Maintaining a Strong Control of the Family-Wise Error Rate in Gatekeeping Group Sequential Designs with Primary and Delayed Secondary Endpoints: A Series of Simulation Studies

By:

Alessandra Iaboni

A biostatistics practicum report submitted to the
Department of Public Health Sciences in conformity with
the requirements for the degree of Master of Science.

Queen's University
Kingston, Ontario, Canada
September 2019

Supervisor: Dr. Keyue Ding

Abstract

Introduction: Group sequential designs provide the benefit of the potential for early stopping when implemented in clinical trials, but the use of interim analyses in these designs raises the issue of multiple comparisons. The assessment of multiple endpoints within a group sequential trial thus also faces the challenge of inflating the probability of observing a false positive result in one or more than one of the endpoints due to the multiple comparisons being made. This family-wise type I error (FWER) is further at risk of inflation when the secondary endpoint of the trial requires a larger sample size than the primary endpoint to detect the desired effect. These delayed endpoints are a subject that is underrepresented in the existing literature.

Objectives: To determine if error spending and error recycling approaches to designing group sequential trial boundaries can strongly control the FWER at a pre-specified nominal α level for gatekeeping procedures in the analysis of multiple endpoints with a delayed secondary endpoint. To assess the relative power levels between the different error spending and error recycling approaches which do strongly control the FWER.

Methods: Simulation studies were conducted by generating treatment and control endpoint values for the required sample sizes. This simulated data was repeatedly generated, and the proportion of type I errors was recorded. These simulations for FWER were repeated with a bias correction applied to the secondary endpoint's test statistic during analysis to assess if methods which previously did not strongly control the FWER could be made to control it. Power was also assessed by simulating the proportion of times a true treatment effect between trial groups was detected by the different testing procedures in both the primary and secondary endpoints.

Results: Before the bias correction, only the error spending scenarios designed for 80% secondary power achieved strong control of the FWER. After applying the bias correction, error spending scenarios designed for both 80% and 90% secondary power were strongly controlled. No error recycling methods were found to strongly control the FWER. Study power was determined to be highest in the error spending studies which had sample sizes originally designed for higher secondary power. There were no detectable differences in power levels between designs which used O'Brien Fleming primary boundaries compared with Pocock primary boundaries.

Conclusions: While these results illustrate the use of error spending designs in combination with a bias correction to analyze delayed secondary endpoint data in group sequential trials, a more general solution is needed to account for more combinations of possible treatment effects in then endpoints, desired power of the study, and correlation between endpoint response values.

Acknowledgements

Firstly, I would like to express my gratitude to my supervisor, Dr. Keyue Ding, for all of his expertise, guidance, and teachings during this practicum process. His patience and encouragement provided me this incredible opportunity to explore and learn about this exciting area of research and the more general field of clinical trials.

I would also like to thank the faculty and staff of the Department of Public Health Sciences at Queens University. Their expertise in the subject matter and unending willingness to assist has made contributions to my education and development throughout this program that have be invaluable.

I would also like to acknowledge my funding support from my supervisor, the Canadian Cancer Trials Group, and the Queen's Graduate Awards.

To my friends and classmates, I must thank them for all of their time spent learning alongside me and their support of me during this program. I have valued their friendships more than I can express and their contributions to my achievements by always pushing me to be better. I am sure that they will all accomplish extraordinary things moving forward their careers.

Finally, I would like to thank my family for their unending support and confidence in me and my goals over the duration of this degree, with particular thanks to my mother and aunt whose personal and professional achievements in their respective health care disciplines have served as sources of pride and motivation for me throughout my entire life. Without them serving as such incredible role models, I would not have been able to accomplish any of what I have today. They will forever be my biggest fans and my biggest inspiration.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

1.1 Multiple Endpoints in Clinical Trials

When establishing efficacy of new treatments through clinical trials, investigators are typically interested in the differences between new and existing therapies for multiple endpoints [5]. Each endpoint can represent a different way that a single therapy may affect a patient's physical, mental or emotional wellbeing. These endpoints of interest could be considered co-primary or primary and secondary endpoints depending on their relative importance to the success of the treatment. Co-primary endpoints could be used if a treatment was developed with the intent of treating two or more symptoms of a similar underlying condition. Secondary endpoints can sometimes be considered additional claims, besides the original purpose for the treatment's development, which could be noted on packaging or drug labels [5].

Typically, the significance of the treatment on secondary endpoints is relevant only if the treatment is also effective on the primary endpoint, for which the therapy was initially designed. In the case of clinical trials assessing the benefits of cancer treatments, usually the primary outcome in studies is overall survival (OS) and another outcome frequently considered is patient's quality of life (QoL), which is a standard level that consists of the expectations of an individual or society for a good life [5]. These two endpoints well fit the traditional roles of primary and secondary endpoints respectively in which OS is of great interest, while the QoL reflects the patient's overall well-being in their prolonged life.

From situations such as this, gatekeeping methodologies for study designs were developed in which significance of a test of the primary endpoint acts as a gate for the testing of secondary endpoints [2]. In the case of multiple secondary endpoints (or tertiary endpoints), significance can be assessed in a similar fashion with significant tests of each secondary endpoint acting as a gate for the next endpoint in a pre-determined sequence [2]. If at any point an endpoint is determined not to have a significant effect from treatment (or to have a significant opposite effect from the direction expected), then testing ceases and any endpoints for which the non-significant test was a gatekeeper will not be tested [2].

When multiple endpoints are being considered in a clinical trial, the typical measurements of validity of the study must also be tweaked in order to include data collected for both endpoints. Sample size calculations for study design, as well as significance level used during study analyses, are both based on what investigators would consider an acceptable rate of type I errors for a study. When multiple endpoints are being investigated, it is important to control, not only the chances of observing a false positive result in each endpoint separately, but the in the overall study. This introduces the concept of the family-wise type I error rate (FWER):

$$FWER = P\{Reject\ at\ least\ one\ true\ null\ H_i\ (i = 1,2)\} \leq \alpha \ [11]$$

An important issue of testing multiple endpoints within a clinical trial is ensuring that the probability of detecting a type I error in any of the outcomes being studied is less than a pre-

specified nominal α level. If the FWER for a particular set of study analyses is always less than α, it is said that the analysis maintains strong control of the FWER [1].

1.2 Use of Group Sequential Testing in Clinical Trials

A good clinical trial design should permit early stopping as soon as the treatment effect on the primary endpoint becomes clear [7]. This could mean clear evidence that the experimental treatment is better than the treatment being received by the control group. It would then be considered unethical to continue the trial and withhold the new treatment from all patients. However, the other possibility is that the experimental treatment displays such poor results that it is extremely unlikely the remaining study participants could display outcomes positive enough to produce an overall positive and significant result for the trial. The study would then be stopped for futility. New experimental treatments could also benefit from periodic re-evaluation for concerns about patient safety when little is known about tolerability of the drug [5]. Group sequential designs were introduced to allow one or more multiple looks at the accumulating data during the duration of the trial [7]. These looks, based on the cumulative evidence, can be used to claim efficacy or futility of the study treatment earlier than at the study's planned final analysis [7]. This is an important possibility for researchers since conducting studies is highly expensive as well as time and resource intensive [7].

For testing of multiple endpoints while controlling the FWER, a class of flexible testing strategies called group sequential designs can be combined with gatekeeping procedures to incorporate interim analyses into the testing of all endpoints of interest (primary and secondary) [5,11]. The time points at which interim analyses are conducted are referred to as information times. Denoted by t_{ij} (for i ranging from 1 to the number of endpoints and j ranging from one to the number of possible analysis times), the information time is defined as the proportion of the required (or fixed) sample size to detect the expected effect at the pre-defined significance and power levels [10]. Possible information times are between values of 0 and 1. Thus, when testing of multiple endpoints is designed as a group sequential design, there are different strategies of implementing hierarchical tests. Consider a possible study with a single interim analysis for the primary endpoint and the secondary endpoint. The two main hierarchical strategies which can be used are:

1) Stagewise hierarchical [2]

   At the time of the primary endpoint's interim analysis, denoted t_{11}, the primary hypothesis is tested against a critical value of c_{11}. If the standardized test statistic (denoted T_{ij}) is found to be significant, $T_{11} > c_{11}$, then the secondary hypothesis is tested at the same time. If $T_{21} > c_{21}$, then it can be claimed that there is a treatment effect on both endpoints. If $T_{21} \leq c_{21}$ then the trial would conclude there is a treatment effect on the primary endpoint only and the trial would stop.

   If $T_{11} \leq c_{11}$, the testing would continue on to the primary endpoint at its fixed sample size where the primary hypothesis would be tested at time t_{12} for $T_{12} > c_{12}$. If this is true, a

treatment effect can be claimed for the primary endpoint and the secondary hypothesis will be tested at time t_{22}. If $T_{22} > c_{22}$, then a treatment effect can be claimed on the secondary endpoint as well; otherwise only a treatment effect on the primary endpoint will be claimed.

If $T_{12} \leq c_{12}$ then the trial would conclude having not rejected either the primary or secondary hypothesis.

2) Overall hierarchical [2]

The same methodology is used as in 1), however the trial does not end if the secondary endpoint, tested at interim analysis time t_{21}, is not significant. If $T_{21} \leq c_{21}$, then the trial will continue to the final analysis and the secondary hypothesis will still be tested at time t_{22} for $T_{22} > c_{22}$ to see if a treatment effect can be claimed for the secondary endpoint.

These strategies could be extended to include more than two endpoints or more interim analysis times as desired.

The popularity of including interim analyses in study design stems from the difficulty of planning a clinical trial well [5]. Interim analyses aid in the potential need to assess the validity of nuisance parameters such as minimum detectable treatment effect, population variance, and expected event rates which come from the "best guesses" of the investigators when planning the study [5]. Due to the difficulty in accurately selecting these parameters, assessment of the study results before the full sample size has accrued is valuable as potential early stopping criteria may result in saved time and resources.

1.3 Study Design Methodologies

1.3.1 Error Spending Boundaries

Two of the most widely utilized methods for analyzing data in group sequential trials are through the use of Pocock boundaries [9] and O'Brien Fleming boundaries [8]. These methods were developed to provide different means of allocating the type I error allotted for the total analysis of a single endpoint across the pre-planned interim analysis times. That is, using the critical boundary values from these types of boundaries, analysis of the study data as it accumulates will still maintain strong control of the overall probability of observing a false positive result. Pocock boundaries are designed so that, at each analysis time, the critical values against which test statistics are compared remain constant for all information times [9]. Dissimilarly, O'Brien Fleming boundaries are designed with almost all of the pre-specified amount of type I error allocated to the final analysis and very little allocated to the interim analyses [8]. This results in the need for a very strong association to be present in order for the hypothesis to be rejected before collecting the full required sample size, resulting in greater power for the analysis compared to using Pocock boundaries of the same sample size [5].

When implementing these types of boundary definitions for a primary and a secondary endpoint, the Lan-DeMets alpha spending function approximations of the original Pocock and O'Brien Fleming boundaries are often used [7]. The Lan-DeMets method was developed to provide a more

flexible process for determining group sequential boundary values that did not require analysis information times to be known in advance of conducting the first analysis [3]. Using this technique, Lan-DeMets versions of the previously used Pocock and O'Brien Fleming boundaries can be found which well-approximate the original boundary types, but do not find it necessary to pre-specify the interim analysis times [3].

1.3.2 Error Recycling Methodologies

Another type of design which can be used to control type I error when testing multiple endpoints is error recycling designs. In these methodologies, Bonferroni-based procedures for multiple testing are implemented to strongly control the FWER at a pre-specified α level [6]. The test mass (or total value of α) can be split among k endpoints to be tested, resulting in respective type I error probabilities $\alpha_1, \dots, \alpha_k$ [6]. After testing the endpoints in a pre-determined sequence, test mass from endpoints which have been determined to be significant at their respective allocated significance levels, can be reused (or recycled) to increase the significance levels allotted to the tests of other endpoints [6]. When the test mass allocations are determined ahead of the analyses, as well as which tests may recycle mass to which subsequent tests (and in what proportion), then these methods have been proven to strongly control the FWER in multiple testing scenarios [6]. Application of these testing procedures to group sequential trials can be performed by altering the amount of test mass able to be recycled to a secondary endpoint based on the information time at which the primary endpoint was deemed significant [6].

1.4 Current Challenges

The literature surrounding multiple testing procedures has gained traction in the last decade. Papers involving the feasibility of implementing gatekeeping procedures while still maintaining strong control of the FWER have investigated the issue in multiple ways, with each publication expanding on the knowledge from the last.

In 2007, Hung et. al. determined that, for a stagewise hierarchical strategy, the naïve approach of testing the secondary endpoint (at the same time as the primary hypothesis is rejected) using a single boundary value of z_α at the full α level would result in an inflated FWER [5]. They also found that increasing the sample size of the secondary endpoint's analysis time past that of the primary endpoint results in inflation of the FWER [5]. While they proposed a conservative approach of using the α level at which the primary hypothesis was rejected for analysis of the secondary endpoint, their recommendation was not employed in future works [5]. Instead, in 2010, Glimm et.al tested their own hypothesis that the Bonferroni approach to testing secondary endpoints was too conservative for these multiple comparisons and posed the question of whether the use of an alpha spending boundary would be better for the analysis of a secondary endpoint [2]. At around the same time, Tamhane et.al. assessed this problem by testing all four combinations of O'Brien Fleming and Pocock boundaries for the primary and secondary endpoint analysis for strong control of the FWER and found all combinations to be well controlled with power increased for combinations that used O'Brien Fleming boundaries [11]. However, this work was restricted to a single interim analysis for each endpoint, and analyses

had to be conducted at common information times [11]. Works in the last two years have focused on these exact problems with a 2018 publication by Tamhane et. al.[10] assessing the possibility of more than two looks at the same information times and a 2019 publication from Gou and Xi [4] investigating different information times of the interim analyses for different endpoints.

The predominant strategy used in the investigation of these issues in the literature is a stagewise hierarchical strategy for testing multiple endpoints (as defined above) [5,10]. It is of interest, however, how these group sequential trial methods perform with the use of the overall hierarchical strategy. This strategy allows more opportunities to analyze the secondary hypothesis since the trial would continue until accrual of the sample size required to power testing of the secondary endpoint is completed. Since this has not been thoroughly investigated, it is unknown how the different strategies differently control error rates in gatekeeping procedures for such group sequential designs [4,5,10,11].

As a culmination of all these works, there is still uncertainty as to whether more complex applications of these problems would have strong control of the FWER, such as more than one possible information time for an interim analysis or the need for a larger sample size in the secondary endpoint resulting in a delayed final analysis time. The previously investigated topics can be amalgamated for the formation of this new problem in which multiple interim looks must be considered, at information times which vary between endpoints, while the secondary endpoint requires a larger sample size to detect its expected effect than is needed for the primary endpoint, with an overall hierarchical gatekeeping strategy. This could be motivated by the potential for secondary endpoints to exhibit more subtle effects from treatment than are expected from the primary endpoint requiring larger sample sizes to achieve the same power. Though Hung et.al. assessed the control of the FWER for delayed secondary endpoints, they concluded that extending the sample size needed for the secondary endpoint past that of the first resulted in inflated type I error rates with no mention of a means to correct this problem [5].

Finally, there has been little work done on the feasibility of implementing error recycling methodologies alongside error spending boundary approaches, making the error recycling scenarios to be considered in this report novel concepts.

1.5 Study Objectives

The series of simulations studies in this report aim to determine which methods of implementing gatekeeping procedures in group sequential trials for a primary and a delayed secondary strongly control the FWER. Assessing control of the FWER will include investigating scenarios which implement traditional α level group sequential boundaries, as well as scenarios with boundaries based on α-recycling methods. If the tested scenarios do not exhibit strong control of the FWER at a pre-specified α level, then a bias correction will be derived to decrease, and ultimately control, the FWER for those scenarios not initially controlled by the group sequential boundaries. Upon determination of which scenarios are found to strongly control the FWER, the

power of these study designs (ability to detect true treatment effects) will be assessed as a means of selecting more favourable designs for use in practice.

2.0 Methods

The study being investigated in this report for strong control of the FWER is a gatekeeping procedure for which a secondary endpoint will only be tested upon finding significance in a related primary endpoint. This set-up is motivated by a situation in which a primary endpoint, such as OS, is expected to have a larger treatment effect from the study's intervention than the expected treatment effect of the same intervention on the secondary endpoint, such as QoL. These treatment effects could differ due to the nature of the endpoint, the biological mechanism of the intervention, or the increased time required to observe the necessary number of events in the study [5]. As a result, the primary endpoint will require a smaller sample size to significantly detect the effect compared to the secondary endpoint, with intended type I error and power levels held constant. The secondary endpoint is then considered to be delayed due to the necessity of accruing more participants than for the primary endpoint before the final analysis can be completed. This inflated sample size for the delayed secondary endpoint is where the potential for uncontrolled type I error rates stems from, as demonstrated in the works of Hung et.al. [5]. As the sample size for the secondary endpoint increases, the FWER also increases farther from the intended α level [5].

These studies will be designed such that the response variables are normally distributed. It has been demonstrated that methods developed with normal data can then be extended to other types of outcome data, such as binary endpoints and time to event endpoints, through asymptotic results and approximations derived from the central limit theorem [10].

2.1 Study Design

Two different study design methods will be tested for their control over the FWER: error spending and error recycling designs. Boundaries for both methods will be calculated using the Lan-DeMets error spending function to approximate the Pocock and O'Brien Fleming boundary definitions [3,8,9]. The primary endpoint will be tested for an expected treatment effect of 0.5 while the secondary endpoint will be tested for an expected treatment effect of 0.35, leading it to require a larger sample size and be considered "delayed" as the trial will need to continue to accrue patients after the sample size needed for the primary endpoint has been reached.

Interim analysis times are defined such that the sample size at the primary interim analysis corresponds to the sample size at the first possible interim analysis for the secondary endpoint, and the sample size at the primary final analysis corresponds to the sample size at the second possible interim analysis for the secondary endpoint. Because the sample sizes for the primary and secondary endpoints differ, the interim analyses for the secondary endpoint are designed to occur at the same number of patients as the primary endpoint analyses; this will lead to identical sample sizes but different information times for the analyses of each of the two endpoints.

Overall hierarchical strategies will be used for analysis in these simulation studies. When working with a delayed secondary endpoint as described above, there is no value in using the stagewise hierarchical strategy since the "delay" would never be reached if the study terminated at the maximum sample size of the primary endpoint. That is, the sample size required for the secondary endpoint to detect its expected effect will be accrued regardless of the time at which the primary endpoint is found to be significant.

2.1.1 Group Sequential Trial Scenario Designs

For the test scenarios in which both endpoints are designed from error spending approaches, a single type I error rate is chosen. Boundaries are then determined according to the different information times at which an analysis could occur for each endpoint. Primary endpoints can be tested at one interim analysis and at the final analysis. Secondary endpoints can only be tested at one interim analysis and then again at the final analysis. Which interim analysis time is used depends on which primary analysis time results in then rejection of the primary endpoint.

