Practical Considerations for Modern Clinical Trials: Three Projects in Clinical Trial Design, Conduct and Analysis

Subodh Selukar

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Susanne May, Chair

Megan Othus, Chair

David Prince

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Practical Considerations for Modern Clinical Trials: Three Projects in Clinical Trial Design, Conduct and Analysis

Subodh Selukar

Co-Chairs of the Supervisory Committee:
Susanne May
Department of Biostatistics

Megan Othus
Fred Hutchinson Cancer Research Center

This dissertation comprises three projects that span the design, conduct and analysis of contemporary clinical trials, presented in individual chapters. The first project extends methods for stratified randomization to account for the possibility that experimental arms of a platform trial differ in eligibility criteria. The second project proposes a novel approach for evaluating cure model appropriateness in studies with long-term survivors. The third project develops a framework for sequential monitoring in one N-of-1 trial and joint analysis of a series of sequentially-monitored N-of-1 trials.

Project 1: We extend methods for stratified randomization to the setting of differing experimental arm eligibility in platform trials. We suggest modifying block randomization by including experimental arm eligibility as a stratifying variable, and we suggest modifying the imbalance score calculation in dynamic balancing by performing pairwise comparisons between each eligible experimental arm and standard of care arm participants eligible to that experimental arm. We also derive a formula to quantify the relative efficiency loss of platform trials with varying eligibility compared to trials with non-varying eligibility.

Project 2: We develop a novel approach for evaluating cure model appropriateness in studies with long-term survivors. We propose a method that assesses the proportion of uncured remaining at the time of analysis. We demonstrate that this method has desirable asymptotic and finite-sample properties with parametric models and that it displays superior performance over existing methods.

Project 3: We propose a framework for the sequential monitoring of one N-of-1 trial and the joint analysis of a series of sequentially-monitored N-of-1 trials. We suggest considering design blocks (repeated units of time with fixed numbers of each treatment allocation) as independent units for use with existing monitoring boundaries when analyzing continuous data with a linear mixed-effects model. To jointly analyze several trials together, we propose computing a bias-adjusted estimate for each trial and then combining the estimates with a random-effects model with inverse-variance weighting. We show that type-1 error can be inflated for N-of-1 trials with few treatment blocks under sequential monitoring, but trials with a substantial number of treatment blocks or with a substantial number of periods per block can have nominal rates. For those settings, our proposed framework for sequential monitoring can support clinicians in providing important decisions earlier, on average, for patients engaged in N-of-1 trials.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

This dissertation would not have been possible without the unwavering support of my parents, Rajesh and Vanashree Selukar, and my fianceé, Amanda Alexander. Their encouragement served as a source of confidence for me when I may otherwise have wavered.

I wish to thank my friends, both those from before graduate school and those whom I met in the program. I was no stranger to challenges in my studies, and my friends always made those times easier.

My experience within the Department of Biostatistics was shaped by many people, and I am greatly appreciative of the opportunities and training I have received by being surrounded by people of such talent and wisdom. I want to give special thanks to Nayak Polissar and Gitana Garofalo in addition to my dissertation committee. Nayak was one of my first contacts in the department, and I am thankful for how he always seems to see and promote my potential. And I thank Gitana for frequently going beyond the duties of graduate program advisor in order to support me. Many thanks go to Marco Carone and Amalia Magaret for providing me with feedback as supervisory committee members and to David Prince for his suggestions as a reading committee member.

Finally, I want to express my deepest thanks to my dissertation co-advisors, Susanne May and Megan Othus. Together, they have shaped how I approach problems and communicate statistics, and they have provided me with multiple valuable opportunities even beyond the dissertation research. I am incredibly grateful for Megan's infectious enthusiasm and insight and for Susanne's keen attention to the scope and directions of my research. Because of these qualities, I am proud of the state of this dissertation, and I hope to emulate these qualities as a researcher in the future.

DEDICATION

To my family, my friends, and my advisors

Chapter 1

STRATIFIED RANDOMIZATION FOR PLATFORM TRIALS WITH DIFFERING EXPERIMENTAL ARM ELIGIBILITY

This research has been published in the journal *Clinical Trials*:
Selukar S, May S, Law D, Othus M. Stratified randomization for platform trials with differing experimental arm eligibility [published online ahead of print, 2021 Aug 21]. Clin Trials. 2021;17407745211028872. doi:10.1177/17407745211028872

1.1 Background

The novel coronavirus COVID-19 spurred an unprecedented effort for rapid discovery of effective therapeutic and preventative agents. Some of these efforts involve the use of platform trials (I-SPY COVID-19, NCT04488081; DisCoVeRy, NCT04315948; RECOVERY, NCT04381936; SOLIDARITY, EudraCT Number 2020-000982-18), a type of master protocol design that compares multiple experimental arms to a common standard of care arm. Authors have previously described how to add and drop arms within a single protocol, making the design appealing for rapidly developing areas. [1, 2]

In oncology, the LEAP trial (NCT03092674) [3] was a platform trial for acute myeloid leukemia (AML), conducted by the NCI-funded SWOG Cancer Research Network. The trial evaluated several therapies with varying mechanisms of action, including immunotherapies and targeted agents. Because of these varying mechanisms, certain participant subpopulations were thought to potentially be harmed by specific therapies, so eligibility necessarily differed across arms. For example, patients with pre-existing autoimmune diseases were precluded from receiving checkpoint inhibitor immunotherapy. However, trial investigators did not want to modify trial-wide eligibility to address this, as a key desire for the trial

was to be as inclusive as possible. Allowing for varying eligibility did not impact trial conduct beyond requiring appropriate treatment assignment and, thus, randomization. Other trial examples have acknowledged varying eligibility, but they did not detail their procedures for treatment assignment, and authors have called for more research on this issue. [2, 4, 5]

Investigators may wish to perform stratified randomization to ensure that important baseline prognostic factors remain balanced across study arms. Further, adjusting for stratification variables may also increase power during analysis. [6] Existing methods for stratified randomization assume that all arms share the same eligibility criteria. As such, utilizing these methods requires applying the most stringent set of eligibility criteria: the intersection across every study arm's eligibility. This may limit trial accrual and may hinder inference to the appropriate target populations. [7, 8]

In this chapter, we propose extensions to existing methods of both block randomization and dynamic balancing to appropriately perform stratified randomization in the setting of varying eligibility. We provide worked examples of both approaches, and we also briefly describe the efficiency (in terms of the size of the common standard of care arm) of platform trials when experimental arms differ in participant eligibility.

1.2 Methods

1.2.1 Extending Existing Methods for Stratified Randomization

In the analysis of a trial with differing experimental arm eligibility, each experimental arm is compared to the subset of participants in the standard of care arm who meet the experimental arm's eligibility criteria. For example, consider a trial with three experimental arms, E1, E2 and E3. A participant randomized to the standard of care arm and eligible to experimental arms E1 and E2 but not E3 could be used in comparisons with E1 and E2 but is not used in a comparison with the experimental arm E3. This raises a difficulty of treatment assignment: how should we appropriately assign participants to maintain a desired allocation ratio, achieve balance across specified stratification factors and also accommodate

this varying eligibility? This section outlines the modifications we propose to allow existing methods in block randomization and dynamic balancing to address this problem.

Throughout this chapter, we consider a platform trial defined as a multi-arm trial with a single, common standard of care arm and at least two experimental arms that may start and/or end at different time points. We want to achieve balance on a prespecified set of stratification variables observable prior to arm assignment that remains the same across experimental arms. We assume all participants are eligible to the standard of care arm and at least one experimental arm for enrollment, and subjects are assigned to exactly one study arm. We describe 1:1 randomization in the examples, but each of the two proposed methods easily accommodates other allocation ratios in a manner corresponding to what would be done for the respective method without differing eligibility.

Block Stratified Randomization

In block stratified randomization, investigators create blocks (of fixed or varying size) for each stratum defined by the stratification variables and fill the blocks with random treatment assignments such that balance is achieved between arms within each block. A newly enrolled participant is matched with the corresponding stratum block and assigned to the next unassigned treatment.

We propose adding arm eligibility as an additional stratification factor to accommodate arms with differing eligibility. In other words, each possible combination of experimental arm eligibility contains its own nested set of the prespecified strata. With three experimental arms and differing eligibility criteria for each arm, a participant could be eligible to only one of the three, two of the three or all three arms, thus representing 7 ($= 2^3 - 1$) experimental arm eligibility strata. Additionally, these eligibility strata would themselves contain strata based on the prespecified stratification variables.

This simple extension addresses the problem of differing eligibility while achieving balance and targeting the desired allocation ratio. However, the number of possible eligibility combinations increases exponentially with the number of experimental arms. A trial with

three experimental arms and just one binary stratification variable would have 14 strata, and many trials may desire more experimental arms and/or stratification factors.

Importantly, while the maximum number of eligibility combinations for K experimental arms is 2^K-1, this does not necessarily represent the number eligibility strata for a given trial, as these depend on exactly how the experimental arms differ in eligibility. To illustrate this, we describe two examples. First, consider a trial with two experimental arms. In addition to any trial-wide eligibility criteria, the first experimental arm (arm A) only recruits subjects over the age of 65 and the second (arm B) only recruits biomarker positive subjects. In this case, subjects can be eligible to both arms (older than 65 and biomarker positive) or exactly one (either older than 65 or biomarker positive but not both), so the number of eligibility combinations is fully $2^2 - 1 = 3$. Suppose, instead, arm B had no additional restrictions beyond the trial-wide eligibility. In that case, while there are three possible combinations, only two eligibility strata are needed: (1) eligible to A and B or (2) only eligible to B, as no subjects would be only eligible to A. To expand on this, consider a different trial with three experimental arms (E_1, E_2 and E_3). Beyond trial-wide eligibility criteria, E_1 only recruits subjects younger than 65, E_2 only recruits subjects younger than 75, but E_3 has no additional restrictions. This results in just three eligibility strata: (1) eligible to E_1, E_2 and E_3 (for subjects younger than 65), (2) eligible to E_2 and E_3 but not E_1 (subjects aged 65-74), or (3) only eligible to E_3 (subjects aged 75 and older).

Dynamic Balancing

A general scheme of dynamic balancing [9] involves calculating the imbalance caused by provisionally assigning a participant to each eligible study arm and then using these imbalance scores to weight the participant's randomization toward the arm that results in the least imbalance. As each new participant enters the study, this procedure is repeated. The method requires specification of how to compute an "imbalance score" and how to use these imbalance scores to weight study arm assignment.

To accommodate differing arm eligibility, we propose that investigators modify an existing

dynamic balancing scheme's imbalance score calculation by considering pairwise calculations for each experimental arm and participants assigned to the standard of care arm who were eligible for that experimental arm. These pairwise calculations can be summarized into a single imbalance score (e.g., the maximum across the pairwise calculations) for adding the new participant to a given eligible study arm analogous to an imbalance score from the underlying dynamic balancing scheme.

For example, consider imbalance measured by differences in counts, defined as $T(\cdot)$, of subjects of the same stratification factor level as a new participant. For each eligible experimental arm E_j $(j = 1, 2, \ldots)$ and corresponding standard of care subset C_{E_j}, we can tally the number who have the same stratification factor level and add to the tally the new participant. Then we conduct pairwise comparisons with the absolute differences $|T(E_j) - T(C_{E_j})|$. The imbalance score for adding the new participant to a given arm could then be the maximum of these differences. (We provide full details for this scheme in the Online Appendix.) Existing methods, which do not allow for varying eligibility, only calculate one count for the standard of care arm $(T(C))$ and that same count is used in all the difference calculations, e.g. $|T(E_1) - T(C)|$.

The calculated imbalance scores may then be mapped to randomization weights as in dynamic balancing without varying eligibility criteria. As these weights can accommodate a desired allocation ratio, this method also resolves the issue of varying eligibility while achieving balance and targeting the desired allocation ratio.

Stratified Randomization when Adding or Dropping Experimental Arms

Dropping an experimental arm does not require modifications to the above procedures. The dynamic balancing algorithm would no longer compute an imbalance score for adding to the dropped arm and would not compute pairwise differences with that arm. The block randomization method would discontinue the strata that include the dropped arm. For example, if an arm E3 were dropped with E1 and E2 continuing, then strata for eligibility to E1 only, E2 only or both E1 and E2 would continue, but other strata that include eligibility

to E3 would not be used for future participants.

However, the above methods do require further specification after the addition of an experimental arm. The methods involve the eligibility and arm assignment of previously-randomized participants, so we must specify how each algorithm accommodates these existing participants but also ensures only concurrently randomized participants are compared to account for potential changes in participant characteristics over time (e.g., more recent participants may have systematic differences in prior care compared to those enrolled earlier).

In order to only analyze concurrently randomized participants (i.e., not include prior standard of care participants in the analysis of a new arm), we recommend not using previously-randomized standard of care participants in randomization calculations for the new arm - even if they would have been eligible for the new arm. But to ensure maximal use of the existing data, calculations involving any continuing arms should continue to include previously-randomized standard of care participants eligible to the continuing arms.

In block randomization, this means, upon the addition of the new arm, the creation of a new stratum for participants only eligible to the new arm and also new strata for participants eligible to the new arm and combinations of the continuing arms. For example, if a new arm E3 is added with continuing experimental arms E1 and E2, then four new strata would be added: one for participants eligible to only E3, one each for E1 and E3 or E2 and E3 and one for E1, E2 and E3 eligibility. The previously-existing strata would continue for new participants eligible to any combination of the continuing arms E1 and E2 but not E3.

This recommendation does not require any significant changes to the dynamic balancing procedure after the addition of the new arm. When a new participant is eligible to continuing experimental arms, the imbalance score calculations will continue to include all participants previously assigned to the continuing arms and those assigned to standard of care but eligible to the continuing arms. Calculations involving the new arm will only include participants assigned since the addition of the new arm and eligible to the new arm.

In the STAMPEDE trial, investigators implemented a different method for randomization after adding a new experimental arm: they restarted the stratified randomization process

after the addition of each arm. [4] While this seems logistically simpler, from a simulation study summarized in the Online Appendix, we conclude that the algorithm we suggest above causes less deviation from the desired allocation ratio in the balancing process. The choice may have practical impacts to sequential monitoring of the study, as deviations from the desired allocation ratio can affect statistical operating characteristics such as type-1 error.

1.3 Results

1.3.1 Worked Examples

Block Randomization

For block randomization, we illustrate the proposed extension with an example with one binary stratification factor of biomarker status (positive or negative). Consider a trial with a standard of care arm, C, and two experimental arms, E1 and E2. Subjects can be eligible for any combination of the experimental arms but all are eligible for C and at least one experimental arm.

