta(n_{0nc}-x_{0nc})+\beta_0-1}$$

so that $\Pr(\pi_0|x_{0c}, x_{0nc}, \Theta)$ ~ *Beta(*$x_{0c} + \Theta x_{0_{nc}} + \alpha_0$, $(n_{0c} - x_{0c}) + \Theta(n_{0_{nc}} - x_{0_{nc}}$*)* +$\beta_0$*).* A non-informative uniform prior was used for the initial prior of *π₀* so that $\alpha_0 = \beta_0 = 1$. The power parameter Θ was set to 0.5 so that the influence of the non-concurrent controls was reduced by half.

4.2.7 Power Prior (Dynamic)

In a dynamic extension of the power prior, the power parameter is considered a random variable and is estimated based on the similarity of the non-concurrent and concurrent control data. Specified by Duan et al. the normalized power prior is as follows[46]:

$$\Pr\left(\pi_0, \theta \middle| x_{0_{nc}}\right) = \frac{1}{C(\theta)} L\left(\pi_0 \middle| x_{0_{nc}}\right)^{\theta} \Pr(\pi_0)\Pr(\theta)$$

Where $\Pr(\pi_0)$ is a *Beta*(α_0, β_0) prior on π_0, $Pr(\theta)$ is a *Beta*$(\alpha_\theta, \beta_\theta)$ prior on the power parameter, and $C(\theta)$ is a normalizing constant that ensures the resulting posterior is consistent with the likelihood principle[46]. The normalizing constant is:

$$\begin{aligned} C(\theta) &= \int L\left(\pi_0 \middle| x_{0_{nc}}\right)^{\theta} \Pr(\pi_0) d\pi_0 \\ &= \int {\pi_0}^{\theta x_{0_{nc}}} (1-\pi_0)^{\theta(n_{0_{nc}} - x_{0_{nc}})} {\pi_0}^{\alpha_0 - 1} (1-\pi_0)^{\beta_0 - 1} d\pi_0 \\ &= \int {\pi_0}^{\theta x_{0_{nc}} + \alpha_0 - 1} (1-\pi_0)^{\theta(n_{0_{nc}} - x_{0_{nc}}) + \beta_0 - 1} d\pi_0 \\ &= \beta(\theta x_{0_{nc}} + \alpha_0, \theta(n_{0_{nc}} - x_{0_{nc}}) + \beta_0) \end{aligned}$$

Where β(x,y) indicates the beta function. Expressed in full the normalized power prior is:

$$\begin{aligned} \Pr\left(\pi_0, \Theta \middle| x_{0_{nc}}\right) &\propto \frac{{\pi_0}^{\theta x_{0_{nc}}} (1-\pi_0)^{\theta(n_{0_{nc}} - x_{0_{nc}})} {\pi_0}^{\alpha_0 - 1} (1-\pi_0)^{\beta_0 - 1} \theta^{\alpha_\theta - 1} (1-\theta)^{\beta_\theta - 1}}{\beta(\theta x_{0_{nc}} + \alpha_0, \theta(n_{0_{nc}} - x_{0_{nc}}) + \beta_0)} \\ &\propto \frac{{\pi_0}^{\theta x_{0_{nc}} + \alpha_0 - 1} (1-\pi_0)^{\theta(n_{0_{nc}} - x_{0_{nc}}) + \beta_0 - 1} \theta^{\alpha_\theta - 1} (1-\theta)^{\beta_\theta - 1}}{\beta(\theta x_{0_{nc}} + \alpha_0, \theta(n_{0_{nc}} - x_{0_{nc}}) + \beta_0)} \end{aligned}$$

After all control data is observed, the joint posterior for π_0 and Θ is then:

$$\Pr\left(\pi_0, \Theta \middle| x_{0_{nc}}, x_{0_c}\right) \propto L\left(\pi_0 \middle| x_{0_c}\right) \Pr(\pi_0, \Theta | x_{0_{nc}})$$

$$\propto {\pi_0}^{x_{0_c}} (1-\pi_0)^{n_{0_c} - x_{0_c}} \frac{{\pi_0}^{\theta x_{0_{nc}} + \alpha_0 - 1} (1-\pi_0)^{\theta(n_{0_{nc}} - x_{0_{nc}}) + \beta_0 - 1} \theta^{\alpha_\theta - 1} (1-\theta)^{\beta_\theta - 1}}{\beta(\theta x_{0_{nc}} + \alpha_0, \theta(n_{0_{nc}} - x_{0_{nc}}) + \beta_0)}$$

$$\propto \frac{{\pi_0}^{x_{0_c} + \theta x_{0_{nc}} + \alpha_0 - 1} (1-\pi_0)^{n_{0_c} - x_{0_c} + \theta(n_{0_{nc}} - x_{0_{nc}}) + \beta_0 - 1} \theta^{\alpha_\theta - 1} (1-\theta)^{\beta_\theta - 1}}{\beta(\theta x_{0_{nc}} + \alpha_0, \theta(n_{0_{nc}} - x_{0_{nc}}) + \beta_0)}$$

Non-informative uniform priors (*Beta*(1,1)) were used for $\Pr(\pi_0)$ and $\Pr(\theta)$. A Markov chain Monte Carlo (MCMC) method was used to sample from the joint posterior distribution and estimate the posterior distribution of π_0.

4.2.8 Multisource Exchangeability Models (MEMs)

The MEM framework considers two possible combinations of exchangeability between the non-concurrent and concurrent control data datasets: either the data are exchangeable and pooled, or the data are non-exchangeable and only concurrent data is used. These two cases are represented by two models denoted Ω_1 and Ω_2 respectively. The posterior distribution of π_0 based on the concurrent and nonconcurrent data will be:

$$\Pr(\pi_0 | x_{0_{nc}}, x_{0_c}) = \sum_{i=1}^{2} w_i \Pr(\pi_0 | \Omega_\mathrm{i}, x_{0_{nc}}, x_{0_c})$$

Where w_i is a posterior weight, based on the data, given by

$$w_i = Pr(\Omega_\mathrm{i} | x_{0_{nc}}, x_{0_c}) = \frac{\Pr(x_{0_{nc}}, x_{0_c} | \Omega_\mathrm{i}) \Pr(\Omega_\mathrm{i})}{\sum_{j=1}^{2} \Pr(x_{0_{nc}}, x_{0_c} | \Omega_\mathrm{j}) \Pr(\Omega_\mathrm{j})}$$

such that $\sum_{i=1}^{2} \mathrm{w_i} = 1$. $\Pr(\Omega_\mathrm{i})$ is the prior probability that model i represents the true model and $\Pr(x_{0_{nc}}, x_{0_c} | \Omega_\mathrm{i})$ is the integrated marginal likelihood of model i. A *Beta*(α_0, β_0) prior on π_0 is assumed for both the concurrent and non-concurrent control data so the integrated marginal likelihood of each model is given by:

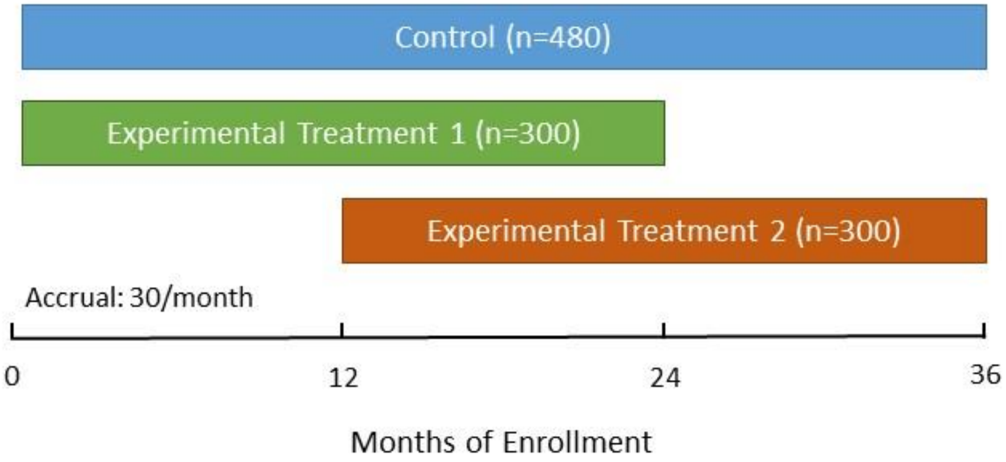
$$\Pr(x_{0_{nc}}, x_{0_c} | \Omega_\mathrm{i}) = \frac{\beta(x_{0_c} + \alpha_0 + s_i x_{0_{nc}}, (n_{0_c} - x_{0_c}) + \beta_0 + s_i(n_{0_{nc}} - x_{0_{nc}}))}{\beta(\alpha_0, \beta_0)} * \left(\frac{\beta(x_{0_{nc}} + \alpha_0, n_{0_{nc}} - x_{0_c} + \beta_0)}{\beta(\alpha_0, \beta_0)}\right)^{1 - s_i}$$

Where s_i is an indicator of whether the non-concurrent data is considered exchangeable or not. In model 1 (Ω_1), $s_1 = 1$ and in model 2 (Ω_2), $s_2 = 0$. This yields a *Beta*$\left(x_{0_c} + \alpha_0 + s_i x_{0_{nc}}, (n_{0_c} - x_{0_c}) + \beta_0 + s_i(n_{0_{nc}} - x_{0_{nc}})\right)$ posterior distribution for each MEM which makes the overall posterior distribution of π_0 a weighted mixture of two Beta distributions. The *calc.MEM.betabin* R function developed by Kaizer et al. was used to estimate the posterior distribution of π_0[35]. Equal prior weights for exchangeability ($\Pr(\Omega_\mathrm{i})$=0.5) were assigned and a uniform *Beta*(1,1) prior was assumed for π_0.

4.2.9 Trial Design

The trial opens with 1:1 randomization to the control arm and the first experimental treatment. The second experimental treatment arm opens to enrollment 12 months after trial launch and patients are randomized 1:1:1 across the three groups until month 24. At 24 months, the first experimental treatment arm closes and randomization continues 1:1 to the control and second experimental treatment arm through 36 months. The accrual per month was set to 30 patients so the total sample size is 300 in each experimental arm and 480 in the control group. (Figure 4.1) Randomization was blocked such that there was equal group allocation every month.

Figure 4.1 Timeline of Enrollment in Simulated Open Platform Trials


Control (n=480)
Experimental Treatment 1 (n=300)
Experimental Treatment 2 (n=300)
Accrual: 30/month
0
12
24
36
Months of Enrollment