Scenario 1A: O'Brien Fleming – O'Brien Fleming Boundaries with 80% secondary power

The study of the primary endpoint will be designed as a one-sided 2.5% significance level test with 90% power for a target effect of 0.5. One interim analysis will be performed at an information time of $2/3$, or when two thirds of the events have occurred, to ensure 50% power at the interim analysis. With an O'Brien Fleming boundary type, this requires a sample size of 170 patients (85 per group).

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.67	114	2.5093	0.00605	0.00605
1	170	1.9929	0.01895	0.02500

The study of the secondary endpoint will be designed as a one-sided 2.5% significance level test with 80% power for a target effect of 0.35. The boundary will be designed for two interim analyses corresponding to the same sample sizes as the analyses in the primary study. For example, 67% of 170 patients is equal to 43% of 262 patients and 65% of 262 patients is the full sample size of 170 patients for the primary endpoint. Adjusting for two interim analyses with an O'Brien Fleming boundary, a sample size of 262 patients (131 per group) is required.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.43	114	3.2140	0.00065	0.00065
0.65	170	2.5636	0.00518	0.00583
1	262	1.9890	0.01917	0.02500

Scenario 1B: Pocock – O'Brien Fleming Boundaries with 80% secondary power

The study of the primary endpoint will be designed as a one-sided 2.5% significance level test with 90% power for a target effect of 0.5. One interim analysis will be performed at 0.6 information time, or when 60% of the events have occurred, to ensure approximately 70% power

at the interim analysis. With a Pocock boundary type, this requires a sample size of 186 patients (93 per group).

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.60	112	2.1035	0.01771	0.01771
1	186	2.2305	0.00729	0.02500

The study of the secondary endpoint will be designed identically to that in Scenario 1A. A sample size of 262 patients (131 per group) will be required to adjust for two interim analyses with O'Brien Fleming boundary type. Only the information times of the interim analyses will change due to the increased sample size from (170 to 186): the secondary interim analyses will be conducted at collection of 43% of events, or at 71% of events, followed by the final analysis.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.43	112	3.2417	0.00059	0.00059
0.71	186	2.4270	0.00761	0.00821
1	262	2.0018	0.01679	0.02500

Scenario 1C: O'Brien Fleming – O'Brien Fleming Boundaries with 90% secondary power

The study of the primary endpoint will be designed identically to that in Scenario 1A.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.67	114	2.5093	0.00605	0.00605
1	170	1.9929	0.01895	0.02500

The study of the secondary endpoint will be designed as a one-sided 2.5% significance level test with 90% power for a target effect of 0.35, requiring a sample size of 350 patients (175 per group). The boundary will be designed for two interim analyses (with an O'Brien Fleming boundary) corresponding to the analysis times in the study of the primary endpoint.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.32	114	3.2900	0.00050	0.00050
0.49	170	3.0181	0.00127	0.00177
1	350	1.9674	0.02323	0.02500

Scenario 1D: Pocock – O'Brien Fleming Boundaries with 90% secondary power

The study of the primary endpoint will be designed identically to that in Scenario 1B.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.60	112	2.1035	0.01771	0.01771
1	186	2.2305	0.00729	0.02500

The study of the secondary endpoint will be designed as one-sided 2.5% significance level with 90% power for a target effect of 0.35, requiring a sample size of 350 patients (175 per group) This is adjusting for two interim analysis (with an O'Brien Fleming boundary) corresponding to the analysis times in the study of the primary endpoint.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.32	112	3.2900	0.00050	0.00050
0.53	186	2.8654	0.00208	0.00258
1	350	1.9716	0.02242	0.02500

Alterations in study power levels do not affect the boundary calculations, but do affect the sample size calculations, and the degree to which the secondary endpoint is delayed [7]. These similarities and differences can be seen by comparing the results of scenarios 1A with 1C and scenarios 1B with 1D.

2.1.2 Error Recycling Scenarios Designs

For error recycling scenarios, the type I error used is defined by the results of the primary endpoint's analysis. When the primary endpoint is rejected at its interim analysis, the full value of α is able to be recycled to the test of the secondary endpoint, resulting in α=0.025 for this investigation. Analyses following the primary endpoint's final analysis will be implement two different methods of determining the recycled test mass amount.

The first two error recycling scenarios will have the alpha level of their secondary test based on the significance level of the test of the primary endpoint at its final analysis.

Scenario 2A: O'Brien Fleming – O'Brien Fleming Boundaries

O'Brien Fleming boundary types will be used for both the primary and secondary decision boundaries. The study of the primary endpoint will be designed as a one-sided 2.5% significance level test with 90% power for a target effect of 0.5. One interim analysis will be performed at 2/3 information time, or when two thirds of the events have occurred, to ensure 50% power at the interim analysis. This requires a sample size of 170 patients (85 per group) adjusting for one interim analysis.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.67	114	2.5093	0.00605	0.00605
1	170	1.9929	0.01895	0.02500

If the primary hypothesis is rejected at the interim analysis, the tests for the secondary endpoint will be designed by recycling of the full type I error. This is a one-sided 2.5% significance level test with 80% power to detect an effect of 0.35. Adjusting an O'Brien Fleming boundary for one interim analysis (at the time of the primary interim analysis) and the final secondary analysis, we require a sample size of 260.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.44	114	3.2003	0.00069	0.00069
1	260	1.9639	0.02431	0.02500

If the primary hypothesis is rejected at the final analysis, the tests for the secondary endpoint will be designed by recycling the test mass found from the significance level of the critical value for

the primary test's final analysis ($P[Z \geq 1.9929] = 0.0231$). This test is designed as a one-sided 2.31% significance level test with 80% power to detect an effect of 0.35. Adjusting an O'Brien Fleming boundary for one interim analysis requires a sample size of 266.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.64	170	2.6122	0.00450	0.00450
1	266	2.0192	0.02050	0.02310

Scenario 2B: Pocock – O'Brien Fleming Boundaries

The study of the primary endpoint will be designed as a one-sided 2.5% significance level test with 90% power for a target effect of 0.5. One interim analysis will be performed at 0.6 information time to ensure 50% power at the interim analysis. With a Pocock boundary type, this requires a sample size of 186 patients (93 per group).

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.60	112	2.1035	0.01771	0.01771
1	186	2.2305	0.00729	0.02500

Rejecting the primary analysis at the interim analysis allows for recycling of the full $\alpha = 0.025$. The results in a secondary test designed the same as the first, secondary test in Scenario 2A requiring a sample size of 260.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.43	112	3.2279	0.00062	0.00062
1	260	1.9636	0.02438	0.02500

Rejecting the primary test at the final analysis will result in a secondary test at 1.286% significance level (recycling 1.286% of the test as $P[Z \geq 2.2305] = 0.01286$) with 80% power to detect an effect of 0.35. Adjusting for one interim analysis requires a sample size of 312.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.60	186	3.0178	0.00127	0.00127
1	312	2.2432	0.01158	0.01286

The final two error recycling scenarios will base the significance level of the secondary tests on the type I error spent on the primary endpoint at the final analysis.

Scenario 2C: O'Brien Fleming – O'Brien Fleming Boundaries

The study of the primary endpoint will be designed identically to that in Scenario 2A.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.67	114	2.5093	0.00605	0.00605
1	170	1.9929	0.01895	0.02500

Rejecting the primary analysis at the interim analysis allows for recycling of the full $\alpha = 0.025$. This results in a secondary test designed the same as the first secondary test in Scenarios 2A and 2B requiring a sample size of 260.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.44	114	3.2003	0.00069	0.00069
1	260	1.9639	0.02431	0.02500

When the primary hypothesis is rejected at the final analysis, the alpha spent at the final analysis will be recycled for the secondary endpoint. This secondary test will be a one-sided 1.895% test with 80% power to detect an effect of 0.35. Adjusting for one interim analysis, a sample size of 282 is required.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.60	170	2.8058	0.00251	0.00251
1	282	2.0936	0.01644	0.01895

Scenario 2D: Pocock – O'Brien Fleming Boundaries

The study of the primary endpoint will be designed identically to that in Scenario 2B.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.60	112	2.1035	0.01771	0.01771
1	186	2.2305	0.00729	0.02500

Rejecting the primary analysis at the interim analysis allows for recycling of the full $\alpha = 0.025$. The results in a secondary test designed the same as the first, secondary test in the previous error recycling scenarios requiring a sample size of 260.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.43	112	3.2279	0.00062	0.00062
1	260	1.9636	0.02438	0.02500

When the primary hypothesis is rejected at the final analysis, the alpha spent at the final analysis will be recycled for the secondary endpoint. This secondary test will be a one-sided 0.73% significance level test with 80% power to detect an effect of 0.35. Adjusting for one interim analysis, a sample size of 356 is required.

Information Time	Sample Size	Critical Value	Alpha Spent	Cumulative Alpha Spent
0.52	186	3.2900	0.00050	0.00050
1	356	2.4463	0.00679	0.00729

2.1.3 Sample Size Calculations

Samples size calculations were conducted in accordance with the methods of Jennison and Turnbull's works on designing group sequential trials [7]. This method applies an inflation factor

$R(K, \alpha, \beta)$ to the number obtained from a standard sample size calculation to accommodate for interim analyses done during the study [7]. The size of the inflation factor is dependent on the desired power of the study, type I error of the study, number of interim analyses, and type of boundary being used (O'Brien Fleming or Pocock) [7].

Since this reference does not provide a full range of inflation constants for all combinations of type I error and power levels, a conservative approach was implemented in which the larger of the two boundary inflation factors was used when the desired values fell between the values provided. That is, to calculate a sample size for an O'Brien Fleming boundary, adjusted for 2 interim analyses, with 80% power and $\alpha = 0.025$, it is only known that the inflation factors are $R(K, \alpha, \beta)$ = 1.007 for $\alpha = 0.01$ and $R(K, \alpha, \beta)$ = 1.017 for $\alpha = 0.05$ [7]. In this situation, the inflation factor of 1.017 would be used to ensure the study is not underpowered.

To calculate samples sizes for the simulations with error spending approaches, two sample size calculations were conducted: one for the primary endpoint and one for the secondary endpoint. The sample size for the primary endpoint is adjusted for one interim analysis, while the sample size for the secondary endpoint is adjusted for two interim analyses.

Below is the sample size calculation for scenario 1A, in which both endpoints will be tested with O'Brien Fleming boundaries and 80% power with $\alpha = 0.025$.

> ***Primary Endpoint:*** $\alpha = 0.05$ *(one-sided 0.025 boundary),* $\beta = 0.2, \delta = 0.5, \sigma^2 = 1$
>
> *This requires a fixed sample size of 85 people per group, or 170 total. Adjusting for one interim analysis using an O'Brien Fleming boundary requires an inflation factor of:* $85 \times R(K, \alpha, \beta) = 85 \times 1.007 \approx 85$ *per group or 170 patients total.*
>
> ***Secondary Endpoint:*** $\alpha = 0.05$ *(one-sided 0.025 boundary),* $\beta = 0.2, \delta = 0.35, \sigma^2 = 1$
>
> *This requires a fixed sample size of 129 people per group, or 258 total. Adjusting for two interim analyses using an O'Brien Fleming boundary requires an inflation factor of:* $129 \times R(K, \alpha, \beta) = 129 \times 1.017 \approx 131$ *per group or 262 patients total*

To calculate sample sizes for the error recycling deigns, three sample size calculations were conducted: one for the primary endpoint and two for the secondary endpoint. Due to the varying levels of type I error able to be recycled upon rejection of the primary hypothesis, multiple sample size possibilities are required for the tests of the secondary endpoint. If the primary endpoint is rejected at the interim analysis and the full type I error of 0.025 can be recycled, then the secondary sample size will be calculated for 0.025 error and adjusted for one interim analysis. However, if the primary endpoint is rejected at the final analysis, then a smaller amount of type I error will be recycled and used for the sample size calculation (which is also adjusted for one interim analysis). Using this smaller type I error after rejecting the primary test at its final analysis results in a larger required sample size, and thus a more delayed secondary endpoint.

Below is the sample size calculation for scenario 2A, in which both endpoints will be tested with O'Brien Fleming boundaries and 80%. Type I error will be 0.025 for the primary endpoint, and either 0.025 or 0.0231 depending on the time of rejection of the primary null hypothesis.

> ***Primary Endpoint:*** $\alpha = 0.05$ *(one-sided 0.025 boundary),* $\beta = 0.2, \delta = 0.5, \sigma^2 = 1$
>
> *This requires a fixed sample size of 85 people per group, or 170 total. Adjusting for one interim analysis using an O'Brien Fleming boundary requires an inflation factor of:* $85 \times R(K, \alpha, \beta) = 85 \times 1.007 \approx 85$ *per group or 170 patients total*
>
> ***If the primary hypothesis is rejected at the interim analysis:***
>
> ***Secondary Endpoint:*** $\alpha = 0.05$ *(one-sided 0.025 boundary),* $\beta = 0.2, \delta = 0.35, \sigma^2 = 1$
>
> *This requires a fixed sample size of 129 people per group, or 258 total. Adjusting for one interim analysis using an O'Brien Fleming boundary requires an inflation factor of:* $129 \times R(K, \alpha, \beta) = 129 \times 1.008 \approx 131$ *per group or 262 patients total.*
>
> ***If the primary hypothesis is rejected at the final analysis:***
>
> ***Secondary Endpoint:*** $\alpha = 0.0462$ *(for a one-sided 0.0231 boundary),* $\beta = 0.2,\ \delta = 0.35,$ $\sigma^2 = 1$
>
> *This requires a fixed sample size of 132 people per group, or 264 total. Adjusting for one interim analysis using an O'Brien Fleming boundary requires an inflation factor of:* $129 \times R(K, \alpha, \beta) = 132 \times 1.008 \approx 133$ *per group or 266 patients total.*

Complete sample size calculations for all test scenarios can be found in Appendix A.

2.2 Simulation Studies

The simulations conducted to investigate the objectives of this report were designed to emulate the way studies would be conducted in practice for ease of comprehension and implementation. Based on the largest sample size required for analysis of the secondary endpoint, patient trial data was generated for the required number of subjects. Data was generated for "test" and "control" arms of the trial separately, with half the total sample size in each group. Each subject had a primary outcome value and secondary outcome value randomly generated from correlated normal distributions. The primary endpoint distributions from which the outcome values were generated displayed a situation in which the "test" group had a larger mean than the control group, intended to illustrate a treatment effect in the primary endpoint. Both groups had secondary endpoints generated from the same distribution with equal means to demonstrate the lack of a treatment effect on the secondary endpoint.

There are four situations to consider for the testing of a primary and a secondary endpoint: Both null hypotheses are true, neither null hypothesis is true, only the primary null hypothesis is true, or only the secondary null hypothesis is true. This report focuses on establishing control of the FWER by examining the case when only the secondary null hypothesis is true: that there is a true primary treatment effect but there is no treatment effect in the secondary endpoint. This is the

only case which needs to be considered to examine the FWER. If neither null hypothesis is true, there is no chance of achieving a type I error in either endpoint. If both null hypotheses are true, then the FWER is already strongly controlled by the boundary definitions for the primary endpoint. Due to the nature of O'Brien Fleming and Pocock boundaries, we know that they both will strongly control the type I error rate for the primary boundary. Since the secondary endpoint can not be tested until the primary endpoint is rejected, any chance of observing a type I error in the secondary endpoint will already have been preceded by a type I error in the primary endpoint thus making the FWER equal to the type I error rate in the primary endpoint which is known to be strongly controlled. Similarly, if there is a true treatment effect in the secondary endpoint only and not in the primary endpoint, the FWER will be strongly controlled simply by the O'Brien Fleming or Pocock boundary used for the primary tests. As a result, the situation which must be investigated to determine if a method has strong control of the FWER is when a type I error can not be found in the primary endpoint but may occur in the delayed secondary endpoint.

Once the simulated study data has been generated, the primary endpoint response values of the "test" and control groups were analyzed at the planned analysis sample sizes with t-tests. At the time the primary endpoint's test statistic is found to exceed the given boundary value, testing of the secondary endpoint can begin. The secondary endpoint was analyzed in the same way until the critical boundary value was exceeded by a test statistic or the final analysis was conducted without finding significance. If the secondary endpoint analysis was reached and one the comparisons of means (either interim or final analysis) was found significant, then the trial was recorded as a type I error. If both the secondary endpoint tests were not deemed significant, or the primary endpoint was never deemed significant, then no type I error was recorded. This was simulation of a clinical trial was repeated for one million iterations to provide a robust estimate of the proportion of times a type I error occurs to report as the estimated FWER for the scenario.

Each scenario was tested for a variety of treatment effects and correlations between primary and secondary outcome values. Correlations between endpoint values were tested at levels of 0.5, 0.75 and perfect correlation, 1. No lower values were assessed as it is previously known that gatekeeping procedures illustrate worsened control of the FWER as correlation between the endpoints increases [5]. Primary treatment effects were assessed at values of 0.1, 0.2, 0.3, 0.4, 0.5 (the target effect for which the study sample sizes were designed), 0.75, and 1. It is imperative to test treatment effects larger than those the study's sample size was designed to detect as overpowering the primary endpoint may have consequences on the probability of observing a type I error in the secondary endpoint [5]. This occurs since a larger treatment effect will cause the primary hypothesis to be rejected at the interim analysis more often, leading to a lower critical value which must be surpassed by the test statistic of the secondary endpoint [5].

As a major concern of this report is the effect on which delaying the secondary endpoint will have on the FWER, simulations will examine testing the secondary endpoint solely using O'Brien Fleming boundaries while the primary endpoint will be tested with a mix of both O'Brien Fleming and Pocock boundary types. This is because, when determining the sample

sizes needed for each test, Pocock boundaries require a larger inflation factor to adjust for the interim analyses and will result in a larger, more delayed, sample size than that required when using O'Brien Fleming secondary boundaries [5].

2.3 Correction of Correlation Bias

As the purpose of these simulations if to determine whether (or not) the FWER is inflated by the testing of a delayed secondary endpoint in a gatekeeping procedure for a primary and a secondary endpoint, if a bias is noted by the initial tests which result in the FWER not being strongly controlled, a method of adjusting for, or correcting, that bias would be valuable.

As the correlation between the two endpoints of interest increases, a difficult situation arises in which a true treatment effect in the primary endpoint is likely to result in extreme values in the primary responses. Despite the lack of knowledge about the of true association between treatment and the secondary endpoint, likelihood of extreme values occurring in the secondary response values increases, which would inflate the false positive rate of the secondary tests [10]. As a result, we need to correct for the bias introduced when extending the sample size for the secondary analysis past that needed for testing of the primary endpoint.