In Table 1.1, we provide sample treatment assignments using blocks of size 6 and 1:1 randomization (other allocations are straightforward to implement). Suppose we have already randomized several participants (assignments struck through), and the next participant to be randomized is eligible to E1 and E2 and is biomarker positive. As shown in the table (in bold and italics), the new participant would be assigned E1.

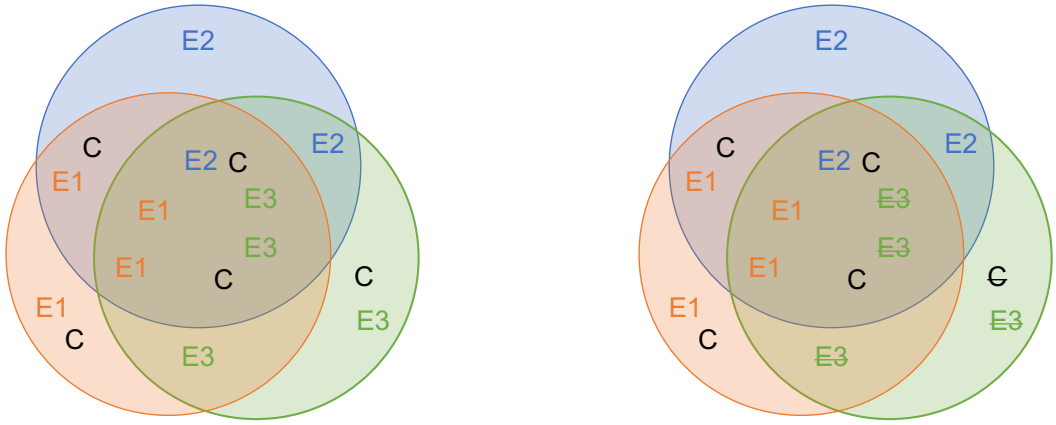
Dynamic Balancing

To illustrate dynamic balancing, we consider a trial with three experimental arms, E1, E2 and E3, a standard of care arm, C, and one binary stratification factor for biomarker status (positive or negative). Again, subjects can be eligible for any combination of the experimental arms but all are eligible for C and at least one experimental arm.

We depict the eligibility and treatment assignments of already-randomized participants with a Venn Diagram (Figure 1.1a). Each point labeled E1, E2, E3 or C represents the

Table 1.1: One block of treatment assignments for participants in each of the strata defined by arm eligibility (each combination of experimental arms E1 and E2) and biomarker status (positive or negative)

Eligible to E1 Only		Eligible to E2 Only		Eligible to E1 and E2	
Positive	Negative	Positive	Negative	Positive	Negative
~~C~~	~~E1~~	~~E2~~	~~E2~~	~~E1~~	~~C~~
~~E1~~	~~C~~	~~C~~	~~E2~~	***E1***	~~E2~~
~~E1~~	E1	C	~~C~~	E2	E2
C	C	C	E2	C	C
E1	C	E2	C	E2	E1
C	E1	E2	C	C	E1
⋮	⋮	⋮	⋮	⋮	⋮



(a) Already-Randomized Subjects Partway through Accrual

(b) Already-Randomized Subjects Partway through Accrual, Accounting for New Subject Eligible to E1, E2 and C but not E3

Figure 1.1: Eligibility and Study Arm Assignment Partway through Accrual. Each currently-randomized participant is represented by their arm assignment, and their location within the circle indicates their experimental arm eligibility prior to randomization.

treatment assignment of an already-randomized participant, and the circle(s) containing that point represent the eligibility of that participant. Points lying in the intersection of multiple eligibility circles indicate the participant is eligible to more than one experimental arm. To use existing randomization methods, it would require that these three circles overlap completely: in other words, all subjects must be eligible for all experimental arms.

Figure 1.1a describes the eligibility and assignments partway through accrual. In this case, for example, the blue (top) circle represents eligibility for E2. One participant randomized to E2 was only eligible for E2, another was eligible for E2 and E3 and one was eligible for all experimental arms. Three participants randomized to C were eligible for E2 (two were eligible for all arms and one was eligible for E2 and E1) and would be used in imbalance calculations made for new participants eligible for E2.

Suppose the trial enrolls a new biomarker positive participant eligible to enroll on arms E1, E2 and C but not E3. In Figure 1.1b, we illustrate how the dynamic balancing algorithm will not incorporate the participants randomized to E3 or standard of care participants who were only eligible to E3 (struck-through). The calculation includes standard of care participants who were eligible to all arms or any combination of E1 and E2.

For this example, we implement an extended Pocock-Simon [9] procedure. We focus on the counts of already-randomized participants of the same biomarker status and compute the imbalance from adding the new participant to each eligible study arm. To calculate the imbalance due to adding to a given arm, we do the following: For each eligible experimental arm, we compute the absolute difference between the count in the experimental arm and those in C who could have been randomized to that experimental arm. The imbalance score for adding to the given study arm is then the largest among these absolute differences. (See Online Appendix for full details.)

In Table 1.2, we cross-tabulate the number randomized to each eligible experimental arm and the number randomized to C who were eligible for that experimental arm by stratification level. The tallies in bold represent the participants with the same level of the stratification factor as the new participant.

Table 1.3 details the full calculations of the modified dynamic balancing procedure by examining the imbalance due to adding the new participant to each eligible study arm. The right column of each row gives the imbalance score due to the addition to a given arm. We see that adding the new participant to E1 or E2 gives an imbalance of 2, while adding to C gives an imbalance score of 0.

With C having the lowest imbalance score after addition of the new participant, we give it the largest weight in a weighted randomization procedure (in which the user could also specify the allocation ratio).

Table 1.2: Tallies by Arm and by Stratification Factor Level†

	E1	C_{E1}
Negative	1	2
Positive	**3**	**2**

	E2	C_{E2}
Negative	1	2
Positive	**2**	**1**

† The two tables provide the counts by biomarker status of those participants randomized to E1, E2 or C_{E1} and C_{E2}, the standard of care arm participants who were eligible for E1 and E2, respectively. The bold numbers represent the counts for participants of the same biomarker status as the new participant.

Table 1.3: Algorithm Procedure for New Study Subject

Add to E1†

E1	C_{E1}	Imbalance
4	2	$\lvert 4-2\rvert$

E2	C_{E2}	Imbalance
2	1	$\lvert 2-1\rvert$

$\max(|4-2|, |2-1|) = 2$

Add to E2†

E1	C_{E1}	Imbalance
3	2	$\lvert 3-2\rvert$

E2	C_{E2}	Imbalance
3	1	$\lvert 3-1\rvert$

$\max(|3-2|, |3-1|) = 2$

Add to C†

E1	$\mathbf{C}_{E1}$	Imbalance
3	**3**	$\lvert 3-3\rvert$

E2	$\mathbf{C}_{E2}$	Imbalance
2	**2**	$\lvert 2-2\rvert$

$\max(|3-3|, |2-2|) = 0$

† Each row indicates the addition of the new participant to a given study arm. The tables show the relevant tallies modified for this addition in bold and the right column gives the imbalance score computed by adding the new participant to that study arm.

1.3.2 Efficiency of Platform Trials with Differing Arm Eligibility

Ventz et al. [2] discuss the potentially dramatic reductions in sample size when conducting a platform trial compared to multiple independent two-arm studies. However, they also observe that the additional flexibility to add arms mid-study incurs an efficiency loss in sample size compared to randomizing all experimental arms initially.

We can describe the efficiency of platform trials with differing arm eligibility by evaluating the proportion of subjects randomized to the standard of care arm. In a two-arm study with 1:1 allocation, the proportion randomized to standard of care is $\frac{1}{2}$. In a multi-arm trial with K experimental arms and equal allocation, it is $\frac{1}{K+1}$. Platform trials with differing arm eligibility are maximally efficient when they resemble a multi-arm trial with equal allocation (corresponding to all shared eligibility) and minimally efficient when resembling a two-arm study (corresponding to each participant being eligible to exactly one experimental arm).

Using the law of total probability, we can fully characterize the proportion randomized to the standard of care arm when we assume balance is reached (see Online Appendix). We perform this calculation under the assumption that that all experimental arms are started and ending at the same time, but a more general framework allowing for adding and dropping arms is possible, if needed. The proportion depends on the probability of each combination of experimental arm eligibility. As expected, if the probability of subjects being eligible to all experimental arms is high, the platform trial has higher efficiency (i.e., more closely resembles a multi-arm trial with equal allocation); as this probability decreases, the efficiency decreases.

As an example, suppose a platform trial has two experimental arms, E_1 and E_2, and a subject's probability of being eligible to both experimental arms is α, and the probability of being eligible for only E_1 is $\frac{1-\alpha}{2}$ and the probability of being eligible for only E_2 is also $\frac{1-\alpha}{2}$. Each experimental arm plans to enroll 100 subjects at a 1:1 allocation with the standard of care arm.

Consider three scenarios: $\alpha_1 = 1$, $\alpha_2 = 0$ and $\alpha_3 = \frac{1}{2}$. The scenario with α_1 is a situation

with no varying eligibility, while α_2 has completely distinct eligibility between experimental arms, and α_3 represents an intermediate scenario. The probability of being randomized to standard of care (under assumptions given in the Online Appendix), is $\frac{1}{3}$, $\frac{1}{2}$ and $\frac{5}{12}$, respectively.

A different way to assess the efficiency of platform trials with differing arm eligibility is by comparing the relative sample sizes of the standard of care arm (differences in the total size of the trial due to differing eligibility would only occur via the size of this arm). In the above example, the sizes of the standard of care arm would be 100, 200 and $\frac{1000}{7} \approx 143$ based on the planned allocation and sample sizes. With these calculations, we reach the same conclusion as above: if the probability of subjects being eligible to all experimental arms is high, the platform trial has higher efficiency.

1.3.3 LEAP Trial Example

In this section, we describe the decision-making and treatment assignment process for the LEAP trial (clinicaltrials.gov ID: NCT03092674), which used the method for dynamic balancing described above. As stated in the background section, the LEAP trial was a platform trial for AML: in particular, it evaluated AML therapies in medically less-fit older adults, a patient population that suffers from therapeutic resistance and reduced chemotherapy tolerance. A key goal of the trial was to be "as unrestrictive as possible and to provide treatment options for the real-world patient." [3]

The trial was designed to begin with a standard of care arm, azacytidine monotherapy, and three experimental arms: (1) nivolumab in combination with azacytidine; (2) midostaurin in combination with azacytidine; and (3) decitabine in combination with cytarabine. [3] (Arm 3 was only going to open after Phase 2 accrual to arms 1 and 2 was complete.) Notably, the nivolumab combination arm used a checkpoint inhibitor to activate an immune response, and the midostaurin combination was known to have possible cardiac toxicities. As such, while some patients could be ineligible to these two experimental arms, they would still be recruited if they were eligible to at least one experimental arm and the standard of

care. Excluding all patients with preexisting autoimmune disease or heart risks would have limited the generalizability of study outcomes.

The study team anticipated that most patients would be eligible to all study arms, so allowing varying eligibility was not expected to substantially increase trial sample size but would allow for better generalizability and provide wider access to trial therapies.

In addition, the investigators also wanted to balance randomization on age, performance status and FLT3 mutation status, which are important prognostic variables for patients with AML. [3] A chance imbalance on these variables would complicate the interpretations of the analysis, so the investigators chose to employ stratified randomization.

They opted to use dynamic balancing, as opposed to block randomization, based on the number of possible eligibility combinations, number of stratifying variables and the planned sample size set. The imbalance score calculations for LEAP were identical to the calculations used in the example for dynamic balancing above.

For randomization weights, the study employed telescoping weights of 0.75 and 0.25, if eligible to 2 arms, or 0.75, 0.1875 and 0.0625, if eligible to 3 arms, with the highest weight given to the arm with the smallest imbalance score. This choice of weights was based, in part, on a paper by Brown et al., [10] who assessed different weighting schemes of dynamic balancing. However, their paper only considered trials with two arms and a limited number of weights, and research on optimal weighting in more general settings has not been done.

The trial was closed to accrual after observing an unexpected safety signal in the nivolumab combination arm. [11] The study randomized 78 subjects: 26 patients were randomized to the standard of care, 26 to the midostaurin combination arm and 26 to the nivolumab combination arm. Seven patients were not eligible for the nivolumab combination arm and one was not eligible for the midostaurin combination arm.

We observe that approximately 10% of randomized subjects (8 of 76) were not eligible to all experimental arms. In this example, we see that that differing eligibility did not substantively impact the efficiency of the trial.

1.4 Conclusions

Platform trials offer a framework to efficiently conduct randomized studies within a developing research area. In some platform trials, if experimental arms differ in eligibility, existing methods for stratified randomization cannot be naively employed. This chapter proposes extensions to existing methods that properly account for differing arm eligibility.

While straightforward, the extensions we outline in this project pave the way for more flexible platform trials using stratified randomization. With current methods for stratified randomization, trials must enroll only participants eligible to all experimental arms, limiting accrual and generalizability. The extensions we describe here address this issue directly and obviate such a requirement.

Our proposed methods are not intended to replace important conversations regarding the increased trial complexity by allowing varying experimental arm eligibility. As shown in this chapter, such trials require adjustment to the treatment assignment process, and the complexity also affects other aspects such as trial logistics. While some scenarios may justify these complications, others may benefit from simpler, common eligibility criteria across arms. It is an important role of biostatisticians to scrutinize proposed eligibility criteria and evaluate whether the advantages of differing eligibility outweigh the increased complexity.

We note that it is valid to implement simple randomization to assign subjects to eligible treatment arms, and simple randomization is straightforward to implement. Researchers can employ stratified randomization to increase power [6] or to prevent imbalance by chance on important prognostic variables, but the increased costs in implementation may not always be justified in this complex setting. Again, it is a role of biostatisticians for a given study to gauge whether the increase in computational complexity justifies the benefits of stratified randomization.

This research outlines one modification for varying eligibility each for block randomization and dynamic balancing, but other possible modifications are possible. For block random-

ization, our recommendation dramatically increases the number of strata as the number of treatment arms increases. As such, this recommendation may only be practical with small numbers of stratification factors and/or few eligibility strata, and, while the dynamic balancing algorithm may be more difficult to implement, it does not suffer from this problem. Separately, our recommendation for dynamic balancing is flexible to many choices, but the efficiency may depend strongly on the underlying chosen dynamic balancing scheme.