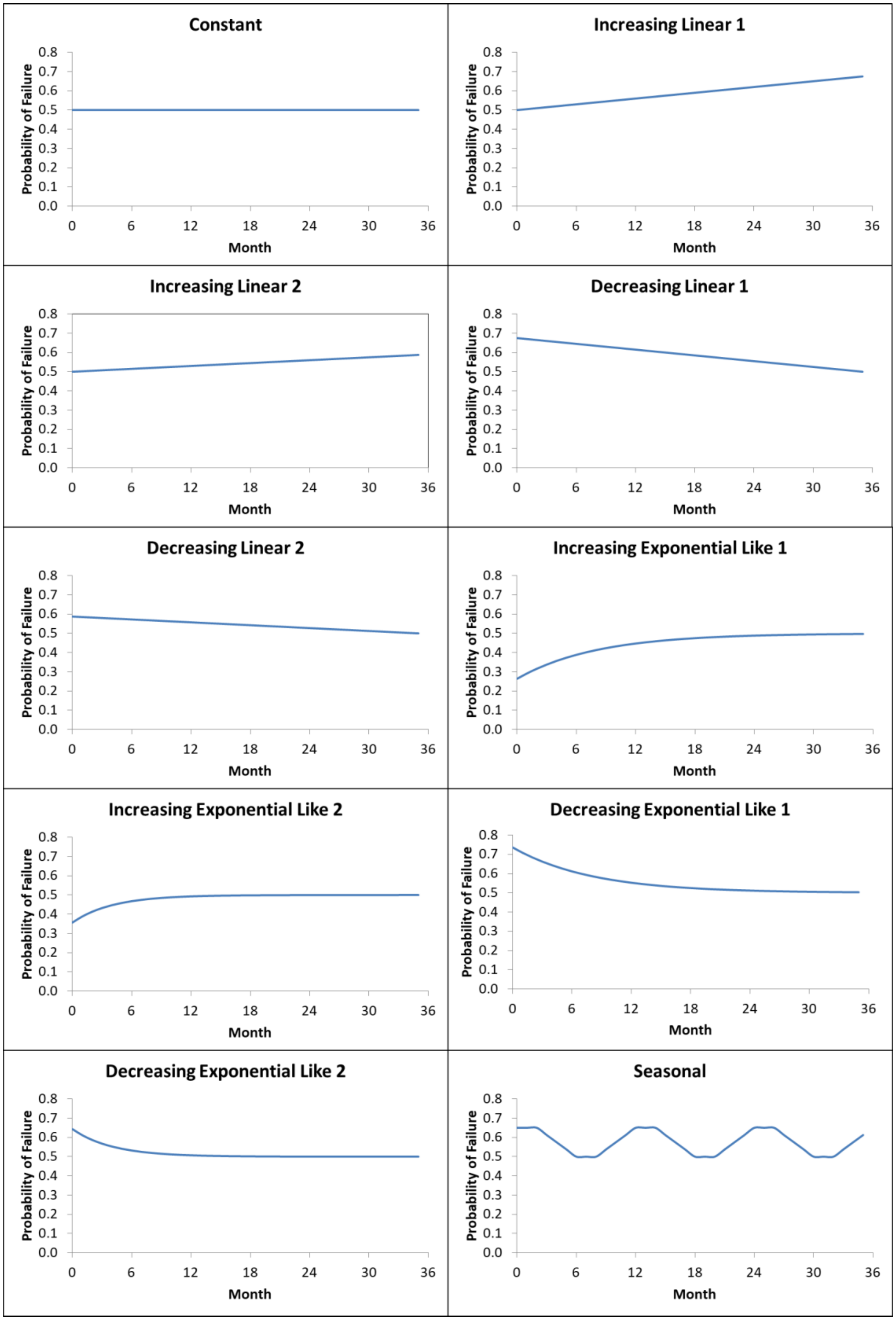
4.2.10 Scenarios

Multiple scenarios with varying temporal effects on failure rates in the control group were explored and are shown in figure 4.2. Some examples of these different scenarios are presented in the discussion. Under the null hypothesis, experimental treatment group 2's failure rate was equivalent to control. Two alternative scenarios were explored in which the failure rate of experimental treatment 2 was 0.8 and 0.75 times that of the control.

4.2.11 Simulation Procedures and Operating Characteristics

For each scenario, 10,000 trials were simulated. For a single trial, all methods were applied on the same set of "subjects" by using the same set of uniform variates and the same randomization scheme. If the uniform variate of a subject was smaller than the "true" failure rate of the subject's assigned group, the subject was deemed to have failed. Monte Carlo sampling was used to evaluate the posterior distributions of π_0 and π_2 for all methods. If the posterior probability that $\pi_0 > \pi_2$ was greater than 0.975, experimental treatment 2 was declared effective. Type I error rates under null scenarios and power under alternative scenarios are reported. Estimated mean, standard deviation and bias for π_0 and the relative risk (π_2/π_0) for each scenario are also reported. All simulations were conducted in R version 3.5.3[31].

Figure 4.2 Simulated Failure Rates Overtime in the Control Group



4.3 Results

The probability of experimental treatments being declared effective under each scenario is shown in table 4.1. As expected, when failure rates are fixed under the *"constant"* scenario, type I error and power are equivalent for the comparison of experimental treatments 1 and 2 to their respective concurrent controls. However, when experimental treatment 2 is compared to all controls, type I error decreases from 0.251 to 0.243 and power increases from 0.695 to 0.782 and 0.877 to 0.931 for true relative risks of 0.8 and 0.75 respectively. Powers for the test-then-pool, power prior and MEM approaches lie in between that of the concurrent only and all pooled approaches. Under the test-then-pool approaches, type I error is higher than both the concurrent only and all pooled approaches (type I error = 0.0263 and 0.0270 for test-then-pool 1 and 2 respectively). Conversely, the type I error of the static power prior, dynamic power prior, and MEM all fall below the all pooled approach (type I error = 0.0209, 0.0216 and 0.226 respectively).

Mean estimates of π_0 under each scenario are shown in table 4.2. Under the *"constant"* scenario, π_0 estimates are equivalent across all analytical approaches. The mean for the concurrent control approach has the highest variability and all pooled the lowest (SD = 0.285 vs. 0.224 under null scenario). In between these two extremes, the test-then-pool, power prior, and MEM approaches have similar standard deviations. Relative risk estimates follow the same patterns with more precise estimates for methods that borrow information from non-concurrent controls. (Table 4.3)

Under an increasing linear trend, the rate of failure in the control group increases monthly at a constant rate. As a result, the estimate of π_0 is higher for concurrent controls compared to all controls pooled (0.623 vs. 0.5873 and 0.562 vs. 0.544 under the null for the *"increasing linear*

1" and *"increasing linear 2"* scenarios respectively, Table 4.2). Consequently, pooling controls randomized prior to the initiation of experimental group 2 results in lower type I error (Type I error = 0.0017 vs. 0.225 and 0.007 vs. 0.0273 for all pooled vs. concurrent only under the *"increasing linear 1"* and *"increasing linear 2"* scenarios respectively). Power is also lower for the pooled analyses as relative risk estimates are inflated towards 1 by the inclusion of the non-concurrent controls with lower failure rates (Table 4.1 and 4.3). The bias of relative risk estimates is shown for all scenarios in table 4.4. In the *"increasing linear 1"* scenario, when all controls are pooled, relative risk estimates are 0.0639, 0.0531, and 0.0492 higher than the true relative risks of 1, 0.8 and 0.75 respectively. In contrast, under this same scenario, bias ranges from only 0.004-0.005 in analyses that include only concurrent controls. When the rate of increasing failure is slower under the *"increasing linear 2"* scenario, estimates from pooled analyses remain upwardly biased at 0.0363, 0.0313 and 0.0294 for true relative risks of 1, 0.8 and 0.75 respectively, compared to values that range from 0.005-0.006 for analyses of concurrent controls.

Under the *"increasing linear 1"* scenario, the test-then-pool, power prior, and MEM approaches all yield less bias then the all pooled analysis. However, bias under these approaches remains higher than the concurrent only analysis. Similarly, type I error and power for these approaches fall in between the two extremes. Under the *"increasing linear 1"* scenario, the test-then-pool approach produced the least biased estimates of the methods that incorporate all controls. In the null scenario, bias is 0.020 and 0.014 for the test-then-pool 1 and 2 approaches respectively versus 0.064 in all pooled (Table 4.4). However, the variability around the relative risk and π_0 estimates for the test-then-pool approaches is the highest of all methods explored (Tables 4.2 and 4.3). Type I error for the second test-then-pool approach is similar to that of concurrent controls only (0.0226 vs. 0.0225). Power is lower but comparable when the true relative risk is 0.75 (power for test-then-pool 2 = 0.839 and 0.960 vs. 0.866 and 0.968 for concurrent controls when relative risk is 0.8 and 0.75 respectively, Table 4.1). The first test-then-pool approach, which has a higher threshold for declaring concurrent and non-concurrent controls different, has lower type I error and power (Table 4.1).

In the *"increasing linear 2"* scenario, where failure rates increase at a slower rate, the difference between concurrent and non-concurrent controls is less extreme than in the *"increasing linear 1"* scenario. As a result, the second test-then-pool approach becomes less similar to the concurrent only analysis in terms of type I error and power (Table 4.1). Type I error for the static power prior, dynamic power prior and MEM approaches remains lower than that of the test-then-pool approaches; but power is relatively comparable across all 5 approaches (range between 0.748-0.769 when true relative risk is 0.8 and 0.923-0.936 when true relative risk is 0.75, Table 4.1). These values are higher than the power of the all pooled analysis, and comparable to the concurrent only analysis when the true relative risk is 0.75. However, when the true relative risk is 0.8, power is slightly lower for these approaches than the concurrent only. Despite the similarity in power when the true relative risk is 0.75, relative risk estimates remain upwardly biased for all 5 approaches (range 0.015-0.020 for test-then-pool, power prior and MEM approaches versus 0.006 in concurrent only analysis. Table 4.4). Relative risk estimates from the second test-then-pool approach remained the least biased of the methods that incorporate all controls. However, the bias of the first test-then-pool approach was comparable to that of the static power prior and greater than that of the MEM and dynamic power approaches. (Table 4.4)

Under the decreasing linear trends, the estimate of π_0 is lower in concurrent controls compared to controls randomized prior to the initiation of experimental treatment 2. As a result, the trends observed under the *"increasing linear 1"* and *"increasing linear 2"* scenarios are switched. Type I error is higher for the all pooled analyses compared to the concurrent only (0.164 and 0.0732 vs. 0.0240 and 0.0235 for the *"decreasing linear 1"* and *"decreasing linear 2"* scenarios respectively, Table 4.1). Type I error is also inflated for the test-then-pool, power prior, and MEM approaches. Although values are lower compared to the all pooled analyses (Type I error ranges from 0.0530-0.0795 and 0.0392-0.0617 for scenarios 1 and 2 respectively), they are still substantially higher than the type I error of the concurrent only analyses (Table 4.1). Of the methods that incorporate all controls, the second test-then-pool approach again has the least biased estimates for relative risk and π_0; however, the variance around the test-then-pool estimates remain the highest of all approaches (Tables 4.2-4.4). Relative risk estimates for the test-then-pool, power prior, and MEM approaches are closer to the values of the concurrent only analyses compared to the all pooled approach. The static power prior, which down weights the non-concurrent data by 50%, has the most biased estimates of the five approaches. The dynamic power prior and MEM yield less biased estimates that are comparable across all relative risk scenarios. However, MEM estimates have higher variability, which corresponds to the dynamic power prior having higher power (Table 4.3 and 4.1).

In the exponential like scenarios, failure rates change quickly early on and stabilize towards the end of the enrollment period. Because most of the change occurs early on, in the *"increasing exponential like 1"* and *"increasing exponential like 2"* scenarios, π_0 estimates are lower for analyses incorporating all controls versus concurrent controls only. As such, patterns of type I error and power deflation are similar for the increasing exponential like and increasing linear scenarios (Table 4.1). Compared to the all pooled approach, the test-then-pool, power prior, and MEM approaches were substantially more powerful, but less powerful than analyses that included concurrent controls only. (Table 4.1) Under the *"increasing exponential like 2"* scenario, power is comparable across the five methods, though highest for the second test-then pool approach. However, under the *"increasing exponential like 1"* scenario, power is not comparable across the five approaches. The static power prior performs the worst and the test-then-pool approaches have the highest power. (Table 4.1).

In the *"increasing exponential like 1"* and *"increasing exponential like 2"* scenarios, relative risk estimates were also more biased for methods that incorporated all controls. Across both scenarios, the second test-then-pool approach remained the least biased of the methods that incorporate all controls (Table 4.4). Interestingly, as observed in the increasing linear scenarios, in the exponential scenario with less extreme growth (*"increasing exponential like 2"*), the bias of the first test-then-pool approach was comparable to that of the static power prior, and higher than that of the MEM and dynamic power approaches. Comparing the MEM and dynamic power prior approaches, in the *"increasing exponential like 1"* scenario, relative risk estimates were more biased for the dynamic power prior; however, in the less extreme *"increasing exponential like 2"* scenario, bias was comparable with the dynamic power prior having slightly less bias. (Table 4.4)

Similar to the decreasing linear scenarios, under the decreasing exponential like scenarios, type I error is inflated for approaches that incorporate all controls. In the more extreme *"decreasing exponential like 1"* scenario, type I error is 0.218 in the all pooled analysis compared to 0.025 in the concurrent only. Among the five historical control approaches, type I error is highest for the static power prior approach (0.098) and lowest for the MEM and test-then-pool 2 approaches (0.053 and 0.056 respectively, Table 4.1). In the *"decreasing exponential like 2"* scenario, type I error is 0.069 in the all pools approaches versus 0.023 in the concurrent only analysis. Of the five historical control approaches, the dynamic power prior and MEM approaches have the lowest type I error (0.038 and 0.041 respectively) while the two test-then-pool approaches have the highest (0.058 and 0.052 for test-then-pool 1 and 2 respectively).