Firstly, we will introduce the notation used in deriving a correction of the bias. Let T_{ij} represent the test statistic found from testing the i^{th} endpoint at the j^{th} possible analysis time. Since these scenarios have two endpoints, i can take on values such that $i = 1,2$. Additionally, j has a range of $j = 1,2,3$. The value of $j = 1$ represents the time of primary interim analysis and first possible interim analysis for the secondary endpoint. Similarly, $j = 2$ represents the final analysis of the primary endpoint and second possible interim analysis for the secondary endpoint. A value of $j = 3$ represents the time of the final analysis for the secondary endpoint. Based on set-ups introduced in Glimm et.al. and Tamhane et.al. [2,10], we then can say that the values of T_{ij} have the following distributions:

$$T_{11} \sim N\left(\sqrt{n_1}\mu_1, 1\right), \qquad T_{12} \sim N\left(\sqrt{n_2}\mu_1, 1\right)$$

$$T_{21} \sim N\left(\sqrt{n_1}\mu_2, 1\right), \qquad T_{22} \sim N\left(\sqrt{n_2}\mu_2, 1\right), \qquad T_{23} \sim N\left(\sqrt{n}\mu_2, 1\right)$$

$$corr\left(T_{1j}, T_{2j}\right) = \rho\,, \qquad j = 1,2$$

where μ_1 and μ_2 represent the true treatment effects in the primary and secondary endpoints respectively and n_1 and n_2 represent the sample sizes as the first and second interim analysis times for the secondary endpoint.

Additionally, the boundary values defined to test for rejection of the primary and secondary endpoints are denoted as $c_{ij},\ i = 1,2,\ j = 1,2,3$ similarly to the notation of test statistics, T_{ij}.

When the primary endpoint is rejected at the interim analysis (i.e. $T_{11} > c_{11}$), we have:

$$E[T_{11} \mid T_{11} > c_{11}] = \sqrt{n_1}\mu_1 + \frac{\phi(c_{11} - \sqrt{n_1}\mu_1)}{[1 - \Phi(c_{11} - \sqrt{n_1}\mu_1)]} \qquad (1)$$

Employing the maximum likelihood estimate of μ_1, it can be said that $\hat{\mu}_1 = \frac{T_{11}}{\sqrt{n_1}}$ or $T_{11} = \sqrt{n_1}\hat{\mu}_1$.

$$E\left[T_{11} - \sqrt{n_1}\mu_1 \middle| T_{11} > c_{11}\right] = \frac{\phi(c_{11} - T_{11})}{[1 - \Phi(c_{11} - T_{11})]} \qquad (2)$$

$$E\left[T_{11} - \sqrt{n_1}\mu_1 \middle| T_{11} > c_{11}\right] = \frac{\phi(c_{11} - T_{11})}{[1 - \Phi(c_{11} - T_{11})]} \qquad (3)$$

We also have:

$$E[T_{21} \mid T_{11}] = \sqrt{n_2}\mu_2 + \rho \cdot E[T_{11} - \mu_1 \mid T_{11} > c_{11}] \qquad (4)$$

$$E[T_{21} \mid T_{11}] = \sqrt{n_2}\mu_2 + \rho\left(\frac{\phi(c_{11} - T_{11})}{[1 - \Phi(c_{11} - T_{11})]}\right) \qquad (5)$$

Thus, the following equations can be derived from (5) when testing the secondary endpoint at:

The first possible interim analysis:

$$\hat{T}_{21} = T_{21} - \rho\left(\frac{\phi(c_{11} - T_{11})}{[1 - \Phi(c_{11} - T_{11})]}\right) \qquad (6)$$

The final analysis:

$$\hat{T}_{23} = \frac{\sqrt{n_1}\left(T_{21} - \rho\left(\frac{\phi(c_{11} - T_{11})}{[1 - \Phi(c_{11} - T_{11})]}\right)\right) + \sqrt{n_2}Y_2}{\sqrt{n_1 + n_2}} \qquad (7)$$

where Y_2 is the test statistic of the data collected only after the interim analysis producing a final result of:

$$\hat{T}_{23} = T_{23} - \sqrt{\frac{n_1}{n}}\ \rho\ \left(\frac{\phi(c_{11} - T_{11})}{[1 - \Phi(c_{11} - T_{11})]}\right) \qquad (8)$$

Thus, for the bias correction simulations, these adjusted test statistics are used for testing the secondary interim analysis and secondary final analysis when the primary endpoint is rejected at the time of its interim analysis.

Similar to the results found above, when the primary endpoint is rejected at its final analysis, we find the corrected test statistics to be:

$$\hat{T}_{22} = T_{22} - \rho\left(\frac{\phi(c_{12} - T_{12})}{[1 - \Phi(c_{12} - T_{12})]}\right) \qquad (9)$$

$$\hat{T}_{23} = T_{23} - \sqrt{\frac{n_2}{n}}\,\rho\left(\frac{\phi(c_{12} - T_{12})}{[1 - \Phi(c_{12} - T_{12})]}\right) \qquad (10)$$

2.4 Power Investigation

Though the sample sizes and timing of interim analyses in each scenario were designed to uphold a minimum power percentage, it is always preferable to have power as large as possible in the conduction of confirmatory clinical trials. As a result, despite the potential for data analysis from multiple scenarios to all strongly control the type 1 error at the desired level of 2.5%, differences in power between these methods may assist in selecting one to use in practice. To assess the relative power of these scenarios, simulations similar to those conducted to determine FWER control will be used. These simulations differ only in that they will be conducted based on data in which both the primary and secondary endpoints are generated with a true treatment effect in the underlying distribution.

When assessing power with multiple endpoints, there are two types of power which could be of interest: primary power and secondary power. Primary power is defined as the probability of observing a statistically significant treatment effect when a true treatment effect exists in the relationship between treatment and the primary endpoint [11]. Similarly, secondary power is the chance of observing a significant treatment effect when a true effect exists between treatment and secondary outcome [11]. Secondary power is equivalent to the chances of observing significant treatment effects in both the primary and secondary endpoints when true treatment effects exist for both since the secondary endpoint cannot be deemed significant until the primary endpoint is significant in a gatekeeping procedure [11].

Secondary power is mainly of interest since, when testing multiple endpoints with multiple interim analyses, it is crucial to ensure that the analyses of both endpoints of interest are sufficiently powered. Additionally, while primary power is determined entirely by the features of the test for the primary endpoint, the ability to correctly reject the secondary endpoint in a gatekeeping methodology will be influenced by features of the secondary test as well as the time point and significance level of the primary tests. These other features are determined in part by the time at which the primary test is deemed significant which is variable for any given analysis, thus making secondary power a function of features of designs of both the primary and secondary analyses. The effects of variation in these nuisance parameters will be assessed in the power simulations by using combinations of primary treatment effects of 0.4, 0.5, and 0.75, secondary treatment effects of 0.3, 0.35, and 0.4, and correlations between the endpoints of 0, 0.25, 0.5, 0.75, and 1. Each power simulation will be repeated for 250,000 iterations.

3.0 Results

3.1 Family-wise Error Control

Rates of type I errors found in the secondary endpoint were reported as a proportion of the total simulations run. Trends due to variation in primary treatment effect observed and correlation

between the endpoints can be seen in the following plots. The FWER can be seen to increase, in all scenarios, as correlation increases between the endpoints. The FWER also increases with the presence of larger primary treatment effects in all scenarios.

3.1.1 Error Spending Designs

Scenario 1A: O'Brien Fleming – O'Brien Fleming Boundaries with 80% secondary power

The use of O'Brien Fleming boundaries for both primary and secondary endpoints does provide strong control of the FWER as seen in Figure 1.

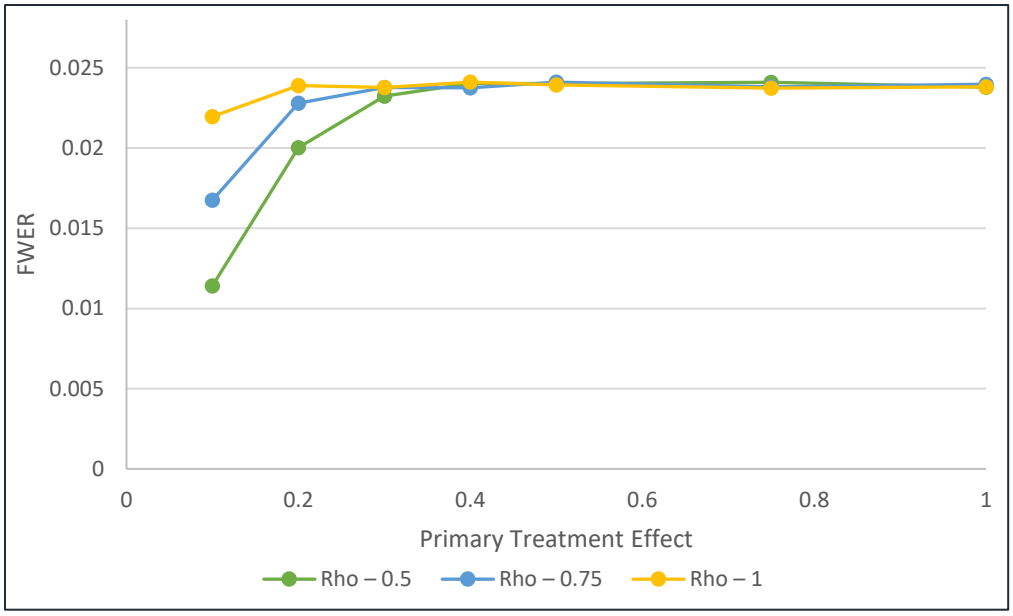

Figure 1: Scenario 1A Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.011418	0.016754	0.021962
0.2	0.020028	0.022803	0.023900
0.3	0.023229	0.023776	0.023773
0.4	0.024033	0.023758	0.024104
0.5	0.024013	0.024105	0.023942
0.75	0.024097	0.023808	0.023726
1	0.023796	0.023983	0.023814

Table 1: Scenario 1A Family-wise Error Rates

Scenario 1B: Pocock – O'Brien Fleming Boundaries with 80% secondary power

Employing a Pocock primary boundary and O'Brien Fleming secondary boundary results in strong control of the FWER with tests designed for 80% power as seen in Figure 2.

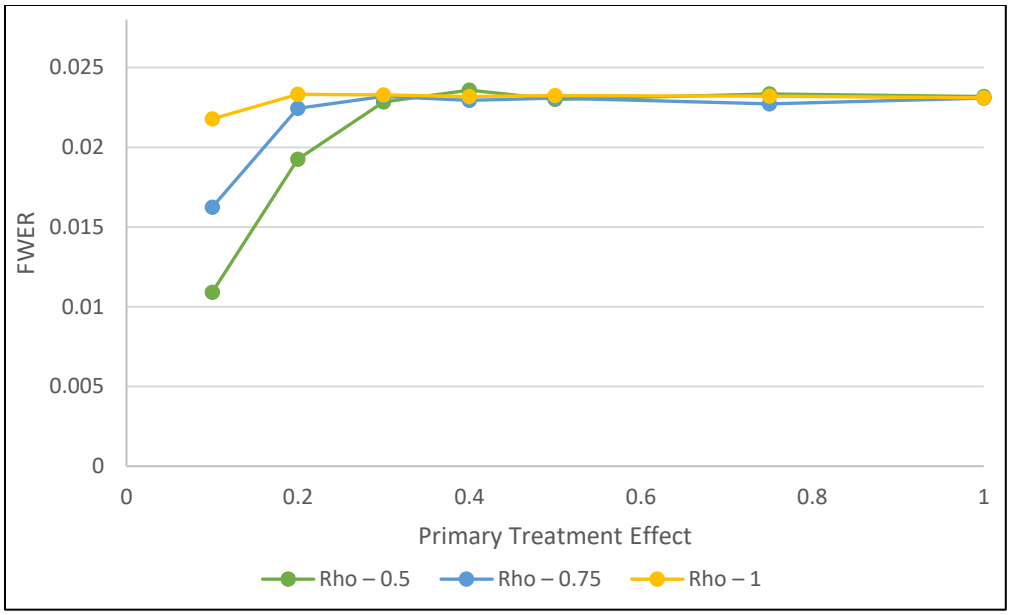

Figure 2: Scenario 1B Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.010909	0.016242	0.021776
0.2	0.019238	0.02244	0.023315
0.3	0.022836	0.023187	0.02329
0.4	0.023582	0.022933	0.023177
0.5	0.023005	0.023082	0.02324
0.75	0.023341	0.022720	0.023207
1	0.023173	0.023085	0.023101

Table 2: Scenario 1B Family-wise Error Rates

Scenario 1C: O'Brien Fleming – O'Brien Fleming Boundaries with 90% secondary power

Strong control of the FWER is not maintained when tests of the secondary endpoint are designed with a further delayed sample size to have 90% power compared to the less delayed sample size for 80% power in Scenario 1A. This can be seen in Figure 3, as the trends reach values marginally greater than the planned 0.025 α level compared to the trends in Figure 1.

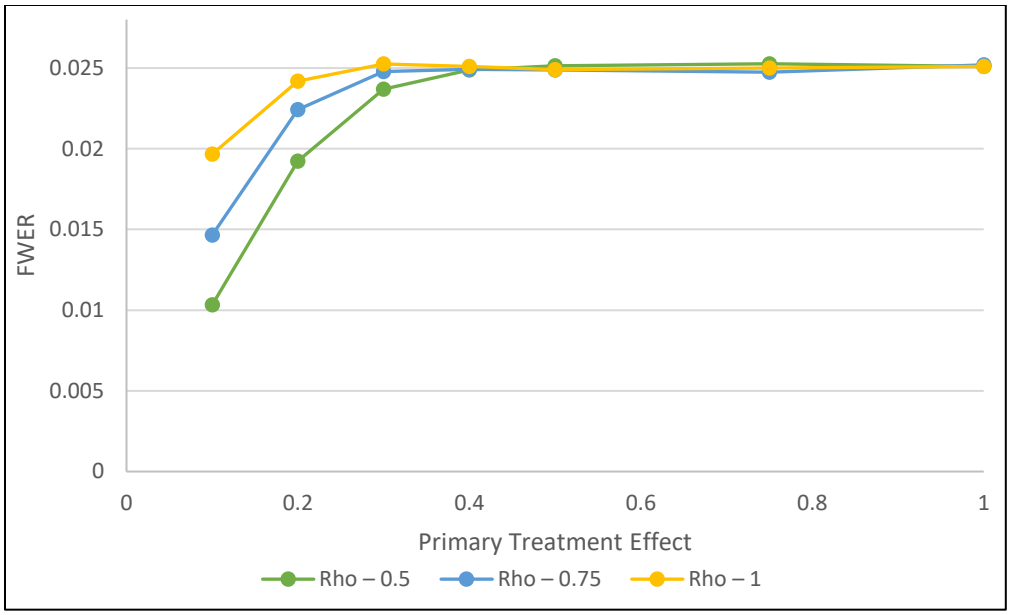

Figure 3: Scenario 1C Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.010330	0.014654	0.019667
0.2	0.019240	0.022427	0.024197
0.3	0.023687	0.024776	0.025252
0.4	0.024888	0.024915	0.025105
0.5	0.025150	0.024891	0.024902
0.75	0.025266	0.024738	0.025001
1	0.025104	0.025196	0.025105

Table 3: Scenario 1C Family-wise Error Rates

Scenario 1D: Pocock – O'Brien Fleming Boundaries with 90% secondary power

Similarly to Scenario 1C, it is shown that increasing the power of the secondary endpoint tests to 90%, further delaying the secondary endpoint results in loss of strong control of the FWER for a Pocock-O'Brien Fleming boundary combination for the primary and secondary endpoints respectively. This can be seen in Figure 4 as the family-wise error rates increase to slightly above the intended 0.025 α level as the primary treatment effect increases.

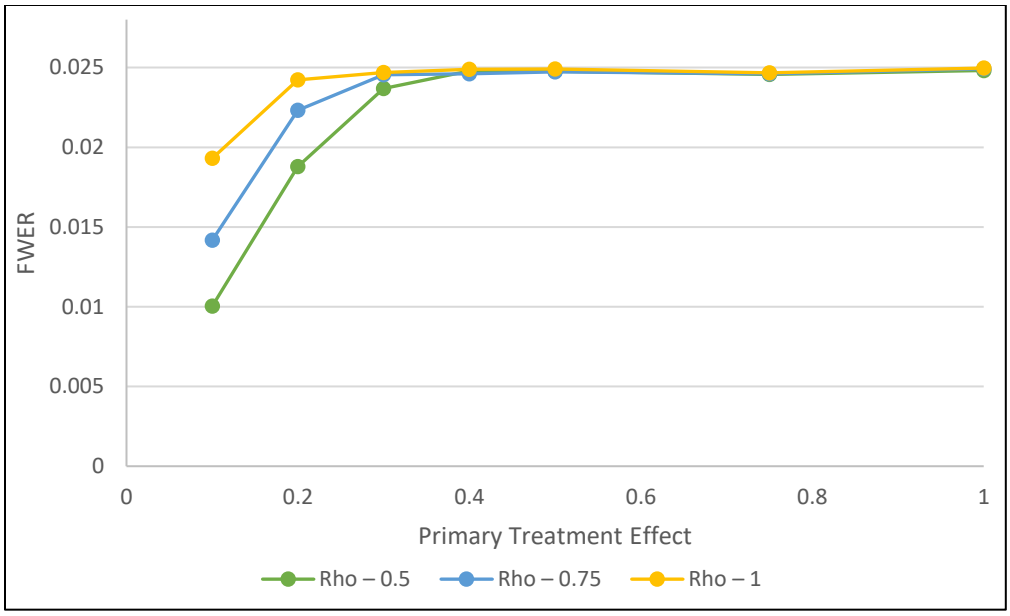

Figure 4: Scenario 1D Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.010032	0.014168	0.019302
0.2	0.018792	0.022310	0.024223
0.3	0.023682	0.024535	0.024688
0.4	0.024793	0.024598	0.024888
0.5	0.024767	0.024721	0.024901
0.75	0.024558	0.024596	0.024672
1	0.024812	0.024919	0.024964

Table 4: Scenario 1D Family-wise Error Rates

3.1.2 Error Recycling Designs

Scenario 2A: Alpha Spent-based O'Brien Fleming – O'Brien Fleming Boundaries

The error recycling methods used in this scenario, combined with O'Brien Fleming boundaries for both endpoints, result in the FWER beginning to exceed the intended 0.025 α level at a primary treatment effect of 0.4 and plateau at that level as seen below in Figure 5.