We also describe the efficiency of platform trials with differing experimental arm eligibility. As the number of participants eligible to all experimental arms increases, the platform trial becomes more efficient: the size of the common standard of care arm decreases to resemble the size in a multi-arm trial with non-varying eligibility across study arms. As such, a trial may be less efficient than expected if the proportion of participants ineligible to arm(s) is higher than anticipated, and this directly impacts the planned sample size and costs for the trial. Researchers can estimate a crude maximum size of the standard of care arm by assuming participants would each be eligible to exactly one experimental arm and refine this estimate with the formula in the appendix based on anticipated participant eligibilities.

Throughout this project, we assume all experimental arms require balance on the same set of prespecified stratification factors. This serves to simplify the presentation, but careful implementation can also allow investigators to employ the proposed extensions with differing subsets of the trial's set of stratification factors across experimental arms. This may be desirable if, for example, one arm has a small sample size that limits the number of allowable stratification factors for that arm. The implementation would require a user to first map each combination of participant arm eligibilities to the joint set of stratification factors needed for the arms of that combination. Then, the user would modify the proposed extensions by using these combinations in place of the full set of stratification factors: in block stratified randomization, use these combinations to form the nested strata within corresponding eligibility strata and, in dynamic balancing, use the combination corresponding to the new participant's eligibility for the balancing calculations.

The implementation of these algorithms to properly address varying arm eligibility can re-

quire non-trivial effort when developing a study. But we believe that the methods we present here will be important in the design of more flexible platform trials that can accommodate differing eligibility criteria.

Chapter 2

RECEUS: RATIO ESTIMATION OF CENSORED UNCURED SUBJECTS, A DIFFERENT APPROACH FOR STUDYING SUFFICIENT FOLLOW-UP IN STUDIES OF LONG-TERM SURVIVORS

2.1 Introduction

2.1.1 Sufficient Follow-Up Time in Cure Models

Researchers have been interested in estimating the fraction of patients cured of cancer for over 70 years. Early studies reported the five-year survival rate as an assessment of cure, but Boag [12] and Berkson and Gage [13] argued against this measure and introduced models that instead analyzed the cure fraction explicitly as the proportion of a cohort not susceptible or "cured" of the event of interest. Since then, cure model literature has grown with many new methods and extensions.

When members of the population are cured, cure models allow researchers to explicitly describe this population heterogeneity. [14] This can facilitate answering questions such as whether subgroups differ in long-term survivorship or how survival differs in those who are not long-term survivors. In health economics, when conducting cost-effectiveness analyses, cure models may improve estimation of mean overall survival (mOS). [15] Calculation of the mOS requires the use of parametric models (unless all subjects within a study have observed events), and Othus et al. (2017) describe how to implement cure models within this framework.

Amico and Van Keilegom [16] describe that a key assumption in all cure models is that there is sufficient follow-up to identify model parameters. To motivate the importance of this sufficient follow-up time, we examine the 3-arm phase 2/3 study, S1117 (clinicialtrials.gov

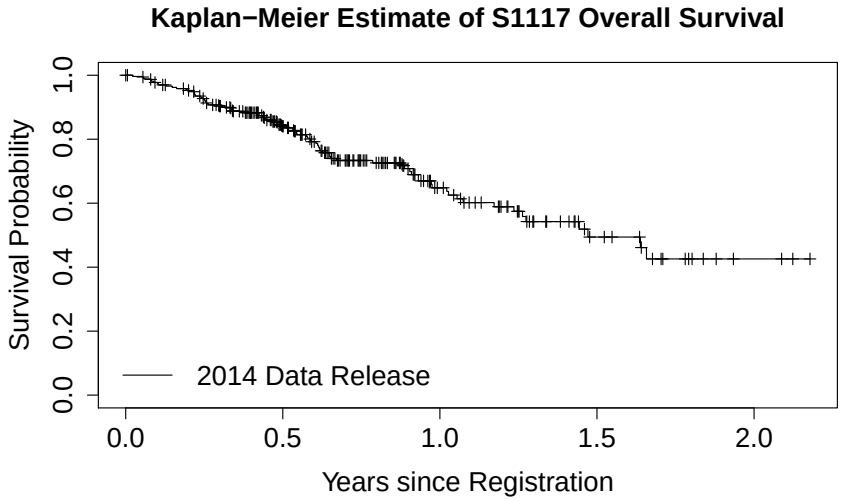


Figure 2.1: Kaplan-Meier estimate for the survival function based on data of trial S1117 ending in 2014.

identifier: NCT01522976 [17]), which investigated combination therapy versus single-agent azacitidine for newly diagnosed myelodysplastic syndrome (MDS) patients. Anderson [18] reported that approximately 40% of patients with MDS may be cured by allogeneic bone marrow transplantation. While the combination arms in S1117 failed to show sufficient efficacy to proceed to phase 3, prior research such as Anderson's paper motivated the clinical leadership to want to evaluate whether some subjects may have been cured of their disease.

Based on the data first released by the data safety monitoring committee, clinical investigators believed that results for overall survival (Figure 2.1) indicated a plateau at the right tail or that a non-zero fraction of subjects may be long-term survivors of the disease. However, there is heavy censoring before the plateau, so it seems there may not be adequate follow-up to fit a cure model.

2.1.2 Existing Methods to Study Sufficient Follow-Up Time

As per Amico and Van Keilegom [16], the mixture cure framework is a popular and well-studied area in the broader cure models literature. This framework considers the population to be a mixture of cured subjects who will not experience the outcome of interest and subjects susceptible to the outcome. As a result, the cumulative distribution function (cdf) of the event times T can be written as $F(t) = (1 - \pi)F_0(t)$, where π is the cure fraction in the population and $F_0(t)$ is the cdf of the event times for uncured subjects. The existing methods for studying sufficient follow-up were developed within this framework.

Maller and Zhou [19] were among the first to study sufficient follow-up time, focusing on the supports of the censoring distribution, with cdf $G(t)$, and the event time distribution for uncured subjects. In particular, they examined $\tau_{F_0} = \min_t\{t : F_0(t) = 1\}$, the earliest time that the event time cdf for uncured subjects reaches 1, and τ_G, the analogous quantity for the censoring distribution. They identified a criterion $\tau_{F_0} < \tau_G$ as a necessary condition for valid assessment of a cure fraction. In words, this states that the longest event times for uncured subjects cannot be unobservable due to censoring.

They proposed a hypothesis test to quantify this approach, called the $\hat{\alpha}_n$ test. The test quantifies the difference between the largest event time (an estimate of τ_{F_0}) and the largest censored time (an estimate of τ_G).

Two studies remarked on poor control of type-1 error by the $\hat{\alpha}_n$ test and each developed a different method to address the issue: one by Maller and Zhou [20] themselves, the q_n test, and one by Shen [21], the $\tilde{\alpha}_n$ test. Both modify the test statistic of $\hat{\alpha}_n$ to improve upon it. The q_n test requires explicit specification of the true cure fraction and censoring distribution in order to calculate critical points, making it difficult to use in practice. The $\tilde{\alpha}_n$ test does not impose such a requirement.

All three of these tests focus on assessing sufficient follow-up time with a hypothesis test of $\tau_{F_0} < \tau_G$. However, in practice, many studies will necessarily have finite follow-up due to cost, so they will rarely have long enough follow-up to allow for the tail of the uncured

subjects' event time distribution to be observed before being censored. This means the premise of testing $\tau_{F_0} < \tau_G$ may itself be unrealistic in many research settings.

Fortunately, Yu et al. [22] describe how cure fraction and median survival estimates improve as follow-up time increases without necessarily reaching the point of all failures being observed. Simulations summarized in the appendix expand on this to show that in Weibull mixture cure models, estimation with low mean-squared error and nominal confidence interval coverage can be achieved with 1% uncured remaining or longer follow-up, depending on the setting. Taken together, this motivates a different approach to quantifying sufficient follow-up time, which we describe in Section 2.2.1.

2.1.3 Evaluating Sufficient Follow-Up in S1117

Returning to the example of S1117, the existing methods for quantifying sufficient follow-up provide mixed evidence for fitting a cure model to the available data ($\hat{\alpha}_n < 0.001$ and $\tilde{\alpha}_n = 0.134$). Thus, these methods do not guide an analyst to a clear conclusion regarding the appropriateness of a cure model. They also do not support the intuition that heavy censoring prior to the plateau may violate the assumption of sufficient follow-up.

This paper aims to provide a different approach for evaluating the appropriateness of a cure model. We describe our proposed statistic in Section 2.2.1. Section 2.2.2 defines the possible classification errors in claiming sufficient follow-up with this approach. We derive asymptotic properties for the proposed statistic estimated by maximum likelihood in Section 2.2.3 and then suggest approaches for addressing sensitivity to model misspecification in Section 2.2.4. We verify finite sample performance in Section 2.3. Then we conclude with two data examples in Section 2.4 and a discussion, Section 2.5.

2.2 A Novel Method for Quantifying Sufficient Follow-Up Time

2.2.1 Targeting the Proportion of Uncured Remaining

Throughout the rest of the paper, in light of the mixture cure framework, we consider the true event times T to have survival function

$$S(t;\pi,\theta) = \pi + (1-\pi)S_{uc}(t;\theta) \tag{2.1}$$

with π the unknown cure fraction and a survival function $S_{uc}(t)$ belonging to common parametric families with unknown parameter vector $\theta \in \Theta \subseteq \Re^p$, $p < \infty$ (we may suppress the dependence of S and S_{uc} on π and θ in the notation for clarity). We also consider independent, uniformly sampled accrual times $A \sim \text{Unif}(0, a)$ (a known) and a known administrative censoring time τ (with $a < \tau$). These are intended to parallel the real-world context of clinical trials that often accrue until a prespecified time a and analyze at a prespecified time τ.

Based on Yu et al. [22] and summaries in the appendix, we observe that estimation and inference improve as the proportion of uncured remaining decreases but may also depend on the underlying cure fraction. This motivates that the proportion of uncured remaining can be used to quantify the sufficiency of follow-up. Based on this, we propose that the quantity

$$r = \frac{S_{uc}(\tau)}{S(\tau)} = \frac{S_{uc}(\tau)}{\pi + (1-\pi)S_{uc}(\tau)} \tag{2.2}$$

can be useful to quantify sufficient follow-up time in a mixture cure setting. We can then use the following estimator (where we replace the population quantities with suitable estimates):

$$\hat{r}_n = \frac{\hat{S}_{n,uc}(\tau)}{\hat{S}_n(\tau)} = \frac{\hat{S}_{n,uc}(\tau)}{\hat{\pi}_n + (1-\hat{\pi}_n)\hat{S}_{n,uc}(\tau)}. \tag{2.3}$$

Heuristically, this ratio statistic quantifies the estimated proportion of uncured remaining at the administrative censoring time and standardizes it to the estimated overall proportion remaining and censored at τ. This standardization incorporates the cure fraction and censoring pattern in the data (as opposed to simply targeting the proportion uncured alone).

We propose to use this statistic in a method we call RECeUS (Ratio Estimation of Censored Uncured Subjects, pronounced "ree-sus") to assess sufficient follow-up as follows:

1. Estimate the quantities $S(\tau)$, $S_{uc}(\tau)$ and π.

2. If $\hat{\pi}_n$ is very small, then either a cure model is likely not a valid model or follow-up is likely insufficient for valid results. We suggest using $\hat{\pi}_n > 0.025$ (or a higher threshold) as a screening procedure.

3. If $\hat{\pi}_n$ is away from 0, then estimate r - small values of $\hat{r}_n$ represent sufficiency of follow-up time. We suggest using $\hat{r}_n < 0.05$ to reasonably conclude sufficient follow-up time. We evaluate the use of this threshold in Section 2.3.

2.2.2 Defining Possible Errors when Concluding Sufficient Follow-Up

In addition to motivating a statistic for quantifying the sufficiency of follow-up, the results of Yu et al. [22] and the appendix also motivate reframing the possible errors when concluding sufficient follow-up. In previous literature, errors have been defined in a manner parallel to hypothesis testing with type-1 and -2 errors. But because we now allow for the claim of sufficient follow-up time with a non-zero (but small) proportion of uncured subjects, we must redefine the errors involved.

We incorrectly claim to not have sufficient follow-up when the method suggests follow-up is not sufficient but a cure fraction exists and few uncured subjects remain at the end of the study. On the other hand, we incorrectly claim sufficient follow-up time for employing a cure model when a method suggests follow-up is sufficient but (a) no cure fraction exists to generate the data, or (b) a cure fraction exists but too high a fraction of uncured subjects remains at the end of the study to reliably identify it.

While these errors do resemble type-1 and type-2 errors in hypothesis testing, they rely on the latent status of subjects at the end of the study in addition to data generating parameters. As such, we distinguish them by adopting the following notation. First, define the random

variable $G = \mathbf{1}\{\text{Declare cure model is appropriate}\}$. Next, let $\pi \in [0,1]$ be the underlying cure fraction, and $u = S_{uc}(\tau) \in [0,1]$ be the fraction of uncured subjects remaining (have not experienced the event of interest) at the end of the study (note that uncured subjects would not be identifiable in practice as cure is considered a latent status). Then define

$$\gamma_\pi(u) \equiv Pr(G = 1; u, \pi). \tag{2.4}$$

A useful method for distinguishing when a cure model is and is not appropriate should thus have the following properties:

1. $\gamma_\pi(u) \to 1$ when $\pi > 0$ and $u \to 0$; and

2. $\gamma_\pi(u) \to 0$ when either

   (a) $\pi = 0$ with any value $u \in [0,1]$, or

   (b) $\pi > 0$ but $u \to 1$.

The first property states that the method increases its probability to 1 in reaching the correct conclusion that the cure model is appropriate when a cure fraction exists and the proportion of uncured subjects remaining at the end of the study decreases to 0.

The second property states that the method decreases to 0 in its probability in reaching the incorrect conclusion that the cure model is appropriate when either (a) no cure fraction exists to generate the data, or (b) a cure fraction exists but the fraction of uncured subjects remaining at the end of the study increases to 1.

2.2.3 Asymptotic Properties when Estimating via Maximum Likelihood

For the estimator $\hat{r}_n$, we require estimation of $S(\tau)$. In this paper, we present asymptotic properties and simulation results when we fit a parametric mixture cure model via maximum likelihood. In general, one may also use more flexible methods to estimate $S(\tau)$.

We accrue each subject $i = 1, \ldots, n$ at an accrual time $A_i \sim \text{Unif}(0, a)$, and we follow these subjects until the administrative censoring time τ. As such, we record an indicator for censoring (Δ_i) along with an observed time as the minimum of either their elapsed study time (T_i) or the time from accrual until administrative censoring time ($\tau - A_i$). In other words, we observe a sample of n independent and identically distributed pairs $\{(Y_i, \Delta_i) : i = 1, \ldots, n\}$ with $Y_i = \min(T_i, \tau - A_i)$ and $\Delta_i = \mathbf{1}\{T_i \leq \tau - A_i\}$. Define $G(t; \pi, \theta) = Pr(T \leq t)$ with derivative $g(t; \pi, \theta)$ based on the parametric model for the event times.