Under both decreasing exponential like scenarios, relative risk estimates from the second test-then-pool approach remained the least biased but had higher variability than the all pooled, power prior, and MEM approaches. In the *"decreasing exponential like 1"* scenario, bias was -0.07 in the all pooled analysis compared to 0.007 in the concurrent only. The static power prior approach produced the most biased estimate of the 5 historical control approaches (-0.044), followed by the dynamic power prior (-0.024) and MEM (-0.012). In the less extreme *"decreasing exponential like 2"* scenario, bias was more comparable across the five historical control approaches ranging from 0.009 for the test-then-pool 2 to -0.0175 for the static power prior (Table 4.4).

In the *"seasonal"* scenario, failure rates cycled up and down annually. As a result, although monthly rates varied, the overall annual rates were constant. Since the second arm was introduced after one year of enrollment, π_0 estimates were comparable across all approaches and results were similar to those observed under the *"constant"* failure rate scenario.

Table 4.1 Probability of Experimental Treatment being Declared Effective

			E2 vs Control						
Scenario	True RR	E1 vs Control	Concurrent Only	All Pooled	Test then Pool (1)	Test then Pool (2)	Power Prior - Static	Power Prior - Dynamic	MEM
Constant	1.00	0.0251	0.0251	0.0243	0.0263	0.0270	0.0209	0.0216	0.0226
	0.80	0.6954	0.6954	0.7815	0.7736	0.7668	0.7516	0.7455	0.7535
	0.75	0.8717	0.8768	0.9307	0.9236	0.9170	0.9171	0.9125	0.9133
Increasing Linear 1	1.00	0.0228	0.0225	0.0017	0.0210	0.0226	0.0041	0.0084	0.0138
	0.80	0.7702	0.8660	0.6776	0.8141	0.8390	0.7697	0.7921	0.7870
	0.75	0.9241	0.9679	0.9051	0.9524	0.9600	0.9428	0.9473	0.9461
Increasing Linear 2	1.00	0.0249	0.0273	0.0070	0.0203	0.0234	0.0114	0.0135	0.0128
	0.80	0.7352	0.7832	0.7192	0.7568	0.7689	0.7512	0.7543	0.7479
	0.75	0.8958	0.9337	0.9153	0.9300	0.9355	0.9279	0.9301	0.9286
Decreasing Linear 1	1.00	0.0215	0.0240	0.1638	0.0795	0.0609	0.0776	0.0548	0.0530
	0.80	0.8713	0.7694	0.9792	0.8131	0.7927	0.9374	0.8735	0.8211
	0.75	0.9710	0.9283	0.9982	0.9395	0.9329	0.9897	0.9680	0.9454
Decreasing Linear 2	1.00	0.0276	0.0235	0.0732	0.0617	0.0554	0.0410	0.0392	0.0429
	0.80	0.7909	0.7257	0.9122	0.8378	0.8113	0.8573	0.8293	0.8193
	0.75	0.9312	0.9012	0.9835	0.9465	0.9321	0.9658	0.9535	0.9432
Increasing Exponential Like 1	1.00	0.0249	0.0237	0.0010	0.0234	0.0235	0.0032	0.0090	0.0162
	0.80	0.5524	0.6679	0.3194	0.6199	0.6433	0.4614	0.5288	0.5509
	0.75	0.7469	0.8479	0.5753	0.7996	0.8207	0.7032	0.7509	0.7556
Increasing Exponential Like 2	1.00	0.0226	0.0240	0.0071	0.0198	0.0218	0.0105	0.0125	0.0136
	0.80	0.6522	0.6961	0.6078	0.6569	0.6743	0.6467	0.6547	0.6455
	0.75	0.8339	0.8688	0.8299	0.8563	0.8642	0.8546	0.8576	0.8520
Decreasing Exponential Like 1	1.00	0.0288	0.0248	0.2181	0.0767	0.0561	0.0983	0.0619	0.0527
	0.80	0.8278	0.7202	0.9794	0.7503	0.7356	0.9295	0.8363	0.7614
	0.75	0.9542	0.8913	0.9976	0.9023	0.8972	0.9849	0.9471	0.9089
Decreasing Exponential Like 2	1.00	0.0259	0.0229	0.0692	0.0576	0.0527	0.0408	0.0377	0.0413
	0.80	0.7443	0.7003	0.9047	0.8227	0.7944	0.8381	0.8109	0.7998
	0.75	0.9097	0.8723	0.9757	0.9275	0.9143	0.9506	0.9347	0.9234
Seasonal	1.00	0.0251	0.0245	0.0248	0.0267	0.0278	0.0222	0.0223	0.0228
	0.80	0.8080	0.8059	0.8760	0.8706	0.8638	0.8538	0.8494	0.8542
	0.75	0.9419	0.9487	0.9759	0.9721	0.9706	0.9684	0.9670	0.9680

Table 4.2 Mean and Standard Deviations of π_0 Estimates

	True	Concurrent Only		All Pooled		Test then Pool (1)		Test then Pool (2)		Power Prior - Static		Power Prior - Dynamic		MEM	
Scenario	RR	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Constant	1.00	0.5001	0.0285	0.5001	0.0224	0.5001	0.0243	0.5001	0.0253	0.5001	0.0234	0.5001	0.0238	0.5001	0.0236
	0.80	0.5000	0.0288	0.5001	0.0228	0.5002	0.0246	0.5001	0.0257	0.5000	0.0238	0.5001	0.0242	0.5001	0.0240
	0.75	0.5005	0.0285	0.5004	0.0228	0.5004	0.0245	0.5005	0.0256	0.5005	0.0237	0.5005	0.0240	0.5004	0.0239
Increasing Linear 1	1.00	0.6230	0.0275	0.5873	0.0222	0.6137	0.0331	0.6174	0.0316	0.6011	0.0229	0.6079	0.0261	0.6112	0.0299
	0.80	0.6223	0.0278	0.5867	0.0222	0.6129	0.0335	0.6167	0.0321	0.6004	0.0231	0.6073	0.0263	0.6106	0.0303
	0.75	0.6222	0.0279	0.5869	0.0222	0.6129	0.0333	0.6165	0.0320	0.6005	0.0230	0.6073	0.0264	0.6106	0.0303
Increasing Linear 2	1.00	0.5616	0.0285	0.5439	0.0227	0.5513	0.0282	0.5541	0.0289	0.5507	0.0236	0.5524	0.0247	0.5517	0.0259
	0.80	0.5613	0.0284	0.5432	0.0224	0.5509	0.0283	0.5537	0.0290	0.5502	0.0234	0.5519	0.0245	0.5513	0.0258
	0.75	0.5612	0.0283	0.5434	0.0225	0.5512	0.0281	0.5540	0.0289	0.5502	0.0234	0.5520	0.0245	0.5515	0.0257
Decreasing Linear 1	1.00	0.5512	0.0286	0.5870	0.0222	0.5609	0.0343	0.5568	0.0328	0.5732	0.0234	0.5663	0.0269	0.5631	0.0310
	0.80	0.5512	0.0287	0.5873	0.0225	0.5608	0.0342	0.5570	0.0329	0.5734	0.0237	0.5664	0.0271	0.5631	0.0311
	0.75	0.5514	0.0282	0.5869	0.0221	0.5615	0.0339	0.5573	0.0324	0.5732	0.0232	0.5665	0.0265	0.5634	0.0305
Decreasing Linear 2	1.00	0.5256	0.0289	0.5436	0.0226	0.5359	0.0286	0.5332	0.0294	0.5366	0.0237	0.5349	0.0249	0.5355	0.0262
	0.80	0.5255	0.0283	0.5435	0.0225	0.5360	0.0281	0.5331	0.0289	0.5365	0.0234	0.5349	0.0245	0.5355	0.0257
	0.75	0.5250	0.0287	0.5431	0.0225	0.5357	0.0283	0.5326	0.0291	0.5362	0.0236	0.5345	0.0248	0.5351	0.0260
Increasing Exponential Like 1	1.00	0.4846	0.0286	0.4417	0.0225	0.4773	0.0347	0.4807	0.0325	0.4582	0.0236	0.4685	0.0280	0.4742	0.0324
	0.80	0.4850	0.0286	0.4420	0.0223	0.4780	0.0345	0.4811	0.0324	0.4586	0.0235	0.4689	0.0280	0.4747	0.0324
	0.75	0.4844	0.0286	0.4416	0.0223	0.4769	0.0347	0.4803	0.0326	0.4581	0.0235	0.4683	0.0280	0.4740	0.0325
Increasing Exponential Like 2	1.00	0.4987	0.0286	0.4798	0.0224	0.4880	0.0287	0.4909	0.0293	0.4871	0.0235	0.4890	0.0247	0.4884	0.0262
	0.80	0.4989	0.0286	0.4804	0.0227	0.4882	0.0286	0.4910	0.0293	0.4875	0.0236	0.4893	0.0248	0.4887	0.0262
	0.75	0.4986	0.0287	0.4800	0.0227	0.4882	0.0286	0.4909	0.0294	0.4872	0.0236	0.4890	0.0248	0.4885	0.0262
Decreasing Exponential Like 1	1.00	0.5153	0.0287	0.5583	0.0226	0.5225	0.0348	0.5192	0.0326	0.5417	0.0237	0.5314	0.0282	0.5256	0.0325
	0.80	0.5158	0.0289	0.5585	0.0226	0.5230	0.0349	0.5198	0.0329	0.5420	0.0238	0.5319	0.0283	0.5262	0.0327
	0.75	0.5154	0.0284	0.5580	0.0223	0.5227	0.0344	0.5195	0.0324	0.5416	0.0234	0.5316	0.0278	0.5260	0.0322
Decreasing Exponential Like 2	1.00	0.5010	0.0284	0.5199	0.0223	0.5117	0.0286	0.5089	0.0293	0.5127	0.0233	0.5108	0.0246	0.5113	0.0261
	0.80	0.5017	0.0287	0.5204	0.0226	0.5123	0.0286	0.5093	0.0293	0.5132	0.0236	0.5113	0.0248	0.5119	0.0262
	0.75	0.5012	0.0287	0.5200	0.0228	0.5119	0.0287	0.5089	0.0294	0.5127	0.0238	0.5109	0.0249	0.5115	0.0263
Seasonal	1.00	0.5744	0.0281	0.5746	0.0224	0.5745	0.0240	0.5745	0.0250	0.5745	0.0233	0.5745	0.0236	0.5745	0.0234
	0.80	0.5747	0.0285	0.5747	0.0227	0.5748	0.0243	0.5748	0.0253	0.5747	0.0237	0.5747	0.0240	0.5748	0.0238
	0.75	0.5751	0.0282	0.5751	0.0223	0.5752	0.0241	0.5753	0.0251	0.5751	0.0233	0.5751	0.0236	0.5752	0.0234