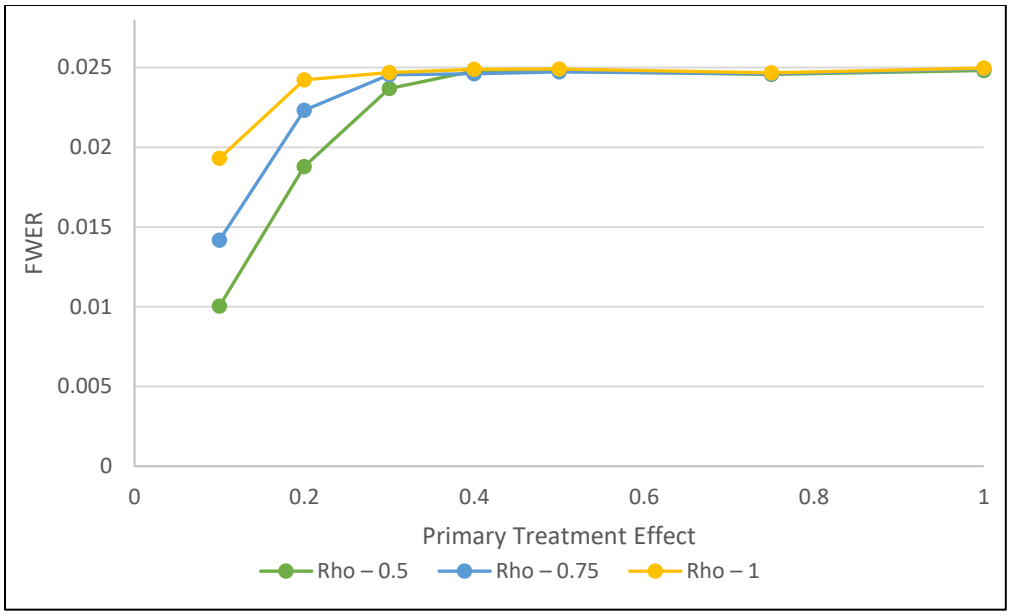

Figure 5: Scenario 2A Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.011123	0.016374	0.021714
0.2	0.019668	0.022884	0.024744
0.3	0.024084	0.024615	0.025173
0.4	0.025253	0.025269	0.025449
0.5	0.025368	0.025407	0.025606
0.75	0.025274	0.025021	0.025315
1	0.025222	0.025599	0.025388

Table 5: Scenario 2A Family-wise Error Rates

Scenario 2B: Alpha Spent-based Pocock– O'Brien Fleming Boundaries

Similarly to Scenario 2A, the FWER for this error recycling method are above the 0.025 α level as the primary treatment effect increases as shown in Figure 6. Though the use of a Pocock boundary for the primary endpoint reduces the primary treatment effect at which the α level is exceeded to 0.3 from 0.4.

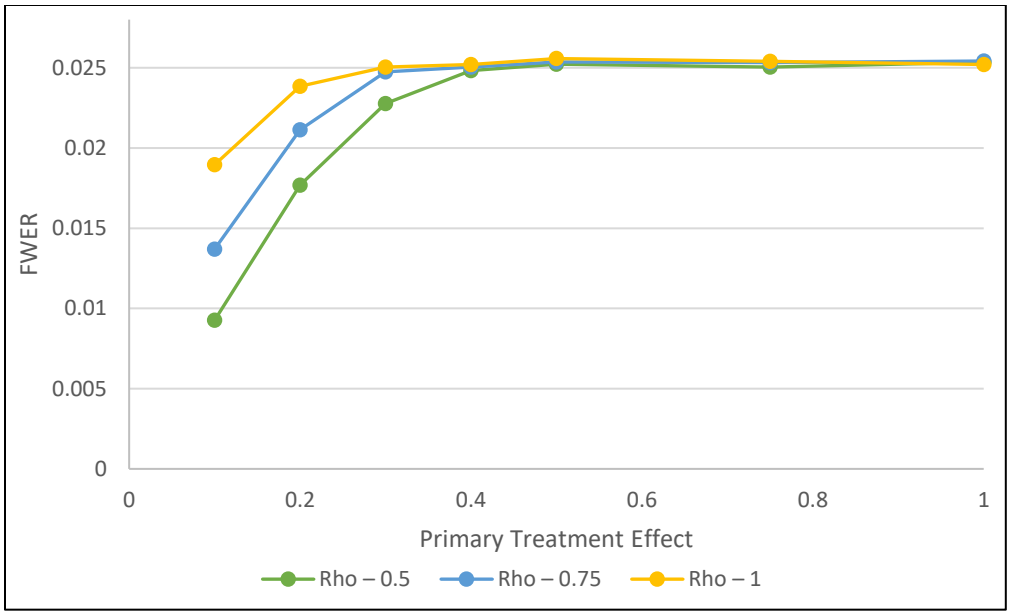

Figure 6: Scenario 2B Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.009267	0.013700	0.018959
0.2	0.017679	0.021128	0.023857
0.3	0.022764	0.024739	0.025038
0.4	0.024829	0.025037	0.025213
0.5	0.025216	0.025356	0.025574
0.75	0.025052	0.025331	0.025403
1	0.025346	0.025421	0.025212

Table 6: Scenario 2B Family-wise Error Rates

Scenario 2C: Significance Level-based O'Brien Fleming – O'Brien Fleming Boundaries

Basing the amount of recycled test mass on the significance level of the tests for the primary endpoint, similar results can be seen for use of both O'Brien Fleming boundaries when compared to basing the recycled test mass on the error spent at the final primary test. Figure 7 shows that the FWER marginally exceeds the intended 0.025 α level for the larger primary treatment effects.

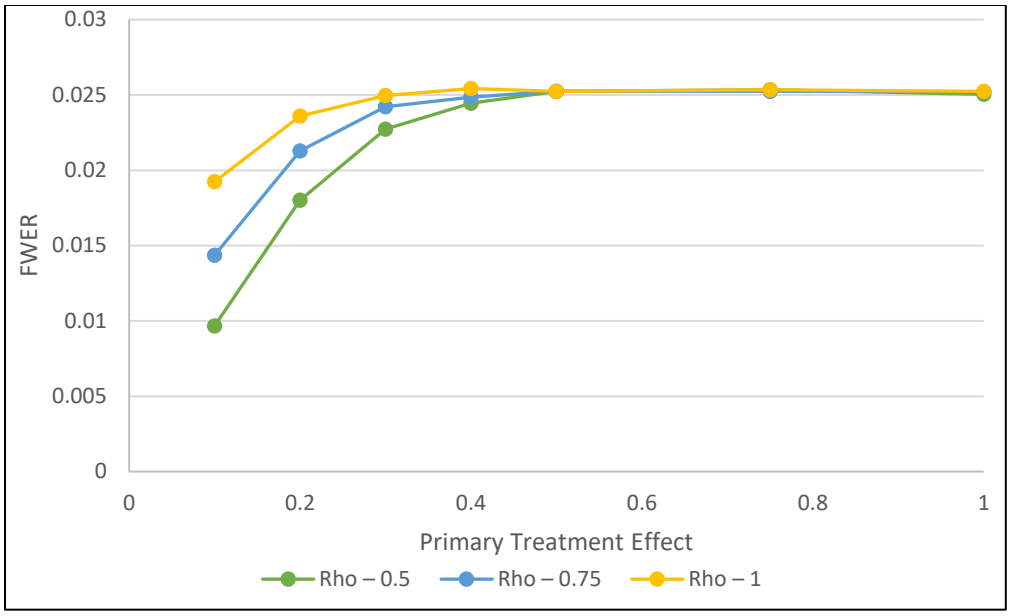

Figure 7: Scenario 2C Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.009674	0.014363	0.019238
0.2	0.018021	0.021279	0.023609
0.3	0.022739	0.024216	0.024952
0.4	0.024453	0.024858	0.025428
0.5	0.025231	0.025245	0.025245
0.75	0.025346	0.025245	0.025346
1	0.025036	0.025212	0.025247

Table 7: Scenario 2C Family-wise Error Rates

Scenario 2D: Significance Level-based Pocock– O'Brien Fleming Boundaries

Just as in the previous error recycling results, the use of a Pocock primary boundary and error recycling based on primary test significance level results in FWER which exceed the intended α level for most primary treatment effects and all correlations, as seen in Figure 8.

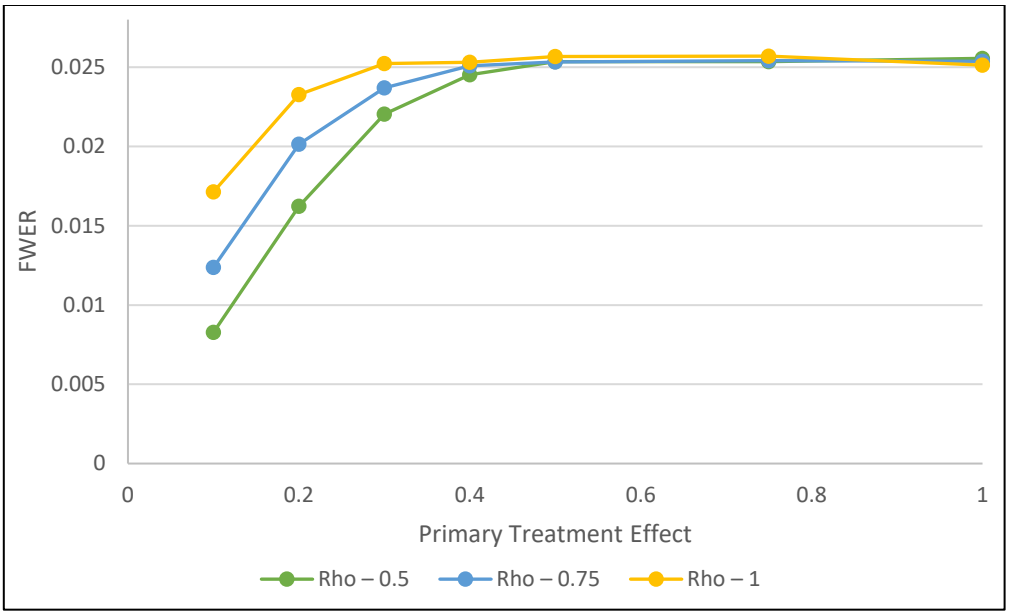

Figure 8: Scenario 2D Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.008270	0.012367	0.017133
0.2	0.016235	0.020151	0.023276
0.3	0.022039	0.023700	0.025238
0.4	0.024525	0.025086	0.025315
0.5	0.025358	0.025332	0.025680
0.75	0.025360	0.025414	0.025699
1	0.025559	0.025365	0.025124

Table 8: Scenario 2D Family-wise Error Rates

3.2 Bias Corrected Family-wise Error Control

Simulations were carried out with identical treatment level and correlation variation as the original tests for FWER control in each for the eight pre-defined scenarios. After adjusting the test statistics for the tests of secondary endpoints with the method of bias correction previously introduced (Section 2.3), the FWER can be plotted for trends associated with changing the true primary treatment effects and the correlation. The FWER in these assessments are expected to be consistently lower as the bias correction method systematically reduces the magnitude of each secondary test statistic before comparison for significance.

3.2.1 Group Sequential Trial Designs

Scenario 1A: O'Brien Fleming – O'Brien Fleming Boundaries with 80% secondary power

The bias correction in this scenario maintains the strong control of the FWER (Figure 9) that was observed in the original simulations.

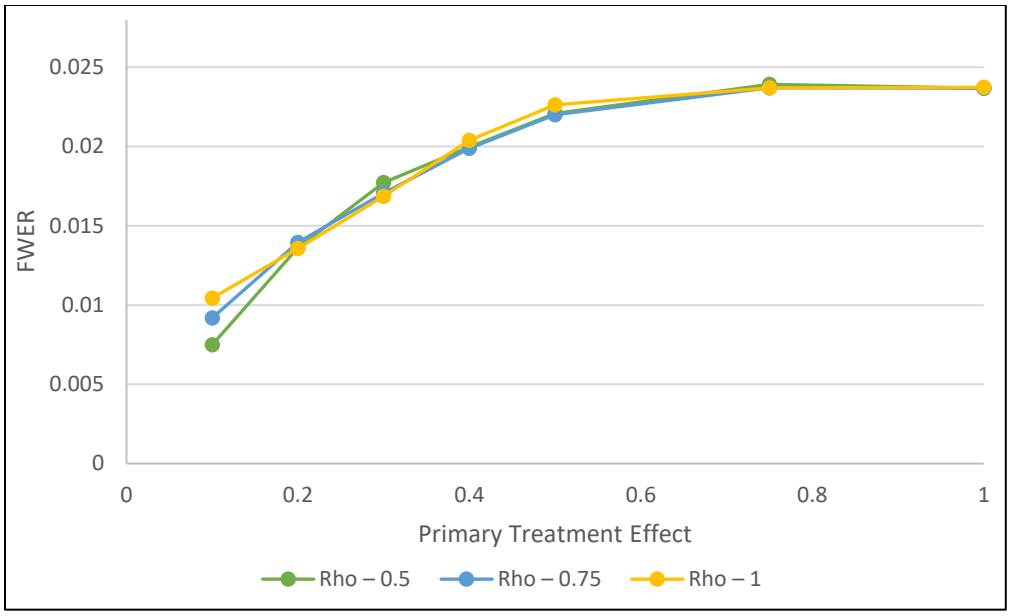

Figure 9: Bias Corrected Scenario 1A Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.007490	0.009181	0.010442
0.2	0.013598	0.013951	0.013573
0.3	0.017730	0.017058	0.016847
0.4	0.019961	0.019878	0.020378
0.5	0.022076	0.022000	0.022619
0.75	0.023916	0.023703	0.023705
1	0.023686	0.023672	0.023731

Table 9: Bias Corrected Scenario 1A Family-wise Error Rates

Scenario 1B: Pocock – O'Brien Fleming Boundaries with 80% secondary power

Just as in the original simulations, the bias corrected simulations result in strong control of the FWER which can be seen in Figure 10.

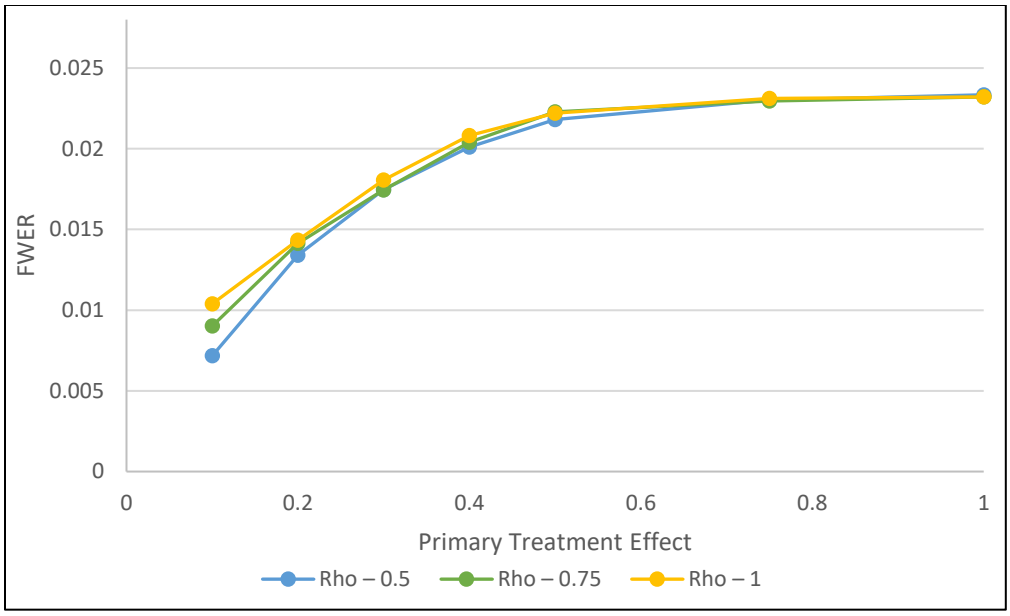

Figure 10: Bias Corrected Scenario 1B Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.007171	0.009016	0.010390
0.2	0.013402	0.014128	0.014326
0.3	0.017468	0.017455	0.018062
0.4	0.020114	0.020402	0.020814
0.5	0.021814	0.022292	0.022198
0.75	0.023010	0.022957	0.023122
1	0.023343	0.023212	0.023211

Table 10: Bias Corrected Scenario 1B Family-wise Error Rates

Scenario 1C: O'Brien Fleming – O'Brien Fleming Boundaries with 90% secondary power

As seen in Figure 11, the FWER stays predominantly below the intended α level and slightly exceeds that level for the treatment effects greater than the estimated primary treatment effect of 0.5 for which the study was designed. These values should be considered as converging to the intended nominal significance level of 0.025.

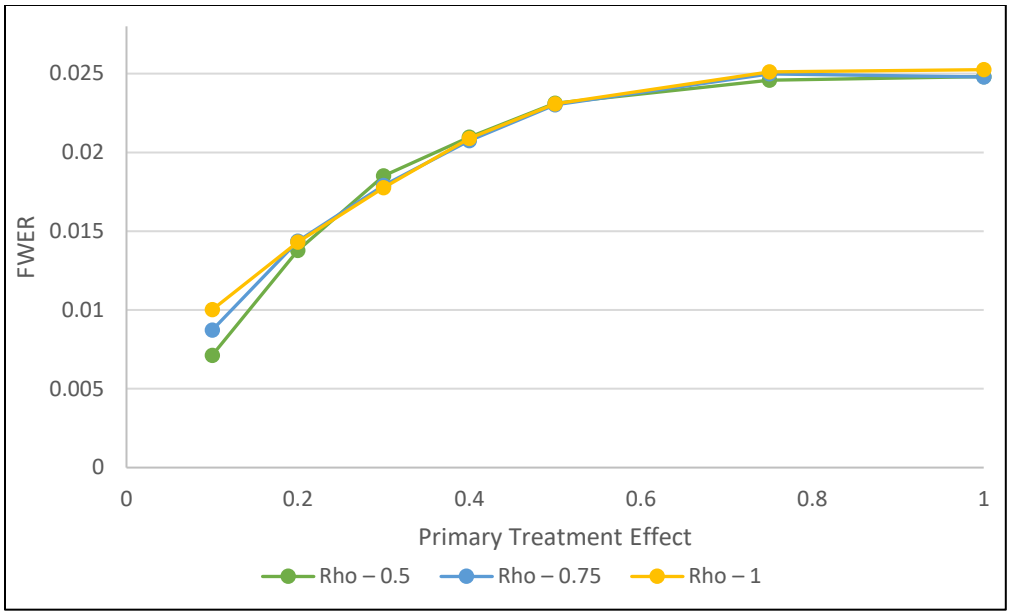

Figure 11: Bias Corrected Scenario 1C Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.007128	0.008723	0.010015
0.2	0.013776	0.014368	0.014313
0.3	0.018505	0.017897	0.017763
0.4	0.020966	0.020743	0.020888
0.5	0.023128	0.023009	0.023084
0.75	0.024580	0.024979	0.025106
1	0.024811	0.024787	0.025250

Table 11: Bias Corrected Scenario 1C Family-wise Error Rates

Scenario 1D: Pocock – O'Brien Fleming Boundaries with 90% secondary power

After the bias correction, the simulated error rates for all combinations of primary treatment effects and correlation values yielded FWER below the intended nominal significance level (Figure 12).

