Sequential Rematched Randomization and Adaptive Monitoring with the Second-Generation p-value to increase the efficiency and efficacy of Randomized Clinical Trials

By

Jonathan Joseph Chipman

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biostatistics

May 10, 2019

Nashville, TN

Approved:

Frank Harrell, Ph.D.

Robert A. Greevy, Jr, Ph.D.

Jeffrey Blume, Ph.D.

Lindsay Mayberry, Ph.D.

To the entire Chipman family:

My parents, siblings and their families, and fiancée Danielle.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Randomized Trials are considered the gold standard in establishing the benefit of an intervention. A review of Phase III trials funded by the National Institute of Neurological Disorders and Stroke systematically found the overall societal benefit of their trials to be much greater than the total costs (Johnston et al. 2006). Societal costs that were that of trials ending pre-maturely – before reaching a clear clinical conclusion when it could have otherwise continued. While the overall societal benefit is great for randomized clinical trials, the face-value cost may be prohibitive for starting a trial. Among agents approved by the Food and Drug Administration between 2015-2016, the median enrollment was 488 patients (IQR, 230 – 740) with a median cost per patient of $41,117 (IQR, $31,802 - $82,362) (Moore et al., n.d.). The long-ago rally cry endures that: "Reducing the costs of trials is absolutely crucial for the public good" (Collier 2009). Regardless of how financially well-endowed is a clinical trial, all investigators seek to reach a clear clinical conclusion while minimizing burderns and resources. From regulatory bodies, to industry, to academia, all can agree on the need to run an efficient and efficacious trial.

In this body of work, we develop new methods that increase the efficiency and balance of clinical trials and the ability to follow studies until reaching clear clinical conclusions. The developments include novel extensions to Sequential Matched Randomization and adaptive monitoring of clinical trials using the Second Generation p-value. While the context focuses on randomized trials, both contain insights that extend into efficient and efficacious use of observational studies.

Randomization removes systematic confounding and, for the Frequentist, further provides the foundation of accurate estimates of uncertainty in estimating treatment effects. While the randomization space provides these overall benefits, any single instance of randomization includes at least some degree of imbalance on baseline covariates. To reduce the risk of imbalances, trialists often turn to Stratified Block Randomization, which allocates treatments within categorical profiles (strata) of select baseline covariates. It is easy to implement yet is limited to balancing on categorical, or categorized, baseline covariates. The number of strata increases multiplicatively for

each baseline covariate used to create strata.

In Chapter 1, we extend Sequential Matched Randomization which has already been shown to provide greater efficiency and balance than Stratified Block Randomization. Matched Randomization is a refined form of Stratified Randomization which finds the set of patients (or clusters) which collectively are most similar to each other based upon a distance matrix. Randomization occurs within each pair. As patients enter a study sequentially, some optimality is lost in determining the best pairing of patients. Our extensions allow for (1) dynamically updating an empirically estimated matching threshold throughout the study and (2) breaking matches when a better match enters the study. These extensions nearly regain the optimality lost when patients all known and matched prior to randomization. We show through the REACH clinical trial case-study that randomization-based inference under our method can achieve nearly the same efficiency as a fully-adjusted linear model. And we are able to increase overall balance of baseline covariates which is important for study that which to explore subgroup analyses for personalizing medicine.

While chapter one makes a case for our method, it also provides insight to other covariate-adjusted randomization schemes and observational studies. We note that minimization schemes vary in allowing a degree of randomization. Some provide little randomization while other have embedded an element of randomization. To the extent that randomization is included with minimization, it is a powerful contender for increasing efficiency and balance of baseline covariates in trials. As for the balance of baseline covariates, the case study sheds insights on how overall balancing can be influenced by the distribution of each baseline covariate.

Randomization itself comes with a great degree of variability that can affect the sample size assumptions made when designing a clinical trial. Traditional sample size estimates are often based upon a minimal size to detect an effect to be statistically significant. However, achieving statistical significance alone does not imply a novel target is beneficial to the scientific community.

In Chapter Two we introduce an adaptive monitoring design grounded on following studies until either ruling out effects deemed trivial to the null hypothesis or until ruling out a highly actionable effect that would change practice. These effects are ruled out making use of the Second Generation p-value which draws inference upon interval hypotheses. Under any inferential paradigm, one calculates the overlap between an estimated interval and the interval hypothesis. Complete overlap is evidence for the

hypothesis, zero overlap is evidence for the alternative, and partial overlap indicates an inconclusive study.

Monitoring a study until ruling out trivial or highly actionable effects endows the study with clinical relevance and many statistically beneficial properties. False discoveries most commonly occur when an effect is close to the null hypothesis (the corresponding p-value is barely less than alpha). Following a study until ruling out trivial effects reduces the false discovery probability by eliminating those effects estimated to be close to the point null. We compare our adaptive monitoring design with monitoring with posterior probabilities which may similarly follow studies until ruling out the set of trivial or highly actionable effects.

To improve operational characteristics, the design waits until a period of time before applying monitoring rules and requires affirming alerts to stop a trial. The wait time reduces errors which may occur before the estimated effect and interval have sufficiently stabilized. And, the affirmation step reduces bias inherent in all adaptive monitoring designs.

The operating characteristics of these adaptive monitoring designs require simulation, which may be a barrier to implementation. We provide, in Chapter Three, an R package, sgpvAM, which simulates adaptive monitoring with the second generation p-value. The function is versatile to allow the user to provide their own generated MCMC replicates with intervals or to generate MCMC replicates with effects and outcomes under any random distribution. The output allows the user to address practical considerations such as the probability of a study ending by a certain number of observations and the probability of study being inconconclusive. An inconclusive study can occur when outcomes are not observed immediately. Observed outcomes may suggest stopping for ruling out trivial or highly actionable effects, yet the study may not rule out the same effects after observing the remaining outcomes.

In this chapter, we conduct extensive simulations to then make practical recommendations. The Type I error is minimized, and less than 0.05, when waiting until the interval width has the length equal to the absolute midpoint between the trivial and highly actionable effects. This holds true even when allowing an unrestricted, infinite sample size. Though not investigated in this work, this open the door in a principled manner for restarting trials or continuing to gather observational data as it accumlates. In many trials, observations are not immediately observed relative to enrollment. By increasing the number of steps required to affirm an alert, one is able to substantively

reduce the risk of an inconclusive study when outcomes are not immediately observed relative to enrollment.

This body of work lays a foundation for many future developments. ...

CHAPTER 2

REMATCHING ON-THE-FLY: SEQUENTIAL REMATCHED RANDOMIZATION AND A CASE FOR COVARIATE-ADJUSTED RANDOMIZATION

2.1 Introduction

Randomization eliminates systematic confounding and has been considered the "gold standard" for clinical trials since World War II(Bothwell et al. 2016). And yet, the most common approaches to randomization — Block and Stratified Randomization — are just as old (Peirce and Jastrow 1884; Hill 1952; Armitage 1982). They are limited in their ability to control the balance of baseline covariates. Covariate-adjusted randomization, which includes Stratified Block Randomization, eliminates from the randomization space singular randomizations with poor baseline covariate balance and consequently reduces the uncertainty in permutation-based inference. It also increases the efficiency of model-based inference by orthogonalizing baseline covariates with treatment assignment.

Balance is an important aim for clinical trials. Poor chance imbalances on key baseline covariates can bring to question the face-validity of a randomized trial(Rosenberger and Sverdlov 2008; Leyland-Jones 2003) especially when estimating the Population Average Treatment Effect from the Sample Average Treatment Effect. Also, greater covariate balance allows trialists to estimate heterogenous treatment effect across key subgroups. Trials with secondary aims for personalizing medicine to key subgroups rely upon overall covariate balance(Diener-West et al. 1989; Fu, Zhou, and Faries 2016). To reduce chance imbalances, the trialist may turn to covariate-adjusted randomization (the focus of this paper) and / or fit a model that adjusts for chance imbalances(Berchialla, Gregori, and Baldi 2018). With model-based inference, the trialist must convince the clinical community their model is fully transparent and sufficiently adequate(D. A. Freedman 2008a, 2008b; Lin 2013).

In non-sequential trials, where all patients or clusters are known before randomization, covariate-adjusted randomization randomizes within optimal partitions of the patients. Common and novel methods (and the imbalance measure they optimize) include: Stratified Block Randomization (exact covariate matches)(Fisher 1935), Matched Randomization (overall sum of paired distances from a distance matrix)(Greevy

et al. 2004), Rerandomization (any imbalance measure though commonly overall difference in covariate distances)(Morgan and Rubin 2012), Propensity-Constrained Randomization (overall difference in propensity score variability)(Loux 2014), and Kernel Randomization (a linear or non-linear function of covariate differences) (Kallus and 2018 2018). Stratified Block Randomization randomizes within exact matches of categorized baseline covariate levels whereas the other methods randomize within near matches of continuous and categorical baseline covariates. Exact matching is frequently limited in the number of matching baseline covariate levels before overwhelming the sample size. Though still under development, some theoretical and empirical evidence points to Kernel Randomization as the preferred method for greatest balance(Kallus and 2018 2018).

With sequential enrollment, the full covariate pattern and optimal partitions are unknown until the end of the study. Many of the non-sequential methods have extensions for sequential enrollment yet lose some optimization compared to non-sequential randomization. Minimization handles the sequential optimization problem by directly allocating treatments to achieve optimal balance on specified imbalance criteria(Taves 1974). Randomization occurs when all treatment would have the same impact on minimizing the imbalance criteria. Without randomization, Minimization schemes may remain subject to systematic confounding. To relax deterministic allocation, treatment may instead be allocated with a biased coin favoring the optimal treatment per the imbalance criteria (Pocock and Simon 1975; Atkinson 1982). A common covariate-adjusted biased coin method includes Urn Randomization(Wei and Lachin 1988). Of the remaining non-sequential randomization schemes, Sequential Matched Randomization and Sequential Rerandomization set a tuning parameter to sequentially optimize balance according to their imbalance measure(Kapelner and Krieger 2014; Zhou et al. 2018).

This work achieves two purposes. First, it extends Sequential Matched Randomization to recover some of the optimal balance achieved in single batch Matched Randomization (when all patients or clusters are known prior to randomization). The extensions allow (1) matches to rematch if a better mate enters the study, (2) the tuning parameter for optimal matches to adjust dynamically according to the number of current matches and covariate distribution, and (3) patients to enroll in blocks. Second, through a two-arm case study (n = 512 patients) it re-emphasizes the value of sequential covariate-adjusted methods, as a whole, compared to Block Randomization. Sequential allocation methods considered include Block Randomization, Stratified Block Randomization,

Urn Randomization, Sequential Matched Randomization, Atkinson's Minimization with and without a biased coin, and Begg and Iglewicz's Minimization. We do not include Sequential Rerandomization as it requires determining an optimal threshold.

2.2 Notation

We denote a study as having enrolled $i = 1, \ldots,$ to N patients throughout $b = 1, \ldots,$ to M batches of sequential enrollment. At the b^{th} batch of enrolling patients, denote the set of unmatched patients as U_b and the number of patients in the set as $||U_b||$. Let $||R_b||$ be the number of expected remaining study entrants. In general, covariate-adjusted randomizations are adjusted to p baseline covariates in which a categorical covariate having q levels is coded by $q - 1$ dummy variables.

In the simulation section, we generate $j = 1, \ldots,$ to 20,000 randomization schedules for each allocation scheme.

2.3 Sequential Matching and Sequential Rematching

Sequential Matched Randomization uses the matching on-the-fly algorithm published elsewhere(Kapelner and Krieger 2014) yet is worth summarizing to then extend. When the first set of patients individually enroll, they are randomized to either treatment or control and form a "reservoir" of patients who have been randomized but, as of yet, have not matched with any other patient. Once a reservoir of pre-specified size has built up, subsequent enrolling patients are compared to reservoir members on baseline covariates using a distance matrix. The entering patient becomes mates with their best reservoir match if they meet a pre-specified distance threshold of similarity, and the entering patient receives the opposite treatment randomized to their mate. Both are excluded from the reservoir of potential mates. If, however, the best reservoir match is not similar enough, the entering patient is randomized and joins the reservoir awaiting a mate. By the end of the study, some patients may not have found a mate, and treatment allocation may be imbalanced. Hereafter, we'll refer to the matching-on-the-fly algorithm as Sequential Matching.

Common practice uses Mahalanobis Distance and builds the initial reservoir to p+2 patients. This distance matrix reduces to euclidean distances when covariates are independent, and p+2 observations are required before distances may be uniquely

estimated. Current literature suggests setting the pre-specified threshold as a quantile from the F(p, i - p) distribution(Kapelner and Krieger 2014); justification lies in the fact that distances of normal covariates, when appropriately scaled, asymptotically follow the F(p, i-p) distribution.

Throughout Sequential Matching, a patient may match before their better matching mate enrolls. We allow mates to break and rematch so long as patients do not match within their treatment group. For example, in a two arm study, controls are not allowed to match with controls and treatments to treatments. We refer to this re-matching on-the-fly algorithm as Sequential Rematching and the randomization scheme as Sequential Rematched Randomization.

2.4 Dynamic and Empirical Threshold

Sequential Matching, as is, uses a fixed quantile of the F-distribution; we formalize the use of a dynamic threshold from an empirically-estimated distribution of randomly matched distances. Lacking omniscience, an improperly selected fixed similarity threshold may be problematic. An overly strict fixed threshold would yield no matches and in turn may be no better than Complete Randomization. In contrast, an overly relaxed threshold would degenerate to a block-two randomization scheme vulnerable to subversion bias. Our proposed dynamic threshold reflects the chance of matching an existing reservoir member out of all potential mates including those yet to enroll. More formally, this is the proportion Q_b where

$$Q_b = \frac{||U_b|| - 1}{||U_b|| + ||R_b|| - 1}.$$

Although the number of possible mates in Rematching is the whole set of entrants, we develop Q_b for use with both Sequential Matching and Sequential Rematching.

Since not all covariates are normally distributed, we empirically estimate a reference null distribution, F, as suggested in the matching on-the-fly paper(Kapelner and Krieger 2014). To estimate F at the b^{th} batch of enrolling patients (F_b), a random set of $i/2$ matches are bootstrap sampled from the upper- (or lower-) triangle of the distance matrix. The dynamic threshold is the averaged Q_b percentile across bootstrap samples of F_b. To achieve equal treatment allocation, the threshold for matches is removed when the reservoir size is the same or less than the number of patients left to enroll.

$$Threshold_b = \begin{cases} \hat{F}_b^{-1}(Q_b) & ||U_b|| < ||R_b|| \\ \text{best match(es)} & ||U_b|| \geq ||R_b|| \end{cases}$$

The threshold may also be removed to keep the reservoir size within a specified maximum tolerated imbalance (Berger, Ivanova, and Deloria Knoll 2003).

2.5 Batch Entry

As is, Sequential Matching requires matching patients one at a time. However, trials will allocate treatments in batches of various enrollment sizes. As when all patients are known at the study outset, batch enrollment increases the chance of finding a better mate. We extend Sequential Matching to allow for batches of multiple patients using an "optimal" algorithm; mates are collectively determined based upon the set of matches that yield the smallest sum of distances. The R package nbpMatching easily finds optimal mates.

2.6 Case Study: REACH Trial

2.6.1 Context, Data Preparation, and Simulation Set up

The Rapid Education/Encouragement And Communications for Health (REACH) randomized clinical trial(Nelson et al. 2018) provides text message-delivered diabetes support for 12 months to help diabetic patients manage glycemic control (as measured by Hemoglobin A1c) and adhere to treatment medication. Patients are randomized into one of three treatment arms with a 2:1:1 allocation ratio: no text message, text message, text message and monthly phone coaching. Now with complete enrollment, 512 patients are currently being followed up through 3, 6, 12, and 15 months.

At baseline, clinical and demographics covariates were collected that may be associated with 12 month Hemoglobin A1c, including baseline Hemoglobin A1c, age at enrollment, gender, years of education, years of diabetes, race / ethnicity, medication type, income, and type of insurance. Refer to supplement for the overall distribution of each baseline covariate.

Randomization often occurred before all baseline covariates could be collected. Most notably, baseline Hemoglobin A1c generally took additional time to process. In this

study, clinical coordinators enrolled patients and sent baseline data to the Data Coordinating Center for weekly batch randomization. Though REACH follow-up continues, all baseline covariates have been obtained, and 418 patients have recorded three month Hemoglobin A1c.

2.6.2 Simulations

With complete baseline covariates, we simulated the ability of various treatment allocation schemes to achieve balance among baseline covariates among treatment groups and compared the efficiency in estimating predicted three month Hemoglobin A1c. Contender allocation schemes focused on Block Randomization, Stratified Block Randomization, Urn$(0, \beta)$ Randomization, single batch Matched Randomization, Sequential Matched Randomization without the proposed extensions (similar to current literature), Sequential Matched Randomization with the dynamic threshold, Atkinson's Minimization Algorithm with a 2/3 biased coin (Efron 1971) and with deterministic allocation, and Begg and Iglewicz Minimization. The minimization schemes aim to reduce the standard error in estimating the treatment effect.