Table 4.3 Mean and Standard Deviations of Relative Risk (RR= π_2/π_0) Estimates

	True	Concurrent Only		All Pooled		Test then Pool (1)		Test then Pool (2)		Power Prior - Static		Power Prior - Dynamic		MEM	
Scenario	RR	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Constant	1.00	1.0070	0.0821	1.0044	0.0737	1.0047	0.0764	1.0051	0.0775	1.0051	0.0750	1.0052	0.0755	1.0050	0.0754
	0.80	0.8065	0.0730	0.8044	0.0672	0.8045	0.0686	0.8049	0.0695	0.8050	0.0680	0.8050	0.0683	0.8049	0.0681
	0.75	0.7558	0.0708	0.7542	0.0655	0.7544	0.0671	0.7546	0.0679	0.7546	0.0662	0.7546	0.0666	0.7546	0.0665
Increasing Linear 1	1.00	1.0041	0.0630	1.0639	0.0621	1.0200	0.0708	1.0137	0.0687	1.0399	0.0609	1.0289	0.0632	1.0241	0.0671
	0.80	0.8051	0.0587	0.8531	0.0589	0.8180	0.0648	0.8129	0.0631	0.8338	0.0577	0.8250	0.0592	0.8211	0.0620
	0.75	0.7547	0.0580	0.7992	0.0578	0.7667	0.0633	0.7620	0.0618	0.7813	0.0569	0.7730	0.0584	0.7695	0.0609
Increasing Linear 2	1.00	1.0054	0.0732	1.0363	0.0684	1.0234	0.0738	1.0184	0.0743	1.0240	0.0686	1.0212	0.0696	1.0226	0.0709
	0.80	0.8060	0.0664	0.8313	0.0642	0.8205	0.0673	0.8165	0.0678	0.8212	0.0639	0.8189	0.0645	0.8200	0.0654
	0.75	0.7559	0.0634	0.7794	0.0613	0.7691	0.0643	0.7653	0.0646	0.7700	0.0609	0.7678	0.0615	0.7687	0.0624
Decreasing Linear 1	1.00	1.0056	0.0739	0.9418	0.0608	0.9887	0.0797	0.9958	0.0784	0.9652	0.0639	0.9780	0.0692	0.9848	0.0753
	0.80	0.8060	0.0670	0.7546	0.0566	0.7927	0.0712	0.7981	0.0703	0.7734	0.0593	0.7839	0.0633	0.7894	0.0679
	0.75	0.7547	0.0643	0.7072	0.0548	0.7415	0.0679	0.7469	0.0670	0.7246	0.0572	0.7340	0.0608	0.7388	0.0648
Decreasing Linear 2	1.00	1.0064	0.0786	0.9706	0.0668	0.9858	0.0768	0.9911	0.0782	0.9838	0.0694	0.9873	0.0714	0.9866	0.0734
	0.80	0.8071	0.0702	0.7785	0.0617	0.7905	0.0687	0.7950	0.0696	0.7891	0.0636	0.7918	0.0650	0.7911	0.0662
	0.75	0.7563	0.0675	0.7292	0.0593	0.7403	0.0658	0.7448	0.0670	0.7392	0.0612	0.7418	0.0625	0.7412	0.0637
Increasing Exponential Like 1	1.00	1.0069	0.0845	1.1026	0.0857	1.0237	0.0969	1.0160	0.0923	1.0633	0.0832	1.0415	0.0873	1.0306	0.0932
	0.80	0.8066	0.0756	0.8834	0.0781	0.8195	0.0843	0.8138	0.0811	0.8519	0.0757	0.8345	0.0781	0.8255	0.0820
	0.75	0.7586	0.0732	0.8305	0.0753	0.7715	0.0821	0.7657	0.0789	0.8009	0.0731	0.7846	0.0757	0.7765	0.0796
Increasing Exponential Like 2	1.00	1.0073	0.0822	1.0446	0.0767	1.0285	0.0833	1.0227	0.0838	1.0296	0.0769	1.0261	0.0781	1.0276	0.0799
	0.80	0.8054	0.0734	0.8346	0.0711	0.8223	0.0746	0.8178	0.0748	0.8229	0.0706	0.8202	0.0712	0.8215	0.0723
	0.75	0.7571	0.0715	0.7845	0.0689	0.7724	0.0724	0.7683	0.0730	0.7735	0.0686	0.7709	0.0692	0.7720	0.0703
Decreasing Exponential Like 1	1.00	1.0071	0.0808	0.9267	0.0643	0.9942	0.0877	1.0001	0.0851	0.9558	0.0685	0.9760	0.0769	0.9882	0.0844
	0.80	0.8053	0.0717	0.7415	0.0590	0.7949	0.0764	0.7996	0.0748	0.7646	0.0624	0.7804	0.0684	0.7900	0.0741
	0.75	0.7570	0.0684	0.6971	0.0571	0.7471	0.0727	0.7515	0.0712	0.7188	0.0600	0.7336	0.0653	0.7425	0.0703
Decreasing Exponential Like 2	1.00	1.0074	0.0810	0.9680	0.0685	0.9853	0.0799	0.9910	0.0814	0.9825	0.0712	0.9866	0.0735	0.9860	0.0760
	0.80	0.8060	0.0736	0.7749	0.0641	0.7884	0.0723	0.7933	0.0733	0.7863	0.0662	0.7896	0.0679	0.7891	0.0696
	0.75	0.7554	0.0728	0.7261	0.0641	0.7388	0.0714	0.7434	0.0724	0.7369	0.0661	0.7399	0.0676	0.7394	0.0690
Seasonal	1.00	1.0053	0.0702	1.0030	0.0632	1.0034	0.0650	1.0037	0.0662	1.0037	0.0643	1.0038	0.0646	1.0036	0.0644
	0.80	0.8058	0.0644	0.8042	0.0591	0.8044	0.0604	0.8045	0.0613	0.8047	0.0600	0.8047	0.0603	0.8045	0.0600
	0.75	0.7545	0.0615	0.7532	0.0572	0.7533	0.0583	0.7532	0.0589	0.7536	0.0578	0.7536	0.0580	0.7535	0.0579

Table 4.4 Bias of Mean Relative Risk (RR= π_2/π_0) Estimates

Scenario	True RR	Concurrent Only	All Pooled	Test then Pool (1)	Test then Pool (2)	Power Prior - Static	Power Prior - Dynamic	MEM
Constant	1.00	0.0070	0.0044	0.0047	0.0051	0.0051	0.0052	0.0050
	0.80	0.0065	0.0044	0.0045	0.0049	0.0050	0.0050	0.0049
	0.75	0.0058	0.0042	0.0044	0.0046	0.0046	0.0046	0.0046
Increasing Linear 1	1.00	0.0041	0.0639	0.0200	0.0137	0.0399	0.0289	0.0241
	0.80	0.0051	0.0531	0.0180	0.0129	0.0338	0.0250	0.0211
	0.75	0.0047	0.0492	0.0167	0.0120	0.0313	0.0230	0.0195
Increasing Linear 2	1.00	0.0054	0.0363	0.0234	0.0184	0.0240	0.0212	0.0226
	0.80	0.0060	0.0313	0.0205	0.0165	0.0212	0.0189	0.0200
	0.75	0.0059	0.0294	0.0191	0.0153	0.0200	0.0178	0.0187
Decreasing Linear 1	1.00	0.0056	-0.0582	-0.0113	-0.0042	-0.0348	-0.0220	-0.0152
	0.80	0.0060	-0.0454	-0.0073	-0.0019	-0.0266	-0.0161	-0.0106
	0.75	0.0047	-0.0428	-0.0085	-0.0031	-0.0254	-0.0160	-0.0112
Decreasing Linear 2	1.00	0.0064	-0.0294	-0.0142	-0.0089	-0.0162	-0.0127	-0.0134
	0.80	0.0071	-0.0215	-0.0095	-0.0050	-0.0109	-0.0082	-0.0089
	0.75	0.0063	-0.0208	-0.0097	-0.0052	-0.0108	-0.0082	-0.0088
Increasing Exponential Like 1	1.00	0.0069	0.1026	0.0237	0.0160	0.0633	0.0415	0.0306
	0.80	0.0066	0.0834	0.0195	0.0138	0.0519	0.0345	0.0255
	0.75	0.0086	0.0805	0.0215	0.0157	0.0509	0.0346	0.0265
Increasing Exponential Like 2	1.00	0.0073	0.0446	0.0285	0.0227	0.0296	0.0261	0.0276
	0.80	0.0054	0.0346	0.0223	0.0178	0.0229	0.0202	0.0215
	0.75	0.0071	0.0345	0.0224	0.0183	0.0235	0.0209	0.0220
Decreasing Exponential Like 1	1.00	0.0071	-0.0733	-0.0058	0.0001	-0.0442	-0.0240	-0.0118
	0.80	0.0053	-0.0585	-0.0051	-0.0004	-0.0354	-0.0196	-0.0100
	0.75	0.0070	-0.0529	-0.0029	0.0015	-0.0312	-0.0164	-0.0075
Decreasing Exponential Like 2	1.00	0.0074	-0.0320	-0.0147	-0.0090	-0.0175	-0.0134	-0.0140
	0.80	0.0060	-0.0251	-0.0116	-0.0067	-0.0137	-0.0104	-0.0109
	0.75	0.0054	-0.0239	-0.0112	-0.0066	-0.0131	-0.0101	-0.0106
Seasonal	1.00	0.0053	0.0030	0.0034	0.0037	0.0037	0.0038	0.0036
	0.80	0.0058	0.0042	0.0044	0.0045	0.0047	0.0047	0.0045
	0.75	0.0045	0.0032	0.0033	0.0032	0.0036	0.0036	0.0035

4.4 Discussion

This study sought to determine whether methods developed to integrate historical controls into two-arm studies could be applied to the analysis of later-entry experimental treatments in open platform trials. Two aspects of implementing these approaches were explored. First their performance in statistical inference under the null hypothesis H_0: $\pi_2 \geq \pi_0$ was assessed by evaluating type I error under the null scenario and power under two alternatives. Second, the performance in estimation was assessed by reporting the bias of relative risk estimates produced using each approach under fixed relative risk values. Simulation results confirm that in the presence of no parameter drift, pooling non-concurrent and concurrent controls yields higher power, lower type I error, and more precise, unbiased estimates. Clearly, if outcome rates are known to be constant over time, pooling is the optimal approach. However, constant outcome rates may not be a realistic assumption.

There are many reasons that outcome rates in the control arm can change overtime. The fact that sites open to enrollment at different times, that there may be a learning curve for protocol procedures, that centers have different experience levels conducting clinical trials, and other logistical issues can all impact trial-wide outcome rates. In a vaccine trial, infection outcome rates could vary seasonally as explored in the *"seasonal"* scenario. Linear changes are also possible in cases where standard of care improves over time. In a rare disease population with limited treatment options, early enrollees are likely to be on the more severe end of the disease spectrum, while the later enrollees may be less severe patients. This could result in outcome trends similar to the exponential-like scenarios explored in this simulation study. Even if constant event rates are a realistic assumption for a disease population, an unexpected change in the treatment landscape could impact the anticipated enrollment population and therefore event rates. For instance, in a cancer platform trial, if a new treatment is found to work on a subset of the disease population with a specific biomarker, future patients with this biomarker will not enroll in the trial. If this subset has different event rates than patients without the biomarker, trial-wide events rates could change depending on how event rates differ in the rest of the disease population, and on how large the subset is relative to the overall study population.

When failure rates increase over time, simulation results show that pooling non-concurrent and concurrent control data results in type I error deflation and upwardly biased relative risk estimates that make the experimental treatment appear less effective than it truly is. Conversely, when failure rates decrease over time, pooling yields inflated type I error and downwardly biased relative risk estimates that indicate a stronger treatment effect than truly exists. These results confirm that naively pooling the data is suboptimal when outcome rates change overtime.

Several approaches between the extremes of pooling all controls and using concurrent controls only were explored. In theory, approaches that down weight the non-concurrent data or dynamically borrow information based on how homogeneous the data are provide a viable alternative to losing power and precision by using concurrent controls only while still protecting against bias in the presence of parameter drift. Across scenarios where outcome rates increased or decreased over time, the test-then-pool 2 approach yielded the least biased relative risk estimates of the methods that incorporated all controls. Compared to the test-then-pool 1 approach, the test-then-pool 2 approach had a lower threshold for declaring the concurrent and non-concurrent controls different, and therefore was more likely to default to a concurrent only analysis. The more extreme the change over time, the less biased estimates were as non-concurrent controls were more likely to be discarded from analyses of experimental treatment 2.

Despite having the least biased estimates among approaches that incorporated all controls, type I error for the test-then-pool 2 approach was higher across many of the scenarios. Notably, when outcome rates were constant, the test-then-pool approaches yielded higher type I error than both the all pooled and concurrent only approaches. This result is consistent with Dejardin et al.'s study that explored incorporating a historical control in the analysis of a two-arm non-inferiority trial and observed inflated type I error for a test-then-pool approach[52]. Under constant failure rates, rejection of the hypothesis that the concurrent and non-concurrent controls have equal failure rates can occur due to random variability. The inflation of type I error above the nominal 0.025 level observed in the concurrent only analysis under the *"constant"* scenario lends the test-then-pool approach unacceptable in cases were strict type I error control is needed.