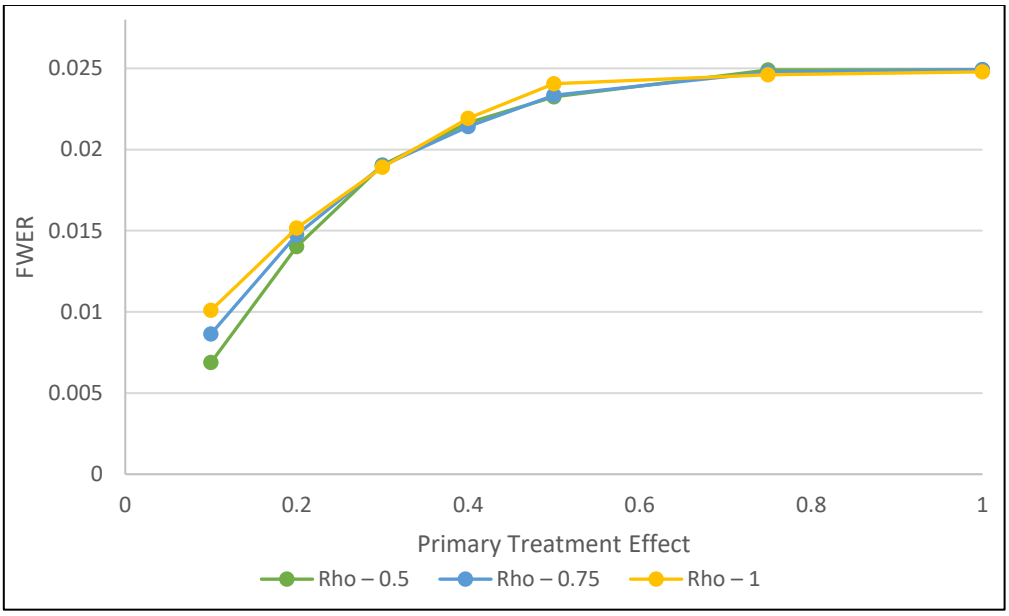

Figure 12: Bias Corrected Scenario 1D Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.006884	0.00865	0.010098
0.2	0.014013	0.014754	0.015160
0.3	0.019039	0.019003	0.018914
0.4	0.021650	0.021412	0.021929
0.5	0.023237	0.023342	0.024056
0.75	0.024898	0.024772	0.024614
1	0.024914	0.024924	0.024783

Table 12: Bias Corrected Scenario 1D Family-wise Error Rates

3.2.2 Error Recycling Designs

Scenario 2A: Alpha Spent-based O'Brien Fleming – O'Brien Fleming Boundaries

After the bias correction, while the FWER for each correlation approaches 0.025 slower across increases in primary treatment effect when compared to the original simulations, the largest primary treatment effects still produce error rates of greater than the intended nominal significance level (Figure 13).

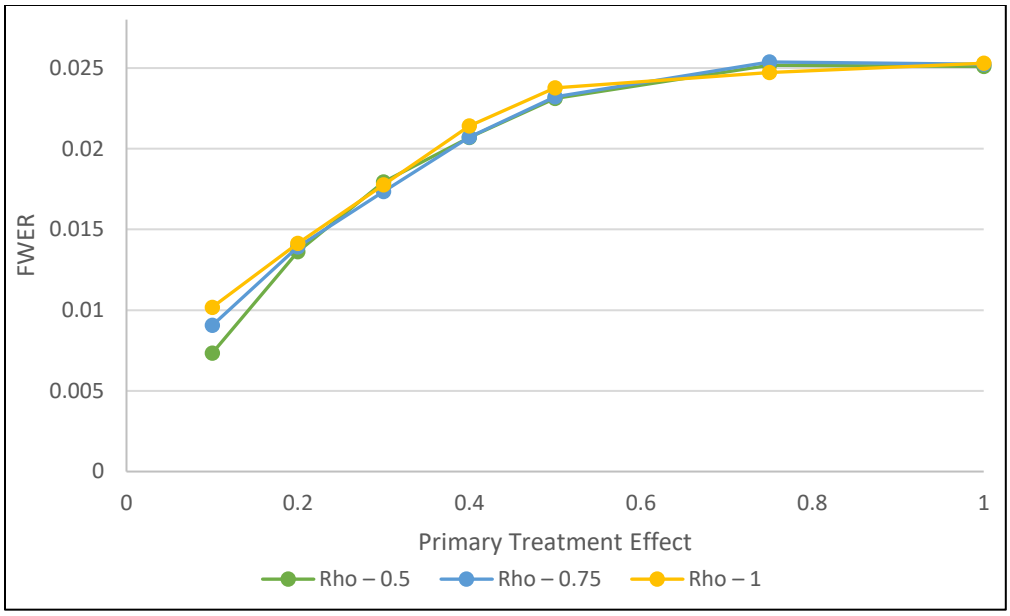

Figure 13: Bias Corrected Scenario 2A Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.007334	0.009053	0.010181
0.2	0.013623	0.013923	0.014142
0.3	0.017944	0.017356	0.017759
0.4	0.020700	0.020724	0.021417
0.5	0.023127	0.023225	0.023778
0.75	0.025174	0.025376	0.024732
1	0.025094	0.025244	0.025299

Table 13: Bias Corrected Scenario 2A Family-wise Error Rates

Scenario 2B: Alpha Spent-based Pocock– O'Brien Fleming Boundaries

As seen with the bias correction for Scenario 2A, implementation of the bias correction in Scenario 2B results in a slower approach to 0.025 than in the original un-corrected simulations but still exceeds the 0.025 threshold. Implementing a Pocock boundary for the primary analysis also slightly increased the FWER at all primary treatment effect levels over the same error recycling method that implemented a primary O'Brien Fleming boundary.

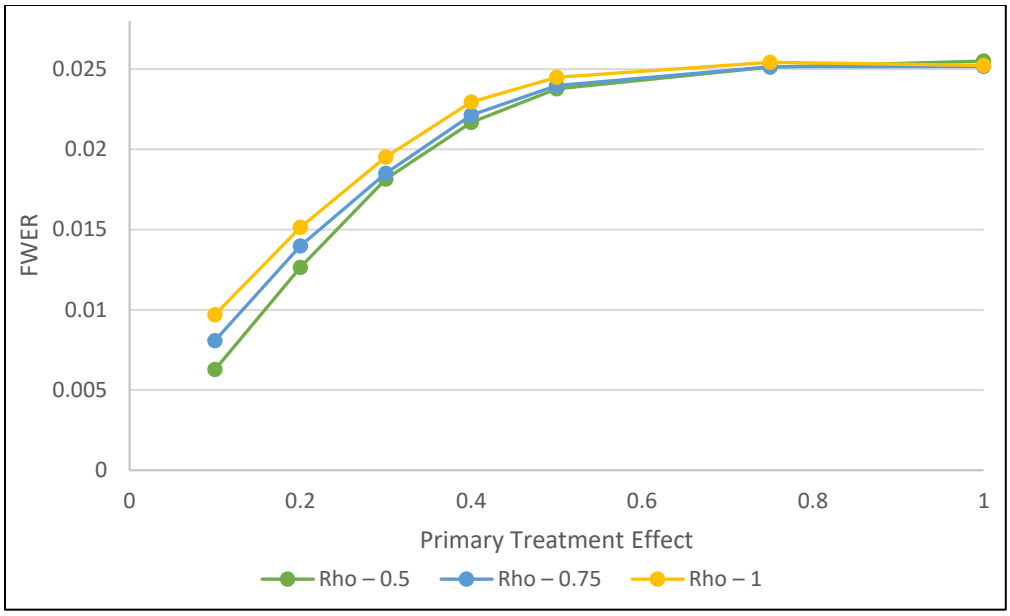

Figure 14: Bias Corrected Scenario 2B Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.006276	0.008082	0.009695
0.2	0.012647	0.013988	0.015145
0.3	0.018151	0.018511	0.019522
0.4	0.021675	0.02214	0.022957
0.5	0.023771	0.023969	0.024485
0.75	0.025123	0.025146	0.025426
1	0.025502	0.025162	0.025222

Table 14: Bias Corrected Scenario 2B Family-wise Error Rates

Scenario 2C: Significance Level-based O'Brien Fleming – O'Brien Fleming Boundaries

Similarly to previous error recycling scenarios, the results from this scenario have improved over the original simulation though, the FWER is still not strongly controlled for the largest primary treatment effects tested as seen in Figure 15.

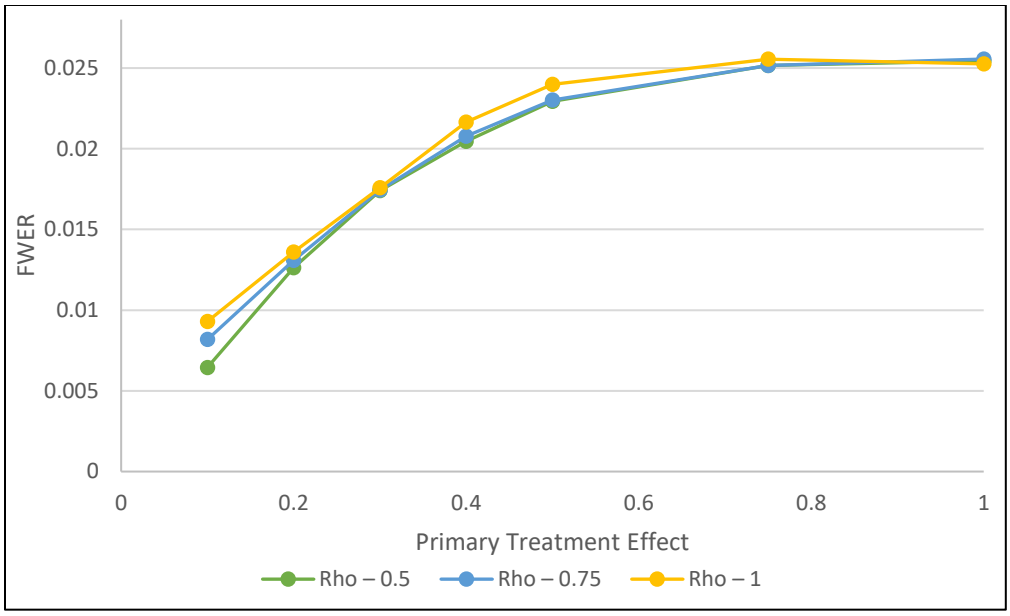

Figure 15: Bias Corrected Scenario 2C Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.006452	0.00818	0.009302
0.2	0.012621	0.013095	0.013599
0.3	0.017402	0.017457	0.017579
0.4	0.020454	0.020786	0.021654
0.5	0.022935	0.023022	0.023986
0.75	0.02517	0.025165	0.025548
1	0.025441	0.025559	0.025261

Table 15: Bias Corrected Scenario 2C Family-wise Error Rates

Scenario 2D: Significance Level-based Pocock– O'Brien Fleming Boundaries

As seen in the previous error recycling scenarios, the rate at which the FWER approaches the intended nominal is slower in this scenario with the bias correction than in the original simulation. The error rates for all correlations exceed the intended nominal significance level for the largest primary treatment effect values.

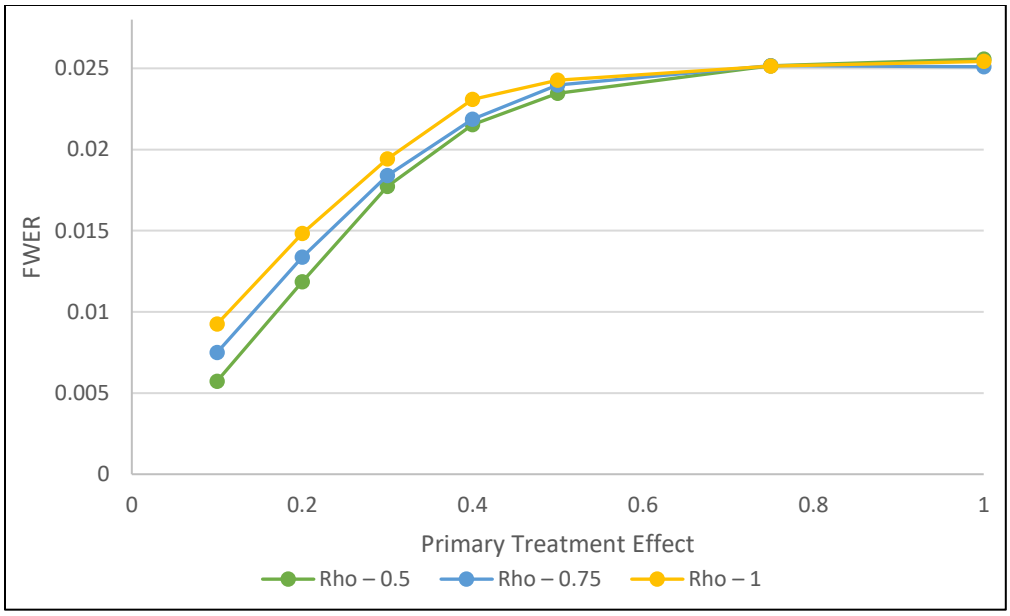

Figure 16: Bias Corrected Scenario 2D Family-wise Error Rate Trends

PE Treatment Effect	Rho – 0.5	Rho – 0.75	Rho – 1
0.1	0.005733	0.007496	0.009248
0.2	0.011851	0.013371	0.014837
0.3	0.017719	0.018397	0.019417
0.4	0.021540	0.021863	0.023081
0.5	0.023456	0.023982	0.024270
0.75	0.025165	0.025167	0.025133
1	0.025580	0.025095	0.025425

Table 16: Bias Corrected Scenario 2D Family-wise Error Rates

3.3 Power Investigation

Simulations the power of the different group sequential trial designs in this report can be used to assess the relative levels of power maintained by each. Though the sample size calculations for the primary and secondary endpoints in each scenario were conducted with at least 80% power, the multiple comparisons required in such gatekeeping procedures do not necessarily result in the secondary power of the combined analysis being maintained at an 80% level. The following results can be used to assess how the power in each scenario compares to other scenarios.

3.3.1 Group Sequential Trial Designs

Scenario 1A: O'Brien Fleming – O'Brien Fleming Boundaries with 80% secondary power

For the expected primary and secondary treatment effects, the power of the studies in this scenario is around 70% across all correlation values. The effects of variations in primary treatment effects, secondary treatment effects, and correlations on power is shown in Figure 18.

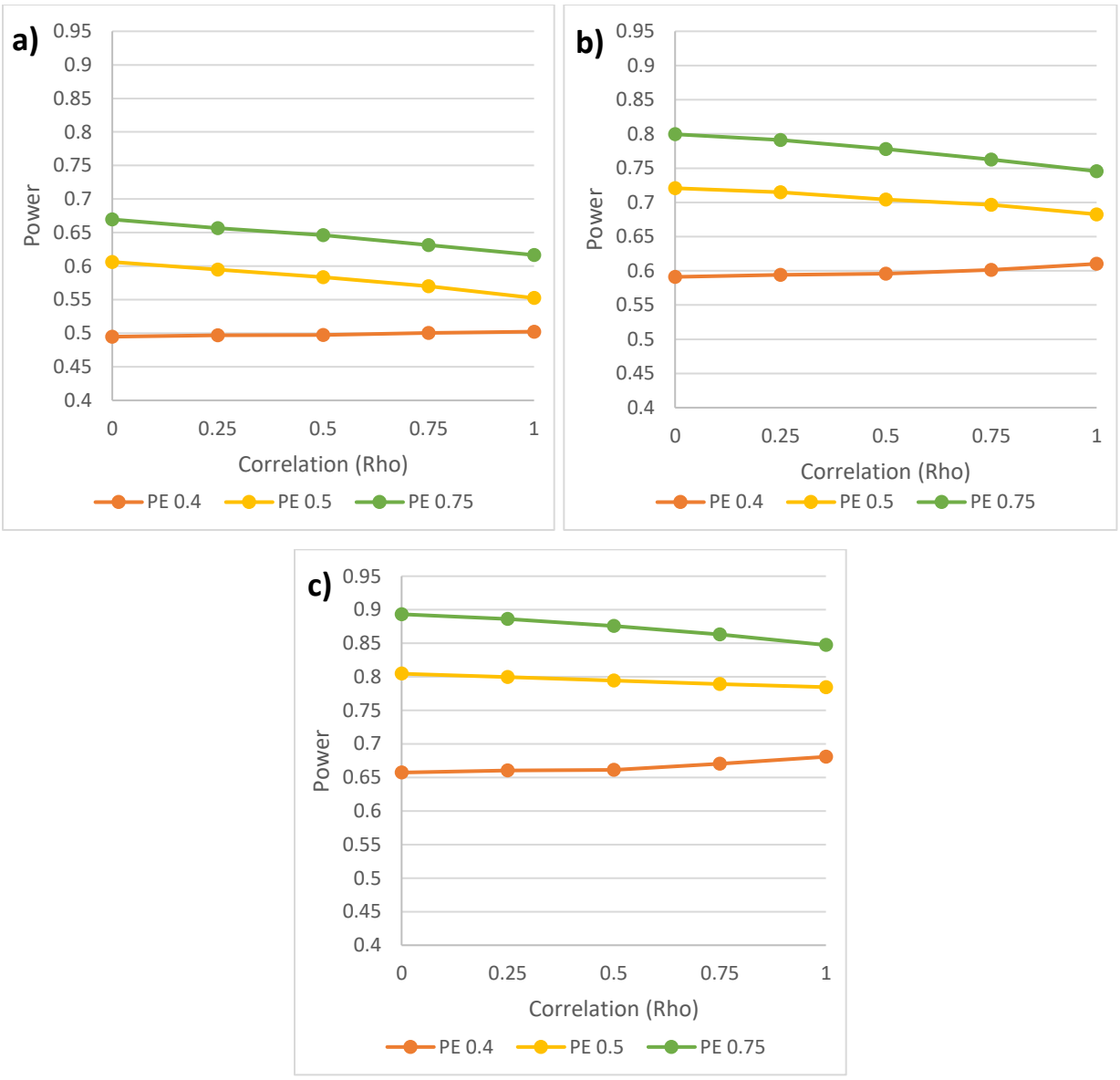

Figure 17: Bias Corrected Scenario 1A Secondary Power Trends (a) Secondary Treatment Effect of 0.3 (b) Secondary Treatment Effect of 0.35 (c) Secondary Treatment Effect of 0.4

	Secondary Treatment Effect								
	0.30			0.35			0.40		
	Primary Treatment Effect								
Rho	0.40	0.50	0.75	0.40	0.50	0.75	0.40	0.50	0.75
0	0.494872	0.606228	0.669636	0.591112	0.720704	0.799520	0.657364	0.804640	0.893276
0.25	0.497044	0.594772	0.656688	0.594144	0.714856	0.791024	0.660368	0.799620	0.886196
0.5	0.497616	0.583768	0.646048	0.595640	0.704032	0.777868	0.661480	0.794372	0.875588
0.75	0.500288	0.569972	0.631284	0.601496	0.696524	0.762528	0.670408	0.789240	0.863072
1	0.502404	0.552540	0.616640	0.610172	0.682576	0.745540	0.680824	0.784608	0.847368

Table 17: Secondary Power Results for Bias Corrected Scenario 1A

Scenario 1B: Pocock – O'Brien Fleming Boundaries with 80% secondary power

The power of this study design is approximately 70% for all correlations when the true treatment values for the primary and secondary endpoints are around their expected values of 0.5 and 0.35 respectively. Variations in primary and secondary treatment effects as well as correlations can be seen in Figure 18.