We carried out Block and Stratified Block Randomization using a block size of two. In practice, blocks of size two are ill-advised for increasing the risk of subversion(Berger, Ivanova, and Deloria Knoll 2003) but provides in simulation the greatest chance of equal treatment allocation especially when stratifying with many levels. As a point of reference we included single batch Matched Randomization as the optimal matching when all patients are known prior to randomization; and, to compare against current Sequential Matched Randomization literature, we included Sequential Matched Randomization with a fixed 20^{th} percentile threshold of the F(p, i-p) distribution. Atkinson's Algorithm determines the most impactful treatment for reducing the estimated treatment variability of a pre-specified model, which we specified as conditioning on all categorical covariates and first through third degree polynomials of continuous covariates. When there were fewer observations than model degrees of freedom, we determined the optimal treatment for the pre-specified model using the generalized inverse(Senn, Anisimov, and Fedorov 2010).

For a realistic comparison to Stratified Block Randomization, each covariate-adjusted allocation was adjusted to site and the most predictive baseline covariates of three month Hemoglobin A1c — baseline Hemoglobin A1c, medication type, and time since diabetes diagnosis (see supplement for how these were determined). And, though

impractical for Stratified Block Randomization, each covariate-adjusted allocation scheme conditioned on all baseline covariates. Stratified Randomization, Urn(0,β), and Begg and Iglewicz Minimization require categorized covariates; for these schemes we conditioned on categorized derivations of baseline Hemoglobin A1c ($< 7, 7-8, 8+$), age ($\leq 60, > 60$), years of education ($\leq 12, > 12$), and time since diabetes diagnosis ($< 10, 10+$).

Though REACH includes three treatment arms, we simplified to two equally allocated arms — receiving no or any text-message intervention. Missing baseline covariates and three month Hemoglobin A1c were multiply imputed using predictive meaning matching via the aregImpute function in the R rms package; and single mean and mode imputation from the multiple imputations was carried out to obtain a single complete dataset. Twenty thousand treatment allocation schedules for each scheme were simulated on the single mean imputed dataset. For each scheme's generated 20,000 treatment schedules, we summarize with boxplots the maximum (worst-case) and average absolute Standardized Mean Difference of all baseline covariates. We also summarize end-of-study treatment allocation difference.

To compare the efficiency of different schemes, we generated outcomes assuming a potential outcomes framework. Twenty thousand datasets were created where a patient's outcome under the control arm equaled predicted three month Hemoglobin A1c plus a random residual. Under the REACH intervention arm, the outcome decreased by 0.5 (a beneficial decrease in this population). By using predicted three month Hemoglobin A1c, we fixed the effect of baseline covariates. The only random components in this framework are the treatment assignment and a random residual of predicted three month Hemoglobin A1c.

For the j^{th} generated potential outcomes dataset, we obtained an observed study for each allocation scheme using the scheme's j^{th} generated allocation schedule. From each observed study, we calculated the permutation-based 95% Confidence Interval Width of the Sample Average Treatment Effect (difference in observed means) and model-based 95% Confidence Interval Width of the estimated treatment effect. The pre-specified Ordinary Least Square Model adjusted for all baseline covariates with restricted cubic splines on each continuous covariate and accounted for multiply imputed baseline covariates. On the j^{th} potential outcomes dataset, we also calculated the relative width of each scheme's Confidence Interval Width compared to the width from Block Randomization. Block Randomization was carried out twice for each generated dataset to compare one instance of Block Randomization to another. From this we calculated

the difference in effective sample size of each scheme relative to Block Randomization.

In the supplement, we further calculate the average size of the reservoir throughout enrollment among matched randomization methods and the expected amount of randomization occurring under each minimization scheme.

2.6.3 Simulation Results

By themselves, the set of most predictive covariates, including site, are highly predictive of three-month Hemoglobin A1c ($R^2 = 0.43$); adjusting for all baseline covariates is slightly more predictive ($R^2 = 0.46$). Atkinson's Minimization only randomized the first patient and is therefore excluded from comparisons regarding permutation-based inference. For this reason, it is also difficult to draw conclusions regarding its average performance balancing baseline covariates.

Balance of Covariates: As a point of reference and across simulations, Block Randomization's median worst-balanced baseline covariate had an absolute Standardized Mean Difference of 0.21 (95% Percentile Confidence Interval: (0.13, 0.32); Figure 1) and average absolute Standardized Mean Difference of 0.09 (0.06, 0.14) (Figure 2).

Adjusting randomization to the most predictive baseline covariates yielded only trivial gain in lessoning the worst-case imbalance (maximum absolute Standardized Mean Difference; Figure 1). Begg and Iglewicz Minimization performed best in reducing the worst-case imbalance to 0.20 (0.12, 0.30). Income level and race / ethnicity remained difficult covariates to balance (supplement).

Worst-case imbalance was better controlled when adjusting to all baseline covariates (Figure 1). Sequenital Matched Randomization, utilizing a dynamic and empirical threshold, reduced the absolute Standardized Mean Difference to 0.14 (0.09, 0.22). This slightly improved upon Sequential Matched Randomization without these extensions 0.16 (0.10, 0.25). Sequential Rematched Randomization had a worst-case imbalance of 0.11 (0.07, 0.20), which was an improvement in recovering the performance of a single batched Matched Randomization (0.09 (0.05, 0.16)). Begg and Iglewicz Minimization performed best among the sequential allocation methods (0.10 (0.05, 0.18)).

Overall baseline covariate imbalance (average absolute Standardized Mean Difference) improved similarly and non-trivially across allocation schemes when adjusting randomization to most predictive baseline covariates (Figure 2). Again, greater improvement occurred when adjusting to all baseline covariates. Adding the dynamic and empirical

threshold improved the performance of Sequential Matched Randomization, and Sequential Rematched Randomization recovered much of the optimal performance of single batch Matched Randomization. Begg and Iglewicz performed essentially as well as single batch Matched Randomization.

Improvement Upon Precision: Block Randomization achieved a treatment effect permutation-based 95% Confidence Interval width of 0.54 (0.50, 0.58) (Figure 3) and a fully-adjusted (all baseline covariates) ordinary least squares 95% Confidence Interval width of 0.38 (0.37, 0.39) (Figure 4). Relative to itself, the 95% Percentile Confidence Interval of change in effective sample size was +/- 20 (Table 1).

The greatest precision gains to permutation-based inference came from schemes adjusting for only the most predictive baseline covariates (Figure 3). An exception was Begg and Iglewicz Minimization, which performed essentially the same on median when adjusting for the most predictive versus all baseline covariates.

Excepting Urn Randomization, all other schemes that adjusted allocation to the most predictive covariates achieved a Confidence Interval width of 0.43 or smaller, with single batch Matched Randomization achieving on median the same efficiency as the model-based efficiency under Block Randomization (Figure 3). Sequential Matched Randomization with the dynamic and empirical threshold yielded a permutation-based Confidence Interval width of 0.41 (0.38, 0.44), an improvement on Sequential Matched Randomization without extensions 0.43 (0.40, 0.47). Sequential Rematched Randomization (0.40 (0.36, 0.45)) further recovered some of the efficiency lost from single batch Matched Randomization (0.38 (0.36, 0.41)). On median, Atkinson's Algorithm with a Biased Coin yielded the greatest permutation-based precision, nearly doubling the effective sample size compared to Block Randomization. That is, Block Randomization would have needed on median 436 (316, 572) additional patients to achieve the same amount of precision. Sequential Rematched Randomization increased the effective sample size by 416 (213, 663).

When adjusting randomization to all baseline covariates, efficiency gains in permutation-based inference were not as pronounced but still substantial (Figure 3). Sequential Rematched Randomization achieved a permutation-based Confidence Interval Width of 0.46 (0.41, 0.51). Atkinson's Minimization with a Biased Coin and Begg and Iglewicz Minimization achieved the greatest permutation-based Confidence Interval Width precision, 0.41 (0.40, 0.42) and 0.42 (0.38, 0.49) respectively. On median, these were both superior to single batch Matched Randomization.

In contrast to permutation-based inference, schemes that adjusted to all baseline covariates achieved the greatest gain in precision from a linear model adjusting to all baseline covariates (Figure 4). While non-trivial, the gains relative to Block Randomization were much less pronounced.

When adjusting to all baseline covariates, the best performing sequential enrollment schemes included Sequential Rematched Randomization, Atkinson's Algorithm with and without a Biased Coin, and Begg and Iglewicz Minimization with Confidence Interval widths of at most, on median, 0.376. Sequential Rematched Randomization yielded an effective increase in sample size of 15 (-4, 37) above and beyond the pre-specified model-based inference under Block Randomization.

End-of-Study Allocation Differences: Block Randomization, Sequential Matched Randomization with a dynamic and empirical thresold, Sequential Rematched Randomization, and single batch Matched Randomization always achieved equal end-of-study allocation (Figure 5). Begg and Iglewicz Minimization was close to achieving equal end-of-study allocation. Atkinson's Algorithm with a Biased Coin, which performed well in achieving balance and efficiency ran a greater risk of ending the study with a treatment imbalance 0.00 (-20.00, 20.00).

Randomization within Minimization: When adjusting Begg and Iglewicz Minimzation to the most predictive covariates, randomization occurred for roughly 20% of patients (7.8% of patients when adjusting to all baseline covariates). See supplement.

Reservoir Size: See supplement for the average reservoir size for Sequential Matched and Rematched Randomization.

2.7 Conclusions

We've introduced extensions to Sequential Matched Randomization that recover much of the optimality lost from single batch Matched Randomization due to sequential entry of patients. These extensions include Sequential Rematching with a dynamic and empirical threshold and the ability to randomize patients in enrollment blocks. Further, our case study re-emphasizes (Ciolino et al. 2011) the value of covariate-adjusted randomization for increasing overall covariate balance and efficiency estimating the Population Average Treatment Effect. The precision of the permutation-based Sample Average Treatment Effect estimator achieved nearly the same efficiency as model-based inference estimating the treatment effect.

Our work is consistent with other findings that there is a balance / efficiency trade-off in choosing which covariates to adjust randomization(Kallus and 2018 2018; **???**). Arguments can be made to prioritize balance, especially when investigating important subgroups and when the strength of relationship between baseline covariates and the outcome is unknown. And, arguments are made for prioritizing efficiency especially when chance imbalances may be adjusted with a model (Medicine and 1999 1999). Weighted distance measures can allow an investigator to design the study prioritizing certain covariates over others(Greevy Jr. et al. 2012).

In introducing Sequential Rematching, we add to a class of look-back allocation schemes. Though patients have matched, their baseline covariates may yet be updated or used again to increase covariate balance and precision. Also in this class are Minimization schemes with and without a biased coin and Sequential Rerandomization. Such schemes are beneficial when some baseline covariates are not initially available at the time of randomization as was the case with Baseline Hemoglobin A1c in the REACH trial.

In this case study, Begg and Iglewicz Minimization performed frequently among the top performing Sequenital allocation schemes and provided equal randomization to, on average, roughly 20% of patients (7% when adjust randomization to all baseline covariates). This provides some reassurance against systematic confounding. Minimization is powerful from a study design perspective, and randomization is critical for eliminating systematic confounding. This work begs the question, what is the extent of randomization necessary to sufficiently reduce systematic confounding?

Many studies set sample size based on the average performance of block randomization (for example by basing of a t-test). This case study provides a cautionary reminder that half of such studies will have a decreased effective sample size. Though all assumptions and operations of a study may work perfectly, a non-adaptive trial may fail to find a treatment effect due to chance alone. Covariate-adjusted randomization helps protect against poor covariate imbalances and increases the chances of a sufficient effective sample size all else the same.

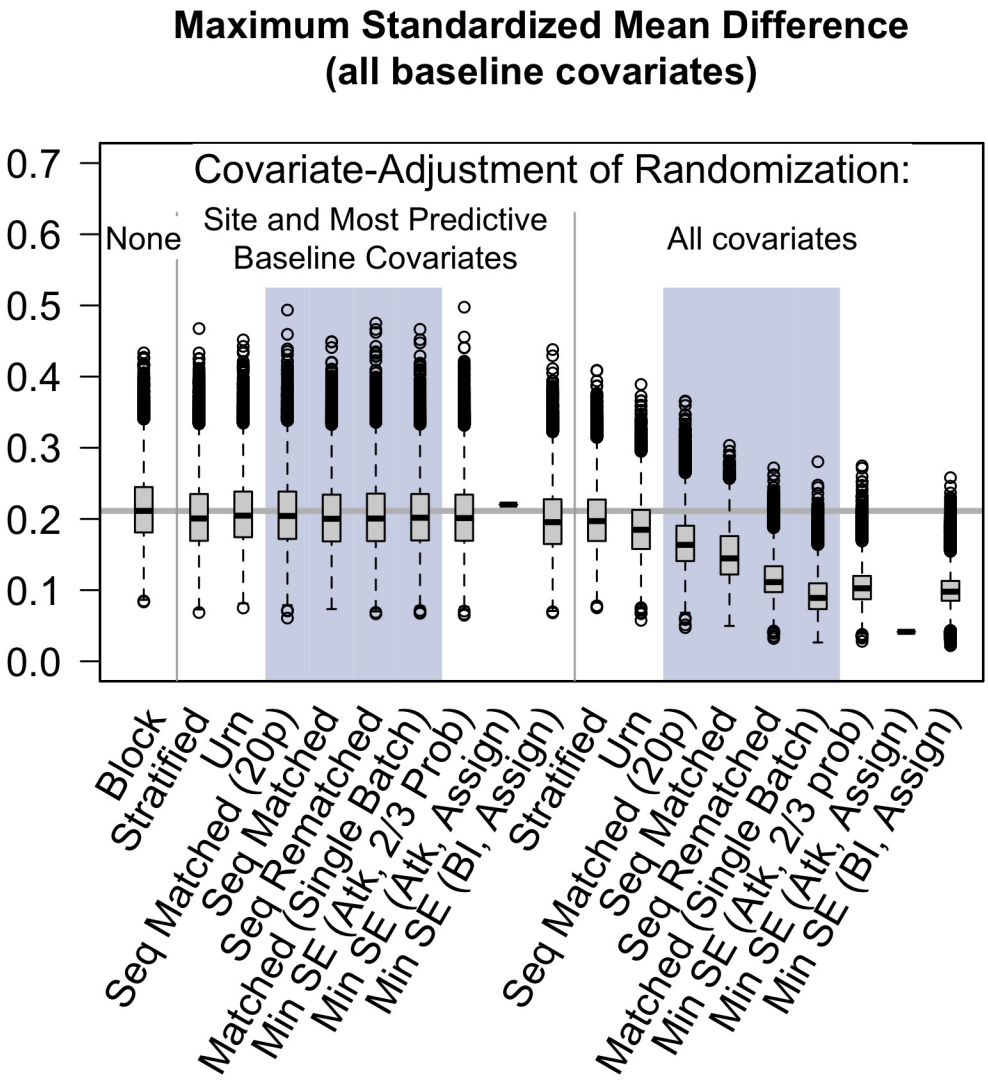

Figure 2.1: Boxplot of the maximum absolute Standardized Mean Difference among all baseline covariates for each of 20,000 simulated allocation sequences per allocation scheme. Matched-based Randomization schemes are shaded in blue. Seq Matched (20p) is Sequential Matching before applying proposed extensions; it uses a fixed 20th percentile of the F(p, i-p) distribution. Min SE schemes are Minimization schemes proposed by Atkinson and Begg and Iglewicz to reduce the standard error of a pre-specified Ordinary Least Squares Model. Min SE (Atk, 2/3 Prob) uses a 2/3 biased coin to randomize to the favorable arm under Atkinson's Minimization.

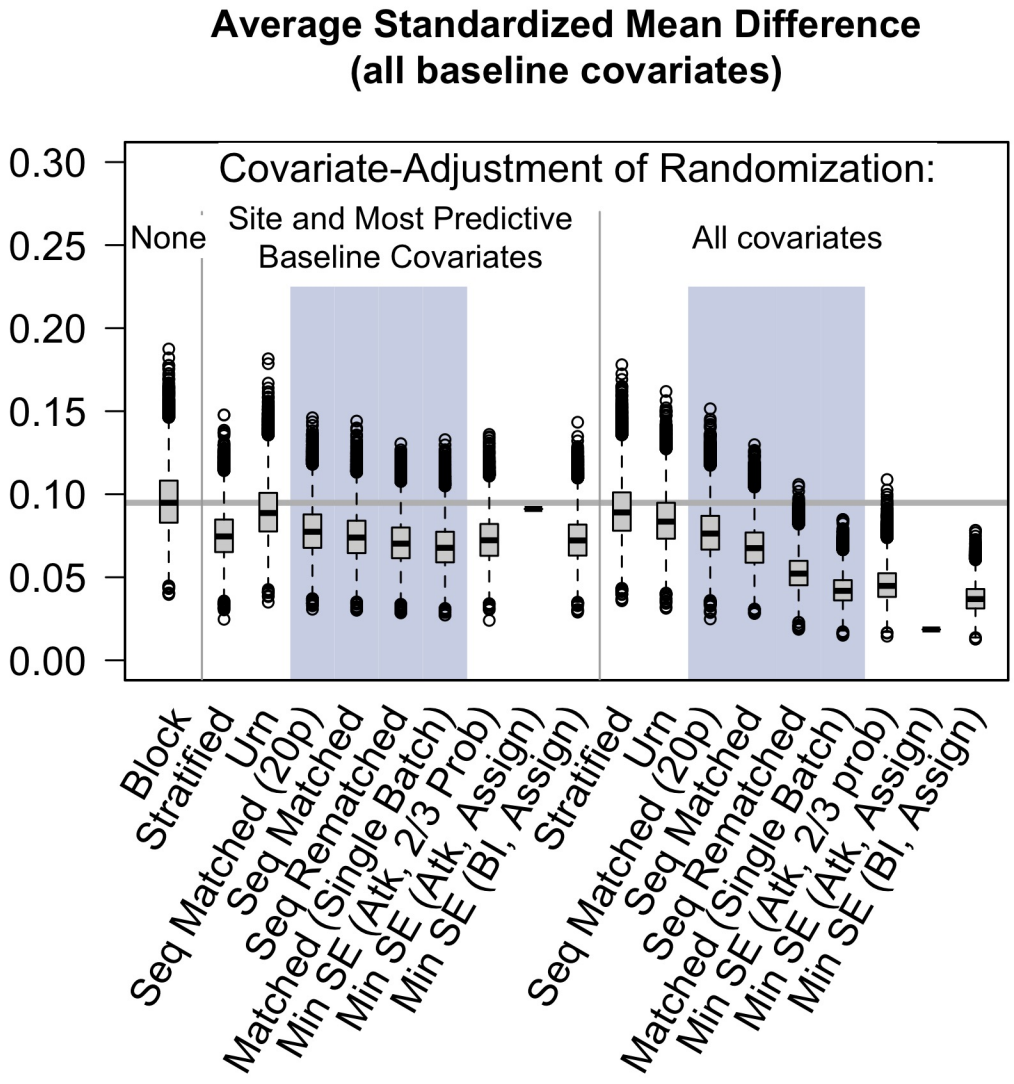

Figure 2.2: Boxplot of the average absolute Standardized Mean Difference among all baseline covariates for each of 20,000 simulated allocation sequences per allocation scheme.