Because of the assumed independence between the event times and accrual, the log-likelihood is proportional to

$$l_n(\pi, \theta; Y, \Delta) \propto \sum_{i=1}^{n} \Delta_i \log g(Y_i; \pi, \theta) + (1 - \Delta_i) \log(1 - G(Y_i; \pi, \theta)).$$

We first estimate the model parameters

$$\eta_n = (\pi_n, \theta_n) = \underset{\pi \in [0,1] \times \theta \in \Theta}{\arg\max}\ l_n(\pi, \theta; Y, \Delta).$$

Now, define

$$\ell(\eta; Y, \Delta) = \Delta \log g(Y; \pi, \theta) + (1 - \Delta) \log(1 - G(Y; \pi, \theta))$$

with a second derivative ($(p+1) \times (p+1)$ Hessian matrix) $\ddot{\ell} = \left[\frac{\partial^2}{\partial \eta_i \partial \eta_j} \ell(\eta; Y, \Delta)\right]$ with $i, j = 1, \ldots, p+1$.

Then we have, with S_C and f_C the survival and probability density functions of $C = \tau - A$,

$$\mathbb{E}[\ddot{\ell}(\eta; Y, \Delta)] = \int_0^\tau \ddot{\ell}(\eta; Y, \Delta = 1) S_C(y) g(y)\, dy + \int_{\tau - a}^{\tau} \ddot{\ell}(\eta; Y, \Delta = 0)(1 - G(y)) f_C(y)\, dy. \quad (2.5)$$

Theorems 1-2 describe the consistency and asymptotic normality of the model parameters and of $\hat{r}_n$ estimated by maximum likelihood.

Theorem 1 (Consistency) *Under regularity conditions R1-R5, listed in the Appendix B.2, we have, by van der Vaart, [23]*

$$\eta_n \to_p \eta_0 = \underset{\pi \in [0,1] \times \theta \in \Theta}{\arg\max}\ \mathbb{E}\left[\ell(\eta; Y, \Delta)\right].$$

Then, by R6 and the continuous mapping theorem, [23]

$$\hat{r}_n = \frac{\hat{S}_{n,uc}(\tau)}{\hat{S}_n(\tau)} \to_p \frac{S_{uc}(\tau)}{S(\tau)} = r.$$

Theorem 2 (Asymptotic Normality) *Let $\mathcal{I} \equiv \mathcal{I}(\eta) = -\mathbb{E}[\ddot{\ell}(\eta; Y, \Delta)]$. Then, by regularity conditions R1-R5 in the Appendix B.2, we have by van der Vaart [23]*

$$\sqrt{n}(\eta_n - \eta_0) \to_d N(0, \mathcal{I}^{-1}).$$

Next, assuming R6 and $r(\eta)$ is differentiable at η_0, denote the $(p+1) \times 1$ gradient vector $D \equiv D(\eta_0) = \left[\frac{\partial}{\partial \eta_i} r(\eta)|_{\eta=\eta_0}\right]$ with $i = 1, \ldots, p+1$. By the Delta method, [23]

$$\sqrt{n}(\hat{r}_n - r) \to_d N(0, D\mathcal{I}^{-1}D^T). \tag{2.6}$$

2.2.4 Addressing Sensitivity to Model Misspecification: RECeUS-AIC

Parametric models can be sensitive to model misspecification. As such, it is common practice in this area to perform model selection when employing parametric models. Researchers can visually inspect various model fits by comparing the fitted survival curve against the Kaplan-Meier estimate, and it has also been suggested to study the profile likelihood function of the cure fraction parameter. [22]

We propose addressing sensitivity to model misspecification in RECeUS by employing AIC. [24] First, select among a class of models with AIC, then use the RECeUS method with the best-fitting model. In this paper, we select from a class of models that includes the non-cure and mixture-cure versions of the Exponential, Weibull, Gamma and Log-Logistic models. Choosing any non-cure model immediately leads to a conclusion that a cure model is not appropriate. This class reflects an array of commonly-used models that can capture a wide variety of behavior in the data.

2.3 Studying Finite Sample Properties via Simulation

In this section, we study the properties of the RECeUS procedure in sample sizes of 100, 250, 500 and 1000 to evaluate the procedure in real-world sample sizes.

We generate data from Gamma, Weibull and Log-Logistic mixture cure distributions with varying parameters and cure fractions (ranging from 0 to 0.8). We additionally vary the administrative censoring time (τ) to represent (approximately) the 75th, 90th, 95th, 99th and 99.9th percentiles of the uncured distribution. Using the notation of Section 2.2.2, this corresponds to using $\pi \in [0, 0.8]$ and $u \in \{0.25, 0.1, 0.05, 0.01, 0.001\}$.

To better represent clinical trial analyses occurring at prespecified times, follow-up times corresponding to the percentiles are rounded to the nearest quarter to generate the data. For example, for the Weibull(2,1) distribution, we use administrative censoring times of 1.25, 1.5, 1.75, 2.25 and 2.75.

For a given distribution, we fix the accrual end at $a = \tau_{0.75}/2$ (half of the 75th percentile of the given uncured distribution) and vary τ as we indicate above. With a fixed a, increasing τ indicates longer study follow-up.

We summarize 5000 simulations in all settings.

For the RECeUS procedure, we evaluate a threshold $\hat{r}_n < 0.05$ and screening with $\hat{\pi}_n > 0.025$. If either criterion is not satisfied, the procedure indicates the data are not adequate for fitting a cure model - due to insufficient follow-up and/or very small or no cure fraction. Larger thresholds for $\hat{r}_n$ mean more frequently concluding a cure model is appropriate - both when this is a correct and an incorrect decision. The opposite applies for smaller thresholds. The $\hat{\alpha}_n$ and $\tilde{\alpha}_n$ tests conclude sufficient follow-up based on both a threshold of 0.05 and the last observed time being a censoring time.

Results in the appendix indicate settings with non-zero cure fraction ($\pi > 0$) and 0.1% uncured remaining ($u = 0.001$) have small mean-squared error and nominal coverage of confidence intervals for cure model parameters, so a cure model seems appropriate. As such, we desire $\gamma_\pi(0.001)$ close to 1. Also, recall that we desire $\gamma_0(u)$ close to 0, a low probability of incorrectly claiming a cure model is appropriate when a cure fraction does not exist to generate the data.

We provide the results from the Weibull(2,1) mixture cure distribution, but patterns are similar across distributions. In Figure 2.2, we present $\gamma_\pi(0.001)$, with varying $\pi \in [0.1, 0.8]$

and $\gamma_0(u)$, with varying $u \in \{0.25, 0.1, 0.05, 0.01, 0.001\}$, for each procedure (RECeUS-AIC, Maller and Zhou's $\hat{\alpha}_n$ and Shen's $\tilde{\alpha}_n$). In Appendix B.3, we also tabulate the results.

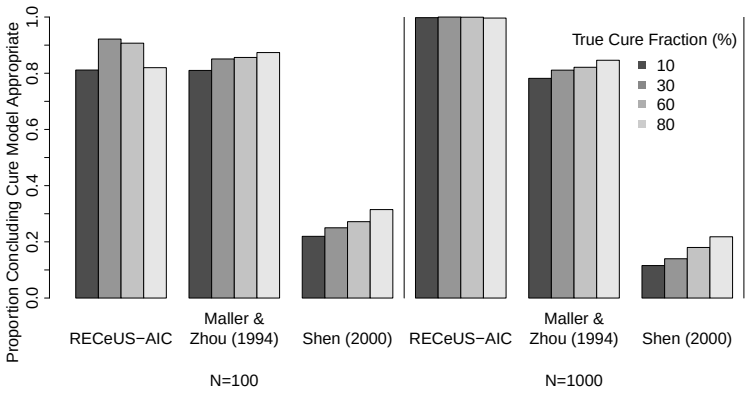
The rate at which the RECeUS-AIC procedure incorrectly claims a cure model is appropriate when $\pi = 0$ decreases with sample size at all values of u. It can have $\gamma_0(u)$ as high as 12.2% in small samples with $n = 100$ and decreases to 1.2% or less by $n = 1000$ for all u.

The Maller and Zhou (1994) $\hat{\alpha}_n$ test has uniformly high $\gamma_0(u)$ in a range from 13.3% to 54.3%. The rates do not decrease with additional sample size as with the RECeUS-AIC procedure. In this Weibull(2,1) example, we find the Shen (2000) $\tilde{\alpha}_n$ statistic has a maximum $\gamma_0(u) = 4.4\%$, and the rates often decrease with additional sample size, but this does not hold for all u. In other distributional settings this also does not always hold.

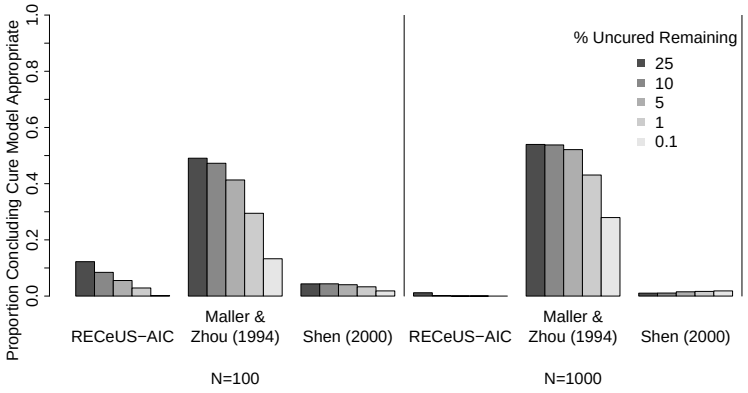
As sample size increases, the RECeUS-AIC procedure more frequently reaches the correct conclusion that a cure model is appropriate with a non-zero cure fraction and 0.1% uncured remaining ($u = 0.001$). The RECeUS-AIC procedure has $\gamma_\pi(0.001) > 80\%$ at $n = 100$ and reaches $\gamma_\pi(0.001) > 99\%$ by $n = 1000$ for all $\pi > 0$.

The $\hat{\alpha}_n$ test has a maximum $\gamma_\pi(0.001) = 0.874$ in settings with a non-zero cure fraction. However, the rates $\gamma_\pi(0.001)$ appear to decrease with sample size and are generally lower than in the analogous settings of the RECeUS-AIC procedure. On the other hand, the $\tilde{\alpha}_n$ test has lower $\gamma_\pi(0.001)$, with a maximum of 0.315, and these rates also appear to decrease with sample size. In these simulations, $\tau_{F_0} > \tau_G$, so any conclusion of sufficient follow-up is technically incorrect based on the construction of the $\hat{\alpha}_n$ and $\tilde{\alpha}_n$ statistics. This explains the decrease with sample size even with a non-zero cure fraction and longer follow-up.

While we claim a cure model is appropriate in settings with non-zero cure fractions and 0.1% uncured remaining, interpreting the results for 25% down to 1% uncured remaining ($u \in [0.01, 0.25]$) is less straightforward. A cure fraction does exist, but these follow-up times may not represent adequate follow-up time for valid cure model results. In other words, we may not know when the value u becomes sufficiently small to desire $\gamma_\pi(u)$ to move from being close to 0 to being close to 1. We describe the behavior of RECeUS-AIC under these settings below.



(a) The proportion of 5000 simulations that conclude a cure model is appropriate under settings with a non-zero cure fraction and 0.1% uncured remaining, $\gamma_\pi(0.001)$



(b) The proportion of 5000 simulations that conclude a cure model is appropriate under settings with no cure fraction ($\pi = 0$) and varying percent uncured remaining, $\gamma_0(u)$

Figure 2.2: Evalulation of RECeUS-AIC, the Maller and Zhou (1994) $\hat{\alpha}_n$ statistic and the Shen (2000) $\tilde{\alpha}_n$ statistic when data are generated with a cure fraction (top) or without a cure fraction (bottom).

With cure fractions of 30% or larger, the probability of concluding a cure model is appropriate increases with sample size when 1% uncured remain ($u = 0.01$). However, with a cure fraction of 10%, this does not hold: this reflects that longer follow-up may be needed to identify smaller cure fractions.

Across all settings with non-zero cure fraction, $\gamma_\pi(u)$ decreases to 0 with sample size when 5% or more of the uncured remain ($u \geq 0.05$). This follows from the use of 0.05 as the threshold for $\hat{r}_n$, which implicitly expresses that a cure model is appropriate with at most 5% uncured remaining and even less with smaller cure fractions.

We conclude that RECeUS has high $\gamma_\pi(0.001)$ with $\pi > 0$, when we desire high rates of concluding a cure model is appropriate, and it has low $\gamma_0(u)$ for all u across many settings. The $\hat{\alpha}_n$ test possesses both high $\gamma_0(u)$ and $\gamma_\pi(0.001)$, while $\tilde{\alpha}_n$ has both low $\gamma_0(u)$ and $\gamma_\pi(0.001)$. At values of $u > 0.001$ with $\pi > 0$, RECeUS has a higher probability of concluding sufficient follow-up earlier with a larger cure fraction.

2.4 Revisiting the Motivating Data Example

SWOG Cancer Research Network, a US NCI-funded clinical trials cooperative group, provides a rare setting for studying cure models because SWOG continues to follow patients after the primary analysis. Therefore, we can directly evaluate whether tests of sufficient follow-up are concordant with results after additional years of follow-up with data examples.

The data in Figure 2.1 represent trial data from 2014. In this case, $\hat{\alpha}_n < 0.001$ calculated on the 2014 data claims sufficient follow-up, but the $\tilde{\alpha}_n = 0.134$ does not. Alternatively, we can employ RECeUS-AIC. For the class given in Section 2.2.4, AIC selects the non-cure Log-Logistic model. As such, by the RECeUS-AIC method, we immediately conclude a cure model is not appropriate for these data.