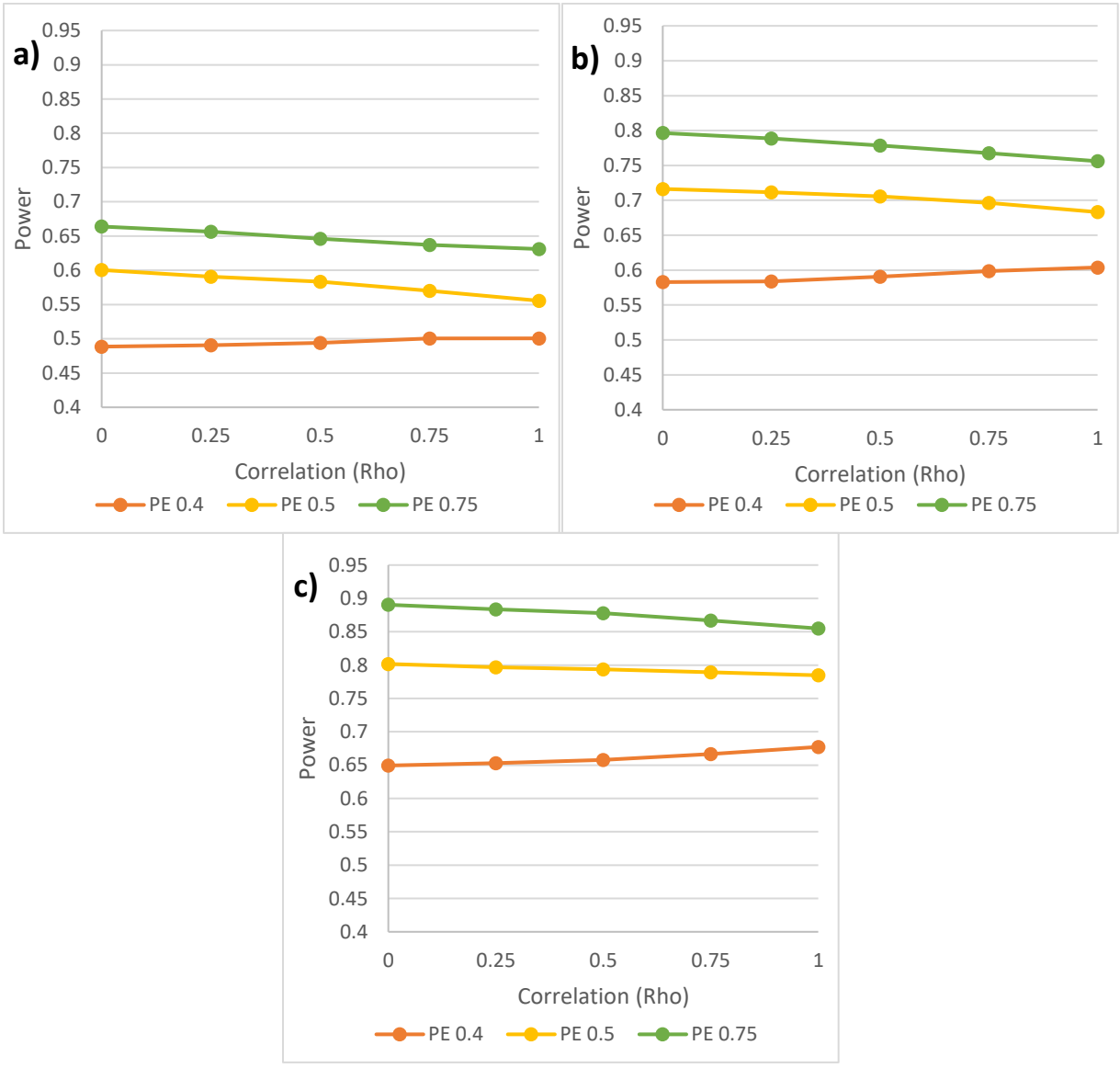

Figure 18: Bias Corrected Scenario 1B Secondary Power Trends (a) Secondary Treatment Effect of 0.3 (b) Secondary Treatment Effect of 0.35 (c) Secondary Treatment Effect of 0.4

	Secondary Treatment Effect								
	0.30			0.35			0.40		
	Primary Treatment Effect								
Rho	0.40	0.50	0.75	0.40	0.50	0.75	0.40	0.50	0.75
0	0.488552	0.600432	0.663892	0.582896	0.716344	0.796492	0.649388	0.801680	0.890472
0.25	0.490800	0.590488	0.656528	0.583928	0.711408	0.788496	0.653020	0.796728	0.883844
0.5	0.494016	0.583464	0.646276	0.590600	0.705604	0.778744	0.657948	0.793708	0.877716
0.75	0.500328	0.570256	0.636940	0.598496	0.696452	0.767592	0.666468	0.789440	0.866648
1	0.500632	0.555588	0.631020	0.603876	0.683172	0.756108	0.677288	0.784684	0.854868

Table 18: Secondary Power Results for Bias Corrected Scenario 1B

Scenario 1C: O'Brien Fleming – O'Brien Fleming Boundaries with 90% secondary power

Power for studies with the true treatment effects at the expected values of 0.5 for the primary treatment effect and 0.35 for the secondary treatment effect is approximately 80%. This is around 10% higher than the power of the same study designed for 80% power in the secondary endpoint found in Scenario 1A. Both the primary and secondary endpoint in this study were designed for 90% power. The effects of variation in treatment effects and correlation on power of this study are shown in Figure 19.

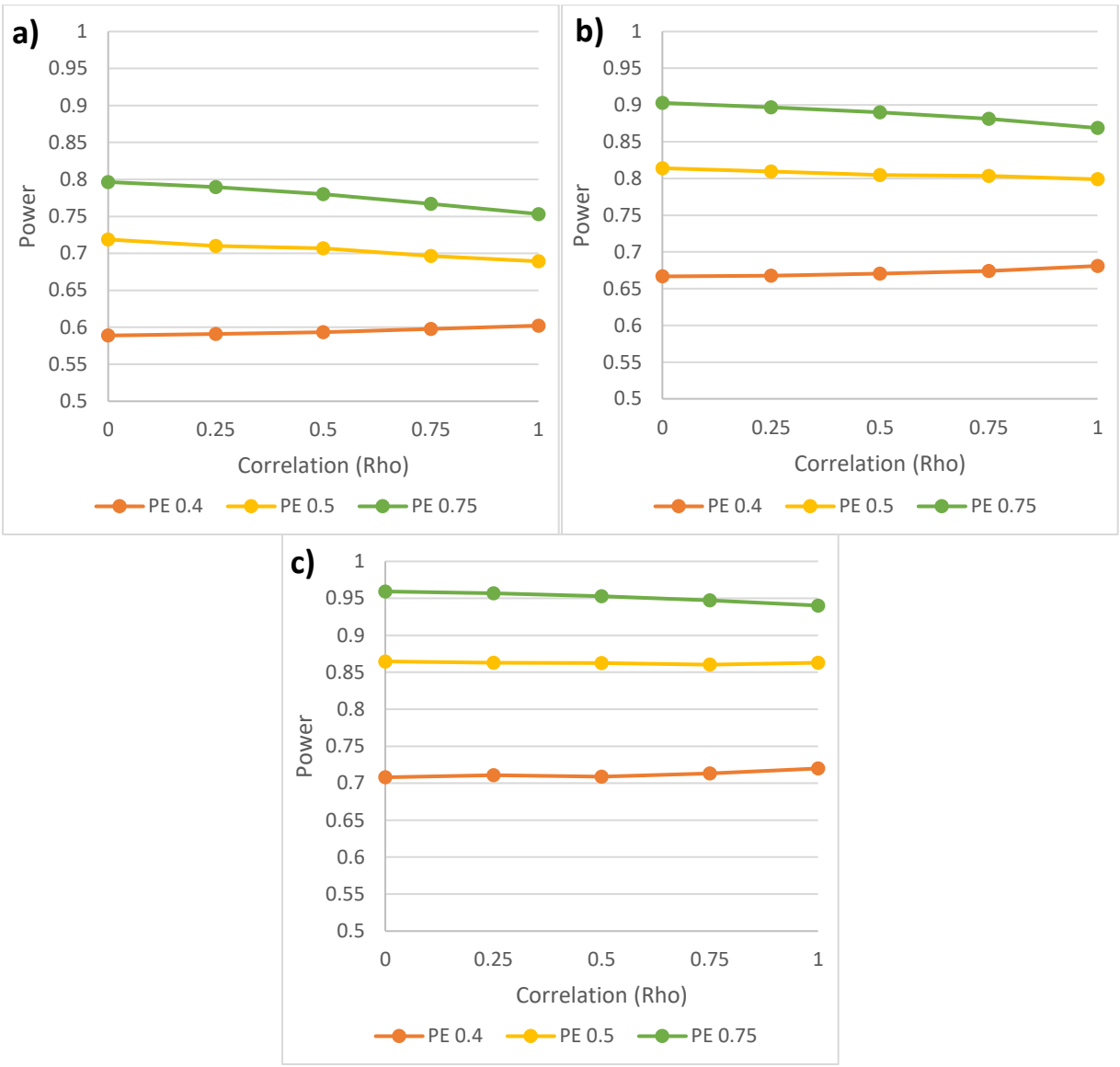

Figure 19: Bias Corrected Scenario 1C Secondary Power Trends (a) Secondary Treatment Effect of 0.3 (b) Secondary Treatment Effect of 0.35 (c) Secondary Treatment Effect of 0.4

	Secondary Treatment Effect								
	0.30			0.35			0.40		
	Primary Treatment Effect								
Rho	0.40	0.50	0.75	0.40	0.50	0.75	0.40	0.50	0.75
0	0.588928	0.718816	0.796560	0.666672	0.813984	0.902936	0.707976	0.864716	0.959332
0.25	0.591000	0.709928	0.789708	0.667568	0.809328	0.896748	0.710812	0.863072	0.956748
0.5	0.593400	0.707008	0.779928	0.670400	0.804644	0.889892	0.708748	0.862584	0.952952
0.75	0.597764	0.696480	0.766832	0.673980	0.803472	0.881100	0.713200	0.860344	0.947340
1	0.602320	0.689276	0.753240	0.680912	0.798904	0.868664	0.719912	0.862916	0.940288

Table 19: Secondary Power Results for Bias Corrected Scenario 1C

Scenario 1D: Pocock – O'Brien Fleming Boundaries with 90% secondary power

Using this study design, the power of this group sequential trial using the expected treatment effects is slightly above 80% when both the primary and secondary endpoint sample sizes were calculated for 90% power. This is greater than the study with the same group sequential boundary types in scenario 1B which was had a secondary endpoint designed for 80% power. Trends in the study power level as treatment effects and correlations are varied can be seen in Figure 20.

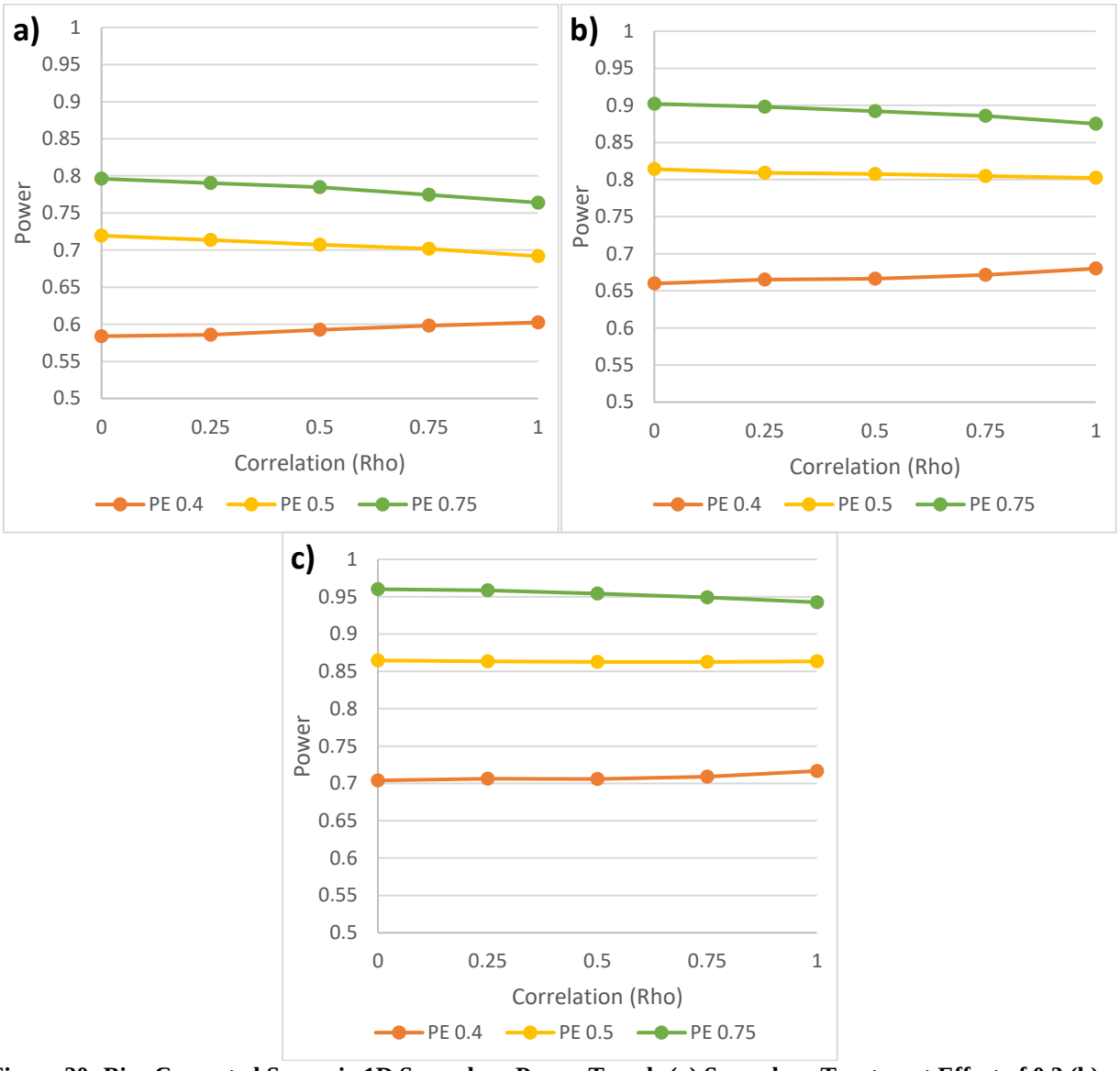

Figure 20: Bias Corrected Scenario 1D Secondary Power Trends (a) Secondary Treatment Effect of 0.3 (b) Secondary Treatment Effect of 0.35 (c) Secondary Treatment Effect of 0.4

	Secondary Treatment Effect								
	0.30			0.35			0.40		
	Primary Treatment Effect								
Rho	0.40	0.50	0.75	0.40	0.50	0.75	0.40	0.50	0.75
0	0.583916	0.719396	0.796136	0.659868	0.814072	0.902016	0.703920	0.864572	0.959984
0.25	0.585764	0.713504	0.790412	0.665020	0.808904	0.898168	0.706244	0.863536	0.958528
0.5	0.592628	0.707424	0.784556	0.666248	0.807380	0.892076	0.705844	0.862708	0.954120
0.75	0.598096	0.701528	0.774376	0.671432	0.804800	0.885636	0.708988	0.862544	0.948900
1	0.602496	0.691708	0.763900	0.680020	0.802180	0.875060	0.716616	0.863336	0.942508

Table 20: Secondary Power Results for Bias Corrected Scenario 1D

3.3.2 Error Recycling Designs

Scenario 2A: Alpha Spent-based O'Brien Fleming – O'Brien Fleming Boundaries

Power for this study for the expected treatment effects for both endpoint (0.5 and 0.35 respectively) is around 70%. Variations in treatment effects for both endpoints and correlations are shown in Figure 21.