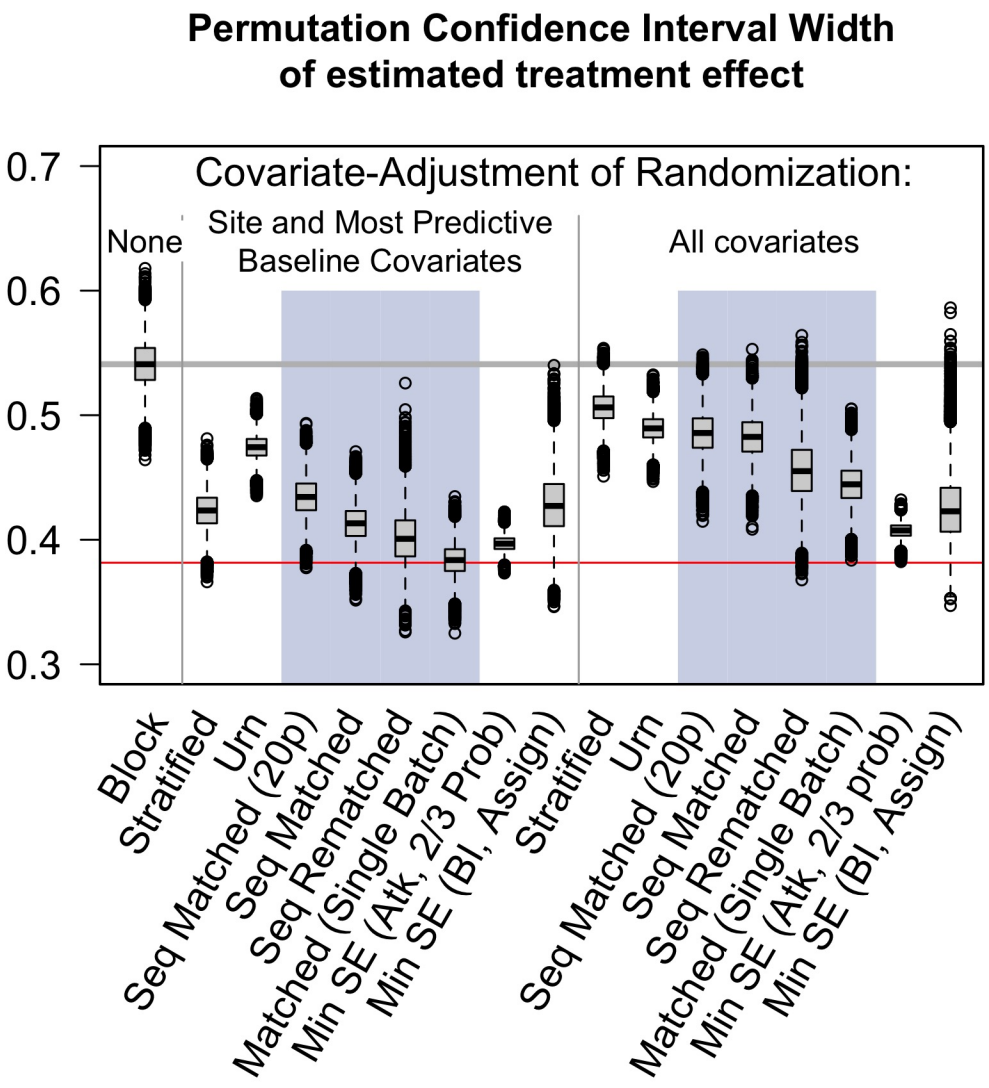

Figure 2.3: Boxplot of the permutation-based 95% Confidence Interval Width estimating the Sample Average Treatment Effect for each of 20,000 simulated observed datasets. Observed outcomes were generated as the predicted three month Hemoglobin A1c plus a treatment effect (if allocated to treatment) and a random residual. The horizontal line in red is the median treatment effect Confidence Interval Width under the pre-specified Ordinary Least Square model and Block Randomization (See Figure 4).

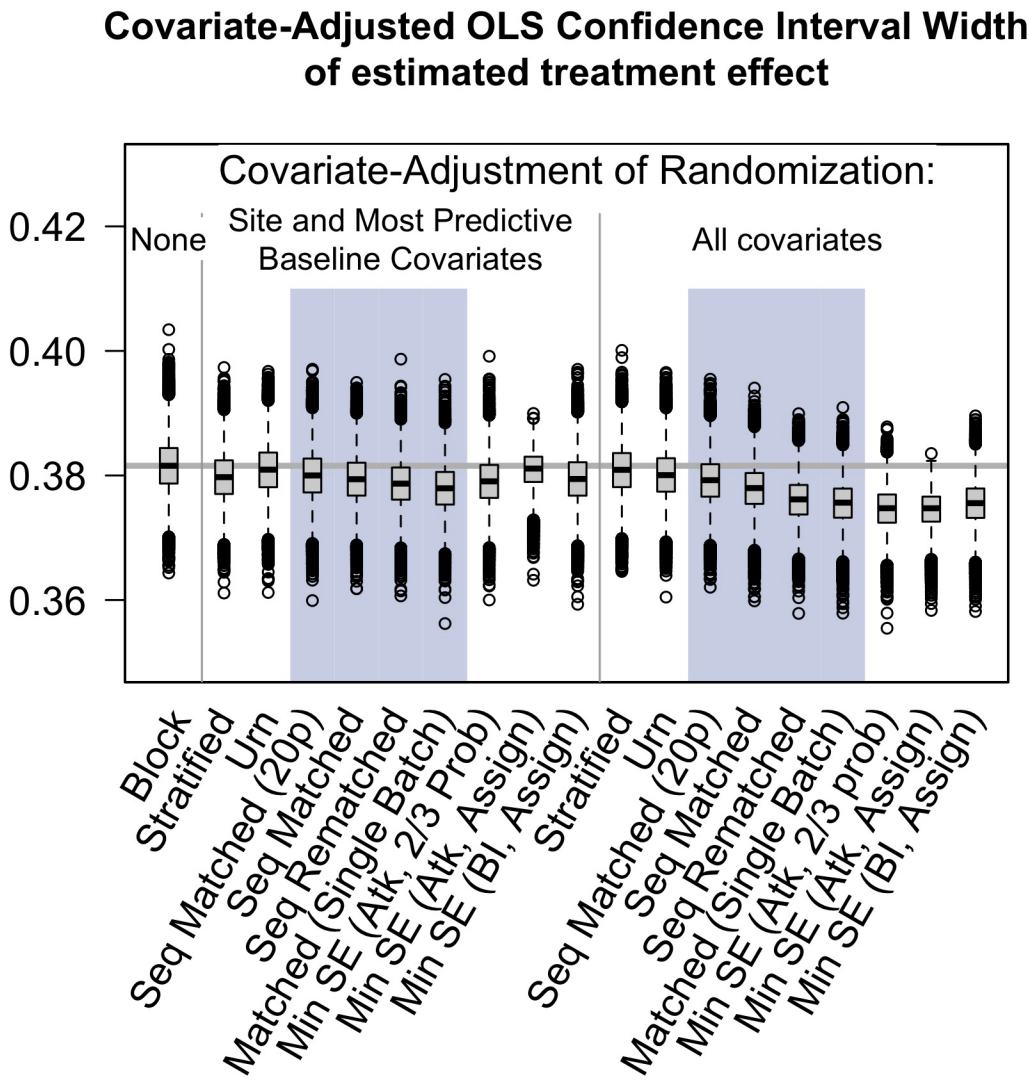

Figure 2.4: Boxplot of the model-based 95% Confidence Interval Width estimating the Sample Average Treatment Effect, adjusted for all baseline covariates, for each of 20,000 simulated observed datasets. The Ordinarly Least Squares (OLS) model was pre-specified to include all baseline covariates with restricted cubic splines on each continuous outcome and accounted for multiple imputations.

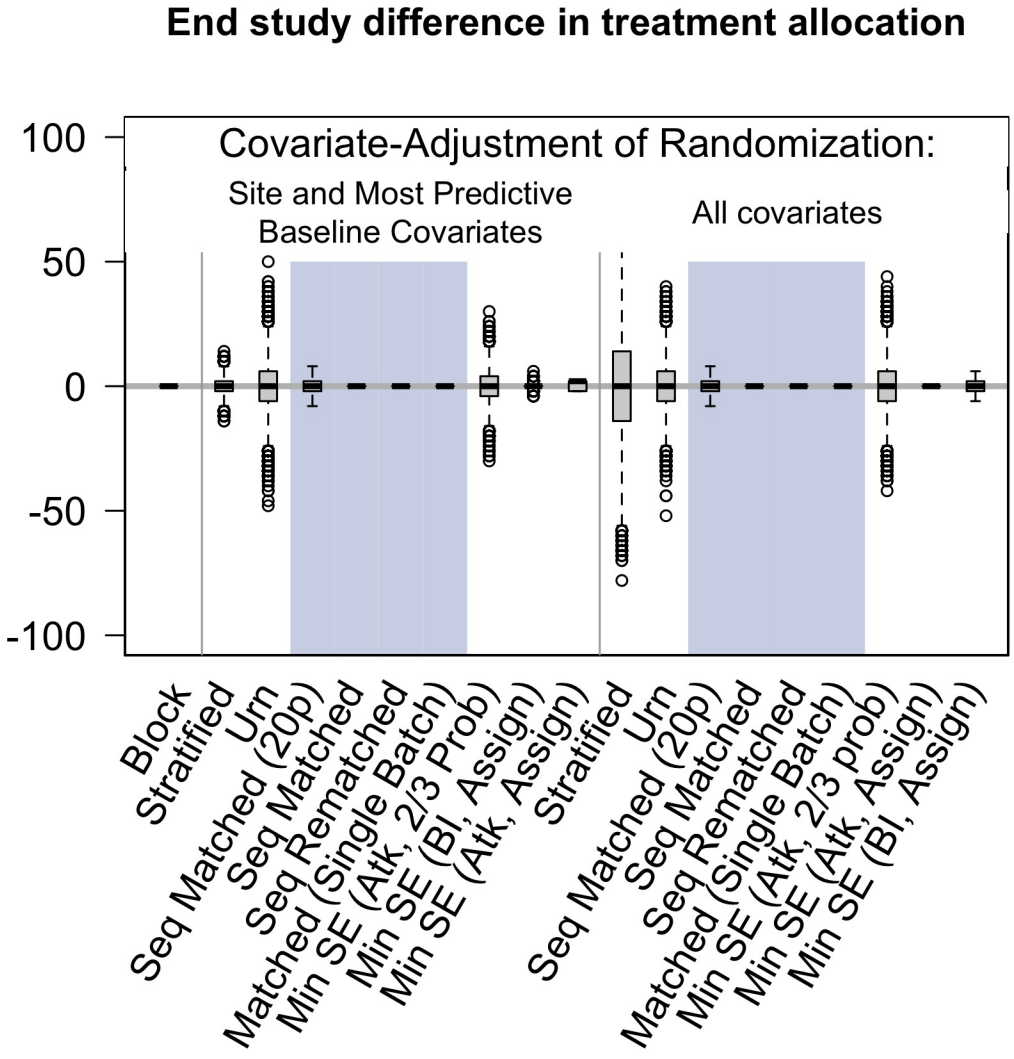

Figure 2.5: Boxplot of the difference in end-of-study treatment allocation for each of 20,000 simulated allocation sequences per allocation scheme.

Table 2.1: Median and 95 Percentile Confidence Interval of difference in effective sample size relative to Block Randomization for each of 20,000 generated observed datasets. Two Block Randomization sequences per generated observed dataset to allow comparing one instance of Block Randomization to another. Schemes are ordered by efficiency gains for permutation-based inference.

		Permutation Estimate		Adjusted Model Estimate	
Covariate-Adjustment	Scheme	Relative Efficiency	Effective Change in N	Relative Efficiency	Effective Change in N
None	Block	1.000	0 (-19, 20)	1.000	0 (-23, 23)
All Baseline Covariates	Stratified	1.067	71 (6, 141)	1.002	2 (-20, 25)
	Urn	1.104	112 (44, 183)	1.004	4 (-18, 27)
	Seq Matched (20p)	1.112	122 (35, 225)	1.006	6 (-15, 29)
	Seq Matched	1.119	129 (37, 242)	1.010	10 (-10, 32)
	Seq Rematched	1.187	209 (62, 385)	1.014	15 (-4, 37)
	Matched (Single Batch)	1.215	244 (128, 388)	1.016	16 (-3, 38)
	Min SE (BI, Assign)	1.277	323 (109, 537)	1.016	16 (-3, 39)
	Min SE (Atk, 2/3 prob)	1.326	388 (278, 508)	1.018	19 (1, 40)
	Min SE (Atk, Assign)	N/A	N/A	1.018	19 (4, 39)
Site and Most Predictive Baseline Covariates	Urn	1.139	152 (79, 229)	1.002	2 (-20, 25)
	Seq Matched (20p)	1.244	280 (156, 431)	1.004	4 (-18, 27)
	Min SE (BI, Assign)	1.263	305 (128, 518)	1.006	6 (-16, 29)
	Stratified	1.275	321 (189, 484)	1.005	5 (-17, 28)
	Seq Matched	1.308	363 (222, 539)	1.006	6 (-16, 28)
	Seq Rematched	1.347	416 (213, 663)	1.008	8 (-13, 31)
	Min SE (Atk, 2/3 Prob)	1.361	436 (316, 572)	1.007	7 (-15, 30)
	Matched (Single Batch)	1.408	503 (331, 711)	1.010	10 (-12, 32)
	Min SE (Atk, Assign)	N/A	N/A	1.001	1 (-13, 20)

CHAPTER 3

ADAPTIVE MONITORING USING SECOND GENERATION P-VALUES DRAFT

3.1 Introduction

Conclusive clinical trials either rule out clinically trivial or clinically actionable treatment effects. Like driving blind, this is a hard ideal to achieve without adaptive monitoring. The costs of ending too soon, when resources and knowledge were available otherwise, can be extraordinary (Pocock and Stone 2016b). On the other hand, a significant yet non-clinically actionable study can also be costly (Pocock and Stone 2016a). For these reasons, and to the extent possible, investigators turn to adaptive monitoring designs because the sample size is "not too big, not too small, but just right (Broglio, Connor, and Berry 2014)." To this we add the imperative: *to the point of being clinically conclusive*.

Unlike many study design aspects, clinical relevance is not an unknown assumption. A-priori scientific relevance informs which treatment effects are trivially null effects and which are clinically actionable enough to change clinical practice. Most sample size estimates already incorporate the latter into the study design. Though some study designs rule out a set of trivially null effects (Kruschke 2013; Hobbs and Carlin 2008; Freedman, Lowe, and Macaskill 1984), more common practice is to test for any difference from the point null.

Trivial effects surround the point null and include, in the least, indistinguishable treatment effects due to rounding error (Blume et al. 2018). They may also include clinically irrelevant changes in a biomarker. This set of effects has many names including but not limited to *Indifference Zone* and *Region of Practical Equivalence* (Blume et al. 2018; Kruschke 2013); we call this set of effects the *Trivial Zone*. Switching from a point null to interval null carries intuitive clinical interpretation and endows a study with desirable statistical benefits.

Interval null hypotheses reduce family-wise Type I Error rates (Blume et al. 2018; Kruschke 2013). False discoveries most frequently occur closest to the point null hypothesis, and an interval null provides a stricter rejection criteria buffer that eliminates many false discoveries. For the same reason, the interval null provides a natural

Type I Error adjustment for multiple looks / comparisons. Inferential approaches to evaluating interval hypotheses include Bayesian, Likelihood, and second-generation p-value inference. For any estimated interval (i.e. Confidence Interval, Credible Interval, Support Interval, etc.), the second-generation p-value draws conclusions from how much of the interval overlaps an interval hypothesis (Blume et al. 2018).