SWOG has additional follow-up data for trial S1117. The data for this trial, in Figure 2.3, represent a scenario with 3 additional years of follow-up, and the data do not seem sufficient for cure model analysis: the tail of the survival probability estimate continues to decrease after 2014. In the figure, we provide the fitted curve based on the 2014 data with the

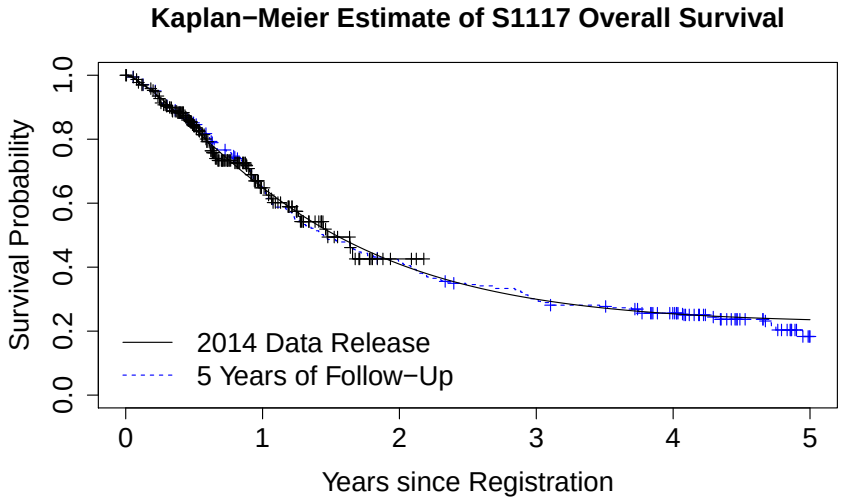


Figure 2.3: Kaplan-Meier estimates for the survival function based on data through trial S1117 end in 2014 (black) and extended follow-up with five years of total follow-up (blue). The best-fitting mixture-cure model (mixture-cure Gamma model) based on 2014 data is overlaid in black.

best-fitting mixture-cure model, the mixture-cure Gamma model. We see the curve seems to fit much of the data closely, but it differs at the right tail. As such, summaries that require accurate estimation of the right tail would be biased.

We see the RECeUS-AIC and $\tilde{\alpha}_n$ methods agree with that conclusion of insufficient follow-up in 2014, while $\hat{\alpha}_n$ claims follow-up was sufficient, which seems inappropriate. This may follow from comments by Maller and Zhou [20] and Shen [21] that the $\hat{\alpha}_n$ test incorrectly concludes a cure model is appropriate too frequently.

As a second example, we examine trial S0106 (clinicialtrials.gov identifier: NCT00085709 [25]). The trial randomized acute myeloid leukemia (AML) patients to either standard therapy or the combination of standard therapy and the drug mylotarg. As with S1117, the promises of allogeneic transplant inducing cure in AML patients in the past prompted investigators to explore cure modeling in S0106. The trial released data in 2011, and $\hat{\alpha}_n =$

0.002 based on 2011 data claims sufficient follow-up, while $\tilde{\alpha}_n = 0.368$ does not. For the RECeUS-AIC method, we select a Weibull mixture-cure model, and we calculate $\hat{\pi}_n = 0.426 > 0.025$ and then $\hat{r}_n = 0.032$. With $\hat{r}_n < 0.05$, we have evidence to conclude sufficient follow-up.

For this trial, SWOG also has additional follow-up in Figure 2.4. We see that even with extended follow-up, a plateau exists at a similar level in the tail of the survival function estimate. This supports the conclusion that follow-up at the trial's data release in 2011 does seem sufficient for cure model analysis. Further, we see that the fitted curve for the best-fitting mixture-cure model, the Weibull mixture-cure model, seems to fit the extended follow-up well.

In this situation, the extended follow-up seems to confirm the conclusions reached by the $\hat{\alpha}_n$ and RECeUS methods calculated based on 2011 data and not the conclusion of the $\tilde{\alpha}_n$ approach. This occurs by construction of the $\tilde{\alpha}_n$ test: it often requires a larger relative difference between the largest observed time and the largest event time than the $\hat{\alpha}_n$ test.

These data examples illustrate other advantages of using the RECeUS approach in assessing sufficient follow-up over other methods: the approach agrees with the results after additional years of follow-up in both examples, while the other methods provide contradictory results.

2.5 Discussion

Othus et al. [26] conclude that cure model results from data with insufficient follow-up time can be biased, so researchers should evaluate the appropriateness of a cure model before disseminating results. Existing tests for assessing sufficient follow-up implicitly conclude that a cure model is inappropriate if any uncured subjects remain at the end of the study. As such, these tests are not calibrated for the real-world setting in which some proportion of uncured subjects do exist at the end of the study: the $\hat{\alpha}_n$ test frequently concludes sufficient follow-up for a cure model when no cure exists, while the $\tilde{\alpha}_n$ test rarely concludes sufficient follow-up even with a non-zero cure fraction and long follow-up.

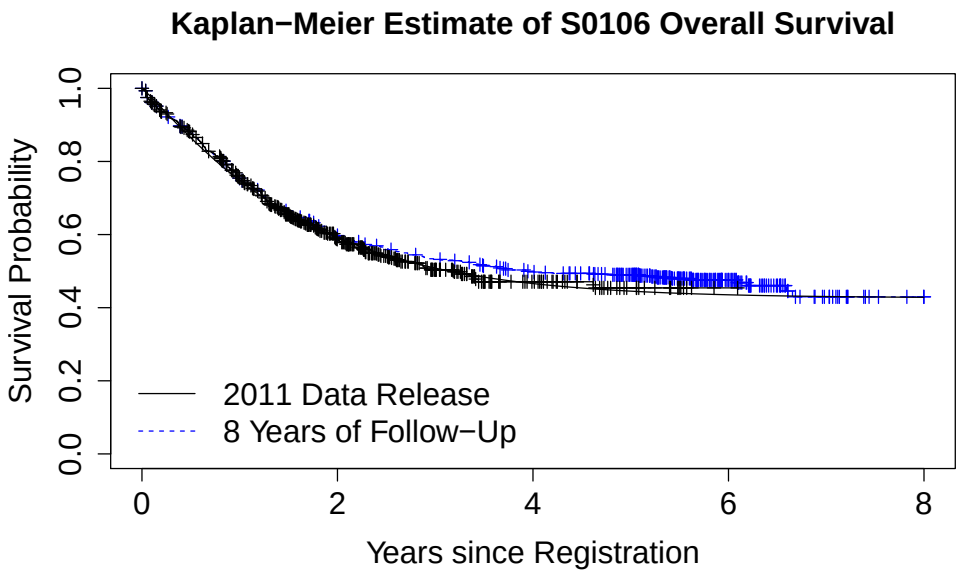


Figure 2.4: Kaplan-Meier estimates for the survival function based on data through trial S0106 end in 2014 (black) and extended follow-up with data from 2018 (blue). The best-fitting mixture-cure model (mixture-cure Weibull model) based on 2011 data is overlaid in black.

From Yu et al. [22] and simulations summarized in the appendix, evidence exists to motivate a different approach. We suggest quantifying sufficient follow-up by studying the proportion of uncured subjects remaining in the study (or a standardized version, r). We demonstrate that, when implemented via maximum likelihood estimation, the estimator $\hat{r}_n$ is consistent and asymptotically normal under standard regularity conditions. In finite samples, the RECeUS procedure frequently concludes sufficient follow-up when a cure fraction exists and the follow-up corresponds to 0.1% uncured remaining or less. The follow-up time needed to identify and reliably estimate a cure fraction decreases as the cure fraction increases away from zero. When the cure fraction is zero, the method also has a low rate for concluding sufficient follow-up.

In this paper, we estimate the ratio r via maximum likelihood estimation following the standard practice in using parametric cure models in applied fields such as health economics, but we do not study the properties in cases where $S(t)$ is estimated in a more flexible manner. This is an important future direction of the work. As the RECeUS method readily permits flexible model specification for $S(t)$, researchers may choose to employ more flexible methods instead or in addition to the suggested model selection procedure.

We believe that this novel method offers a promising new approach to sufficient follow-up time in cure models, and we hope that the interpretability and flexible implementation may lead to the widespread use of RECeUS in scientific practice.

Chapter 3

A FRAMEWORK FOR SEQUENTIAL MONITORING OF ONE N-OF-1 TRIAL AND COMBINING RESULTS ACROSS A SERIES OF SEQUENTIALLY-MONITORED N-OF-1 TRIALS

3.1 Introduction

Randomized controlled trials (RCTs) are the gold standard for evidence in biomedical science in settings where designed experiments are ethical and tractable. However, authors have described several limitations to this approach: (1) standard RCTs study a group-level treatment effect that may fail to describe heterogenous treatment effects across patients; (2) they may be logistically intractable in conditions with a small population; and (3) the long timeline from developing the study to reporting results could be a detriment in the early stages of rapidly-developing settings such as emerging pandemics. [27, 28, 29, 30]

One possible approach to address these limitations is to perform an N-of-1 trial. In an N-of-1 trial, an investigator randomly assigns the subject to a sequence of crossovers between two or more treatment options. These trials are frequently organized with a block structure such that each treatment block contains all treatment options a fixed number of times. [31] For example, a study comparing treatment A against treatment B with 4 periods per block would have repeated blocks with treatment assignments of AABB, ABBA, ABAB, BAAB, BABA or BBAA. One possible sequence for a trial with five blocks is shown in Figure 3.1.

N-of-1 trials provide personalized evidence for the subject and typically have a smaller start-up cost compared to standard RCTs. Additionally, researchers have studied how to combine results from a series of N-of-1 trials to assess the group-level treatment effect. [32] These trials are typically most applicable in settings with a chronic condition and to study treatment options with a fast uptake and limited sustained effect after discontinuation. N-of-

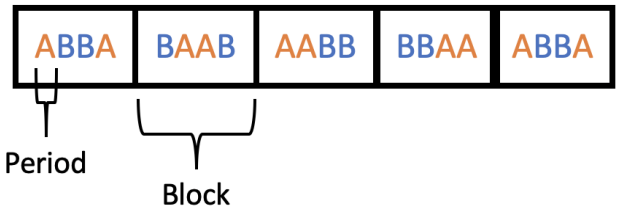


Figure 3.1: Possible sequence of treatment assignments for an N-of-1 trial with two treatment options (A or B), four periods per block and five blocks.

1 trials typically do not evaluate novel interventions and instead consider well-established or relatively safe options. [33] A draft FDA guidance was developed in January 2021 for investigational N-of-1 trials studying antisense oligonucleotide products for rare genetic diseases, [34] but, for this project, we focus on a framework for N-of-1 trials assessing the effectiveness of well-established or relatively safe treatment options.

One illustrative example can be seen in a report from March et al. (1994) that describes 25 N-of-1 trials assessing the effect of paracetamol or a non-steroidal anti-inflammatory drug on patient-reported pain (measured by visual analog score) in patients with osteoarthritis. They were motivated by a lack of consensus from standard RCTs and an understanding that large variability in treatment effect may preclude extrapolating previous results to individuals. Each constituent trial within the series of 25 aimed to choose an appropriate individualized treatment, and the series as a whole aimed to evaluate treatment options in the patient population. [35]

In standard RCTs, discontinuation of the study for one subject does not generally cause the termination of the study as a whole. However, in N-of-1 trials, if an investigator or subject chooses to discontinue their involvement in the study, then the study ends. If this choice occurs after looking at unblinded data or after considering perceived treatment differences, then the study results could be biased. This can be an even larger issue for trials that aim to incorporate blinding in their design - and Punja et al. found that the majority of N-of-1 trials do incorporate blinding. [33] This scenario may occur, for example, if the clinician

and patient assess the patient's outcome at every visit and choose to stop upon observing a single dramatic response. Such an approach inflates type-1 error.

Sequential monitoring methods - methods for validly analyzing accruing data - have been well-studied in standard RCTs, but, to our knowledge, they have not yet been implemented in the N-of-1 trial setting. [36, 37] In this article, we describe the statistical challenges for sequential monitoring in this setting, we develop a framework for sequentially monitoring one N-of-1 trial and jointly analyzing a series of sequentially monitored N-of-1 trials, and we evaluate this framework via simulation. Because sequential monitoring has been most commonly employed within a frequentist paradigm, we focus on frequentist approaches in this work.

3.2 Methods

3.2.1 A Strategy for Sequential Monitoring with N-of-1 Trials

Existing approaches for sequential monitoring cannot naively be applied to the N-of-1 setting because they rely on an "independent increments" structure to the data: contributions to the accumulating test statistic in the sequential trial should be independent. [38] It is important to confirm this holds for N-of-1 trials in which we study only one subject. Additionally, even under an independent increment structure, a concern for studies with correlated data is that the information may not monotonically increase, which also affects properties of existing sequential monitoring methods. [39]

One strategy to employ the existing sequential monitoring boundaries is to identify independent units that satisfy the independent increments assumption. We use this strategy to propose how to implement sequential monitoring in Section 3.2.2 by utilizing the block structure of N-of-1 trials. This relies on two assumptions of crossover trials: (1) no carryover effect: the outcome cannot depend on the previous time period's treatment effect and (2) no outcome drift: the underlying process generating the outcome must not be changing over time outside of the effect of the treatments under study. Design and analysis methods exist

to mitigate the impact of deviation from these assumptions. [31]

For the remainder of this article, we assume the above two conditions hold and that there are no missing data. We consider N-of-1 trials that evaluate exactly two treatment options with a continuous outcome (for a series of trials, we assume all constituent trials measure the same outcome) for a planned number of treatment blocks. The number of treatment periods per block, J, is the same for all blocks. For the ease of presentation, we also assume that treatment blocks are "balanced" with equal numbers of both treatment options within each block. For other aspects of trial conduct and integrity, such as blinding, we refer readers to texts such as Friedman et al., as the considerations are common to all RCTs. [40]

3.2.2 Sequential Monitoring for One N-of-1 Trial

The goal of N-of-1 trials is to decide which treatment option works best for the patient under study, and monitoring accruing data could allow researchers to make the decision sooner than the planned end of the trial. Sequential monitoring methods facilitate valid early stopping for RCTs, but they have not been implemented in N-of-1 trials. While some articles and guidances have stated that sequential stopping rules could be used in N-of-1 trials, [31, 41, 42, 43] we did not find details for implementing these rules, and others have stated the need for researching sequential stopping rules for this setting. [36, 37]

In the current state of N-of-1 literature, no uniformly adopted approach exists for analyzing the data, [33, 44] but authors have suggested using a mixed-effects model as a standard approach. [45] As such, we propose monitoring the Wald test statistic from a linear mixed-effects model. We specify the model with a random intercept term for each treatment block, a fixed intercept and a coefficient for the treatment variable. The Wald test statistic evaluates if the coefficient for the treatment variable is equal to 0.