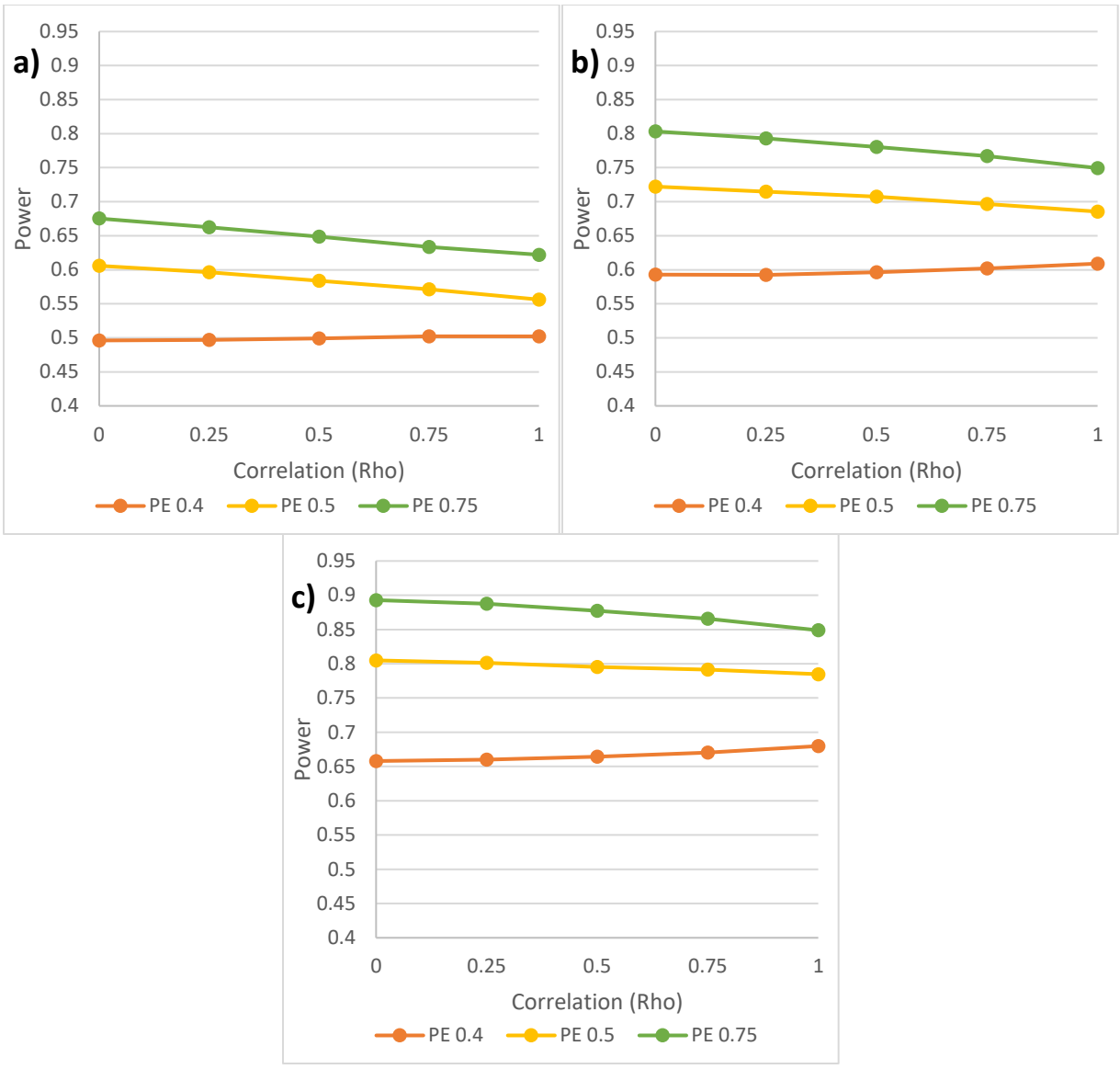

Figure 21: Bias Corrected Scenario 2A Secondary Power Trends (a) Secondary Treatment Effect of 0.3 (b) Secondary Treatment Effect of 0.35 (c) Secondary Treatment Effect of 0.4

	Secondary Treatment Effect								
	0.30			0.35			0.40		
	Primary Treatment Effect								
Rho	0.40	0.50	0.75	0.40	0.50	0.75	0.40	0.50	0.75
0	0.496140	0.605812	0.675284	0.592812	0.722212	0.803232	0.658020	0.804904	0.892884
0.25	0.497164	0.596212	0.662448	0.592444	0.714724	0.792928	0.659880	0.801356	0.887812
0.5	0.499200	0.583788	0.648856	0.596276	0.707524	0.780608	0.664448	0.795296	0.877084
0.75	0.502264	0.571128	0.633432	0.601872	0.696756	0.767184	0.670212	0.791224	0.865576
1	0.502056	0.556228	0.622056	0.609036	0.685440	0.749412	0.679852	0.784768	0.848904

Table 21: Secondary Power Results for Bias Corrected Scenario 2A

Scenario 2B: Alpha Spent-based Pocock– O'Brien Fleming Boundaries

When the true treatment effects, this study design results in power slightly greater than 70%. This is marginally higher power than the using the same test mass recycling method with an O'Brien Fleming primary boundary as in scenario 2A. The effect on power of varying treatment effects for both endpoints and correlations are shown in Figure 22.

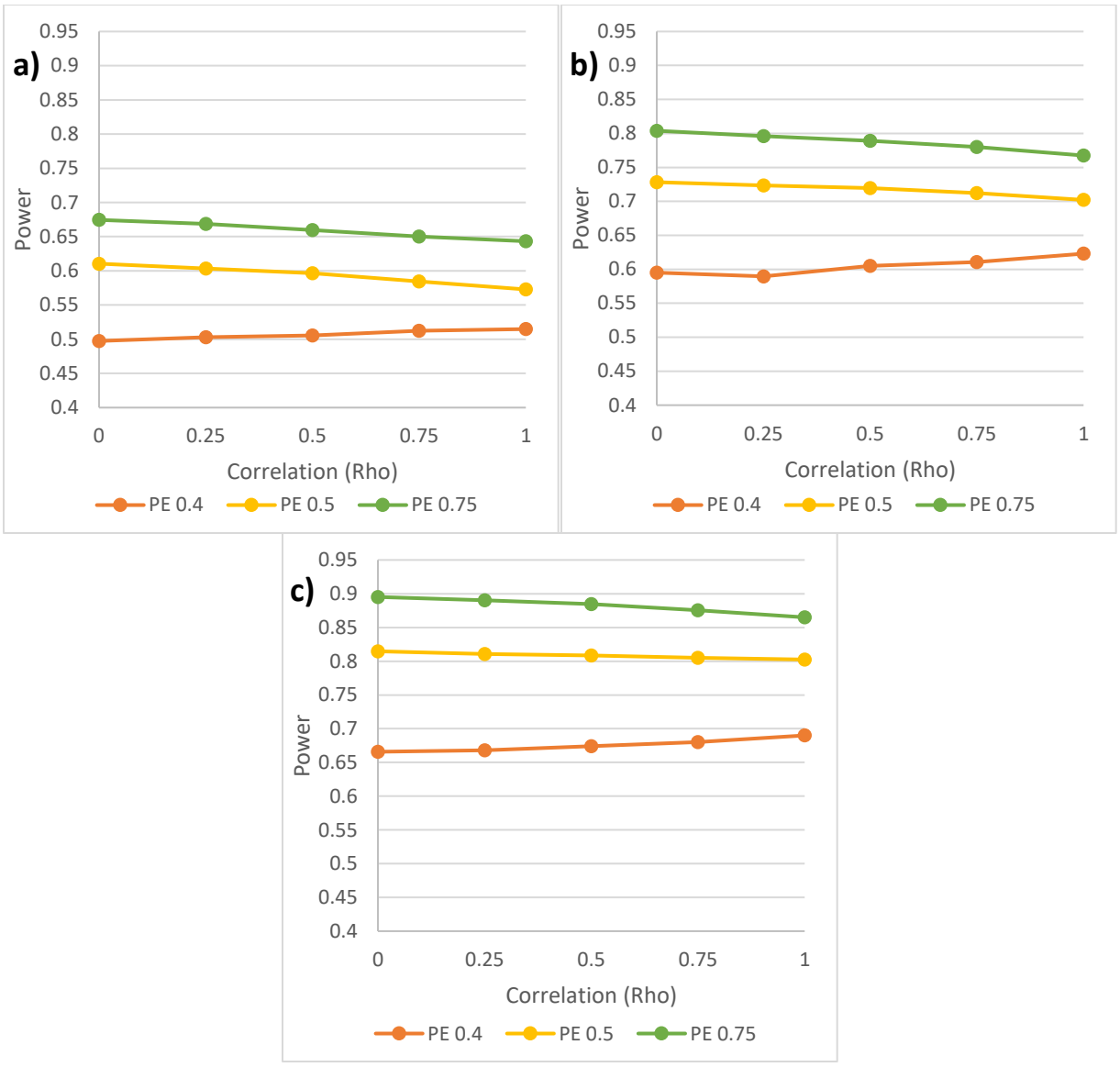

Figure 22: Bias Corrected Scenario 2B Secondary Power Trends (a) Secondary Treatment Effect of 0.3 (b) Secondary Treatment Effect of 0.35 (c) Secondary Treatment Effect of 0.4

	Secondary Treatment Effect								
	0.30			0.35			0.40		
	Primary Treatment Effect								
Rho	0.40	0.50	0.75	0.40	0.50	0.75	0.40	0.50	0.75
0	0.497428	0.610492	0.674512	0.595076	0.728248	0.803896	0.665824	0.814704	0.895192
0.25	0.503156	0.603336	0.668568	0.589600	0.723640	0.796168	0.667980	0.810828	0.890144
0.5	0.505388	0.596564	0.659788	0.605232	0.719628	0.789284	0.673932	0.808716	0.884632
0.75	0.512448	0.584412	0.650248	0.610708	0.712000	0.780192	0.680324	0.805252	0.875316
1	0.514872	0.572832	0.643320	0.623008	0.702140	0.767436	0.690024	0.802452	0.865020

Table 18: Secondary Power Results for Bias Corrected Scenario 2B

Scenario 2C: Significance Level-based O'Brien Fleming – O'Brien Fleming Boundaries

Basing the recycled test mass on the significance level of the primary test produces a secondary power of around 70% when the true treatment effects match those used in the study's sample size calculations. The effects on secondary power of varying these treatment effects and correlation between the endpoints can be seen in Figure 23.

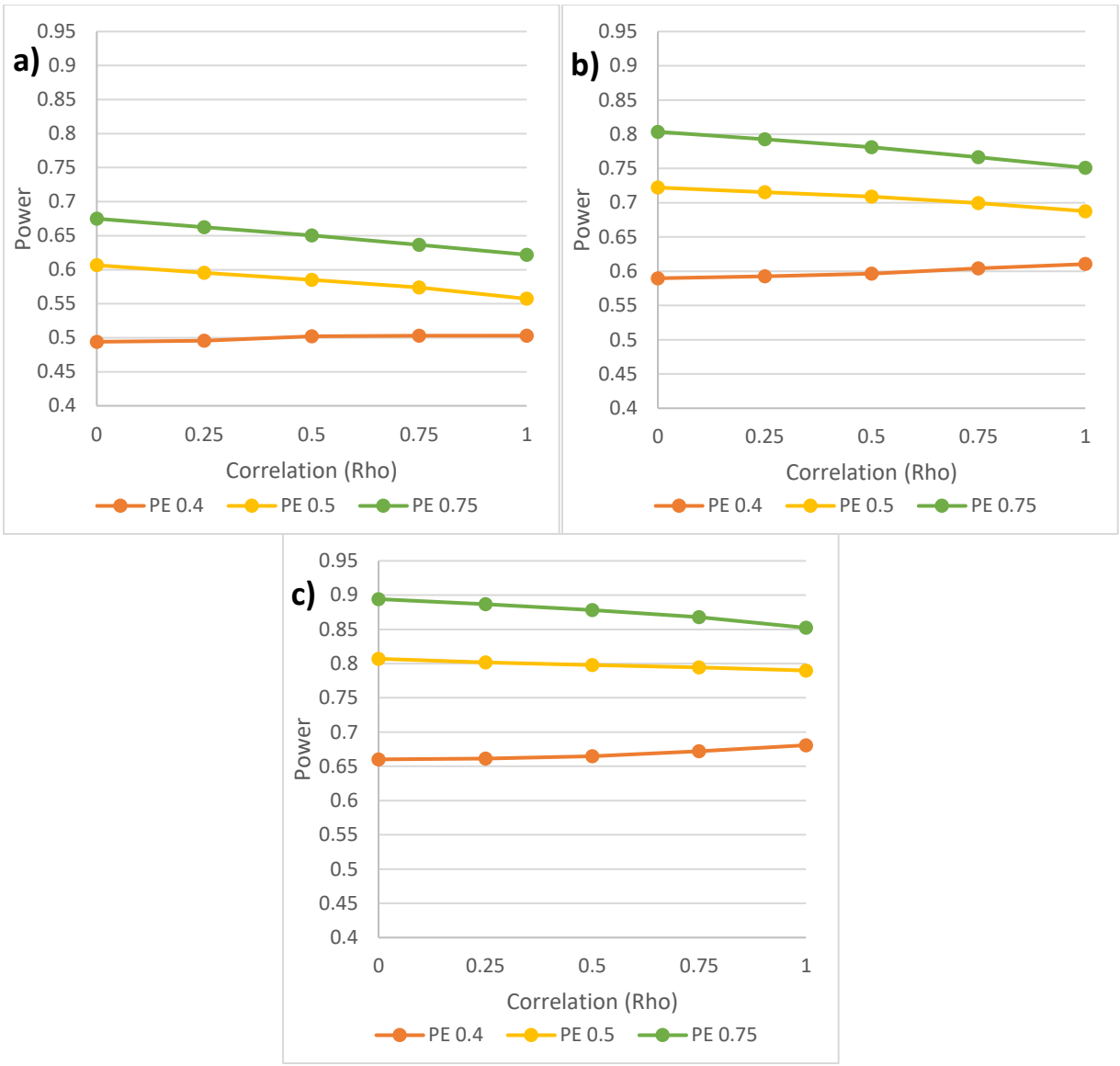

Figure 23: Bias Corrected Scenario 2C Secondary Power Trends (a) Secondary Treatment Effect of 0.3 (b) Secondary Treatment Effect of 0.35 (c) Secondary Treatment Effect of 0.4

	Secondary Treatment Effect								
	0.30			0.35			0.40		
	Primary Treatment Effect								
Rho	0.40	0.50	0.75	0.40	0.50	0.75	0.40	0.50	0.75
0	0.494124	0.606560	0.674980	0.589788	0.722232	0.803612	0.660184	0.807064	0.894172
0.25	0.495864	0.595296	0.662548	0.592520	0.715232	0.792912	0.661480	0.801768	0.886908
0.5	0.501928	0.585040	0.650560	0.596568	0.709084	0.781232	0.664920	0.797920	0.878240
0.75	0.502988	0.574072	0.636568	0.604408	0.699432	0.766340	0.672104	0.794380	0.867924
1	0.503076	0.557324	0.622056	0.610468	0.687740	0.751060	0.680788	0.789916	0.852292

Table 23: Secondary Power Results for Bias Corrected Scenario 2C

Scenario 2D: Significance Level-based Pocock– O'Brien Fleming Boundaries

The power of the study designed in this scenario falls slightly above 70% for all correlation values when 0.5 and 0.35 are the primary and secondary treatment effects respectively. This is slightly above the secondary power found when an O'Brien Fleming boundary was used for the primary endpoint in Scenario 2C. Changes in power resulting from different treatment effects or correlations in the study are shown in Figure 24.

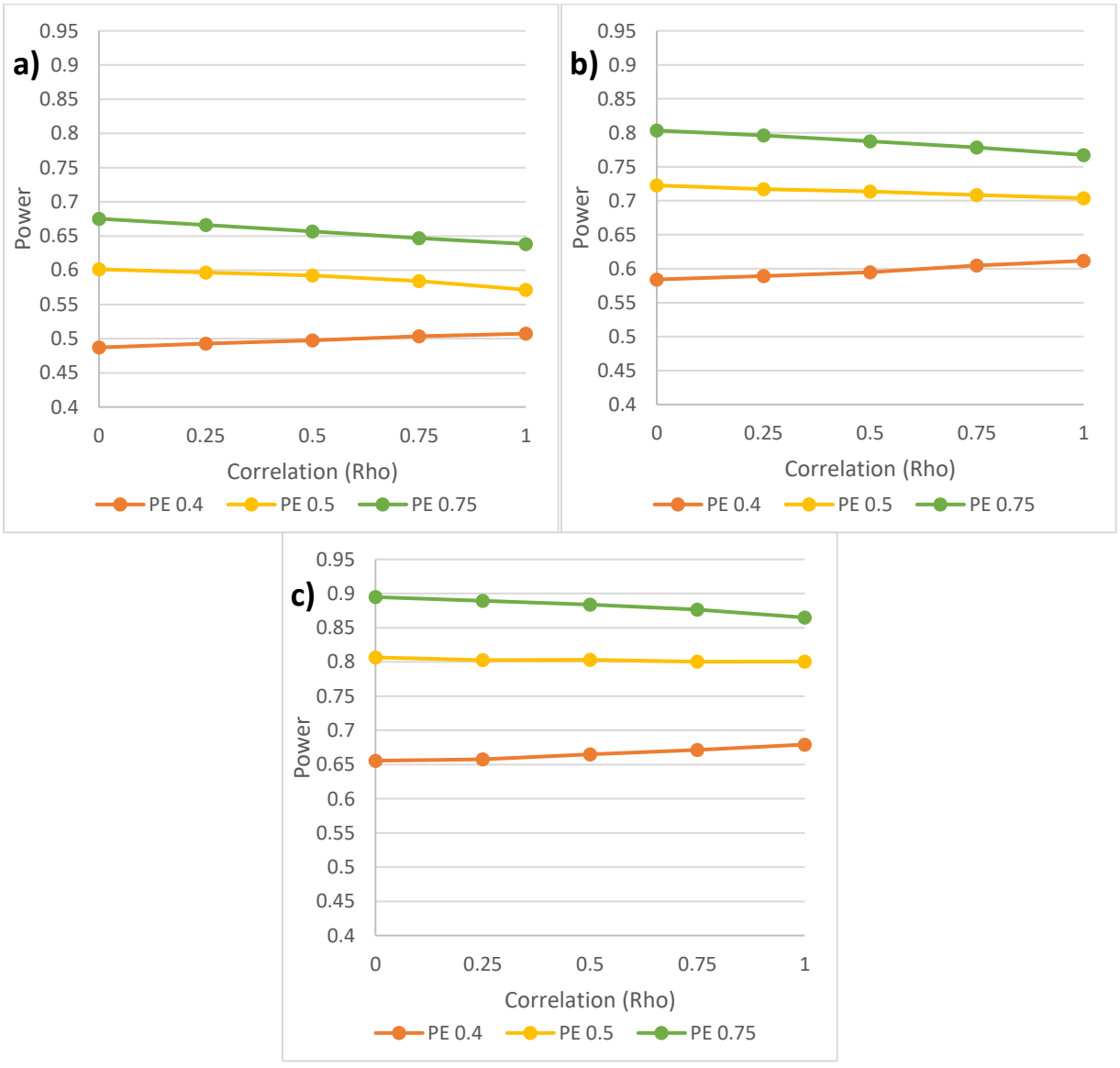

Figure 24: Bias Corrected Scenario 2D Secondary Power Trends (a) Secondary Treatment Effect of 0.3 (b) Secondary Treatment Effect of 0.35 (c) Secondary Treatment Effect of 0.4

	Secondary Treatment Effect								
	0.30			0.35			0.40		
	Primary Treatment Effect								
Rho	0.40	0.50	0.75	0.40	0.50	0.75	0.40	0.50	0.75
0	0.487020	0.601480	0.675312	0.584148	0.722624	0.803408	0.655660	0.806656	0.894916
0.25	0.492672	0.596572	0.666176	0.589296	0.717240	0.796144	0.657536	0.802544	0.889588
0.5	0.497252	0.592424	0.656632	0.594884	0.713472	0.787564	0.664980	0.803168	0.883804
0.75	0.503540	0.584396	0.646772	0.604836	0.708568	0.778372	0.671372	0.800404	0.876772
1	0.507264	0.571500	0.638416	0.611628	0.703744	0.767388	0.679212	0.800768	0.865016

Table 24: Secondary Power Results for Bias Corrected Scenario 2D

4.0 Discussion

4.1 Outcomes from Family-wise Error and Power Simulations

The investigations into the FWER of gatekeeping group sequential trial designs provides insight into the performance of error spending methods and error recycling methods and their respective abilities to strongly control the FWER when facing the problem of multiple comparisons found in all group sequential trial designs. The application of a bias correction to the test statistics for analysis in all of these methods was also able to compensate somewhat for the inflation of the FWER from the multiple comparisons.