Many Bayesian adaptive trial designs incorporate interval null hypotheses (Freedman, Lowe, and Macaskill 1984; Berry et al. 2010; Hobbs and Carlin 2008; Kruschke 2013), and we develop an analogous Second Generation p-value adaptive design. The rest of the paper follows as: establishing a-priori clinically relevant guideposts, introducing the second generation p-value, adaptively monitoring with with the second generation p-value, comparing the second generation p-value and Bayesian adaptive monitoring with the REACH clinical trial data, and drawing final conclusions.

3.2 Clinically Relevant Guideposts

In a two-sided study, four boundaries provide clinical guideposts. Highly actionable treatment effects are of a magnitude of at least of δ_E for benefit or of δ_H for harm. Trivial treatment effects are between δ_{TE} and δ_{TH} and encompass the point null (Figure 1). The remaining effects are moderately actionable; moderate treatment benefits are between δ_{TE} and δ_E and moderate treatment harms are between δ_{TH} and δ_H. In any study, the implementation of an intervention takes into account secondary measures such as side effects, costs, etc.. Hence, a highly actionable effect is likely to outweigh the off-setting benefits / costs. A moderately actionable effect has greater equipoise with offsetting benefits / costs.A one sided study omits δ_H and δ_{TH} to focus only on $\delta_T E$ and δ_E.

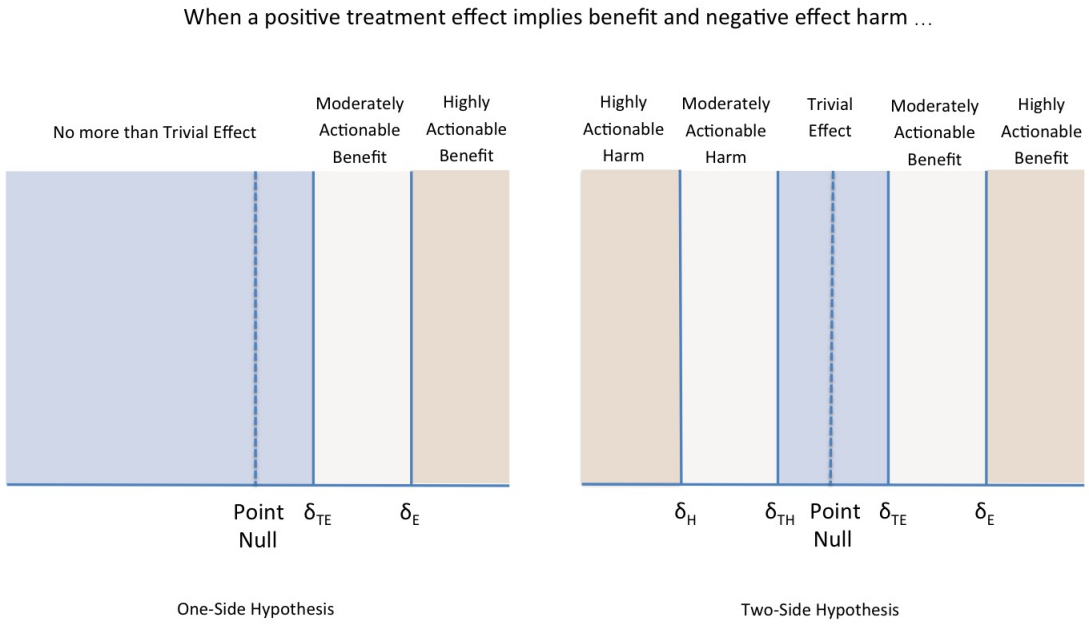

Figure 3.1: Clinically relevant guideposts are determined during study design, based on scientific context, and ought to be incorporated into the final study inference. In a one-sided study (left figure), the clinical guideposts create the regions: no more than trivial effect, moderately actionable effect, and highly actionable effect. In a two-sided study (right figure), three regions are created: trivial effects, moderately actionable effects, and highly actionable effects.

A similar set of clinical guideposts, the *Region of Equivalence*, uses only two boundaries(Freedman, Lowe, and Macaskill 1984; Hobbs and Carlin 2008). Treatment effects outside the *Region of Equivalence* deem the novel intervention as either clinically superior or inferior to the standard of care. Under certain conditions (a one-sided study with a buffer around the point null), these guideposts match the one-sided guideposts we propose. The *Trivial Zone* is necessary to rule out effects trivial to the point null and receive the benefits of an interval null such as reducing the False Discovery Probability.

Setting a-priori clinical guideposts brings transparency to the study design. Most study designs already incorporate δ_E when determining an adequate study sample size. Yet, in many instances, end trial inference does not incorporate δ_E (or other clinical guidepost decisions). For example, without knowing δ_E, none of a Confidence, Credible, or Support Interval inform original study design intentions.

Excellent references are available for helping establish clinical guideposts (Freedman, Lowe, and Macaskill 1984; Spiegelhalter, Freedman, and Parmar 1994; Kruschke 2018;

Blume et al. 2018).

3.3 Second Generation p-value

An inferential metric, the second-generation p-value indicates when the data are compatible with the alternative hypothesis, the *Trivial Zone* null hypothesis, or when the data are inconclusive (Blume et al. 2018). More generally, it may be used for any interval hypothesis. The second-generation p-value calculates the overlap between an interval I (any interval including but not limited to a Confidence, Credible, Support Interval, etc.) and the set of effects Δ_H in the hypothesis H. The interval includes [a, b] where a and b are real numbers such that $a < b$, and the length of the interval is b - a and denoted $|I|$. The overlap between the interval and the set Δ_H is $|I \bigcap \Delta_H|$. The second generation p-value is then calculated as

$$p_H = \frac{|I \bigcap \Delta_H|}{|I|} \times max\left\{\frac{|I|}{2|\Delta_H|}, 1\right\}.$$

The multiplicative factor, $max\left\{\frac{|I|}{2|\Delta_H|}, 1\right\}$, provides a small sample size correction – setting p_H to 0.5 when an interval overwhelms Δ_H by at least twice the length. For a *Trivial Zone* null hypothesis T, trivially null effects are ruled out when the $p_T = 0$ whereas non-trivial effects are ruled out when $p_T = 1$. The data are inconclusive when $0 < p_T < 1$.

Based on the four clinical guideposts, we define and focus on a Highly Actionable Hypothesis, HA, and Trivial Hypothesis, T .

- Hypothesis HA: The treatment effect lies within a *Region of Clinically Highly Actionable Effects* $\Delta_{HA} = (-\infty, \delta_H] \cup [\delta_E, \infty)$ for a two-sided study and $[\delta_E, \infty)$ for a one-sided study where a positive benefit is beneficial. Where a negative effect is beneficial, the one-sided study sets $\Delta_{HA} = (-\infty, \delta_E]$.
- Hypothesis T: The treatment effect lies within a *Region of Clinically Trivial Effects* $\Delta_T = [\delta_{TH}, \delta_{TE}]$ for a two-sided study. In a one-sided study where a positive effect is beneficial, $\Delta_T = (\infty, \delta_{TE}]$ and is called a *Region of At Most Clinically Trivial Effects*. The bounds are mirrored when a negative effect is benificial.

Neither of these two hypotheses include moderately actionable treatment effects. When applied to the four clinical guideposts, nine conclusions may be drawn from the

second-generation p-value and the Regions of Clinically Highly Actionable Effects and Clinically Trivial Effects (Figure 2). To motivate our adaptive monitoring design, we reduce to three conclusions:

- When $p_{HA} = 0$, the treatment effect is not clinically highly actionable.
- When $p_T = 0$, the treatment effect is not trivial different from the point null.
- Otherwise, the treatment effect is inconclusive

Again, the interpretation of these regions closely relate to the interpretations of the *Region of Equivalence*. They match exactly when in the case of a one-sided study with a *Trivial Zone* around the point null.

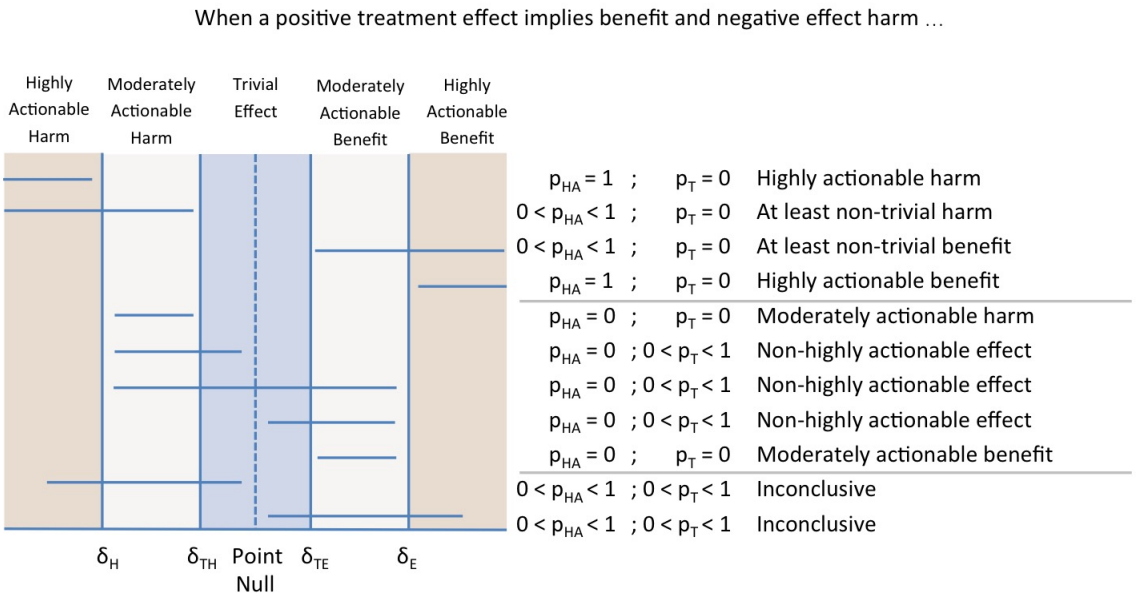

Figure 3.2: Final study inference ought to incorporate a-priori clinical guideposts. The second generation p-value draws inference based on hypothesized sets of treatment effects. We focus on two sets of hypotheses that form the Region of Clinically Trivial Effects and the Region of Clinically Highly Actionable Effects. (This figure focuses on two-sided studies yet similar conclusions are drawn for one-sided studies). With the two sets of hypotheses, the second-generation p-value can rule out trivial effects (the top four conclusions), rule out non-highly actionable effects (the middle five conclusions), or declare the study yet inconclusive. Confidence Intervals that correspond to a p-value close to but not exceeding 0.05 would be declared inconclusive.

3.4 Adaptive Monitoring Rules / Guidance

A study implements the following rules for adaptive monitoring with the second-generation p-value:

- Design: Investigators a-priori determine the four clinical guideposts (two guideposts in the case of a one-sided study).
- Wait to Monitor: Enroll B patients before applying monitoring.
- Monitor: Calculate the second-generation p-value using an inferential interval of choice. Raise an alert when $p_{HA} = 0$ or $p_T = 0$. Continue monitoring until affirming the *same* alert (i.e. that again $p_{HA} = 0$ or $p_T = 0$) K patients later.
- Stop: Stop once affirming an alert or at the end of resources.
- Report: Report only the interval at stopping.

Finding the appropriate B and K is done through simulations in the study design stage; both help protect against Type I Errors and bias. Chapter 3 provides practical guidance for determining B and K. Type I Errors more commonly occur early in adaptive monitoring while estimated statistics are yet unstable, and they may occur randomly in the discrete Brownian motion of the Monitoring Interval. Bias is inherent in all adaptive monitoring schemes that stop at the first instance an alert is raised. Requiring K patients to affirm an alert allows regression to the true treatment effect and improves the reported interval's coverage rate. To draw emphasis, we call intervals used in the monitoring phase Monitoring Intervals as they are not to be interpreted.

When the true effect is moderately actionable, the study may stop for concluding the effect to be either not-trivial or not-highly actionable. This may bring consertnation; however, it is an important study design feature. Without a region of clinically moderately actionable effects, a highly actionable effect would border trivial effects and have a 50-50 chance of stopping for being actionable or trivial. A study that stops for concluding a non-trivial effect may include moderately actionable effects in the inferential interval. And similarly, moderately actionable effects may be included in inferential intervals when stopping to conclude a non-actionable effect. This behavior reflects the greater degree of equipoise between the moderately actionable effects and the off-setting harms/benefits.

Moderately actionable effects have a greater chance of raising conflicting alerts – for example to raise an alert for a non-trivial effect then K patients later alert for a

non-highly actionable effect. For this reason, we require the affirmation alert to be the same as the alert it affirms. Only the final interval's operating characteristics are relevant.

With only adapting the monitoring rules, the remaining rules apply to monitoring with Bayesian Credible Intervals. An alert for a non-highly actionable effect when P(Treatment Effect $\notin \Delta_{HA}$ | Data) > 1 – $\alpha_{criteria-not-actionable}$ or for a non-trivial effect when P(Treatment Effect $\notin \Delta_T$ | Data) > 1 – $\alpha_{criteria-not-trivial}$. $\alpha_{criteria-not-actionable}$ and $\alpha_{criteria-not-trivial}$ are study design tuning parameters based on simulations to achieve a desired end of study Type I Error and Power.

3.5 Adaptively Monitoring the REACH clinical trial

3.5.1 Context

The Rapid Education/Encouragement And Communications for Health (REACH) randomized clinical trial (Nelson et al. 2018) is designed to help patients with diabetes better manage glycemic control (as measured by Percent Hemoglobin A1c) and adhere to medication. Patients randomized to the intervention receive text message-delivered diabetes support for over 12 months. Now with complete enrollment, 512 patients are currently being followed up through 3, 6, 12, and 15 months.

In this population, lower Hemoglobin A1c reflects improved glycemic control. A change in Hemoglobin A1c of +/- of 0.15 (median REACH baseline Hemoglobin A1c of 8.20 [IQR of 7.20, 9.53]) is clinically trivial, whereas a decrease of Hemoglobin A1c of 0.5 is highly actionable to the point of adopting this novel intervention. While we anticipate the intervention to improve Hemoglobin A1c, we designed the study to be two-sided. That is, δ_E = -0.50, δ_{TE} = -0.15, δ_{TH} = 0.15, and δ_H = 0.50.

3.5.2 Simulation

The performance of adaptive monitoring using the second generation p-value and posterior probabilities was observed from 20,000 bootstrap samples REACH patients. With block two randomization half of the patients were randomized to receive the REACH intervention. For outcomes, we used three month predicted Hemoglobin A1c from an ordinary least square model adjusting for baseline covariates including Hemoglobin A1c, age at enrollment, gender, years of education, years of diabetes, race

/ ethnicity, medication type, income, and type of insurance. All continuous baseline covariates flexibly allowed for non-linear associations through restricted cubic splines. Predicted Hemoglobin A1c changed by a treatment effect for those randomized to the REACH intervention. We simulated multiple treatment effect settings {Treatment Effect: -1, -0.75, -0.50, -0.375, -0.15, 0}. Only negative effects were investigated but by symmetry, increases in Hemoglobin A1c of the same magnitude perform similarly. In these simulations we assumed instantaneous outcomes.

Monitoring began after the 40th enrolled patient and continued every 20th patient until either affirming an alert 40 patients later or until reaching the end of resources (512 patients). For each second-generation p-value Monitoring Interval, we fit a marginal ordinary least squares regression of outcome Hemoglobin A1c given treatment assignment and obtained the 95% Confidence Interval for the treatment effect. And, for adaptive monitoring using posterior probabilities, we used a Bayesian alternative to the t-test: two-sample difference of means where the intervention group mean, Y_I, and control group mean, Y_C, were both distributed $t(\mu, \sigma, \nu)$ (Kruschke 2013).

For Bayesian priors, we set $\mu \sim N(\bar{y}$, 1 / ϕ-1(.9)) as a fairly flat prior centered on the average mean Hemoglobin A1c of all observations with a 0.10 probability of observing an absolute change in Hemoglobin A1c greater than 1. For a skeptical prior on μ, we mixed the flat prior 1:1 with $\mu \sim N(\bar{y}$, 0.15 / ϕ-1(0.95)). We set $\sigma \sim$ Gamma(s_y / 1000, s_y 1000) where s_y was the non-pooled standard deviation of all Hemoglobin A1c outcomes, and $\nu \sim Gamma(1/600, 30) + 1$. The ν parameter places a prior on the skewness of the data, and was chosen such that with the flat prior yielded group means similar to raw means.

The two Bayesian adaptive designs (one with a flat and the other skeptical prior) were calibrated to the second generation p-value design by finding the $\alpha_{criteria-not-actionable} = \alpha_{criteria-not-trivial}$ that achieve the same Type I Error given the true effect is zero and that achieve the same probability of stopping for a not-trivial effect given the true effect is -0.15 (the border of trivial effects). The later calibration was then used to compare adaptive designs in terms of reasons for stopping, average trial sample size, bias, and coverage of the reported interval. Operating characteristics are estimated for given treatment effects. When allowing for distributional assumptions of observing the treatment effect, a correctly specified prior has zero bias.