Unexpectedly, under constant failure rates, the power prior and MEM approaches yielded type I error rates below that of both the concurrent only and all pooled analyses. One would expect type I error for these approaches to fall between the concurrent only and all pooled approaches; however, simulation results were consistent with type I error rates observed for the dynamic power prior, robust mixture prior, and commensurate prior explored in Dejardin et al.'s study[52]. Based on the *"constant"* scenario alone, any of these three methods have acceptable type I error; however, in scenarios where failure rates decreased overtime, although type I error was lower than that observed in the all pooled analyses, rates were still inflated relative to the concurrent only analysis and above the nominal 0.025 level.

Relative risk estimates for the power prior and MEM approaches were all less biased than for the all pooled approach. In scenarios with increasing or decreasing failure rates, the static power prior yielded more biased estimates compared to the dynamic power prior. This matches intuition that of the two approaches, the dynamic power prior would be the preferred method as it does not rely on a subjective choice for the power parameter and instead treats the power parameter as a random variable dependent on the similarity between the concurrent and non-concurrent data. Compared to an alternative dynamic approach, the dynamic power prior was comparable to the MEM across many scenarios. With the exception of the *"constant"* scenario, standard deviations around relative risk estimates were slightly higher for the MEM approach compared to the dynamic power prior. In the scenarios with faster rates of change in outcome rates, the MEM demonstrated lower bias in relative risk estimates. However in less extreme scenarios, the dynamic power prior demonstrates slightly less bias. Neither method stands out as clearly better. Of the two approaches, the MEM is easier to implement and requires less computation time.

Although superior to naively pooling the data when outcome rates are non-constant, none of the five approaches explored guarantee type I error control in the presence of parameter drift. It is possible that at the design stage, decision criteria and operating characteristics could be fine-tuned to ensure type I error control using one of these approaches over a variety of temporal trends; however there is no guarantee that the simulated scenarios will cover the true trend. Protecting type I error would also not protect against bias in estimates. Further, in the event there is no parameter drift, this process could result in an ultimately inefficient design.
Where strict error control is needed, such as in a confirmatory drug trial, these methods are likely unacceptable as primary analyses, but could serve as sensitivity analyses to supplement a concurrent only analysis. These methods may be acceptable for early phase screening platforms. Medical device trials, serious conditions and rare diseases may also provide populations where these approaches are acceptable, as FDA guidance for these already encourages the use of historical controls[38, 39]. Extending such guidance to incorporating non-concurrent controls in

analyses of open platforms would be plausible, as long as extensive simulation studies are done across various drift scenarios. This will ensure stakeholders are aware of the possible extent of type I error inflation and bias. However, such simulation studies would be challenging, as uncertainty both in drift and the timing of new arm entries would need to be taken into account.

In conclusion, in the presence of parameter drift, methods that partially borrow non-concurrent data, either through a static weighting mechanism or through methods that allow the heterogeneity between non-concurrent and concurrent data to determine the degree of borrowing, are superior to naively pooling the data in the presence of parameter drift.   However, compared to using concurrent controls only, these approaches cannot guarantee type I error control and may still produce biased estimates. If strict error control is required, using only concurrent controls in the analyses of experimental treatments is the most conservative approach.

CHAPTER 5

Applications to the Neuroprotection Trial

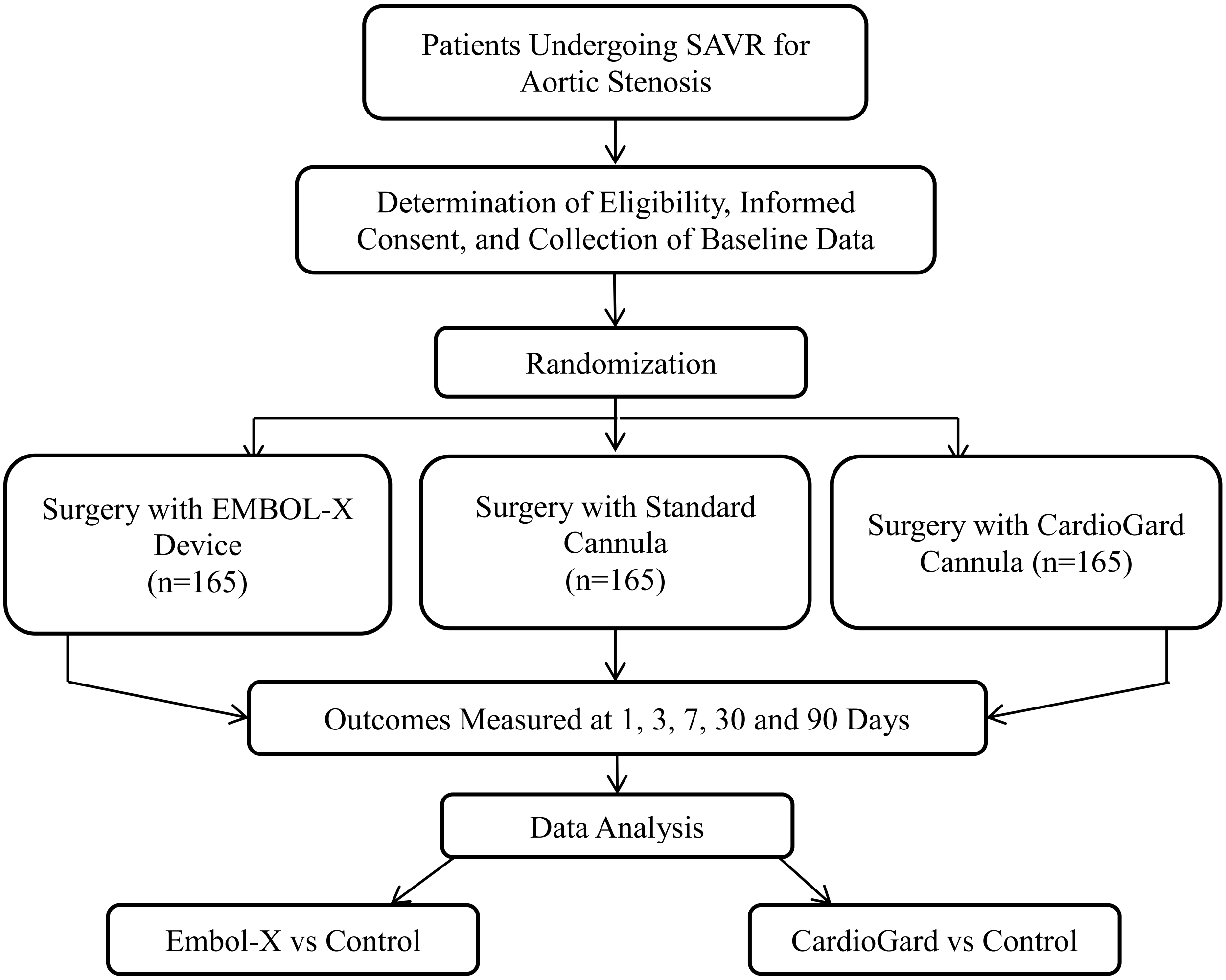
5.1 Introduction

In the motivating example of this work, the CTSN neuroprotection trial evaluated the efficacy of two embolic protection devices for reducing post-operative infarcts after surgical aortic valve replacement (AVR). In the United States, approximately 100,000 patients undergo AVR each year[54, 55]. Prospective studies of surgical AVR patients report post-operative clinically apparent stroke rates between 5 and 17%[56-58] and rates of any cerebral infarction, as measured by post-operative magnetic resonance imaging (MRI), as high as 60%[59, 60]. Most of these infarcts are caused by emboli that the bloodstream carries from the surgical site to the brain[4]. Embolic protection devices that collect debris during surgery aim to prevent infarction.

The two devices evaluated in the neuroprotection trial were an intra-aortic filtration device (Embol-X (Edwards Lifesciences, Irvine, CA)) and a suction based extraction device (CardioGard (Cardiogard, Or-Yehuda, Israel)). Rather than conduct two independent, two-arm controlled trials to evaluate each devices’ efficacy, both device manufacturers agreed to share a platform where their devices would be evaluated simultaneously against a shared control group. For CTSN, this provided an efficient alternative to conducting two trials, both because fewer patients would need to be randomized, and because of the substantial resource savings associated with only needing to build the infrastructure and protocol for a single trial versus two.

The primary outcome of the neuroprotection trial was a composite of death, clinically apparent stroke, or presence of post-operative emboli on diffusion-weighted MRI by post-operative day 7. As described in Chapter 1, under the original design, all three arms were to open to enrollment simultaneously with 1:1:1 randomization (Figure 5.1); however, due to an unexpected delay in 510(k) approval by the FDA, the CardioGard device was not available at the anticipated launch date. Rather than delay launching the trial, the protocol was amended so that the trial would open to enrollment with 1:1 randomization to Embol-X and the control, with a plan to introduce the third arm and 1:1:1 allocation once the CardioGard device was cleared.

Figure 5.1 Original Trial Design Schematic


Patients Undergoing SAVR for Aortic Stenosis
Determination of Eligibility, Informed Consent, and Collection of Baseline Data
Randomization
Surgery with EMBOL-X Device (n=165)
Surgery with Standard Cannula (n=165)
Surgery with CardioGard Cannula (n=165)
Outcomes Measured at 1, 3, 7, 30 and 90 Days
Data Analysis
Embol-X vs Control
CardioGard vs Control

In the analytical plan, each device was to be compared to controls that were concurrently randomized. To ensure adequate power for both devices, the maximum sample size of the control arm was increased so that 165 controls would be concurrently randomized with each device. The Embol-X arm would close after enrolling 165 patients, at which point randomization would continue 1:1 to CardioGard and the control until the CardioGard arm reached 165 patients and trial enrollment closed.

When the neuroprotection trial was designed, few methodological papers were available on designing a multi-arm trial where not all treatment arms open simultaneously. Statistical issues such as whether adjusting for multiple testing was necessary, or whether analyses of arms that opened after initial launch could incorporate information from non-concurrent controls were not well established. This work sought to explore these issues and determine whether the approaches implemented in the neuroprotection trial were appropriate and/or whether alternatives may have been more efficient

5.2 Multiple Testing in the Neuroprotection Trial

In the neuroprotection trial, either or both of the embolic protection devices could be declared effective relative to the shared control. Since determining the efficacy of each device was the primary aim, there was no intent to compare device to device. As a result, both comparisons of device versus control were conducted at the 0.05 significance level. No adjustment was made to control the familywise type I error rate as each comparison was viewed as separate. As explored in Chapter 3, this view is not uncommon.[28, 29] However, as discussed, sharing a control group will induce a dependency between the comparisons since the test statistics for each comparisons will be positively correlated due to the shared control data. This means that in the neuroprotection trial, under the global null hypothesis of no difference between the control and either device, if one device was erroneously declared effective, there would have been a heightened conditional probability that the other device would have also been erroneously declared effective.

As expected, when no multiple testing adjustment is applied, simulation studies in Chapter 3 demonstrated that this dependency does result in conditional type I error rates higher than the nominal significance level. However, simulation studies show that applying a Bonferroni correction would not have controlled the conditional type I error inflation. Although the correction would control the familywise error rate, conditional type I error rates would remain above the nominal significance level. Interestingly, simulation studies demonstrated that even when a multiple testing correction is not applied, although conditional probabilities are inflated, the marginal probability of both treatments being declared effective simultaneously is minimal under the global null.

In light of these results, the decision to not adjust for multiple testing in the neuroprotection trial was appropriate. Applying a correction would have required a larger sample size to maintain power in exchange for a modest improvement in conditional type I error. Since the marginal probability of both devices being declared effective under the global null is minimal, the inflation in conditional type I error probabilities should be seen, in this case, as an acceptable tradeoff with the efficiency gained from sharing a control group. Further, since the two devices were developed by independent companies and the underlying mechanism of each was different (the Embol-X device uses a filtration system to capture debris, while the CardioGard device uses suction to collect debris), the devices are unrelated and their comparisons should be viewed separately.