The error spending approaches all demonstrated the same trends with respect to variation in primary treatment effects and correlation. As expected from the work of Hung et.al., as correlation between endpoints increases, so does the FWER [5]. The increase of the primary treatment effect also results in the inflation of the FWER because of the earlier rejection of the primary hypotheses on average (more frequently occurring at the interim analysis) [5]. Among these approaches, those designed with 80% power for the secondary hypothesis resulted in strong control of the FWER both before and after the bias correction. For the scenarios designed with 90% power for the secondary hypothesis, using both a Pocock and O'Brien Fleming boundary for the primary tests reached barely below the pre-specified nominal α level in the bias corrected simulations but exceeded the α level without the bias correction. This improvement in the FWER can be considered as proof of strong control of the FWER by these methods since increasing the primary treatment effect to 0.75 or 1 while holding all other values constant results in over 99% power to detect a primary treatment effect of 0.5 in the sample sizes used. Since further increases to the primary treatment effect would not increase the chance of rejecting the primary hypothesis over this 99%, and these results were consistent for all correlations including perfect correlation of $\rho=1$, it can be concluded that any further increases of the primary treatment effect size will cause the FWER to continue to plateau at this level not result in a FWER of greater than 0.025.

For the error recycling methodologies, the same trends were observed across the varying primary treatment effects and correlations as found in the error spending approaches: FWER increased with primary treatment effect and with correlation. For all four error recycling methods, the

simulated FWER approaches and slightly exceeds the pre-specified nominal α level for primary treatment effects which greatly exceed the expected treatment effect used in the study design (0.75 and 1) after implementing the bias correction. Prior to the bias correction, the original simulations also resulted in a FWER which approached and slightly exceeded the nominal α level but did so much faster along the increases of the primary treatment effect. It still can not be said that any of these error recycling methods strongly controls the FWER as the pre-specified nominal α level is exceeded for all correlation values. It is possible that a different method of correcting for the effect of multiple comparisons or the implementation of a sample size inflation proportional to the treatment effects and endpoint correlation would better control the FWER than the method of bias correction used in this report. However, a sample size inflation would negate some of the intended benefits of using a group sequential design such as decreased need participant recruitment and resources due to the potential for early stopping.

Assessment of the power levels of the group sequential trials designs demonstrated some overarching trends among both the error spending and error recycling scenarios. As expected from the nature of study power calculations, as the primary or secondary treatment effects are increased, the power of the study also increases. Correlation had differing effects based on the primary treatment effect used in the study. For studies which tested a primary effect lower than the effect level used for the study design (using 0.4 when the sample size was calculated for 0.5), increases in correlation resulted in increases in power. For primary treatment effects at or above the value used in the design of the study, increases in correlation resulted in decreases in power. These conflicting trends occur as a result of the bias correction that was applied to the secondary test statistics during analysis in the simulations. Overall, this is not the case as higher correlations typically increase study power. However, when implementing the bias correction, over-estimation of the primary treatment effect will result in larger values being subtracted from each secondary test statistic since the bias correction amount is dependent on the primary test statistic used to reject the primary hypothesis. This will causes a reduction in the secondary power of the study and a reversal of the trend between correlation and power.

When considering only the designs used in the error spending scenarios, the power values for Scenarios 1A and 1B are nearly identical with Scenario 1A being only marginally higher for low correlation values and Scenarios 1C and 1D having virtually identical power values. As shown in Figure 25, we can not detect a difference in power between using a primary O'Brien Fleming boundary compared to a primary Pocock boundary.

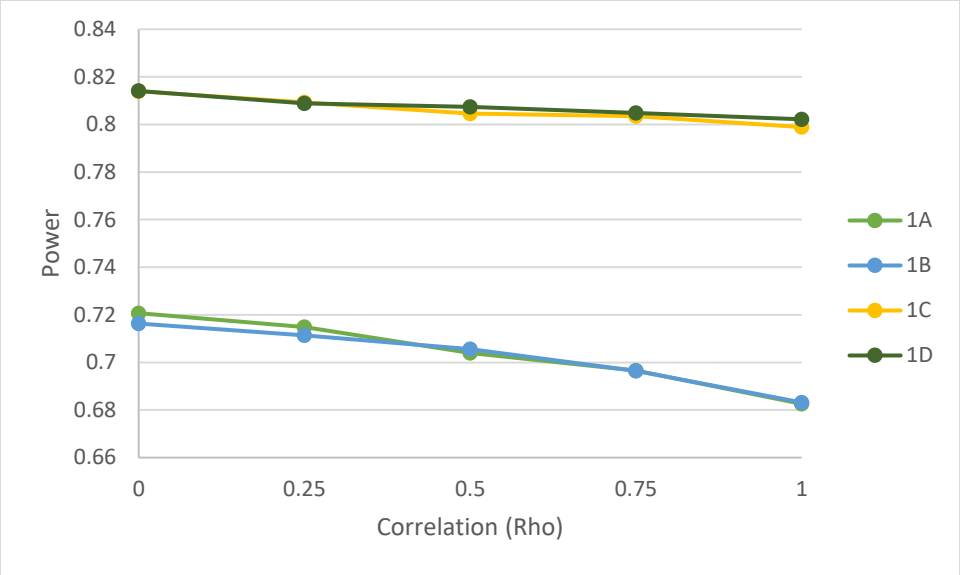

Figure 25: Error Spending Power Results for Expected Treatment Effects (μ_1=0.5 and μ_2=0.35)

When considering only the designs from the error recycling scenarios, we observe that Scenarios 2A and 2C which implemented O'Brien Fleming boundaries had lower power than their counterparts Scenarios 2B and 2D, shown in Figure 26.

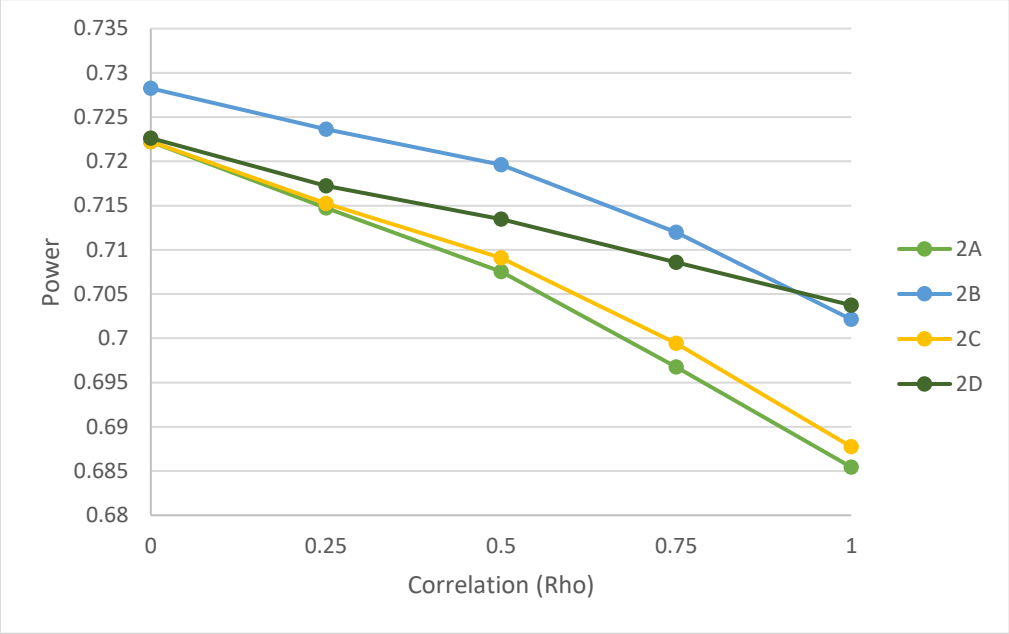

Figure 26: Error Recycling Power Results for Expected Effects (μ_1=0.5 and μ_2=0.35)

Overall, the error recycling methods in which recycled test mass was based on the significance level of the primary tests did not have significantly higher or lower power than then methods in which recycled test mass was based on the error spent in the primary tests.

4.2 Trial Design Recommendations

This investigation of type I error and power has been able to demonstrate the varying abilities of different gatekeeping group sequential trial designs to strongly control FWER in the presence of delayed secondary endpoints. As it was found that only the methods built on error spending approaches strongly controlled this FWER, these methods should be chosen for use over those involving error recycling. Among the error spending methodologies, if the intention is to improve power of the study, the sample size of both the primary and secondary endpoints should be designed for 90 percent. While an O'Brien Fleming boundary should be used for the secondary endpoint, either an O'Brien Fleming or Pocock boundary can be implemented for the primary endpoint with similar results.

4.3 Future Works

While this investigation provided insight into the possibility of strong control of the FWER for multiple group sequential design scenarios, there remain unanswered questions which would be relevant for future research and implementation in this field. Firstly, while the bias correction implemented here displayed a significant improvement in reducing the FWER in the scenarios which did not originally demonstrate strong control of the FWER, definitive strong control was not shown for the scenarios implementing error recycling methods. A different means of adjusting the test statistic or of construction of the group sequential boundary would be of value in order to allow for error recycling methodologies to be used in the testing of delayed secondary endpoints.

Next, there are many elements which come into play when designing a study. Possible adjustments to the anticipated treatment effects, correlation between endpoints, or desired power level of a study affect the potential for the design to achieve strong control of the FWER. This report does not derive a general formulation of the problem in which the dynamics of relative changes in these variables affect the results of the FWER control. A generic equation such as this would be beneficial in determining what power levels, treatment effects, and correlations would be tolerated by error recycling methodologies for them to achieve strong control of the FWER. These conditions could also be used to determine an appropriate bias correction or sample size inflation necessary to strongly control the FWER for the given parameters of a specific study. The alternative to developing these equations is an inefficient repletion of trial and error checks of different parameters in simulations until one is found which strongly controls the FWER.

A final piece of this topic which this report did not investigate was the use of Pocock boundaries for secondary endpoints. As this was done to avoid further delaying the secondary endpoint's sample size before it was known if group sequential designs would strongly control the error, the use of Pocock boundaries could be valuable for use in clinical trials as it provides increased opportunity for early stopping either due to early detection of effects or futility. Investigating scenarios involving secondary Pocock boundaries would be unnecessary, however, if the initial scenarios based on secondary O'Brien Fleming boundaries do not strongly control the family-

wise error. Since Pocock boundaries require larger sample size inflation factors than O'Brien Fleming boundaries, the Pocock boundary would produce a more delayed secondary endpoint than the O'Brien Fleming and could be hypothesized to have less control over the family-wise error rate.

References:

[1] Demets, D. L., & Lan, K. G. (1994). Interim analysis: the alpha spending function approach. *Statistics in medicine*, *13*(13-14), 1341-1352.

[2] Glimm, E., Maurer, W., & Bretz, F. (2010). Hierarchical testing of multiple endpoints in group-sequential trials. *Statistics in Medicine*, *29*(2), 219-228.

[3] Gordon Lan, K. K., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. Biometrika, 70(3), 659-663.

[4] Gou, J., & Xi, D. (2018). Hierarchical testing of a primary and a secondary endpoint in a group sequential design with different information times. *Statistics in Biopharmaceutical Research*, (just accepted), 1-26.

[5] Hung, H. J., Wang, S. J., & O'Neill, R. (2007). Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *Journal of biopharmaceutical statistics*, *17*(6), 1201-1210.

[6] Huque, M., & Mushti, S. (2015). *Alpha-recycling for the Analyses of Primary and Secondary Endpoints of Clinical Trials.* Retrieved May 16, 2019, from https://www.bassconference.org/tutorials/BASS 2015 Huque Mushti.pdf

[7] Jennison, C., & Turnbull, B. W. (1999). *Group sequential methods with applications to clinical trials*. Chapman and Hall/CRC.

[8] O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 549-556.

[9] Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*(2), 191-199.

[10] Tamhane, A. C., Gou, J., Jennison, C., Mehta, C. R., & Curto, T. (2018). A gatekeeping procedure to test a primary and a secondary endpoint in a group sequential design with multiple interim looks. *Biometrics*, *74*(1), 40-48.

[11] Tamhane, A. C., Mehta, C. R., & Liu, L. (2010). Testing a primary and a secondary endpoint in a group sequential design. *Biometrics*, *66*(4), 1174-1184.

[12] Xi, D., & Tamhane, A. C. (2015). Allocating recycled significance levels in group sequential procedures for multiple endpoints. Biometrical Journal, 57(1), 90-107.

Appendix A – Sample Size Calculations

GST Sample Size Calculations

Scenario 1A)

Primary: $\alpha = 0.05$ (this corresponds to a 0.025 one-sided boundary), $\beta = 0.1$, $\delta = 0.5$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.007$

$$85\ subject \times 1.007 \approx 85\ per\ group \rightarrow 170\ total\ people$$

Secondary: $\alpha = 0.05$ (this corresponds to a 0.025 one-sided boundary), $\beta = 0.2$, $\delta = 0.35$, $\sigma^2 = 1$, adjusted for 2 IA, $R(K, \alpha.\beta) = 1.017$

$$129\ subjects \times 1.017 \approx 131\ per\ group \rightarrow 262\ total\ people$$

Scenario 1B)

Primary: $\alpha = 0.05$ (this corresponds to a 0.025 one-sided boundary), $\beta = 0.1$, $\delta = 0.5$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.100$

$$85\ subjects \times 1.100 \approx 93\ per\ group \rightarrow 186\ total\ people$$

Secondary: $\alpha = 0.05$ (this corresponds to a 0.025 one-sided boundary), $\beta = 0.2$, $\delta = 0.35$, $\sigma^2 = 1$, adjusted for 2 IA, $R(K, \alpha.\beta) = 1.017$

$$129\ subjects \times 1.017 \approx 131\ per\ group \rightarrow 262\ total\ people$$

Scenario 1C)

Primary: $\alpha = 0.05$ (this corresponds to a 0.025 one-sided boundary), $\beta = 0.1$, $\delta = 0.5$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.007$

$$85\ subjects \times 1.007 \approx 85\ per\ group \rightarrow 170\ total\ people$$

Secondary: $\alpha = 0.05$ (this corresponds to a 0.025 one-sided boundary), $\beta = 0.9$, $\delta = 0.35$, $\sigma^2 = 1$, adjusted for 2 IA, $R(K, \alpha.\beta) = 1.016$

$$173\ subjects \times 1.016 \approx 175\ per\ group \rightarrow 350\ total\ people$$

Scenario 1D)

Primary: $\alpha = 0.05$ (this corresponds to a 0.025 one-sided boundary), $\beta = 0.1$, $\delta = 0.5$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.100$

$$85\ subjects \times 1.100 \approx 93\ per\ group \rightarrow 186\ total\ people$$

Secondary: $\alpha = 0.05$ (this corresponds to a 0.025 one-sided boundary), $\beta = 0.9$, $\delta = 0.35$, $\sigma^2 = 1$, adjusted for 2 IA, $R(K, \alpha.\beta) = 1.016$

$$173\ subjects \times 1.016 \approx 175\ per\ group \rightarrow 350\ total\ people$$

Error Recycling Sample Size Calculations

Scenario 2A)

Primary: $\alpha = 0.05$ (this corresponds to a 0.025 one-sided boundary), $\beta = 0.1$, $\delta = 0.5$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.007$

$$85\ subjects \times 1.007 \approx 85\ per\ group \rightarrow 170\ total\ people$$

Reject at IA

Secondary: $\alpha = 0.05$ (corresponds to a 0.025 one-sided boundary), $\beta = 0.2$, $\delta = 0.35$, $\sigma^2 = 1$, adjusted for 1 IA, $R(\mu, \sigma.\rho) = 1.008$

$$129\ subjects \times 1.008 \approx 130\ per\ group \rightarrow \ \ 260\ total\ people$$

Reject at Final Analysis

Secondary: $\alpha = 0.0462$ (this corresponds to a 0.0231 one-sided boundary), $\beta = 0.2$, $\delta = 0.35$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.008$

$$132\ subjects \times 1.008 \approx 133\ per\ group \rightarrow 266\ total\ people$$

*** $c_2 \ = \ 2.2305$; $\alpha = P(Z \geq 1.9935) = 0.0231$ ***

Scenario 2B)

Primary: $\alpha = 0.05$ (corresponds to a 0.025 one-sided boundary), $\beta = 0.1$, $\delta = 0.5$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.100$

$$85\ subjects \times 1.100 \approx \ 93\ per\ group \rightarrow 186\ total\ people$$

Reject at IA

Secondary: $\alpha = 0.05$ (corresponds to a 0.025 one-sided boundary), $\beta = 0.2$, $\delta = 0.35$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.008$

$$129\ subjects \times 1.008 \approx 130\ per\ group \rightarrow \ \ 260\ total\ people$$

Reject at Final Analysis

Secondary: $\alpha = 0.02572$ (corresponds to a 0.01286 one-sided boundary), $\beta = 0.2$, $\delta = 0.35$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.008$

$$155\ subjects \times 1.008 \approx 156\ per\ group \rightarrow 312\ total\ people$$

***$c_2 \ = \ 2.2305$; $\alpha = P(Z \geq 2.2305) = 0.01286$ ***

Scenario 2C)

Primary: $\alpha = 0.05$ (corresponds to a 0.025 one-sided boundary), $\beta = 0.1$, $\delta = 0.5$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.007$

$$85\ subjects \times 1.007 \approx 85\ per\ group \rightarrow 170\ total\ people$$

Reject at IA

Secondary: $\alpha = 0.05$ (corresponds to a 0.025 one-sided boundary), $\beta = 0.2$, $\delta = 0.35$, $\sigma^2 = 1$, adjusted for 1 IA

$$129\ subjects \times 1.008 \approx 130\ per\ group \rightarrow \ \ 260\ total\ people$$

Reject at Final Analysis

Secondary: $\alpha = 0.03764$ (corresponds to a 0.01882 one-sided boundary), $\beta = 0.2$, $\delta = 0.35$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.008$

$$140\ subjects \times 1.008 \approx 141\ per\ group \rightarrow 282\ total\ people$$

*** α spent at final analysis of primary test is 0.01882 ***

Scenario 2D)

Primary: $\alpha = 0.05$ (corresponds to a 0.025 one-sided boundary), $\beta = 0.1$, $\delta = 0.5$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.100$

$$85\ subjects \times 1.100 \approx 93\ per\ group \rightarrow 186\ total\ people$$

Reject at IA

Secondary: $\alpha = 0.05$ (corresponds to a 0.025 one-sided boundary), $\beta = 0.2$, $\delta = 0.35$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.008$

$$129\ subjects \times 1.008 \approx 130\ per\ group \rightarrow \ \ 260\ total\ people$$

Reject at Final Analysis

Secondary: $\alpha = 0.01458$ (corresponds to a 0.00729 one-sided boundary), $\beta = 0.2$, $\delta = 0.35$, $\sigma^2 = 1$, adjusted for 1 IA, $R(K, \alpha.\beta) = 1.008$

$$178\ subjects \times 1.008 \approx 178\ per\ group \rightarrow 356\ total\ people$$

*** α spent at final analysis of primary test is 0.00729 ***