3.5.3 Results

The flat and skeptical prior Credible Interval adaptive monitoring designs were calibrated to have the same Type I Error as adaptive monitoring with the second-generation p-value (Type I Error = *0.036*; Figure 3.3). While this was close to 0.05, the Type I Error changes depending on the wait time until monitoring and the number of looks. The flat and skeptical priors as specified both favor, at least some degree the null hypothesis which is reflected as a decrease in Power.

The probability of stopping for being not-trivial provides an adaptive monitoring analog for a Type I error under an interval null hypothesis. When the true treatment effect was truly -0.15 (i.e. the boundary of trivial effects), and when adaptively monitoring with second-generation p-values, the probability of stopping for concluding not-trivial was *0.055* (3.4). *Posterior probability monitoring designs were again calibrated to this same probability for this and the remaining figure results.* The probability of stopping for being not-trivial was slightly higher when monitoring using posterior probabilities conditioned on the flat prior. However, for a clinically meaningful treatment effect of -0.50, monitoring on posterior probabilities conditioned the flat had a higher probability of stopping for concluding a non-highly-actionable effect than monitoring with the second generation p-value. The two designs that monitored on Credible Intervals were more likely to stop for concluding not-highly actionable effects than monitoring on the second-generation p-value; this is consistent with both priors being at least slightly favorable to the point null.

Trials monitored with the second-generation p-value using monitoring conifence intervals lasted longest when the treatment effect was moderately-actionable (Figure 3.5). As posterior probabilities condition on priors more strongly favoring the null, the longer trials occur for effects deemed moderately- to highly-actionable effects.

The bias was well mitigated when adaptively monitoring with the second generation p-value and a posterior conditioned on a flat prior (Figure 3.6). The flat prior Credible Interval pulls estimates toward the point null, which tended to slightly help when the treatment effect was clinically meaningful and slightly hurt when the trivially null. Neither the flat nor skeptical prior are biased if they are each reflective of the true distribution of observing treatment effects. Coverage rates were increasingly better when monitoring with posterior probabilities compared to the second generation p-value. Of the treatment effects investigated, the worst coverage was *0.93* when monitoring with the second generation p-value.

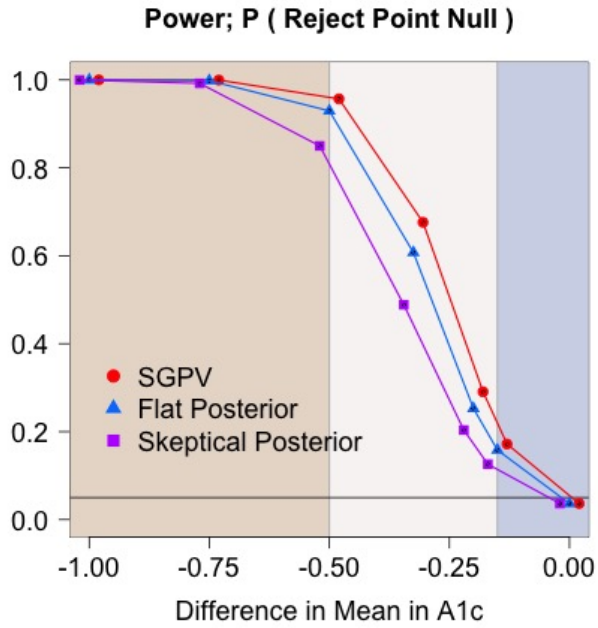

Figure 3.3: In the REACH target population, a decrease in Hemoglobin A1c reflects desireable improvement in glycemic control. The Power curve (i.e. the probability of rejecting the point null of 0), was estimated from 20,000 adaptive monitoring simulations when monitoring using the second generation p-value (red circle) and posteriors conditioning on a flat prior (blue triangle) and skeptical prior (purple square). The intervention was simulated to have an effect of -1 (highly beneficial), -0.75, -0.50, -0.375, -0.15, to 0 effect (the point null). Bayesian adaptive monitoring designs were calibrated to have the same Type I Error as the second generation p-value adaptive design.

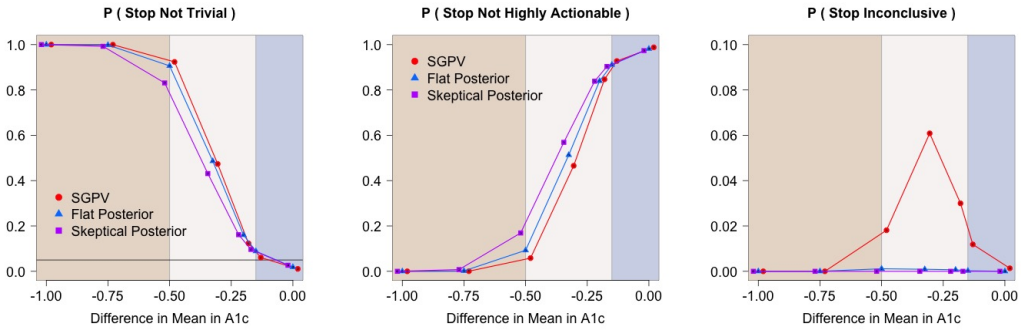

Figure 3.4: In adaptive monitoring, a study can end in one of three states: concluding non-trivial effect, concluding non-highly-actionable effect, or ending at the end of resources. Above are estimated probabilities of ending in each of these states for each design and treatment effect. The probability of concluding non-trivial is an interval null analog to Power. The Bayesian adaptive monitoring designs were calibrated to have the same probability of concluding a non-trivial effect as adaptive monitoring with the second generation p-value. All following results are based off this calibration.

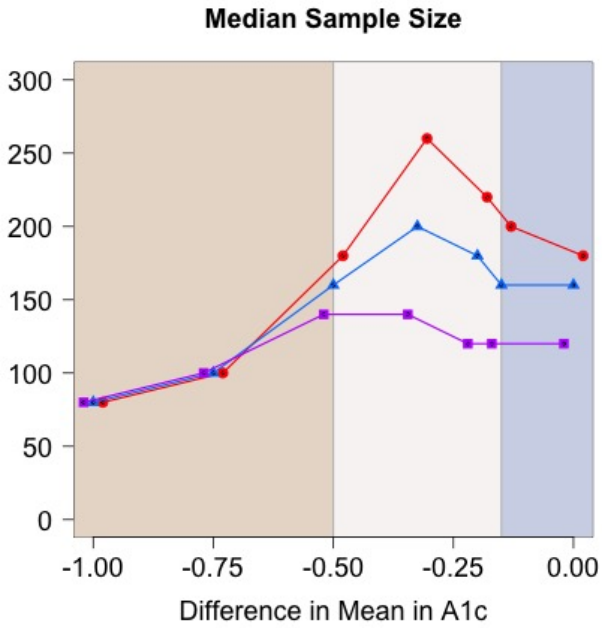

Figure 3.5: Across simulations, the average time to stopping for the three adaptive designs is shown above. In these simulations the earliest possible stopping time was the 80th patient. Monitoring began at the 40th patient and continued at every 20th patient. Stopping required raising an alert and affirming the alert 40 patients later. Moderately actionable treatment effects required the greater sample size, while highly actionable and trivial effects required smaller sample sizes to suggest stopping.

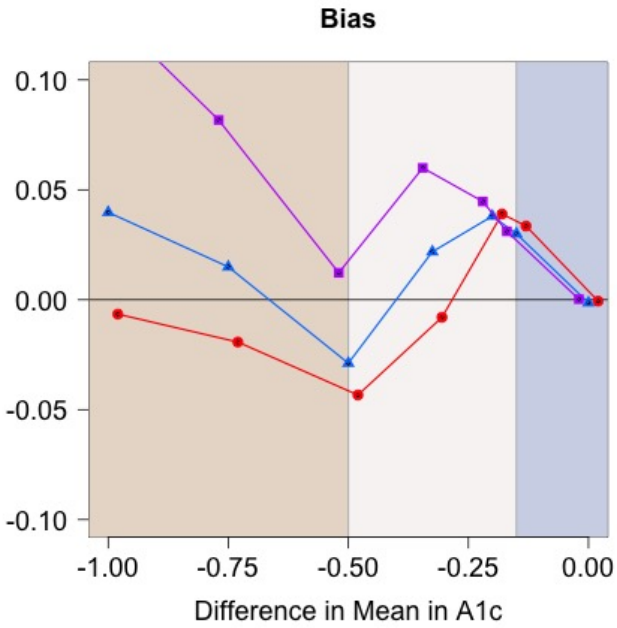

Figure 3.6: At the study end, the final estimate and interval are reported. Across simulations, the bias was well mitagated when adaptively monitoring using the second generation p-value and posterior probability conditioned on a flat prior. A positive bias occurs when being pulled toward the null, as happens with the skeptical prior.

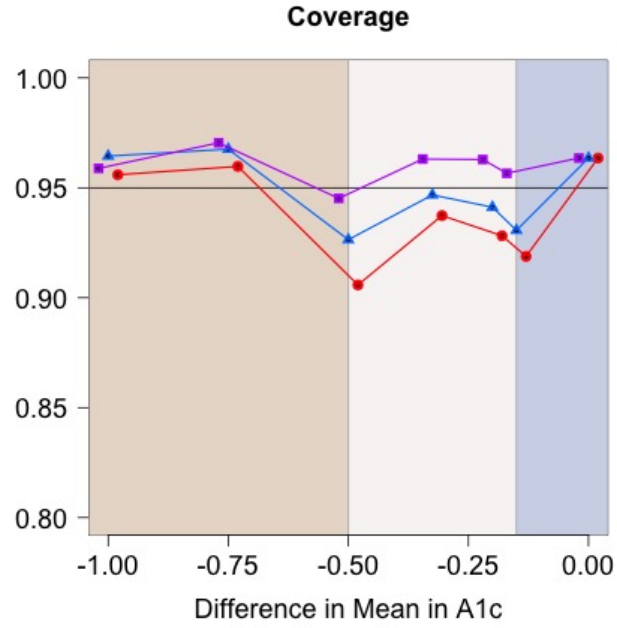

Figure 3.7: At the study end, the final estimate and interval are reported. In these simulations, a 95% Confidence Interval was reported when adaptively monitoring with the second-generation p-value (though any other interval could have been used for monitoring and reporting). And, Credible Intervals are reported for the Bayesian adaptive designs.

3.6 Discussion

We developed an adaptive monitoring scheme, using the second generation p-value, to follow studies until either ruling out trivially null or clinically highly actionable treatment effects. Two major contributions come from this scheme: (1) the monitoring scheme incorporates clinically relevant guideposts that may help trivially null and clinically highly actionable studies stop quickly while still reaching a clear clinical conclusion and (2) the easy-to-calculate second generation p-value allows an investigator to adaptively monitor with their inferential interval of choice (including but not limited to Confidence Intervals, Credible Intervals, and Support Intervals). While not explored explicity in this paper, this design is well suited for following multiple outcomes and following subgroups until they reach clear clinical conclusions.

In any adaptive monitoring implementation, the adoption of a novel agent depends on multiple outcomes and reaching concluding an effect is highly actionable may not mean the agent should be adopted. Data Safety Monitoring Boards are encouraged to think of stopping rules more as guidelines when weighing the totality of evidence.

As seen in other similar designs, the use of an interval null decreases the False Discovery Rate, making the results more reproducible (Blume et al. 2018; Kruschke 2013). Incorporating clinical relevance into the inference brings study design transparency that otherwise may get lost. For example, a Confidence, Credible, or Support Interval alone do not inform of the targeted treatment effect when choosing a sample size.

With data from the REACH randomized clinical trial, we simulated adaptive monitoring using rules based on the second-generation p-value compared to rules based on the posterior probabilities. Monitoring with the second-generation p-value and posterior probabilities conditioned on a flat prior performed similarly in terms of Power, probability of stopping for efficacy, average sample size, bias, and reported interval coverage. Monitoring using a flat prior Credible Interval had a slight edge in these metrics but also a slightly worse probability of concluding highly-actionable effects were not-highly actionable. The flat prior was not uniformly flat, it slightly favored the point null.

All of the adaptive designs studied may be modified to achieve a desired Type I Error and Power. When adaptively monitoring with the second generation p-value, one may change the wait time before monitoring and number of looks. We encourage clinical guideposts to be anchored on clinical relevance and not change solely for the purpose of achieving statistical properties. The same encouragement to not change anchors

denoting clinical relevance hold true for monitoring with posterior probabilities.

CHAPTER 4

THE SGPVAM PACKAGE AND PRACTICAL RECOMMENDATIONS FOR ADAPTIVE MONITORING WITH THE SECOND GENERATION P-VALUE

4.1 Introduction

In Chapter 2 we present an adaptive monitoring scheme that follows studies until evidence supports either a non-trivial or non-highly actionable treatment effect. The design is very easy to implement. It can done by anyone who can think about the clinical interpretation of possible effect sizes and calculate an interval estimate for their effect, such as a credible, support, or confidence interval (CI). However, estimating the operating characteristics of a given study design is not as easy. Without tools to assist them, it could be a barrier to implementation of the method.

In practice, the trialist will want to know the operating characteristics under the following adaptive monitoring design features.

- Frequency of looks at the data, i.e. recalculate CI at every jth subject.
- Minimum precision requirement before applying monitoring rules, i.e. check monitoring rules only if |CI| < w.
- Required observations between an alert and affirmation to stop, i.e. evaluate stopping rule j*k subjects after alert.
- Anticipated maximum amount of data that could be collected, maxN, i.e. the sample size at which the study will cease collection regardless of the stopping rules.
- Lag time between enrolling a subject and observing their outcome measured in the number of subjects recruited during the lag, i.e. m additional subjects will be recruited in the time between one subject being recruited and that one subject's outcome being observed.

The traditional trialist will mainly be looking for the point null Type I error probability and Power, i.e. the probability of concluding an effect is non-trivial for a given true effect size. They may also want some simple summary statistics for the potential sample sizes. We hope to provide access to these and to a more rich set of operating characteristics. Operating characteristics we can potentially estimate for a given set

of design features include the following.

- Distribution of potential sample sizes.
- Point Null Type I error, when the point null is true, i.e. probability of excluding the point null from the final interval estimate. Classical trialists will want reassurance this is $< 5\%$.
- Interval Null Type I error, when the point null is true, i.e. probability of excluding the entire trivial zone from the final interval estimate. This will be less than or equal to the point null Type I error.
- Power vs the point null, i.e. probability the final interval estimate excludes the point null for a given true effect size. This is akin to classical statistical power.
- Power vs the interval null, i.e. probability the final interval estimate excludes the null zone for a given true effect size. This will be less than or equal to the point null power, but is conceptually the preferable quantity. In better terms, this is the probability of concluding the effect is non-trivial for given true effect sizes.
- Probability of concluding effect is non-highly actionable for given true effect sizes.
- Probability of an inconclusive finding at the end of resources. Note a clinically inconclusive finding is a possibility whenever maxN is finite and/or m > 0.
- Bias, MSE, and interval coverage probability from a frequentist perspective.
- False confirmation probability under a specified prior distribution. To address these needs, we provide an R function, sgpvAM, that simulates the above design operating characteristics for normal and binomial outcomes, and we offer practical advice setting the minimum wait time (in terms of inferential interval width) and the number of looks before affirming an alert.

4.2 sgpvAM Package

The sgpvAM package allows the user to obtain study design operating characteristics under a variety of settings for adaptive monitoring using the second generation p-value.

4.2.1 MCMC Replicates

The user may use the sgpvAM function to generate MCMC replicates of outcomes and intervention assignments along with an estimate of the effect and a lower- and upper- interval bound; replicates are generated using parallel computing. Alternatively,

the user may provide their own generated data together with an estimated effect and interval bounds. When using the sgpvAM function, the user specifies the data generation function (any of the r[dist] such as rnorm) along with arguments to the function. Similarly, the user specifies effect generation. Currently, only fixed effects have been thoroughly tested. However, by specifing a distribution for the effects, the user may explore False Discovery Probabilities and other operating characteristics, such as bias, dependent on distributional assumptions of the effect.

4.2.2 One- vs Two-Sided Hypotheses

Clinical Guideposts defining regions of Trivial and Highly Actionable Effects must be provided though may be one- or two-sided. The point null must be within and not a boundary of the Trivial Region. For general nomenclature, inputs to define the regions are: deltaL2 (the Clinically Highly Actionable Boundary less than the point null), deltaL1 (the Trivial Region Boundary less than the point null), deltaG1 (the Trivial Region Boundary greater than the point null), and deltaG2 (the Clinically Highly Actionable Boundary greater than the point null). See Chapter 2 for a thorough discussion of these regions.

4.2.3 Tuning study parameters

To maximize performance of operating characteristics under a given sample size the sgpvAM function allows the user to specify multiple wait time settings, frequency of looks, and number of steps before affirming a stopping rule. (The wait time is the time until the expected Margin of Error achieves a certain length or less). Here we define the Margin of Error as one-half the interval width.

4.2.4 Operating characteristics under normal outcomes

After generating the operating characteristics under a fixed normal outcome, the user may use the locationShift function to obtain operating characteristics under a range of fixed treatment effects. The function uses the saved MCMC replicates and adds to them if needed for additional monitoring.

4.2.5 ECDF of sample size and bias

Once a study design has been selected based on average performance (sample size, bias, and error probabilities), the user may use the ecdf.sgpv function to see the empirical cumulative distribution across the MCMC replicates for sample size and bias under a specific design. This provides an estimate of the probabilitiy a study does not exceed a certain maximum sample size.