5.3 Incorporating Non-concurrent Controls in the CardioGard Efficacy Analysis

In March 2015, the trial launched with Embol-X and control. The following May, the CardioGard arm opened to enrollment. Twelve controls were enrolled prior to the introduction of CardioGard. In July 2016 at a pre-specified interim analysis, after 132 patients had been randomized to the control, 133 to Embol-X, and 118 to CardioGard, the DSMB recommended halting enrollment for futility of both experimental devices. In the final analysis, Embol-X was compared to all enrolled controls, and CardioGard was compared to its 119 concurrently randomized controls.

As determined in Chapter 4, the decision to compare the CardioGard device to concurrently randomized controls exclusively was conservative and avoided any potential bias in efficacy estimates or inflation in type I error due to parameter drift. In the neuroprotection trial, concerns about parameter drift are valid. Overall, 18 clinical sites randomized patients; however, site startup was staggered (Figure 5.2). Differences in patient populations and surgeons across sites, as well as staff learning curves for protocol procedures at each site could all impact trial wide outcome rates overtime. In addition, a protocol revision finalized in October 2015 that widened the eligibility criteria from ≥65 years of age to ≥60 could also have potentially impacted outcome rates since existing infarcts, associated with age, are known to be associated with higher incidence of post-operative infarction.[61] Observed treatment failure rates in the control arm by quarter are shown in figure 5.3. Changes by quarter can be observed; however, in earlier quarters the relative number of controls randomized is small and observed changes could be due to random chance rather than temporal drift.

Figure 5.2 Cumulative Number of Sites Open to Enrollment by Quarter

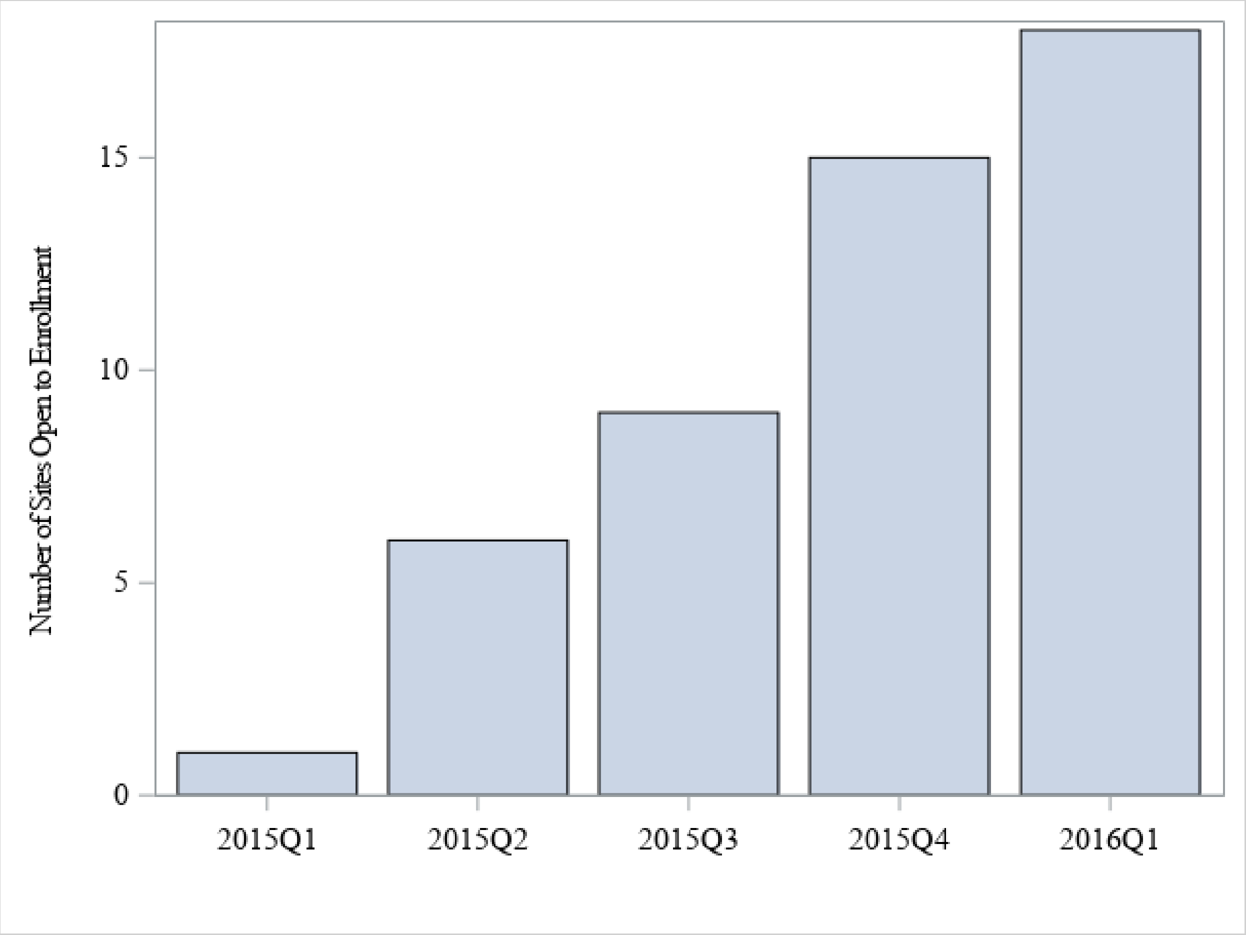
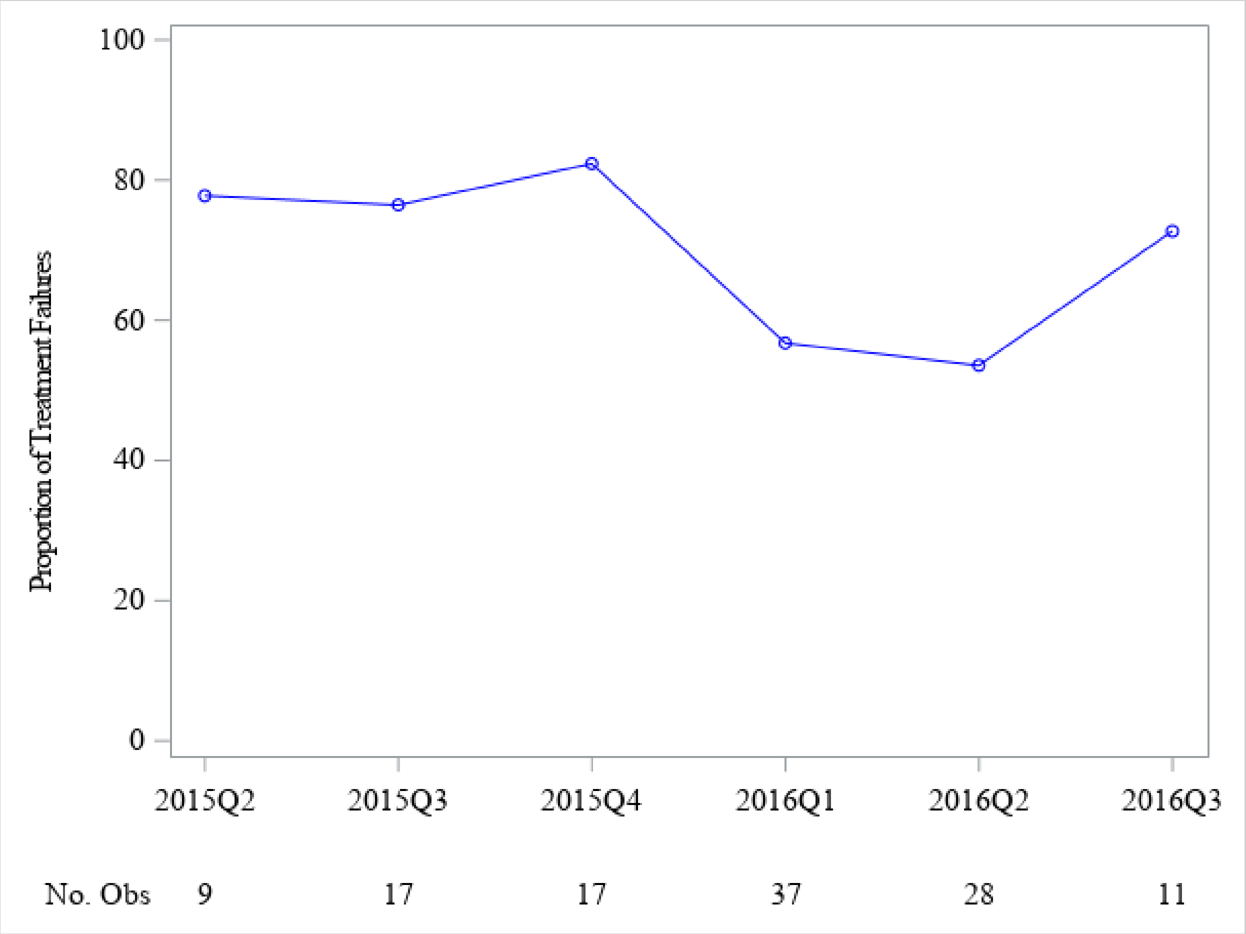


Figure 5.3 Observed Treatment Failure Rates in the Control Arm by Quarter



Baseline and operative characteristics for each device arm and its respective control are shown in table 5.1. Unsurprisingly, because only 12 controls were randomized prior to the initiation of the CardioGard arm and the remaining 119 controls are shared, the characteristics of the control groups for Embol-X and CardioGard are similar.

Table 5.1 Baseline and Operative Characteristics

		Embol-X		CardioGard
	Embol-X	Control[b]	CardioGard	Control[b]
Baseline Characteristics[a]	**(N=133)**	**(N=132)**	**(N=118)**	**(N=120)**
Male sex	81 (60.9)	86 (65.2)	69 (58.5)	77 (64.2)
White	126 (94.7)	118 (89.4)	108 (91.5)	107 (89.2)
Age – yrs	73.6 ±6.6	73.6 ±6.7	74.6 ±6.8	73.4 ±6.7
Medical and surgical history				
Diabetes	36 (27.1)	37 (28.0)	48 (40.7)	36 (30.0)
Renal insufficiency	18 (13.5)	14 (10.6)	15 (12.7)	13 (10.8)
Myocardial infarction	15 (11.3)	10 (7.6)	16 (13.6)	8 (6.7)
Atrial fibrillation	13 (9.8)	16 (12.1)	14 (11.9)	16 (13.3)
Stroke or TIA	11 (8.3)	8 (6.1)	16 (13.6)	8 (6.7)
SF-12[c]				
Physical Health Composite Score	40.1 ±11.0	40.2 ±11.2	41.4 ±10.6	40.5 ±11.2
Mental Health Composite Score	52.9 ±9.6	52.9 ±9.4	53.2 ±9.3	52.9 ±9.3
White Matter Lesion Volume (mm^3)	6303 (2686, 10027)	4704 (2265, 9776)	4592 (2433, 8377)	4719 (2201, 9776)
Maximum Atheroma Grade[f]	2.3 ±0.7	2.3 ±0.6	2.5 ±0.7	2.4 ±0.6
Operative Characteristics[a]				
Surgical Procedure				
Isolated AVR	77 (57.9)	80 (60.6)	67 (56.8)	73 (60.8)
AVR & CABG	51 (38.3)	52 (39.4)	50 (42.4)	47 (39.2)
AVR & MV Repair ± CABG	5 (3.8)	0	1 (0.8)	0
Concomitant procedures	23 (17.3)	20 (15.2)	17 (14.4)	19 (15.8)
Cardiopulmonary bypass time – min	109.1 ±42.4	101.7 ±39.8	104.9 ±39.6	102.2 ±40.2

[a] Categorical measures are presented as the number of patients and (%). If the denominator is not equal to the group sample size, data is presented as the number of patients/the number observed (%). White matter lesion volume is presented as median (IQR) and all other continuous measures are presented as mean (standard deviation).

[b] The first 12 control patients served as controls for Embol-X only and 120 patients were common to both control groups.