This model can be written as

$$\mathbb{E}[Y_{ij}|\gamma_i, \alpha, \beta] = \gamma_i + \alpha + Z_{ij}\theta,$$

With Y_{ij} and Z_{ij} the outcome measurement and treatment variable, respectively, at the jth period ($j = 1, \ldots, J$) of the ith treatment block ($i = 1, \ldots, B$ planned treatment blocks), γ_i the random intercept of the ith block, α the fixed intercept and θ the coefficient for the treatment variable. We assume the γ_i are independently drawn from the mean-zero Normal distribution with variance g^2. The use of random intercepts permits some correlation among treatment periods of the same block, so this model may accommodate some deviation from the assumptions described in Section 3.2.1.

Without loss of generality, suppose $Z = 0$ corresponds to treatment A and $Z = 1$ corresponds to treatment B. By this specification, the estimate of the coefficient for the treatment variable, $\hat{\theta}$, is an estimate of the average treatment effect (ATE) - comparing the change from A to B and averaging across treatment blocks - specific to the subject of the N-of-1 trial. If there are L planned looks at the data, we calculate this estimate after the prespecified blocks $b_1, \ldots, b_L$ ($2 \leq b_1 < \ldots < b_L = B$). In practice, some additional measurements may be available between the time when an estimate is computed and the decision of trial stopping, but we assume these measurements are not considered for the trial stopping decision.

In the Appendix, Section C.1, we verify that this approach satisfies monotonic information growth (the independent increments condition is satisfied by the treatment blocks as independent units and use of the linear mixed-effects model [46]). In particular, we demonstrate that the information for the Wald test statistic increases linearly with the number of treatment blocks when using balanced treatment blocks. Taken together, we can validly employ classical sequential monitoring methods.

In Section 3.3, we evaluate different choices of sequential monitoring boundaries when monitoring the Wald test statistic from a linear mixed-effects model.

3.2.3 Jointly Analyzing a Series of Sequentially Monitored N-of-1 Trials

We assume that, when jointly analyzing a series of trials, the constituent trials have been completed and only the summary point estimate and its estimated standard error are avail-

able. Additionally, for this project, we assume that the constituent trials had the same boundary shape and planned number of looks.

Previous researchers have described how point estimates from a sequentially monitored trial are biased. [47] The bias depends on the monitoring shape and timing for each trial, [48] and this impacts the properties of a combined point estimate from a series of sequentially monitored trials. In addition, early stopping reduces the statistical information for the point estimate compared to stopping at the planned end.

Several options for bias-adjusted point estimators have been proposed in the literature, which we list in Section 3.2.4. To jointly analyze a series of sequentially monitored N-of-1 trials, we propose adjusting the bias of each constituent point estimate and combining the adjusted estimates with an inverse-variance weighted linear mixed-effects model (where each variance is the square of the constituent estimated standard error). [49] We specify the model with a random intercept term for each constituent trial and a fixed intercept. This approach accounts for the impact of early stopping on bias and information within the constituent trials.

The model can be written as

$$\mathbb{E}[T_c|\omega_l, \theta^*] = \omega_c + \theta^*,$$

With T_c the bias-adjusted point estimate of the cth study ($c = 1, \ldots, C$ total constituent trials in the series), ω_c the random intercept for the cth constituent trial and θ^* the fixed intercept. We assume the ω_c are independently drawn from the mean-zero Normal distribution with variance g_S^2.

The estimate of the fixed intercept, $\widehat{\theta^*}$, represents the estimated average - across subjects - of the ATEs. An alternative meta-analysis approach for combining estimates would be to employ a fixed-effects model, but authors have pointed out that such a model would assume that a single common ATE would exist across all members of the study population. [50] The mixed-effects meta-analysis approach, instead, supposes that each subject could have a different ATE and it targets the average across these ATEs. This mixed-effects approach is

more compatible with the personalized medicine setting of N-of-1 trials.

In Section 3.3, we evaluate different point estimators to be used in the joint analysis. We also assess the properties of the combined point estimator.

3.2.4 Candidate Point Estimators

We consider three types of point estimators (their specific forms are provided in the appendix): (1) naive estimator (no bias adjustment), (2) bias-adjusted mean, [51] (BAM), and (3) median-unbiased estimator (MUE). The MUE requires specifying an ordering, and we include the following orderings: analysis-time (AT), sample mean (SM) and likelihood ratio (LR). [52]

By construction, all of these point estimators are members of the class of Z-estimators. With this observation, we derive expressions for standard error of each of the estimators based on the asymptotic variance of members of this class. [53]

3.3 Results

In this section, we evaluate the properties of the proposed methods. First, we assess the type-1 error when sequentially monitoring one N-of-1 trial to evaluate under what scenarios the sequential monitoring has well-calibrated type-1 error. We then study the power, probability of early stopping and average number of blocks at stopping under varying effect sizes. Next, we evaluate the bias and mean-squared error (MSE) of point estimators from one sequentially monitored N-of-1 trial. Finally, we assess the bias and mean-squared error for the proposed combined point estimator for jointly analyzing a series of sequentially monitored N-of-1 trials.

3.3.1 Simulation Setup

We summarize the results from a simulation study in which we simulate N-of-1 trials in each of several settings. We suppose these trials emulate real-world studies that have 1-week treatment periods and the outcome is continuous and has mean 0 under the baseline

treatment (i.e., the treatment for which we assign $Z = 0$ - this can be considered the standard of care or placebo). We set the error variance and the random intercept (within a block) variance each as 1. Treatment blocks are independent, and the order of treatments within blocks is generated randomly.

We vary the planned number of blocks from 3 to 26 with 2 to 6 periods per block, and we vary the treatment effect from 0 to 5.

We study three different sequential monitoring boundary shapes. The "boundary" is the sequence of critical values to which we compare the monitored test statistic - and, while each sequence controls type-1 error at a prespecified level, different sequences, or shapes, induce different stopping rules. [54] Boundary shapes differ by their likelihood of stopping the trial at an earlier analysis. We construct either (1) symmetric O'Brien-Fleming boundaries (OBF), [55] a common conservative-early boundary shape (i.e., need stronger evidence to stop at an earlier analysis), (2) symmetric Pocock boundaries, [56] a less conservative-early boundary shape, or (3) asymmetric boundaries with OBF for efficacy and Pocock for futility, each aimed at controlling the stopping probability under the null (type-1 error) at two-sided 5%. The designs allow for the conclusion that neither treatment option has displayed convincing evidence.

We vary the number of planned looks - prespecified instances at which the trial may be stopped based on an unblinded analysis - with either (1) one look (i.e., no sequential monitoring), (2) two equally-spaced looks, (3) four equally-spaced looks, or (4) looks after every treatment block. For example, with 13 treatment blocks and two looks, an unblinded analysis occurs after 7 treatment blocks and after 13 blocks, and the trial may be stopped at either analysis.

We also assess the effect of 5 versus 10 constituent trials in a series of N-of-1 trials. Previous authors have suggested five or more units as a guideline for random-effects models. [57] For computational efficiency, we simulate 10,000 independent trials for each setting, and then we simulate multiple series of N-of-1 trials by collecting these generated trials into groups of size 5 or 10. For each group, we perform the analysis as described in Section 3.2.3.

We fit all mixed-effects models with the REML procedure using R 4.0 with the packages `lme4` and `metafor`, and we construct sequential monitoring boundaries and compute bias-adjusted estimates with the package `RCTdesign` in R 3.2.

3.3.2 Properties for One Sequentially-Monitored N-of-1 Trial

Type-1 Error

Existing monitoring boundaries assume normality for their construction, but, when conducting larger RCTs, monitoring an asymptotically normal statistic can be appropriate. [58] However, we may expect to see different behavior in the setting of N-of-1 trials. In this section, we assess the effect of the number of looks, the number of blocks and periods per block and shape of the monitoring boundary on the type-1 error.

In Figures 3.2-3.3, we illustrate the type-1 error rates across 10,000 N-of-1 trials for two and six periods per block, OBF or Pocock boundary shapes and varying numbers of looks and numbers of blocks.

We see that type-1 error decreases with larger numbers of blocks, and the type-1 error increases with more looks. The O'Brien-Fleming (OBF) boundary shape has consistently lower type-1 error compared to the Pocock shape, and the asymmetric shape (not pictured) lies between the two.

The type-1 error decreases dramatically from two periods per block to six periods per block. With two periods per block, even one look at the data - no sequential monitoring - does not reach the nominal 5% type-1 error rate by 26 blocks, and more looks results in higher type-1 error inflation. However, with six periods per block, and sequential monitoring with up to four looks at the data can be achieved with nominal type-1 error rate if using an OBF boundary shape.

From these results, we conclude that type-1 error can be very large (inflation of 2-7 times the nominal 0.05 rate) with 13 or fewer treatment blocks and 2 periods per block. With 6 periods per block, the type-1 error ranges from 0.04 to 0.10 for 3-26 treatment blocks, which

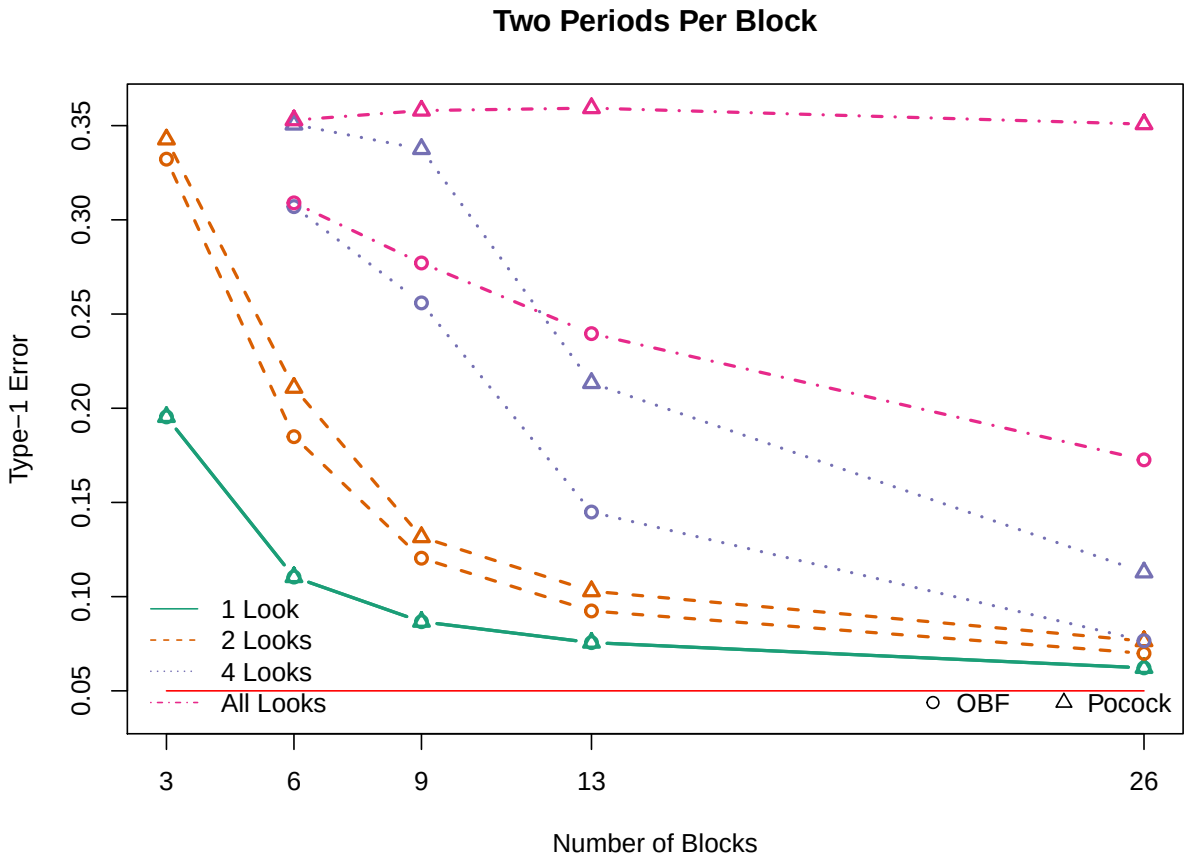


Figure 3.2: Type-1 error for sequentially monitoring one N-of-1 trial with two periods per block and varying boundary shapes, blocks and looks. Nominal 5% type-1 error indicated in red.

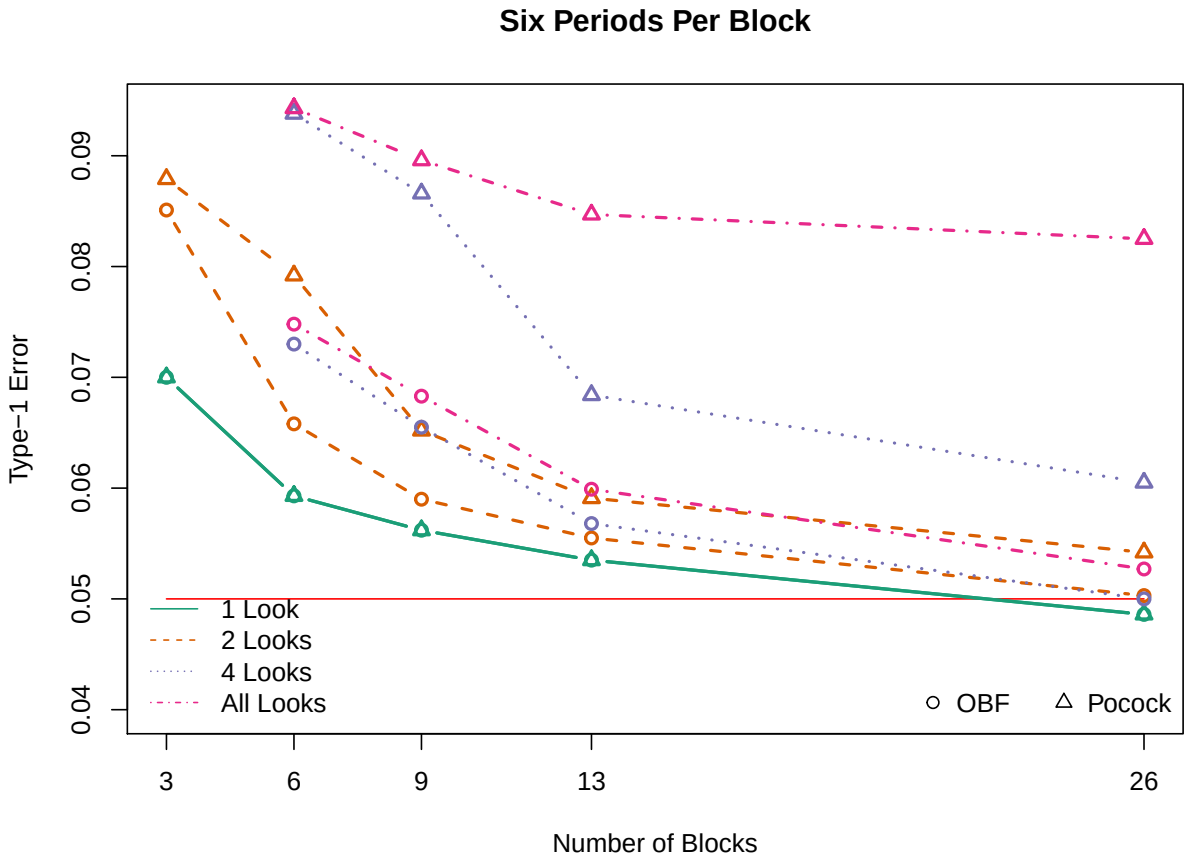


Figure 3.3: Type-1 error for sequentially monitoring one N-of-1 trial with six periods per block and varying boundary shapes, blocks and looks. Nominal 5% type-1 error indicated in red.

may be acceptable, especially with 13 or 26 treatment blocks.