4.2.6 General suggestions

Computations may be time consuming. It is recommended to start with 1000 replicates to get a general sense of average sample size and error probabilities under a variety of investigated wait times and affirmation steps. Investigating many wait times increases the computational burden. When generating data that allows for a location shift, it is recommended to generate MCMC replicates in the (or one of the) mid point(s) between the Clinically Trivial and Highly Actionable Regions. This is the region with greatest expected sample size and reduces the burden of the locationShift function generating additional data when necessary.

4.2.7 Inputs

mcmcData Previously generated data. Default (NULL) uses MCMC generation inputs to generate new or additional data.

nreps Number of MCMC replicates to generate

waitWidths Wait time, in terms of Margin of Error (one half the confidence interval width), before monitoring data.

dataGeneration Function (such as rnorm) to generate outcomes.

dataGenArgs Arguments for dataGeneration function. This includes, in the least, 'n' observations to generate. If 'n' is insufficient for unrestricted adaptive monitoring, addtional data will be generated.

effectGeneration Function (such as rnorm) or fixed value to generate intervention effect (theta).

effectGenArgs Arguments for effectGeneration function (if any)

modelFit An existing or user-defined function to obtain intervals. Two existing functions are provided: 1) lmCI which obtains a confidence interval from a linear model and has class 'normal' indicating normal data and 2) lrCI obtains Wald Confidence Interval from logistic regression model and has class 'binomial' indicating binomial data.

pointNull Point null.

deltaL2 Clinical guidepost less than and furthest from point null.

deltaL1 Clinical guidepost less than and closest to point null.

deltaG1 Clinical guidepost greater than and closest to point null.

deltaG2 Clinical guidepost greater than and furthest from point null.

lookSteps The frequency data are observed (defaults to 1 – fully sequential).

kSteps Affirmation steps to consider range from 0 to maxAlertSteps by kSteps.

maxAlertSteps Maximum number of steps before affirming an alert.

maxN Total enrolled patients equals maxN observed patients plus lagOutcomeN.

lagOutcomeN Total enrolled patients equals MaxN observed patients plus lagOutcomeN. lagOutcomeN are number of observations enrolled but awaiting to observe outcome.

monitoringIntervalLevel Traditional (1-alpha) used in monitoring intervals.

printProgress Prints when adding more data for MCMC replicates to have sufficient observations to monitor until a conclusion. Defaults to TRUE.

outData Returns the MCMC generated data. This can result in an out object with large memory. Yet, with location shift data, can be re-used to obtain operating characteristics of shifted effects.

getECDF Returns the ECDF of sample size and bias for each wait width and number of steps before affirming end of study.

cores Number of cores used in parallel computing. The default (NULL) does not run on parallel cores.

fork Fork clustering, works on POSIX systems (Mac, Linux, Unix, BSD) and not Windows. Defaults to TRUE.

socket Socket clustering. Defaults to TRUE yet only applies if FORK = FALSE.

4.2.8 Return values

The sgpvAM function returns a list with three elements:

1. mcmcMonitoring – the mcmcReplicates when outData is TRUE,
2. mcmcEndOfStudy – operating characteristics on average and ECDF for each combination of wait time and number of steps before affirming a stopping rule
3. inputs – Inputs in to the sgpv function

As supporting material for the package, we have developed an extensive vignette that illustrates using sgpvAM to estimate and explore the impact of study design choices on Point Null Type I error, Power for a range of true effect sizes, average sample sizes, the distribution of possible sample sizes, and more. These include the types of figures and calculations that will be of particular interest to the traditional trialist. The full vignette may be found at https://github.com/chipmanj/sgpvAM or by loading the sgpvAM package and calling Vignette(package = "sgpvAM", topic = "README"). Three of the example figures are presented briefly below.

Figure 4.1 is a classical power curve which shows the probability the final CI will exclude the point null at various true effect sizes. The power when the true effect size is equal to the point null is the classical point null Type I error probability. The figure illustrates the impact of various wait times on the power curve. The second generation p-value adaptive design allows for designing trials with finite and infinite sample sizes in mind. We recommend presenting both. Figure 1 illustrates the infinite sample size. Notice the point null Type I error is bounded below 5% even when the maximum sample size is theoretically infinite. Under this framework, a study that stopped for reaching a maximum sample size could be restarted without concern of controlling point null Type I error.

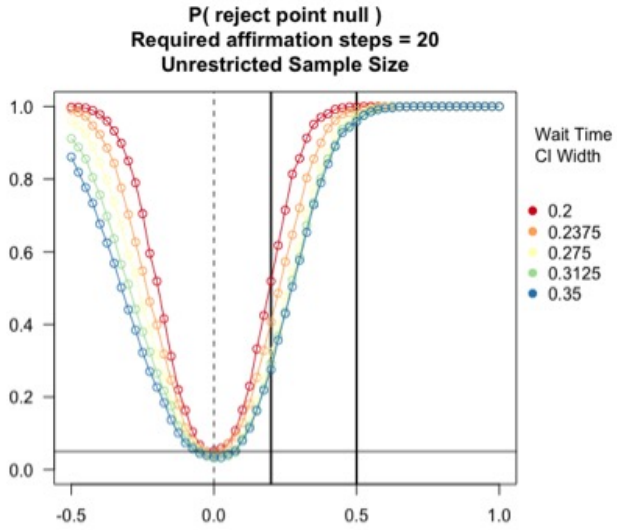

Figure 4.1: Power Curve across treatment effects for rejecting the point null in a one-sided study when requiring different wait times before monitoring. The horizontal line is at 0.05 to indicate the alpha level corresponding to the final reported confidence interval. The first vertical line denotes the upper boundary of At Most Trivial Effects, and the second vertical line denotes the boundary of the Highly Actionable Effects. The Wait times are the expected sample size for achieving a confidence interval width. In this figure there is no restriction on sample size nor a lag in observing outcomes.

Figure 4.2 displays the impact of increasing the affirmation steps on the average sample size. Increasing the affirmation steps has benefits in reducing bias, increasing interval coverage probabilities, and increasing the stability of conclusions particularly in the presence of a lag between recruitment and outcome observation. These benefits come with an increase in average sample size, particularly in between the bounds of the Trivial and Highly Actionable regions, i.e. where erroneous conclusions due to stopping too early are most likely.

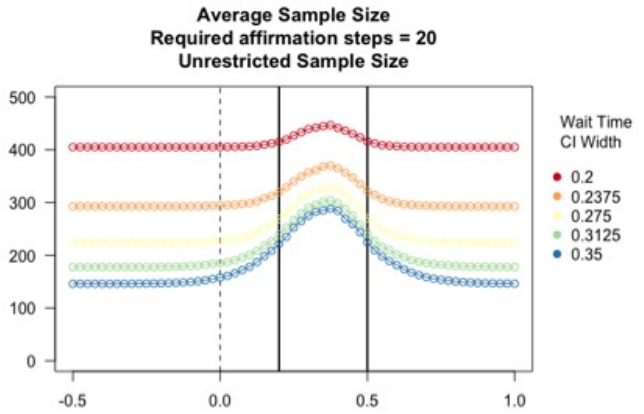

Figure 4.2: The average sample size across treatment effects in a one-sided study when requiring different wait times before monitoring. The first vertical line denotes the upper boundary of At Most Trivial Effects, and the second vertical line denotes the boundary of the Highly Actionable Effects. The Wait times are the expected sample size for achieving a confidence interval width. In this figure there is no restriction on sample size nor a lag in observing outcomes.

Non-adaptive studies are typically designed a with high probability of an inconclusive finding. Consider a non-adaptive study designed to have 80% power at a clinically highly actionable effect size. Such a study will include the point null in its final interval 20% of the time. It will include values from the null region with an even higher probability. Although designed to only stop when a clinically highly actionable conclusion has been found, second generation p-value adaptive monitored trials may yield a clinically inconclusive finding if the trial has a fixed maximum sample size and/or a lag between recruitment and outcome observation. This probability is highest in between the bounds of the Trivial and Highly Actionable regions.

Figure 4.3 illustrates the control of this probability provided through increasing the affirmation steps in the presence of a 50 subject lag time. Even with this large lag, a relatively small affirmation step requirement bounds the probability of an inconclusive finding below 20%.

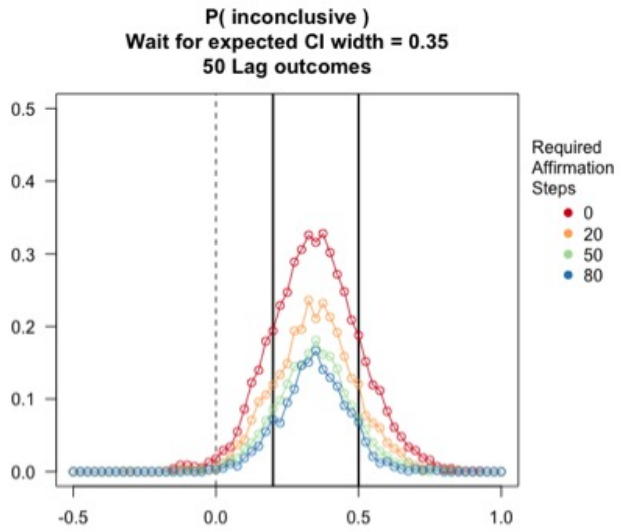

Figure 4.3: The probability of an inconclusive study (i.e. not ruling out Trivial or Highly Actionable effects) when outcomes have a lag time until being observed relative to enrollment. The study stops based upon drawing conclusions from observed data, yet after the remaining lagged outcomes (50 in this figure), the study would suggest more observations are needed to rule out Trivial or Highly Actionable effects. The risk of being inconclusive is greatest in the midpoint between the Trivial and Highly Actionable Regions (the boundary of the regions are respectively denoted by the two vertical lines). Requiring fifty observations to affirm an alert reduces the worst-case risk of being inconclusive to 20% in this setting.

4.3 Practical Recommendations

Of key importance to classical trialists is controlling point null Type I Error and achieving a high probability of excluding the point null when the true effect size is clinically highly actionable. To the medical researcher, key importance is to complete the study with a clinically highly actionable finding. We show that focusing on the later will achieve the former. When a study is not able to observe outcomes immediately, care should be taken to reduce the risk of stopping and then being inconclusive after observing the remaining observations.

To control error rates, one may change the clinical guideposts, reduce the number of times monitoring the study, and/or increase the number of steps before affirming a stopping rule. Ideally, the clinical guideposts chosen for their clinical interpretation. We discourage altering them for the sake of operational characteristics. Instead, we encourage optimizing operating characteristics through the waiting a period of time before monitoring and requiring a number of observations to pass until affirming an

alert to stop the study.

We consider various wait times based upon the expected confidence interval width under an assumed outcome standard deviation. Twenty thousand MCMC replicates of a study with standard normal outcomes are generated using the sgpvAM package under five one-sided hypotheses settings and five symmetric two-sided hypotheses. The settings reflect situations where (Setting 1) the Trivial and Highly Actionable Effect bounds are both close to the point null of zero, (Settings 2-4) the Trivial Effect bound is close to the null and the Highly Actionable Effect is far from the point null, and (Setting 5) situations where the bound for Highly Actionable Effects is far from the null and the bound for Trivial Effects is increasingly close to the Highly Actionable Effect bound. These results generalize to normal outcomes with clinical guideposts relative to the standard deviation.

For both one- and symmetric two-sided hypotheses, the probability of a Type I Error is minimized by waiting until the Margin of Error equals the midpoint between the Trivial and Highly Actionable Zones (Figure 4.4). For example, a one-sided study with At Most Trivial Effects defined as $(-\infty, 0.1]$ and Highly Actionable Effects as $[0.2, \infty)$ reduces the probability of a Type I Error by waiting until the Margin of Error is 0.15. In these simulations, the probability of a Type I Error remained less than or equal to 0.05 when waiting longer, until the Margin of Error equaled the positive boundary of the Trivial Effects.

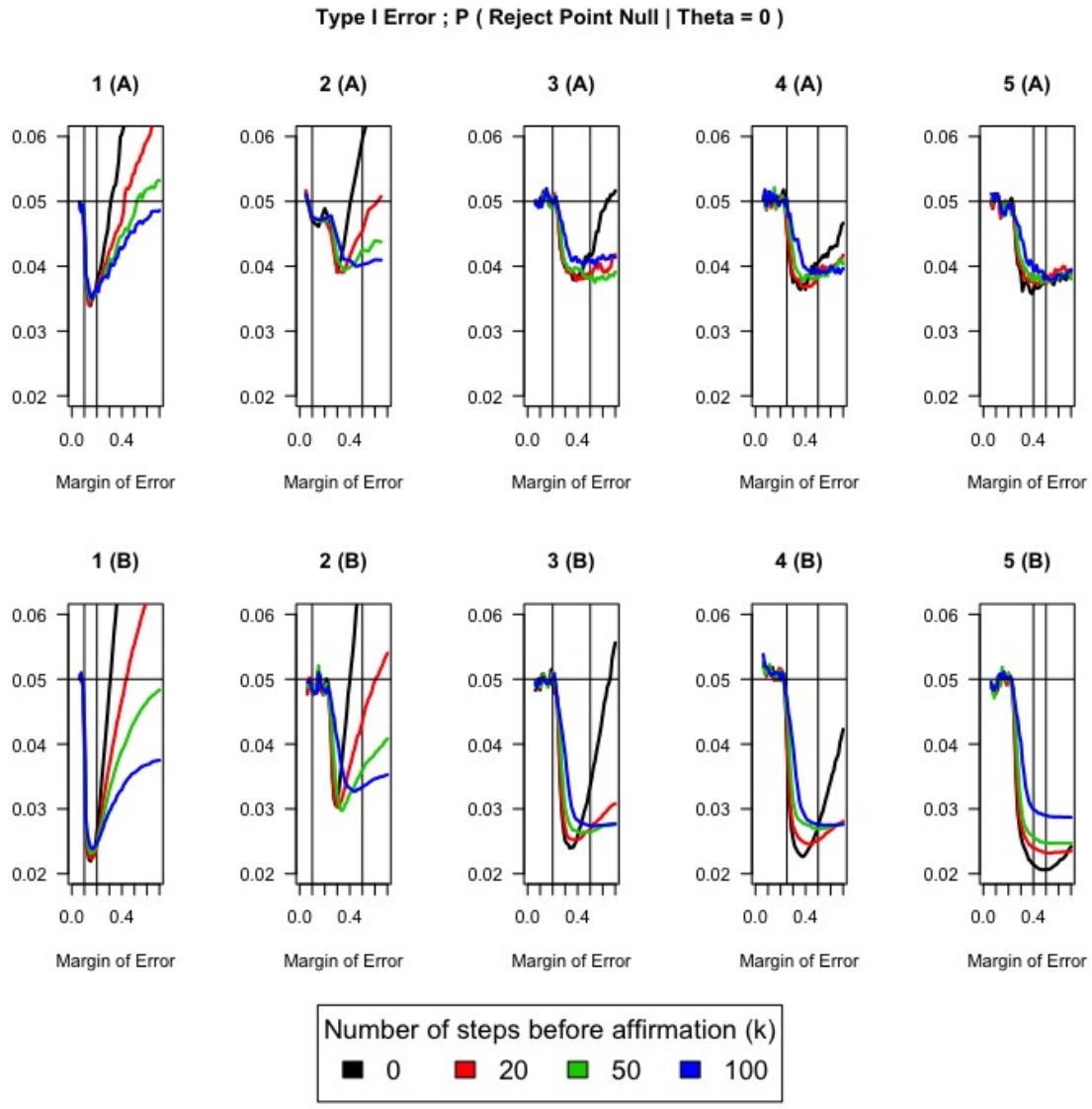

Figure 4.4: Impact of different wait times upon Type I Error. Studies observe a minimum sample size to achieve an expected minimum Margin of Error (half-width of a 95% confidence interval) under an assumed outcome standard deviation of 1. Five one-sided (A) and five symmetric two-sided (B) hypotheses are investigated. The upper bound for the Trivial Effect is denoted by the first vertical line, and the second vertical line denotes the minimal highly actionable effect greater than zero. Five combinations of effect size boundaries are shown. The lower bounds for the symmetric two-sided hypotheses are not shown. For each minimum sample size / minimum CI width, operating characteristics are provided with varying steps before affirming the stopping rule.

The following operating characteristics benefit from a longer wait time (i.e. waiting until the Margin of Error is more narrow): power (Figure 4.5); an interval null equivalent to Type I Error (Figure 4.8 and 8), Power (Figure 4.7), and Type Two Error (Figure 4.9); and the probability of stopping for conclusive observed outcomes yet becoming inconclusive after observing the remaining unobserved outcomes (Figure

4.10). On the other hand, a shorter wait time yields smaller average sample sizes (Figure 4.6). The gains in sample size diminishes once the wait time is the midpoint between the Trivial and Highly Actionable Zones.