Although the choice to compare devices to concurrently randomized controls was appropriate, because the experimental treatments were medical devices, it is possible that regulatory and oversight bodies would have considered any of the alternative methods presented in Chapter 4 viable for the primary analysis plan. Here, the seven analytical approaches explored in Chapter 4 are applied to the neuroprotection trial data. The probability models assumed for all seven approaches and the methods for estimating posterior distributions are the same as presented in Chapter 4. For each approach, estimates of the probability of treatment failure in the control (π_0) and estimates of the relative risks of failure for CardioGard versus control (π_2/π_0) are reported. To explore the possible impact of CardioGard's introduction being delayed farther, these estimates are also reported assuming the CardioGard launch was pushed back 3, 6 and 9 months after the true launch date. In these cases, CardioGard patients randomized before the hypothetical launch date are removed from analysis and control patients randomized prior are considered non-concurrent. All analyses were conducted using R version 3.5.3[31].

Observed primary outcome data under each of the four launch scenarios explored are shown in table 5.2. For all analytical approaches, estimated probabilities of treatment failure in the control arm and the relative risk of failure for CardioGard versus Control are shown in tables 5.3 and 5.4 respectively. Unsurprisingly, because of the small number of patients randomized prior to the actual initiation of CardioGard, relative risk estimates across the seven approaches under the actual scenario are equivalent. Notably, the variance around the estimated probability of failure for the control is comparable across all approaches except for the MEM which has a lower estimate relative to the others (0.033 versus ~0.044), indicating the MEM may underestimate variability in this case.

When the launch of CardioGard is artificially delayed 3 months, the observed proportion of treatment failures in non-concurrent and concurrent controls is similar (68.4% vs. 65%) and relative risk estimates remain similar across the seven approaches. Although estimates are similar, confidence intervals are slightly narrower among the methods that incorporate non-concurrent controls. Interestingly, if the launch is delayed 6 months, the observed rate of treatment failures becomes higher in the non-concurrent controls (76.3% versus. 60.5%) and there is more variability in the relative risk estimates across the seven methods. In the concurrent only analysis, the estimated relative risk is 1.08 versus 0.99 when CardioGard is compared to all controls pooled. Estimates from the test-then-pool approaches differ as the second approach, which has a more liberal threshold for declaring the non-concurrent and concurrent controls unequal, rejects the null that the two control groups are equivalent and defaults to the concurrent only analysis. Alternatively, the first test-then-pool approach, which has a more conservative threshold, defaults to the all pooled analysis. The power prior and MEM approaches each have comparable relative risk values that fall in between the concurrent only and all pooled analyses. When launch is pushed back 9 months, observed failure rates in the non-concurrent versus concurrent controls remain different (75.8% vs. 54.4%). With a greater disparity between the control groups, both test-then-pool approaches default to the concurrent only analysis. The dynamic power prior and MEM approaches have relative risk values closer to the concurrent only analysis whereas the static power prior is closer to the all pooled analysis.

As observed in Chapter 4, the all pooled, static power prior, dynamic power prior, and MEM approaches yield more precise estimates across each scenario. However, simulation studies in Chapter 4 also demonstrated that these approaches are prone to bias when parameter drift occurs. Because the true relative risk of CardioGard versus control is unknown, which approach yields the least biased estimate cannot be determined. The concurrent control analysis is the most conservative, but could be more sensitive to random highs and lows compared to pooling all controls. In the actual case, because only a small number of non-concurrent controls were randomized, ultimately the choice of approach does not affect estimates. However, had more of a delay occurred, the six alternative approaches could provide a set of sensitivity analyses that supplement the concurrent only analyses in combination with an exploration of possible parameter drift.

Table 5.2 Observed Primary Outcomes in the CardioGard Arm and in CardioGard's Non-Concurrent and Concurrent Controls for Varying Launch Scenarios

	Non-Concurrent Control			Concurrent Control			Cardiogard		
	No. Randomized	No. Observed	Treatment Failures No./No. Obs (%)	No. Randomized	No. Observed	Treatment Failures No./No. Obs (%)	No. Randomized	No. Observed	Treatment Failures No./No. Obs (%)
Actual	12	7	5/7 (71.4)	120	112	73/112 (65.2)	118	101	68/101 (67.3)
3 Month	25	19	13/19 (68.4)	107	100	65/100 (65)	104	87	57/87 (65.5)
6 Month	44	38	29/38 (76.3)	88	81	49/81 (60.5)	91	77	50/77 (64.9)
9 Month	70	62	47/62 (75.8)	62	57	31/57 (54.4)	67	54	34/54 (63)

Table 5.3 Estimates for Probability of Failure in Control Arm for Varying Launch Scenarios and Analytical Approaches

	$\hat{\boldsymbol{\pi}}_{\mathbf{0}} \pm \mathbf{var}(\hat{\boldsymbol{\pi}}_{\mathbf{0}})$						
	Concurrent	All	TTP1	TTP2	PPS	PPD	MEM
Actual	0.649 ±0.044	0.653 ±0.043	0.653 ±0.043	0.653 ±0.043	0.651 ±0.044	0.651 ±0.044	0.652 ±0.033
3 Month	0.647 ±0.047	0.653 ±0.043	0.653 ±0.043	0.653 ±0.043	0.650 ±0.045	0.650 ±0.044	0.651 ±0.035
6 Month	0.603 ±0.053	0.653 ±0.043	0.653 ±0.043	0.603 ±0.053	0.632 ±0.047	0.629 ±0.050	0.628 ±0.034
9 Month	0.543 ±0.064	0.653 ±0.043	0.543 ±0.064	0.543 ±0.064	0.617 ±0.051	0.593 ±0.065	0.564 ±0.053

Table 5.4 Estimates for the Relative Risk of Failure in CardioGard versus Control for Varying Launch Scenarios and Analytical Approaches

	Relative Risk (95% Credible Interval)						
	Concurrent	All	TTP1	TTP2	PPS	PPD	MEM
Actual	1.04 (0.85, 1.25)	1.03 (0.85, 1.24)	1.03 (0.85, 1.24)	1.03 (0.85, 1.24)	1.03 (0.85, 1.25)	1.03 (0.85, 1.25)	1.03 (0.86, 1.22)
3 Month	1.01 (0.82, 1.24)	1.00 (0.81, 1.22)	1.00 (0.81, 1.22)	1.00 (0.81, 1.22)	1.01 (0.82, 1.23)	1.01 (0.81, 1.22)	1.01 (0.82, 1.20)
6 Month	1.08 (0.84, 1.37)	0.99 (0.80, 1.21)	0.99 (0.80, 1.21)	1.08 (0.84, 1.37)	1.03 (0.82, 1.28)	1.03 (0.81, 1.29)	1.04 (0.84, 1.25)
9 Month	1.17 (0.85, 1.59)	0.96 (0.74, 1.21)	1.17 (0.85, 1.59)	1.17 (0.85, 1.59)	1.02 (0.77, 1.31)	1.07 (0.78, 1.44)	1.13 (0.84, 1.49)

5.4 Conclusions

As neither of the Embol-X or CardioGard devices demonstrated efficacy for reducing post-operative infarcts after surgical AVR, the need for embolic protection in the AVR population remains. If new, alternative devices or procedures become available, a trial design similar to that of the CTSN neuroprotection trial could provide an efficient platform for evaluating the efficacy of multiple new experimental therapies. Simulation studies in Chapter 3 support the decision to not correct for multiple testing in the neuroprotection trial. As long as devices included in a new platform trial are unrelated and there is no intention to compare devices, a similar approach to multiple testing can and should be implemented for a more efficient design. Simulation studies in Chapter 4 demonstrated that using concurrent controls only in the analyses of experimental treatments is likely the best choice for confirmatory trials with staggered experimental arm entry. This approach avoids potential bias in estimates and is the only way to guarantee error control. However, as explored above, alternative approaches that incorporate all controls should be pre-specified as sensitivity analyses.

CHAPTER 6

Conclusion and Future Directions

Motivated by the recent CTSN neuroprotection trial, this work sought to examine two statistical issues in platform trials of multiple experimental treatments with a shared control. First, whether comparisons of each experimental treatment and the shared control need adjustment for multiplicity, and second, whether methods developed to incorporate historical controls in two-arm trials can be applied to incorporate non-concurrent controls in analyses of experimental treatments in open platform trials to maximize efficiency.

Simulation studies to evaluate the operating characteristics of three-arm platform designs, adjusted and not adjusted for multiple testing, relative to two independent two-arm trials were discussed in Chapter 3. These studies confirmed that FWER is comparable between a platform design that does not adjust for multiple testing and a set of equivalent, independent, two-arm trials. These studies also demonstrated that although multiple testing adjustments control FWER, they do not substantially decrease the conditional type I error rate of an experimental treatment being declared effective given another was also erroneously declared effective in a platform trial. If the trial design includes flexible features, such as arms entering at varying time points or exiting early for efficacy or futility, conditional type I error rates decrease as the overlap between the concurrent control groups of each experimental treatment decreases. In the scenario with the highest conditional error rates, a closed platform with no interim analyses, the marginal probability of both experimental treatments being declared effective simultaneously under the global null was minimal (~0.5%). Therefore, the drop in power from adjusting for multiple testing is likely not worth the modest improvement in conditional error rates. If the experimental treatments being evaluated in a platform trial are unrelated, multiple testing adjustments are unnecessary.

Chapter 4 explored five methods of incorporating non-concurrent controls into the analysis of an experimental treatment arm that opened after the launch of an open platform trial. The five approaches included one static approach (fixed power prior) and four dynamic approaches that allowed the similarity between non-concurrent and concurrent controls to determine the amount of information used from non-concurrent controls (two test-then-pool approaches, a dynamic power prior, and a MEM approach). These methods' performance across both inference on the null hypothesis that the later entry treatment was equivalent or worse than the control and estimation of the relative risk of treatment failure in the later entry arm compared to control were evaluated. Type I error, power, and bias of relative risk estimates under each of these methods were compared to the extremes of using concurrent controls only and naively pooling all controls under varying scenarios of parameter drift. Simulation results confirmed that if there is no parameter drift, naively pooling all control data yields the highest power, lowest type I error, and most precise, unbiased estimates compared to all other approaches. However, if event rates change over time, naive pooling results in type I error inflation or deflation depending on the direction of drift, as well as biased treatment effect estimates. Although superior to naive pooling, none of the five alternative approaches guarantee type I error control or unbiased estimates in the presence of drift. Thus, only concurrent controls should be used as comparators in the primary analysis of confirmatory studies. This conservative approach will guarantee type I error control and protect against bias. Even if investigators are confident that response rates will not change during the course of their trial, unanticipated protocol changes, logistical issues (such as clinical sites closing), or events external to the trial (such as the approval of a new therapy), make the assumption of fixed event rates unrealistic. However, approaches that incorporate all controls can be included as secondary sensitivity analyses in combination with an analysis of possible parameter drift.

Chapter 5 applied the findings of Chapters 3 and 4 to the CTSN neuroprotection trial. Based on the findings of Chapter 3, the decision to not adjust for multiple testing was appropriate. The devices evaluated by the trial were developed independently and captured debris by two different mechanisms. Adjusting the type I error would have required a larger sample size to maintain power and would not have substantially decreased conditional type I error rates. In re-analyses of the trial data, the choice of whether to use concurrent controls only, naively pool all controls, or use any of the five alternative approaches explored in Chapter 4 ultimately did not affect estimates as only a small number of non-concurrent controls were randomized. However, had the introduction of CardioGard occurred later, the use of concurrent controls only was the most appropriate and conservative approach. For future trials in the neuroprotection space, the design of the CTSN trial provides an efficient platform for evaluating the efficacy of multiple new experimental therapies.

When multiple experimental treatments are available, platform trials with a shared control group offer an efficient alternative to the “gold standard” two-arm, randomized controlled trial. While this work has elucidated the effects of adjusting versus not adjusting for multiple testing and evaluated approaches to incorporating all control data in the analyses of open platforms, there remain areas in need of future research. The sample sizes selected in both simulation studies were somewhat arbitrary. Additional studies assessing the impact of smaller and larger sample sizes should be explored. Extensions of the binary outcome studies conducted in Chapter 3 and 4 to continuous and time-to-event outcomes should also be explored. Further, although we recommend the use of concurrent data only in confirmatory analyses, alternative approaches that leverage all control data may still be appropriate in earlier phase studies. Simulation studies exploring the use of these methods and incorporating interim analyses with futility and efficacy-stopping rules should also be conducted.