Power, Probability of Early Stopping and Average Number of Blocks at Stopping

This section summarizes three other properties of sequential monitoring for one N-of-1 trial: power, probability of early stopping and the average number of blocks at stopping. We summarize the results for each of the three measures across 10,000 N-of-1 trials under different effect sizes, number of blocks and periods per block.

From this section onward, we only summarize the settings with 13 or 26 treatment blocks and 1-4 looks and the OBF boundary shape, excluding the setting with 13 blocks, 2 periods per block and 4 looks. These are settings with type-1 error less than 10%, based on the results from Section 3.3.2.

From Figure 3.4, we see the power does not appreciably differ when comparing no sequential monitoring against 2 or 4 looks at the data based on the overlapping power curves. For all of the reported settings, the power approaches 1 with increasing effect size. We also see that the power approaches 1 with smaller effect sizes with increasing numbers of blocks and increasing numbers of periods per block.

Figure 3.5 illustrates that the probability of stopping early increases with more looks at the data, and the probability increases to 1 with larger effect sizes. And, similar to power, the probability approaches 1 with smaller effect sizes as the number of blocks or number of periods per block increases.

We see the results for average number of blocks at stopping, in Figure 3.6, mirror the patterns from the probability of early stopping. As the (planned) number of blocks and the number of periods per block increases, the average number of blocks at stopping decreases. The average also decreases with more looks at the data.

Bias and Mean-Squared Error of Point Estimators

In this section, we compare the candidate point estimators based on bias and mean-squared error (MSE) across 1,000 sequentially monitored N-of-1 trials. Similar to the construction

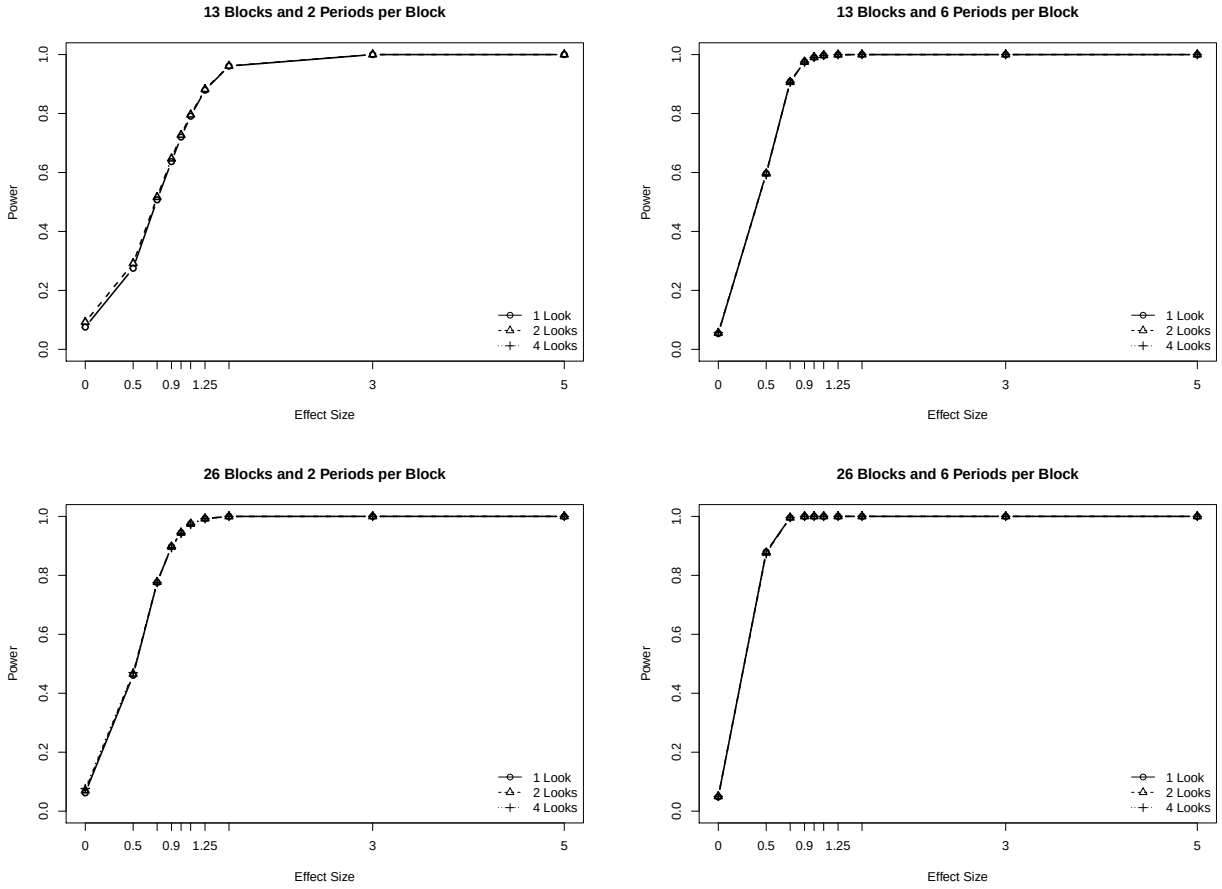


Figure 3.4: Power for sequentially monitoring one N-of-1 trial with varying numbers of blocks, periods per block, looks and an OBF boundary shape. This excludes the setting with 13 blocks, 2 periods per block and 4 looks due to large type-1 error.

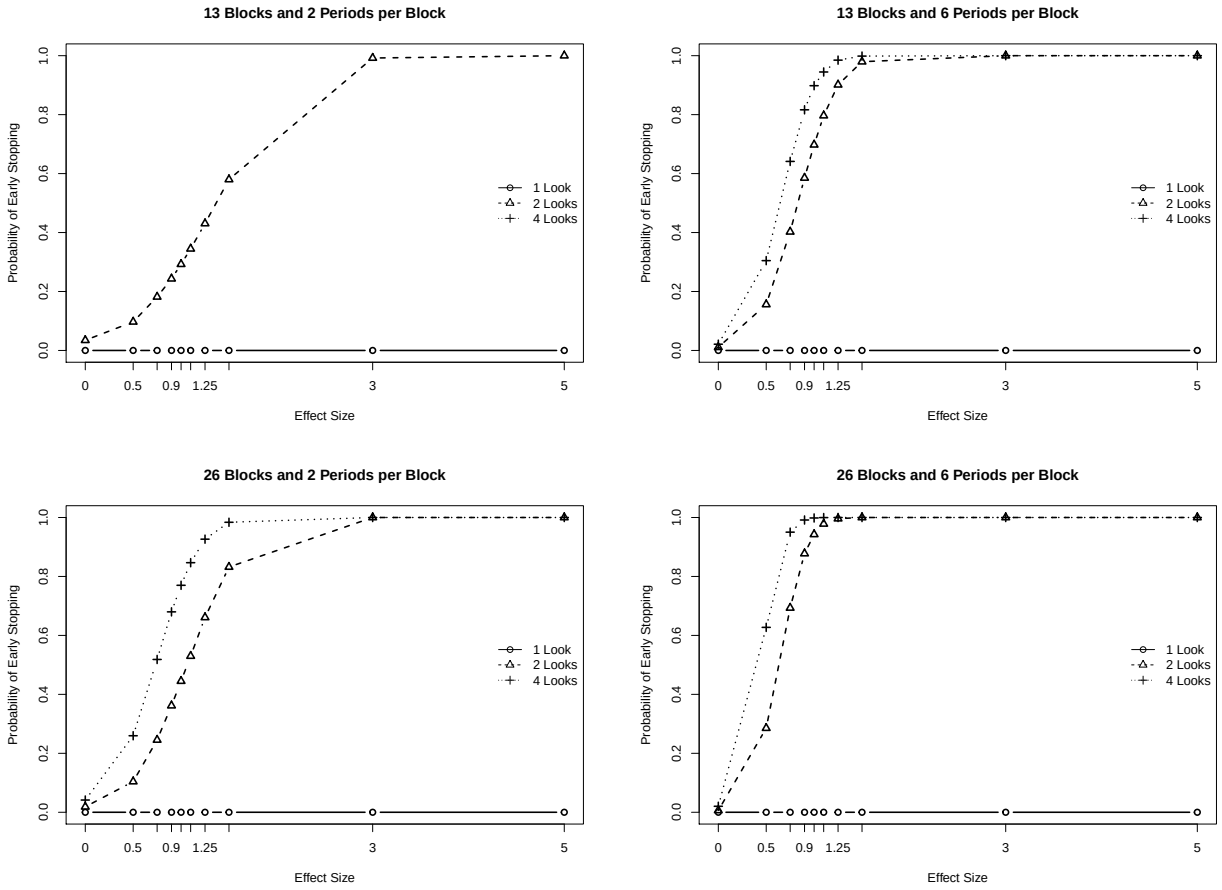


Figure 3.5: Probability of early stopping for sequentially monitoring one N-of-1 trial with varying numbers of blocks, periods per block, looks and an OBF boundary shape. This excludes the setting with 13 blocks, 2 periods per block and 4 looks due to large type-1 error.

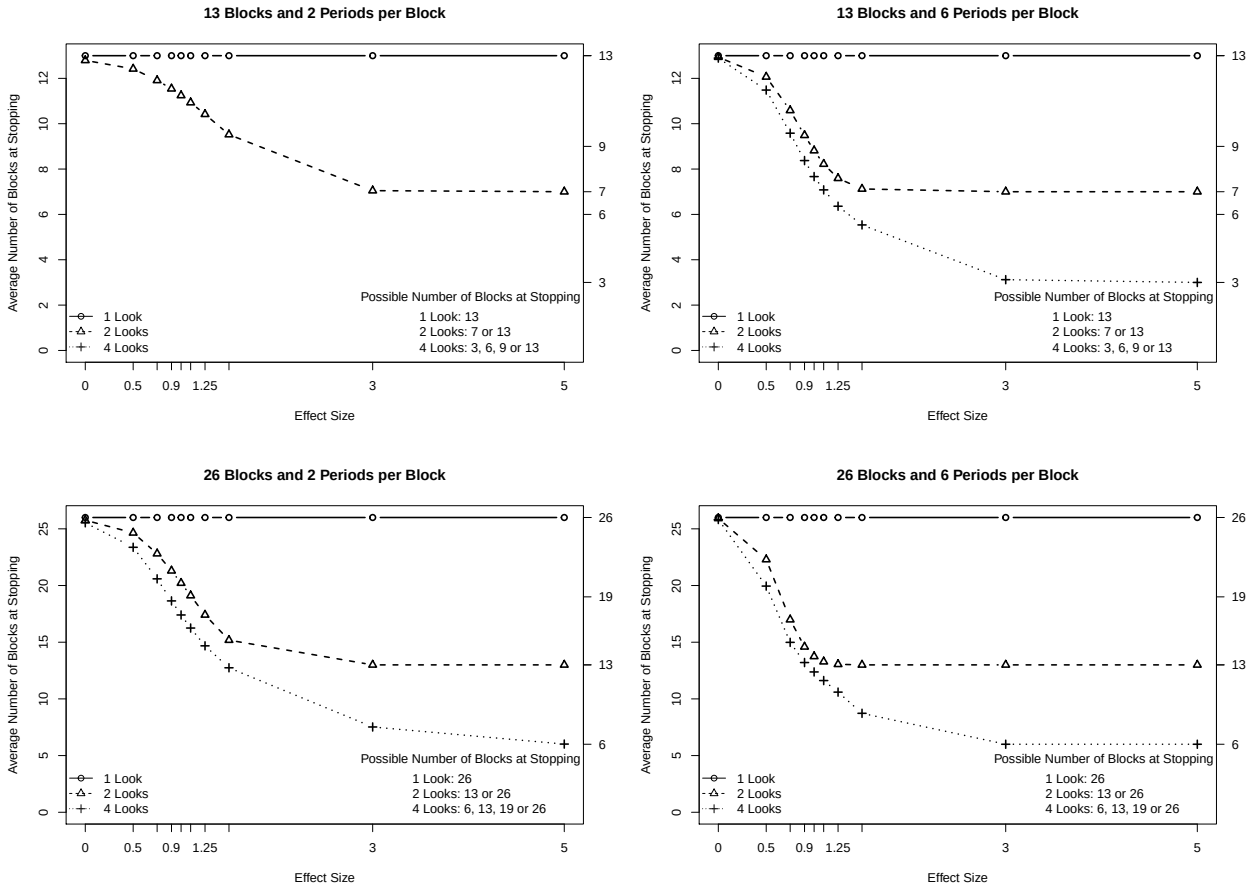


Figure 3.6: Average number of blocks at stopping for sequentially monitoring one N-of-1 trial with varying numbers of blocks, periods per block, looks and an OBF boundary shape. This excludes the setting with 13 blocks, 2 periods per block and 4 looks due to large type-1 error.

of the monitoring boundaries, the estimation of the candidate point estimators also employs normality. The results of this section assess whether the bias-adjusted estimators continue to maintain desirable properties under N-of-1 settings and compare bias and MSE between the candidate estimators.

We present the results for trials monitored with the OBF boundary shape - estimators had consistently lower mean-squared error with the OBF shape compared to the Pocock shape - and 2 looks at the data.

With large effect sizes, the point estimates become identical, so we provide the results for effect sizes smaller than 5 in this article. With 1 look at the data (no sequential monitoring), all point estimators are identical under all settings.

We display the bias and MSE of these candidate estimators in Figure 3.7 for 13 or 26 blocks and 2 or 6 periods per block with 2 looks. (We provide a table with the full results in the Appendix, Section C.3.1.)