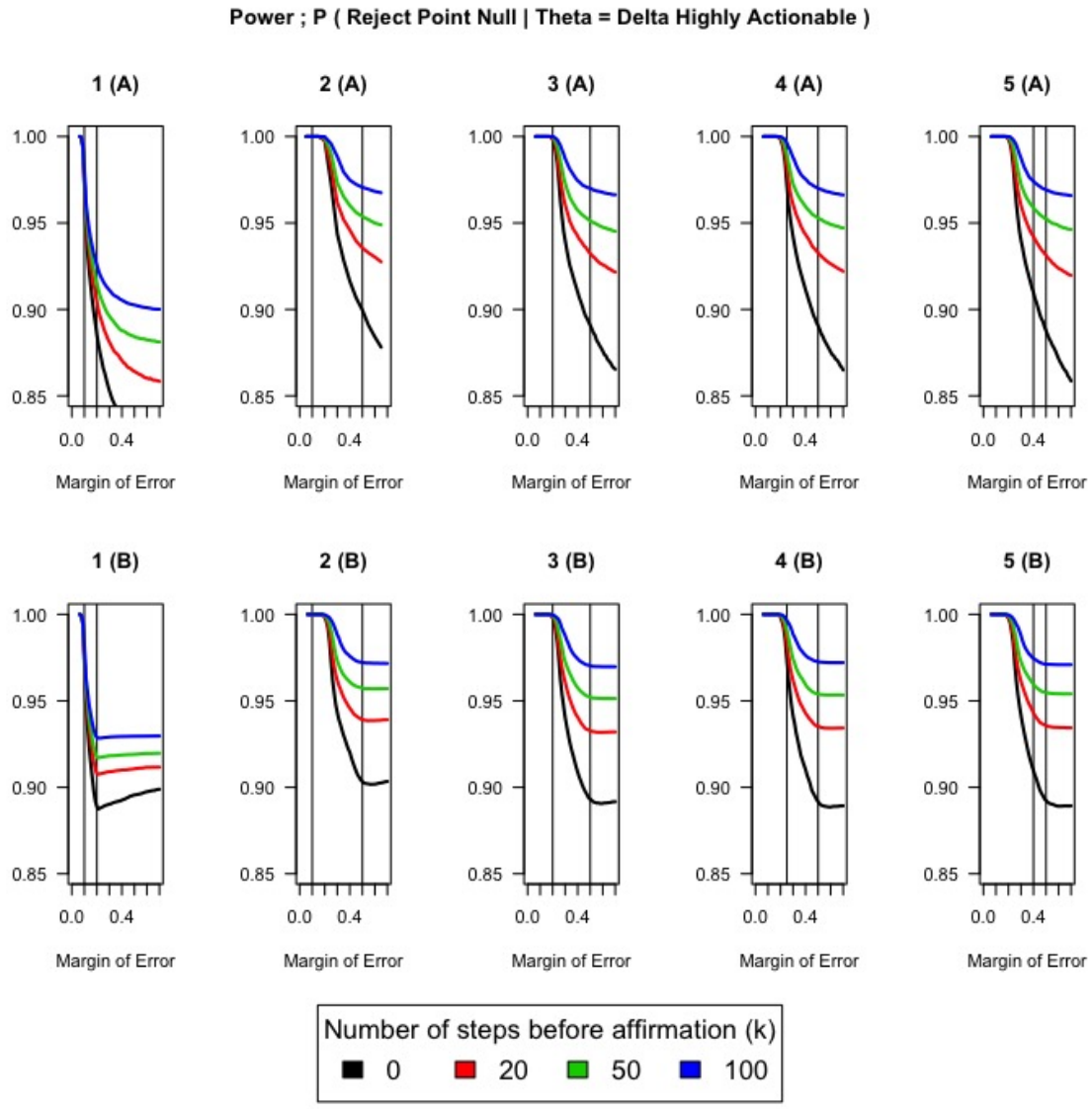

Figure 4.5: Impact of different wait times upon power, i.e. probability of rejecting the point null when the true effect is equal to the boundary of the highly actionable effect zone, i.e. the vertical line on the right. All other features of the figure mirror those of Figure 4.4.

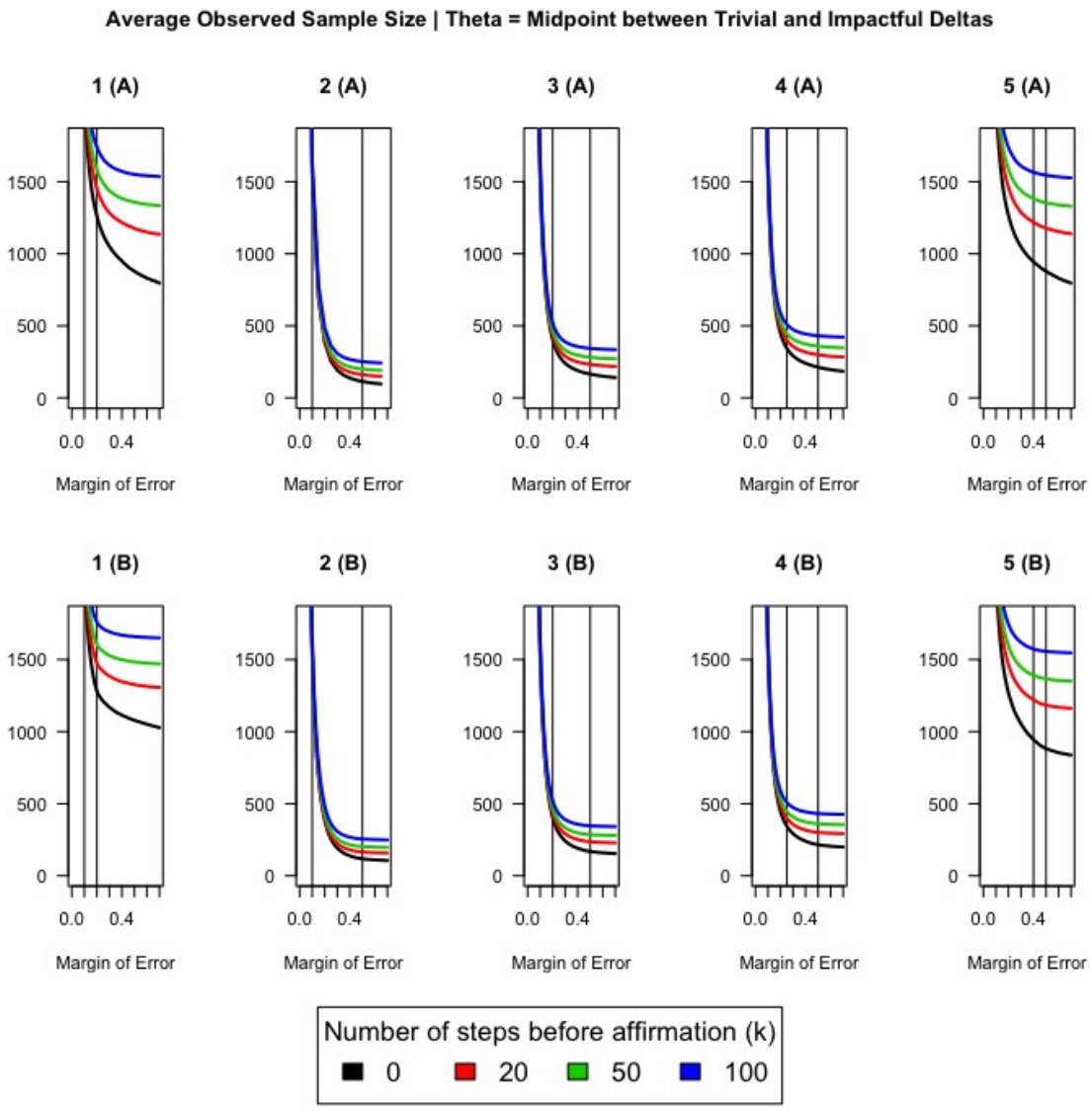

Figure 4.6: Impact of different wait times upon the average observed sample size given the true treatment effect equals the average of the absolute trivial and highly actionable effect boundaries, i.e. the middle of the two vertical lines. When the observation lag is zero, the observed sample size equals the total sample size. All other features of the figure mirror those of Figure 4.4.

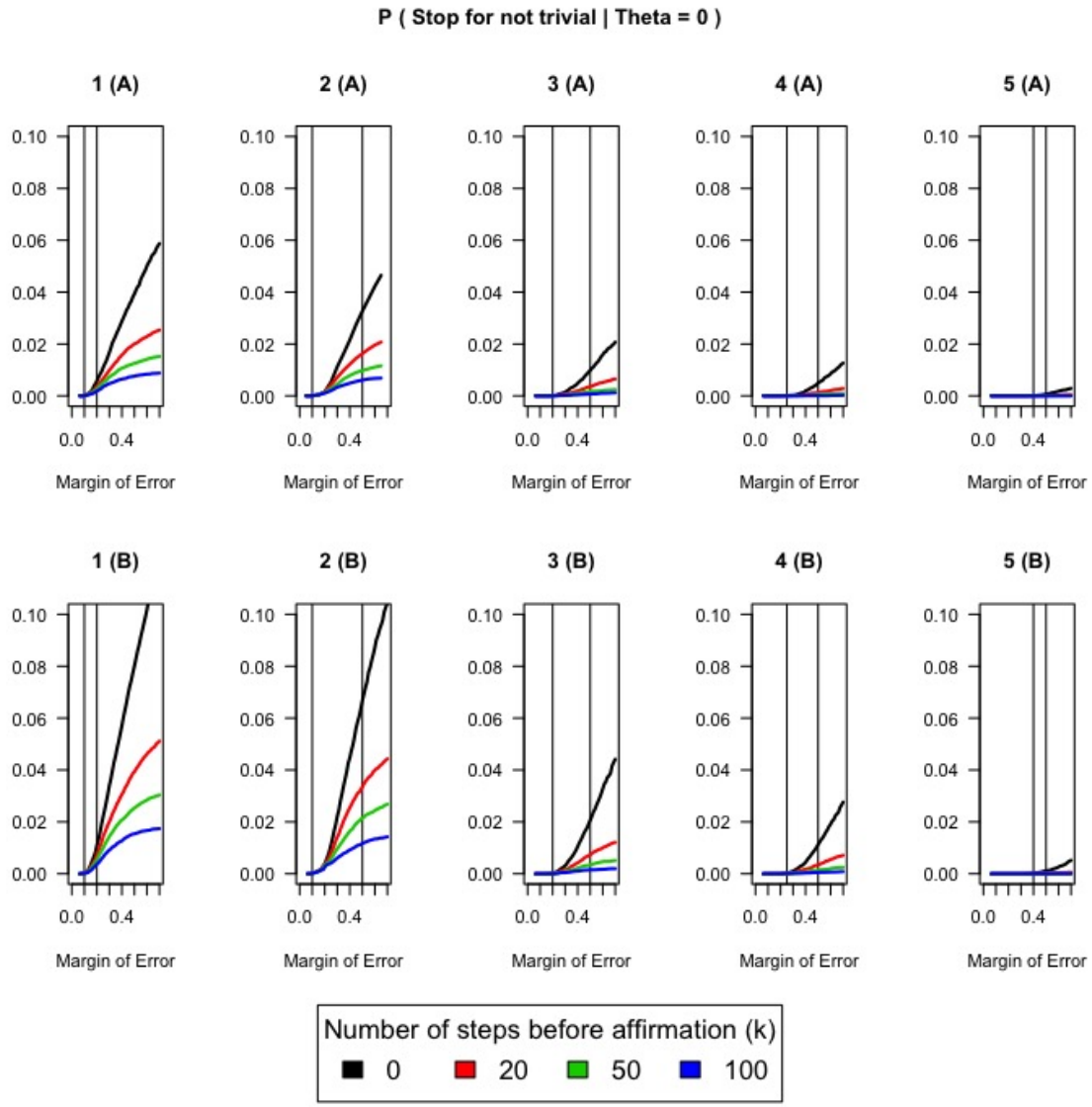

Figure 4.7: Impact of different wait times upon the probability of concluding the treatment effect is not trivial given the null true treatment. This an interval null analog to the Type I Error for a given effect. All other features of the figure mirror those of Figure 4.4.

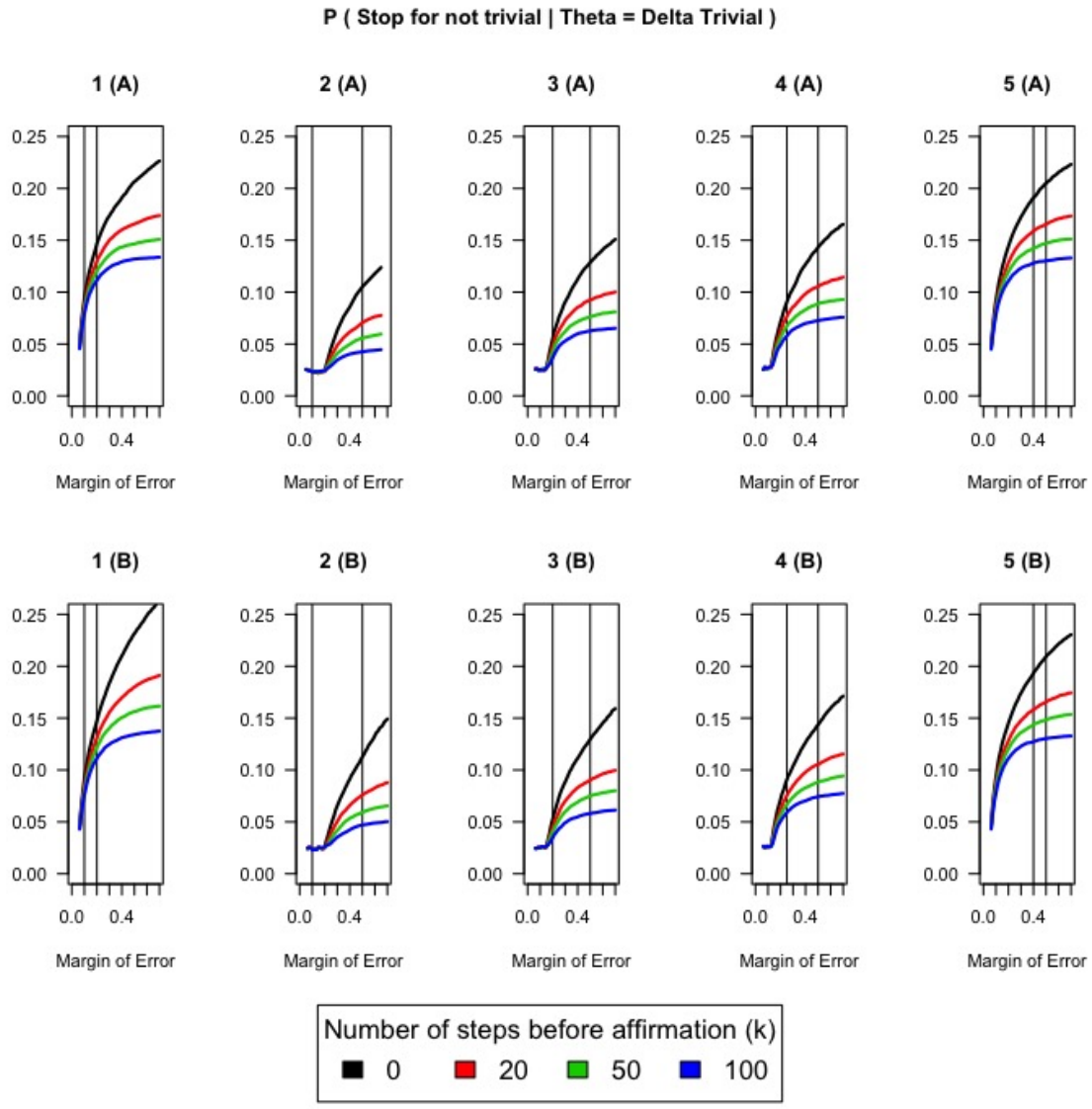

Figure 4.8: Impact of different wait times upon the probability of concluding the treatment effect is not trivial given the true effect equals the upper bound defining Trivial effects. This is an interval null analog to Type I error for a given effect. This figure focuses on the boundary of the Trivial effects where the error probability is greatest. All other features of the figure mirror those of Figure 4.4.