BIBLIOGRAPHY

1. Thomas DW BJ, Audette J, Carroll A, Dow-Hygelung C, Hay M. *Clinical Development Success Rates 2006-2015. A BIO Industry Analysis* 2016.

2. Golub HL. The need for more efficient trial designs. *Stat Med* 2006; 25: 3231-3235. DOI: 10.1002/sim.2629.

3. Orloff J, Douglas F, Pinheiro J, et al. A GUIDE TO DRUG DISCOVERY - OPINION The future of drug development: advancing clinical trial design. *Nat Rev Drug Discov* 2009; 8: 949-957. DOI: 10.1038/nrd3025.

4. Selim M. Current concepts - Perioperative stroke. *New Engl J Med* 2007; 356: 706-713. DOI: DOI 10.1056/NEJMra062668.

5. Mack MJ, Acker MA, Gelijns AC, et al. Effect of Cerebral Embolic Protection Devices on CNS Infarction in Surgical Aortic Valve Replacement: A Randomized Clinical Trial. *JAMA* 2017; 318: 536-547. 2017/08/09. DOI: 10.1001/jama.2017.9479.

6. Parmar MKB, Carpenter J and Sydes MR. More multiarm randomised trials of superiority are needed. *Lancet* 2014; 384: 283-284. DOI: Doi 10.1016/S0140-6736(14)61122-3.

7. James ND, Sydes MR, Clarke NW, et al. Systemic therapy for advancing or metastatic prostate cancer (STAMPEDE): a multi-arm, multistage randomized controlled trial. *BJU Int* 2009; 103: 464-469. 2008/11/08. DOI: 10.1111/j.1464-410X.2008.08034.x.

8. Barker AD, Sigman CC, Kelloff GJ, et al. I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Clin Pharmacol Ther* 2009; 86: 97-100.

9. Saville BR and Berry SM. Efficiencies of platform clinical trials: A vision of the future. *Clin Trials* 2016; 13: 358-366. DOI: 10.1177/1740774515626362.

10. Dunnett CW. Selection of the best treatment in comparison to a control with an application to a medical trial. *Design of Experiments: Ranking and Selection: Essays in Honor of Robert E Bechhofer*. New York: Marcel Dekker, 1984.

11. Thall PF, Simon R and Ellenberg SS. 2-Stage Selection and Testing Designs for Comparative Clinical-Trials. *Biometrika* 1988; 75: 303-310. DOI: Doi 10.2307/2336178.

12. Stallard N and Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Stat Med* 2003; 22: 689-703. DOI: 10.1002/sim.1362.

13. Stallard N and Friede T. A group-sequential design for clinical trials with treatment selection. *Stat Med* 2008; 27: 6209-6227. DOI: 10.1002/sim.3436.

14. Kelly PJ, Stallard N and Todd S. An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *J Biopharm Stat* 2005; 15: 641-658.

15. Magirr D, Jaki T and Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* 2012; 99: 494-501.

16. Jaki T and Magirr D. Considerations on covariates and endpoints in multi-arm multi-stage clinical trials selecting all promisingtreatments. *Stat Med* 2013; 32: 1150-1163.

17. Cheung YK. Simple sequential boundaries for treatment selection in multi-armed randomized clinical trials with a control. *Biometrics* 2008; 64: 940-949.

18. Follmann DA, Proschan MA and Geller NL. Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* 1994; 50: 325-336. 1994/06/01.

19. Hobbs BP, Chen N and Lee JJ. Controlled multi-arm platform design using predictive probability. *Stat Methods Med Res* 2018; 27: 65-78. DOI: 10.1177/0962280215620696.

20. Elm JJ, Palesch YY, Koch GG, et al. Flexible Analytical Methods for Adding a Treatment Arm Mid-study to an Ongoing Clinical Trial. *J Biopharm Stat* 2012; 22: 758-772. DOI: 10.1080/10543406.2010.528103.

21. Ventz S, Alexander BM, Parmigiani G, et al. Designing Clinical Trials That Accept New Arms: An Example in Metastatic Breast Cancer. *J Clin Oncol* 2017; 35: 3160-+. DOI: 10.1200/Jco.2016.70.1169.

22. Dunnett CW. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *J Am Stat Assoc* 1955; 50: 1096-1121.

23. Demets DL and Lan KKG. Interim Analysis - the Alpha-Spending Function-Approach. *Stat Med* 1994; 13: 1341-1352.

24. Levin B and Robbins H. Selecting the Highest Probability in Binomial or Multinomial Trials. *P Natl Acad Sci USA* 1981; 78: 4663-4666.

25. Lee JJ and Chu CT. Bayesian clinical trials in action. *Stat Med* 2012; 31: 2955-2972.

26. Berry SM, Petzold EA, Dull P, et al. A response adaptive randomization platform trial for efficient evaluation of Ebola virus treatments: A model for pandemic response. *Clin Trials* 2016; 13: 22-30. DOI: 10.1177/1740774515621721.

27. Brittain EH and Proschan MA. Comments on Berry et al.'s response-adaptive randomization platform trial for Ebola. *Clin Trials* 2016; 13: 566-567. DOI: 10.1177/1740774516654440.

28. Wason JMS, Stecher L and Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials* 2014; 15. DOI: Artn 364

10.1186/1745-6215-15-364.

29. Freidlin B, Korn EL, Gray R, et al. Multi-arm clinical trials of new agents: Some design considerations. *Clin Cancer Res* 2008; 14: 4368-4371. DOI: 10.1158/1078-0432.Ccr-08-0325.

30. Proschan MA and Follmann DA. Multiple Comparisons with Control in a Single Experiment Versus Separate Experiments - Why Do We Feel Differently. *Am Stat* 1995; 49: 144-149. DOI: Doi 10.2307/2684628.

31. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018.

32. Snow G. blockrand: Randomization for block random clinical trials. R package version 1.3. 2013.

33. Sydes MR, Parmar MKB, James ND, et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. *Trials* 2009; 10. DOI: Artn 39

10.1186/1745-6215-10-39.

34. Yuan Y, Guo BB, Munsell M, et al. MIDAS: a practical Bayesian design for platform trials with molecularly targeted agents. *Stat Med* 2016; 35: 3892-3906. DOI: 10.1002/sim.6971.

35. Kaizer AM, Hobbs BP and Koopmeiners JS. A multi-source adaptive platform design for testing sequential combinatorial therapeutic strategies. *Biometrics* 2018; 74: 1082-1094. DOI: 10.1111/biom.12841.

36. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat* 2014; 13: 41-54. DOI: DOI 10.1002/pst.1589.

37. Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov* 2006; 5: 27-36. DOI: 10.1038/nrd1927.

38. US Food and Drug Administration. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. 2010.

39. US Food and Drug Administration. Guidance for Industry Expedited Programs for Serious Conditions – Drugs and Biologics. 2017.

40. Lewis CJ, Sarkar S, Zhu JW, et al. Borrowing From Historical Control Data in Cancer Drug Development: A Cautionary Tale and Practical Guidelines. *Stat Biopharm Res* 2019; 11: 67-78. DOI: 10.1080/19466315.2018.1497533.

41. Gamalo-Siebers M, Savic J, Basu C, et al. Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation. *Pharm Stat* 2017; 16: 232-249. DOI: 10.1002/pst.1807.

42. Taylor JM, Braun TM and Li ZG. Comparing an experimental agent to a standard agent: relative merits of a one-arm or randomized two-arm Phase II design. *Clin Trials* 2006; 3: 335-348. DOI: 10.1177/1740774506070654.

43. Grossman SA, Schreck KC, Ballman K, et al. Point/counterpoint: randomized versus single-arm phase II clinical trials for patients with newly diagnosed glioblastoma. *Neuro-Oncology* 2017; 19: 469-474. DOI: 10.1093/neuonc/nox030.

44. Pocock SJ. Combination of Randomized and Historical Controls in Clinical-Trials. *J Chron Dis* 1976; 29: 175-188. DOI: Doi 10.1016/0021-9681(76)90044-8.

45. Ibrahim JG, Chen MH, Gwon YJ, et al. The power prior: theory and applications. *Stat Med* 2015; 34: 3724-3749. DOI: 10.1002/sim.6728.

46. Duan YY, Ye KY and Smith EP. Evaluating water quality using power priors to incorporate historical information. *Environmetrics* 2006; 17: 95-106. DOI: 10.1002/env.752.

47. Hobbs BP, Carlin BP, Mandrekar SJ, et al. Hierarchical Commensurate and Power Prior Models for Adaptive Incorporation of Historical Information in Clinical Trials. *Biometrics* 2011; 67: 1047-1056. DOI: 10.1111/j.1541-0420.2011.01564.x.

48. Hobbs BP, Sargent DJ and Carlin BP. Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. *Bayesian Anal* 2012; 7: 639-673. DOI: 10.1214/12-Ba722.

49. Schmidli H, Gsteiger S, Roychoudhury S, et al. Robust Meta-Analytic-Predictive Priors in Clinical Trials with Historical Control Information. *Biometrics* 2014; 70: 1023-1032. DOI: 10.1111/biom.12242.

50. Neuenschwander B, Roychoudhury S and Schmidli H. On the Use of Co-Data in Clinical Trials. *Stat Biopharm Res* 2016; 8: 345-354. DOI: 10.1080/19466315.2016.1174149.

51. Kaizer AM, Koopmeiners JS and Hobbs BP. Bayesian hierarchical modeling based on multisource exchangeability. *Biostatistics* 2018; 19: 169-184. 2017/10/17. DOI: 10.1093/biostatistics/kxx031.

52. Dejardin D, Delmar P, Warne C, et al. Use of a historical control group in a noninferiority trial assessing a new antibacterial treatment: A case study and discussion of practical implementation aspects. *Pharm Stat* 2018; 17: 169-181. DOI: 10.1002/pst.1843.

53. Gelman A, Carlin J, Stern H, et al. 2.1 Estimating a probability from binomial data. *Bayesian Data Analysis*. Third Edition ed. Boca Raton, FL: CRC Press, 2014.

54. *Adult Cardiac Surgery Database Executive Summary 10 Years - STS Period Ending 3/31/2016*. February, 2017. STS National Database.

55. Grover FL, Vemulapalli S, Carroll JD, et al. 2016 Annual Report of The Society of Thoracic Surgeons/American College of Cardiology Transcatheter Valve Therapy Registry. *Ann Thorac Surg* 2017; 103: 1021-1035. DOI: 10.1016/j.athoracsur.2016.12.001.

56. Messe SR, Acker MA, Kasner SE, et al. Stroke After Aortic Valve Surgery Results From a Prospective Cohort. *Circulation* 2014; 129: 2253-+. DOI: 10.1161/Circulationaha.113.005084.

57. Kapadia SR, Kodali S, Makkar R, et al. Protection Against Cerebral Embolism During Transcatheter Aortic Valve Replacement. *J Am Coll Cardiol* 2017; 69: 367-377. DOI: 10.1016/j.jacc.2016.10.023.

58. Leon MB, Smith CR, Mack MJ, et al. Transcatheter or Surgical Aortic-Valve Replacement in Intermediate-Risk Patients. *New Engl J Med* 2016; 374: 1609-1620. DOI: 10.1056/NEJMoa1514616.

59. Floyd TF, Shah PN, Price CC, et al. Clinically silent cerebral ischemic events after cardiac surgery: Their incidence, regional vascular occurrence, and procedural dependence. *Ann Thorac Surg* 2006; 81: 2160-2166. DOI: 10.1016/j.athoracsur.2006.01.080.

60. Alassar A, Soppa G, Edsell M, et al. Incidence and Mechanisms of Cerebral Ischemia After Transcatheter Aortic Valve Implantation Compared With Surgical Aortic Valve Replacement. *Ann Thorac Surg* 2015; 99: 802-808. DOI: 10.1016/j.athoracsur.2014.09.054.

61. Massaro A, Messe SR, Acker MA, et al. Pathogenesis and Risk Factors for Cerebral Infarct After Surgical Aortic Valve Replacement. *Stroke* 2016; 47: 2130-2132. DOI: 10.1161/Strokeaha.116.013970.