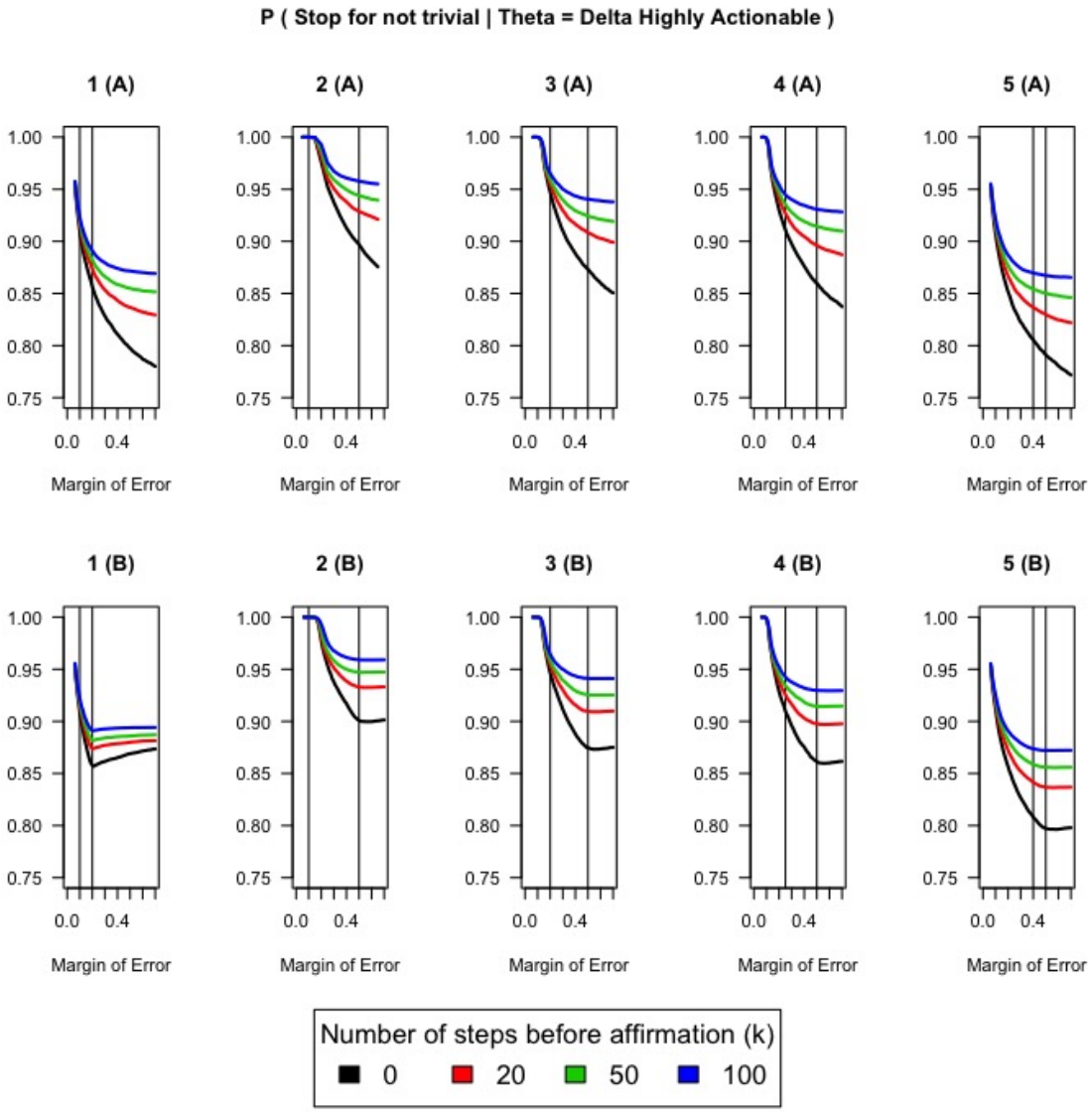

Figure 4.9: Impact of different wait times upon the probability of concluding the treatment effect is not trivial given the true effect equals the boundary of the highly actionable effect zone, ie the vertical line on the right. This is an interval null analog to power for a given effect. All other features of the figure mirror those of Figure 4.4.

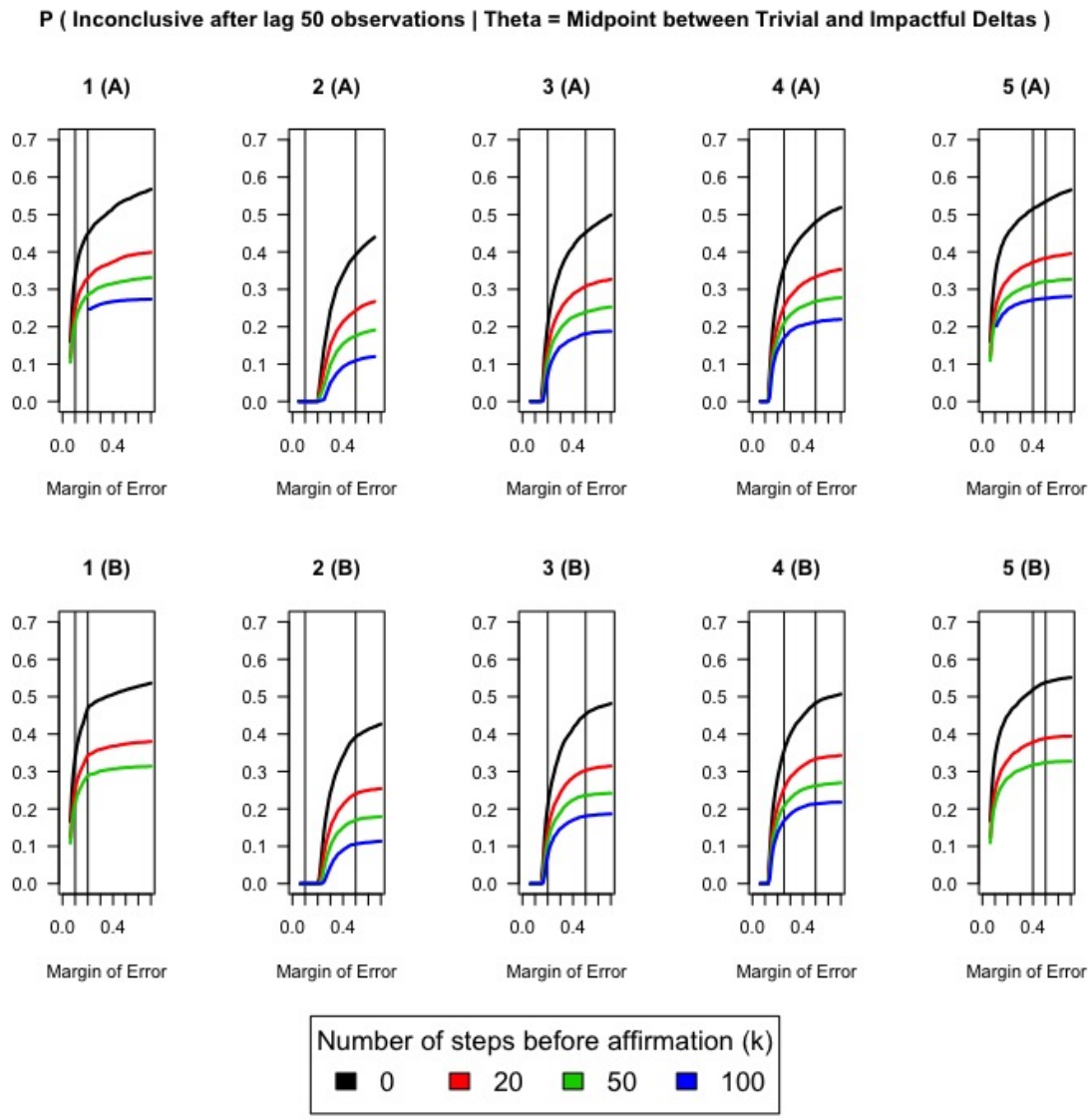

Figure 4.10: Impact of different wait times upon the probability of being inconclusive after stopping for observed outcomes and then observing remaining observations (50 observations in these simulations). In this figure, the true effect is given to be the absolute average of the Trivial and Highly Actionable regions (i.e. the mid point between the two regions); this is the treatment effect with the worst probability of being inconclusive. Refer to figure 3 for the probability of being inconclusive across the range of treatment effects under setting 3 (A). All other features of the figure 11 mirror those of Figure 4.4.

On whole, this suggests waiting no longer than when the Margin of Error achieves a length equal to this midpoint and may motivate waiting slightly longer (i.e. for a smaller Margin of Error). In certain settings, when the boundary of the trivial effects is close to half the boundary of the highly actionable boundary (settings 3 and 4), waiting for a slightly smaller Margin of Error has little added burden on the

average sample size (Figure 4.6, compare settings 3 and 4 to setting 2). Across these simulations, waiting for a slightly tighter Margin of Error balances well maximizing operating characteristics without a substantive increase in average observed sample size.

As a separate practical recommendation, we suggest increasing the number of affirmation steps before stopping when observations are not immediately observable relative to enrollment. For effects that are neither Trivial nor Highly Actionable, there's a very plausible chance the study ends conclusively for the outcomes observed yet becomes inconclusive after observing the outcomes of remaining study observations (Figure 4.10). To reduce this risk, we recommend increasing the number of steps required to affirm stopping rules.

4.4 Conclusion

We provide the sgpvAM package to provide greater ease of access to develop and study the operating characteristics of adaptive monitoring with the second generation p-value. Based on simulations of normally distributed data, we recommend waiting until the expected Confidence Interval Width is one quarter the absolute distance between the Trivial and Highly Actionable Region boundaries. When outcomes do not occur immediately, there is a risk of the study stopping then being inconclusive once all observations are observed. This risk is reduced by requiring an increased number of steps before affirming an alert to stop the study.

CHAPTER 5

CONCLUSIONS

This body of work introduces study design methods and to improve the effective sample size, allow for greater personalization of intervention to subgroups, and adaptively follow studies until reaching a clear clinical conclusion. Software and practical recommendations are provided for developing adaptive monitoring study designs.

First, we provide novel extensions to Sequential Matched Randomization which achieve greater covariate balance and efficiency than existing Sequential Matched Randomization and traditional methods such as Stratified Block Randomization. A dynamic and empirically-estimated matching threshold allows all patients to match and relaxes an assumption that baseline covariates are normally distributed. We allow matches to break and rematch if a better matches enrolls in the study. Under our method, randomization-based inference achieves nearly the same efficiency as fitting an adjusted linear model to adjust for baseline covariates. And, with greater covariate balance, an investigator may better investigate the response of subgroups to intervention.

Second, we introduce adaptive monitoring with the second-generation p-value which allows for following a study until ruling out effects deemed trivial or until ruling out effects highly actionable to change clinical practice. This grounds the clinical trial not only on statistical significance but also clinical relevance, and can help reduce the risk of a trial ending with inconclusive findings. We provide a case study through the REACH, a Vanderbilt University Clinical Trial, aimed to help patients with diabetes increase glycemic control and better adhere to medications.

Finally, we develop statistical software and provide practical recommendations for designing an adaptive monitoring trial using the second generation p-value. The R package sgpvAM simulates data to estimate operating characteristics for these trials; the required simulation may be a barrier of entry to use this adaptive trial design. We recommend a wait time before applying monitoring rules to control the classical Type I error. When outcomes are not immediately observed relative to enrollment, we recommend increasing the number of observations before affirming a stopping alert. This reduces the risk of stopping based on observed observations then finding a study

inconclusive when observing the remaining observations.

REFERENCES

Armitage, Peter. 1982. "The role of randomization in clinical trials." *Statistics in Medicine* 1 (4): 345–52.

Atkinson, A C. 1982. "Optimum Biased Coin Designs for Sequential Clinical Trials with Prognostic factors." *Biometrika* 69 (1): 61.

Berchialla, Paola, Dario Gregori, and Ileana Baldi. 2018. "The Role of Randomization in Bayesian and Frequentist Design of Clinical Trial." *Topoi* 76 (4): 479.

Berger, Vance W, Anastasia Ivanova, and Maria Deloria Knoll. 2003. "Minimizing predictability while retaining balance through the use of less restrictive randomization procedures." *Statistics in Medicine* 22 (19): 3017–28.

Berry, Scott, Bradley Carlin, J Lee, and Peter Müller. 2010. "Bayesian Adaptive Methods for Clinical Trials." CRC Press.

Blume, Jeffrey D, Lucy DAgostino McGowan, William D Dupont, and Robert A Greevy Jr. 2018. "Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses." *PloS One* 13 (3): e0188299 EP.

Bothwell, Laura E, Jeremy A Greene, Scott H Podolsky, and David S Jones. 2016. "Assessing the Gold Standard–Lessons from the History of RCTs." *New England Journal of Medicine* 374 (22): 2175–81.

Broglio, Kristine R, Jason T Connor, and Scott M Berry. 2014. "Not too big, not too small: a goldilocks approach to sample size selection." *Journal of Biopharmaceutical Statistics* 24 (3): 685–705.

Ciolino, Jody, Wenle Zhao, Renee Martin, and Yuko Palesch. 2011. "Quantifying the cost in power of ignoring continuous covariate imbalances in clinical trial randomization." *Contemporary Clinical Trials* 32 (2): 250–59.

Collier, Roger. 2009. "Rapidly rising clinical trial costs worry researchers." *CMAJ* 180 (3): 277–78.

Diener-West, M, T W Dobbins, T L Phillips, and D F Nelson. 1989. "Identification of an optimal subgroup for treatment evaluation of patients with brain metastases using RTOG study 7916." *International Journal of Radiation Oncology, Biology, Physics* 16 (3): 669–73.

Fisher, Ronald A. 1935. "The design of experiments. 1935." *Oliver and Boyd, Edinburgh*.

Freedman, David A. 2008a. "On regression adjustments to experimental data." *Advances in Applied Mathematics* 40 (2): 180–93.

———. 2008b. "On regression adjustments in experiments with several treatments." *The Annals of Applied Statistics* 2 (1): 176–96.

Freedman, L S, D Lowe, and P Macaskill. 1984. "Stopping rules for clinical trials incorporating clinical opinion." *Biometrics* 40 (3): 575–86.

Fu, Haoda, Jin Zhou, and Douglas E Faries. 2016. "Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies." *Statistics in Medicine* 35 (19): 3285–3302.

Greevy, R, B Lu, J H Silber, and P Rosenbaum. 2004. "Optimal multivariate matching before randomization." *Biostatistics (Oxford, England)*.

Greevy Jr., Robert A, Carlos G Grijalva, Christianne L Roumie, Cole Beck, Adriana M Hung, Harvey J Murff, Xulei Liu, and Marie R Griffin. 2012. "Reweighted Mahalanobis distance matching for cluster-randomized trials with missing data." *Pharmacoepidemiology and Drug Safety* 21 (May): 148–54.

Hill, A Bradford. 1952. "The Clinical Trial." *New England Journal of Medicine* 247 (4): 113–19.

Hobbs, Brian P, and Bradley P Carlin. 2008. "Practical Bayesian design and analysis for drug and device clinical trials." *Journal of Biopharmaceutical Statistics* 18 (1): 54–80.

Johnston, S Claiborne, John D Rootenberg, Shereen Katrak, Wade S Smith, and Jacob S Elkins. 2006. "Effect of a US National Institutes of Health programme of clinical trials on public health and costs." *Lancet (London, England)* 367 (9519): 1319–27.

Kallus, N, and 2018. 2018. "Optimal a priori balance in the design of controlled experiments." *Journal of the Royal Statistical Society Series B* 80: 85–112.

Kapelner, Adam, and Abba Krieger. 2014. "Matching on-the-fly: Sequential allocation with higher power and efficiency." *Biometrics* 70 (2): 378–88.

Kruschke, John K. 2013. "Bayesian estimation supersedes the t test." *Journal of Experimental Psychology. General* 142 (2): 573–603.

———. 2018. "Rejecting or Accepting Parameter Values in Bayesian Estimation:" *Advances in Methods and Practices in Psychological Science* 1 (2): 270–80.

Leyland-Jones, B. 2003. *Breast cancer trial with erythropoietin terminated unexpectedly.* The lancet oncology.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *The Annals of Applied Statistics* 7 (1): 295–318.

Loux, Travis M. 2014. "Randomization, matching, and propensity scores in the design and analysis of experimental studies with measured baseline covariates." *Statistics in Medicine* 34 (4): 558–70.

Medicine, AC Atkinson Statistics in, and 1999. 1999. "Optimum biased-coin designs for sequential treatment allocation with covariate information (With Discussion)." *Journal of the Royal Statistical Society Series B* 18 (14): 1753–5.

Moore, T J, H Zhang, G Anderson JAMA internal, and 2018. n.d. "Estimated costs of pivotal trials for novel therapeutic agents approved by the US Food and Drug Administration, 2015-2016." *Jamanetwork.com*.

Morgan, K L, and D B Rubin. 2012. "Rerandomization to improve covariate balance in experiments." *The Annals of Statistics*.

Nelson, Lyndsay A, Kenneth A Wallston, Sunil Kripalani, Robert A Greevy Jr., Tom A Elasy, Erin M Bergner, Chad K Gentry, and Lindsay S Mayberry. 2018. "Mobile Phone Support for Diabetes Self-Care Among Diverse Adults: Protocol for a Three-Arm Randomized Controlled Trial." *JMIR Research Protocols* 7 (4): e92.

Peirce, C S, and J Jastrow. 1884. "On small differences in sensation." *Memoirs of the National Academy of Sciences* III: 73–83.

Pocock, Stuart J, and Richard Simon. 1975. "Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial." *Biometrics* 31 (1): 103–15.

Pocock, Stuart J, and Gregg W Stone. 2016a. "The Primary Outcome Fails - What Next?" *New England Journal of Medicine* 375 (9): 861–70.

———. 2016b. "The Primary Outcome Is Positive - Is That Good Enough?" *New England Journal of Medicine* 375 (10): 971–79.

Rosenberger, W F, and O Sverdlov. 2008. "Handling covariates in the design of

clinical trials."

Senn, Stephen, Vladimir V Anisimov, and Valerii V Fedorov. 2010. "Comparisons of minimization and Atkinson's algorithm." *Statistics in Medicine* 29 (7-8): 721–30.

Spiegelhalter, David J, Laurence S Freedman, and Mahesh K B Parmar. 1994. "Bayesian Approaches to Randomized Trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 157 (3): 357.

Taves, Donald R. 1974. "Minimization: A new method of assigning patients to treatment and control groups." *Clinical Pharmacology &Amp; Therapeutics* 15 (5): 443–53.

Wei, L J, and John M Lachin. 1988. "Properties of the urn randomization in clinical trials." *Controlled Clinical Trials* 9 (4): 345–64.

Zhou, Quan, Philip Ernst, Kari Lock Morgan, Donald B Rubin, and Anru Zhang. 2018. "Sequential Rerandomization." *arXiv.org*, April, 1–23.