Bias decreases with increasing numbers of blocks and periods per block. The MUE candidates have similar bias at all effect sizes, and all estimators have similar bias at effect sizes of 0 and 3. The naive estimator has the highest bias at all settings, and the BAM has the lowest bias. For the BAM, a negative bias with 26 blocks and 6 periods per block results in a larger absolute bias compared to MUE candidates at effect sizes of 1-1.25 - however, the discrepancy is within simulation error, approximately 0.01.

The MSE also decreases with increasing numbers of blocks and periods per block. The MUE candidates have similar MSE at all effect sizes. With 13 blocks and 2 periods per block, the BAM has smaller MSE across effect sizes compared to other candidates and the naive estimator has larger MSE (this pattern reverses at effect size 3), but the largest discrepancies were small, within 0.02. By 26 blocks and 6 periods per block, MSE is nearly identical across point estimators.

The MSE increases for all estimators with increasing effect size. Authors have previously explained this as due to studies more frequently ending earlier, with less information and, thus, larger variances. [59]

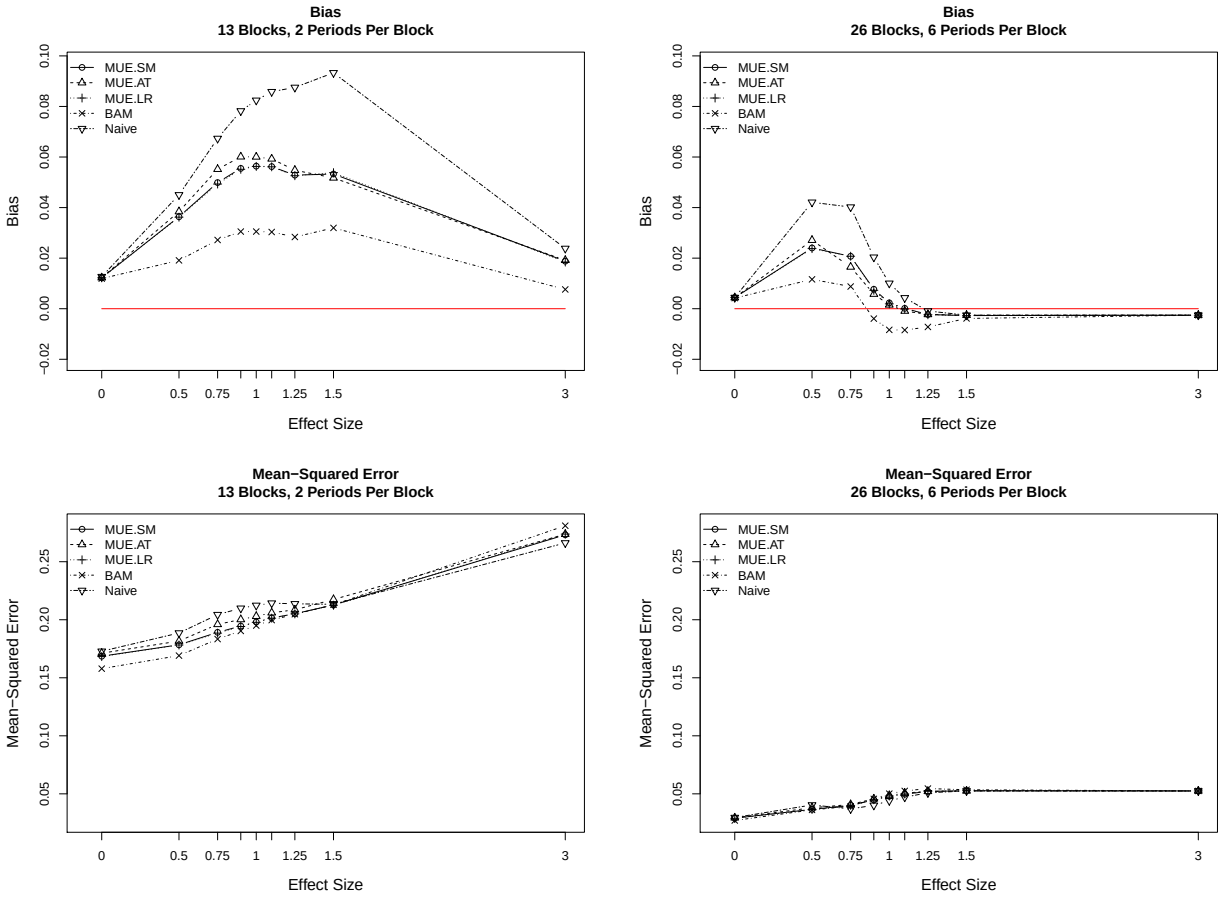


Figure 3.7: Bias (top) and mean-squared error (bottom) for candidate point estimators with 13 treatment blocks and 2 periods per block (left) or 26 treatment blocks and 6 periods per block (right) and 2 looks at the data with an OBF boundary shape. MUE.SM, MUE.AT and MUE.LR refer to the median-unbiased estimators with sample-mean, analysis-time and likelihood-ratio orderings, respectively, BAM refers to the bias-adjusted mean, and Naive refers to the naive estimator. Bias of 0 indicated in red.

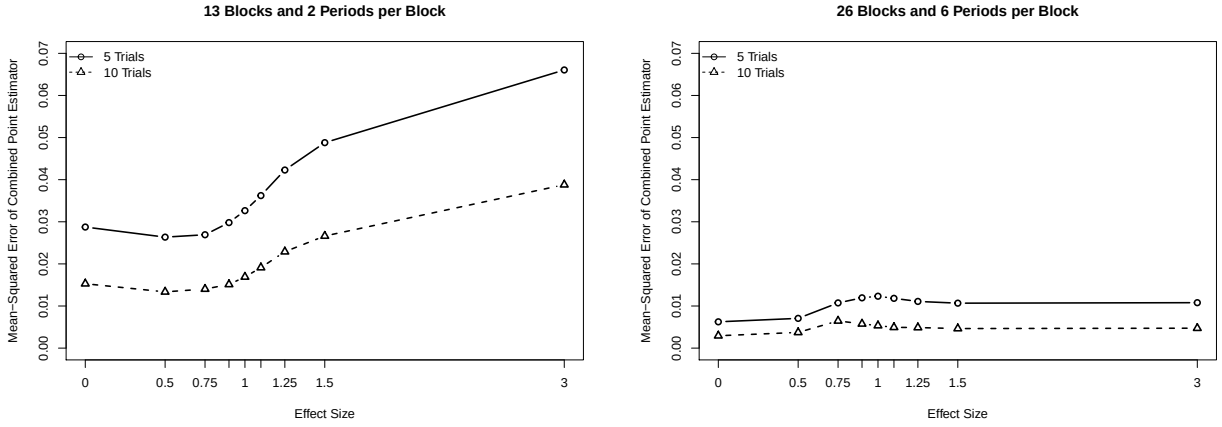


Figure 3.8: Mean-squared error for the combined point estimator from a series of 5 or 10 sequentially-monitored N-of-1 trials with varying numbers of blocks and periods per block and an OBF boundary shape with 2 looks at the data.

3.3.3 Properties for a Series of Sequentially Monitored N-of-1 Trials

Bias and Mean-Squared Error of the Combined Point Estimator

This section presents the results of the bias and MSE for the combined point estimator across 200 series of N-of-1 trials with 5 trials per series or 100 series with 10 trials per series. Based on the results of Section 3.3.2, we employ the bias-adjusted mean as the estimate from each constituent trial within a series for computing the combined point estimator.

We focus on the effect of number of trials on the bias and mean-squared error of the combined point estimator across differing effect sizes, numbers of blocks and periods per block. All constituent trials in a series have the same planned number of blocks and the same number of periods per block. We present the results of the setting in which all trials used an OBF boundary shape with 2 looks at the data.

We see that the MSE decreases with increasing numbers of trials within the series in Figure 3.8. The patterns of MSE for the combined point estimator across different effect sizes and blocks or periods per block resemble the patterns we observe for point estimators

for one N-of-1 trial.

We do not have strong evidence for changes in bias with increasing numbers of trials within a series as evidenced by Figure C.1 in the Appendix, Section C.3.2. However, this result is limited by the small number of simulations.

3.4 Discussion

Sequential monitoring permits statistically valid early stopping of studies for efficacious (or futile) treatments. For N-of-1 trials, sequential monitoring has the potential to deliver rapid, optimal care for individual patients. In this article, we develop a framework for sequential monitoring of one N-of-1 trial and for jointly analyzing a series of already-completed sequentially-monitored N-of-1 trials. We propose that the treatment block design commonly used in N-of-1 trials combined with a mixed-effects model for analysis facilitates the use of existing sequential monitoring methodology in one N-of-1 trial. To combine results from a series of N-of-1 trials, we suggest researchers combine bias-adjusted point estimators from the constituent studies with a random-effects meta-analysis model. [49]

In registrational trials, type-1 error control can be of central importance. However, N-of-1 trials frequently study treatments with minimal risks, so the possibility of early stopping for a promising treatment may outweigh the risks. As such, while we find several settings with large type-1 error, trial-specific considerations may take precedence and strict type-1 error control may not be the highest priority.

In settings with acceptable type-1 error rates - larger numbers of blocks and periods per block and an O'Brien-Fleming boundary shape - we find sequential monitoring can have benefits in the power, probability of early stopping and average number of blocks at stopping. These gains can be realized even with modest effect sizes if the planned number of blocks or number of periods per block is large: for example, with 26 planned blocks and 2 periods per block, the average number of blocks at stopping is nearly half the planned number at an effect size of 1.25.

We compare the bias and mean-squared error (MSE) of several candidate point estimators

for sequential monitoring in one N-of-1 trial, and our results indicate bias-adjusted estimators continue to provide improvement over a naive estimator in the N-of-1 setting. We find that the bias-adjusted mean (BAM) has smaller bias and similar MSE compared to other estimators, so we recommend choosing the BAM for combining point estimates across a series of sequentially monitored N-of-1 trials. When jointly analyzing a series of N-of-1 trials, we observe that the MSE decreases with a larger number of constituent trials, as expected, but we do not have evidence to suggest that the bias of the point estimator decreases with more constituent trials.

In this work, we assume that constituent trials within a series employ the same boundary shape (along with the same number of looks). In the Appendix, Section C.3.3, we present preliminary results for allowing constituent trial boundary shape to differ: we vary the number of trials within a series having either an OBF shape or a Pocock shape. The results suggest that series with more OBF trials have smaller mean-squared error of the combined point estimator, with this effect diminishing with more periods per block or more blocks. We do not have clear evidence for a pattern with bias. In the future, we could expand these results and address a broader range of trial designs.

An important limitation of this work is the relatively small number of simulations for assessing a series of N-of-1 trials. Due to computational burden, we only provide results for 200 series with 5 constituent trials and 100 series with 10 constituent trials, and we also could not study series with more than 10 constituent trials. Because of this, patterns may not be detectable due to simulation variability or due to assessing a maximum of 10 constituent trials: in particular, we could not detect patterns in bias, but such patterns may be apparent with a larger number of simulations.

One avenue of future work is assessing the impact of violations to the assumptions of (1) no carryover effect and (2) no outcome drift. Such violations may occur due to issues such as unplanned unblinding to treatment assignment (possibly causing a carryover effect through a placebo effect) or disease progression (possibly causing outcome drift). Methods exist to minimize the impact of these violations, but they may need modification for the sequential

monitoring setting. For example, methods to address a time trend in the outcome may affect our results for the properties of our framework due to changes in the analysis models.

For this work, we assume that the number of treatment periods per block is fixed and equal across all treatment blocks within an N-of-1 trial. This may not hold, for example, when certain blocks have missing data. Under this setting, our results for monotonic information growth may not apply (even if the data are missing completely at random), so further research is needed to evaluate the impact of differing numbers of treatment periods.

We also assume constituent trials of the N-of-1 series have been completed at the time of analysis, a retrospective approach. In contrast, a cumulative meta-analysis would take a prospective approach by analyzing the series of N-of-1 trials as data from each trial accrue. This approach would require specifying a sequential monitoring procedure at the level of the constituent trials (possibly, in addition to monitoring within each trial), but it could allow investigators to reduce the number of N-of-1 trials needed. A future direction is to evaluate whether the benefits of the cumulative meta-analysis approach warrant the increased complexity relative to the retrospective approach.

The mixed-effects model for analyzing one sequentially monitored N-of-1 trial easily extends to more than two treatment options and to other types of outcomes such as binary or count data. However, outside of the continuous outcome setting, the target estimand for the trial may no longer be an average treatment effect (ATE), and the proposed procedure for combining results may no longer target an average across ATEs. Future work will aim to develop a unified framework to address these concerns.

Tools for sequential monitoring have been well-studied and are widely available for standard RCTs, and the proposed framework allows researchers to utilize these existing resources for the setting of N-of-1 trials. For shorter N-of-1 trials (i.e., those with a small planned number of treatment blocks or with few periods per block), sequential monitoring is not appropriate due to highly inflated type-1 error, but, for longer trials, sequential monitoring under the proposed framework can assist clinicians in making important decisions sooner, on average, for their patients.

BIBLIOGRAPHY

[1] Akihiro Hirakawa, Junichi Asano, Hiroyuki Sato, and Satoshi Teramukai. Master protocol trials in oncology: Review and new trial designs. *Contemporary Clinical Trials Communications*, 12:1:8, 2018.

[2] Steffen Ventz, Brian M. Alexander, Giovanni Parmigiani, Richard D. Gelber, and Lorenzo Trippa. Designing clinical trials that accept new arms: An example in metastatic breast cancer. *Journal of Clinical Oncology*, 35(27):3160:3168, 2017.

[3] Roland B. Walter, Laura C. Michaelis, Megan Othus, Geoffrey L. Uy, Jerald P. Radich, Richard F. Little, Sandi Hita, Lalit Saini, James M. Foran, Aaron T. Gerds, and et al. Intergroup leap trial (s1612): A randomized phase 2/3 platform trial to test novel therapeutics in medically less fit older adults with acute myeloid leukemia. *American Journal of Hematology*, 93(2), 2017.

[4] Masters L Rauchenberger M Van Looy N Diaz-Montana C Gannon M James N Maughan T Parmar MKB Brown L Sydes MR; STAMPEDE Hague D, Townsend S and FOCUS4 investigators. Changing platforms without stopping the train: experiences of data management and data management systems when adapting platform protocols by adding and closing comparisons. *Trials*, 20(294), 2019.

[5] Pistone T Onaisi R Sitta R Journot V-Nguyen D Peiffer-Smadja N Crémer A Bouchet S Darnaud T Poitrenaud D Piroth L Binquet C Michel JF Lefèvre B Lebeaux D Lebel J Dupouy J Roussillon C Gimbert A Wittkop L Thiébaut R Orne-Gliemann J Joseph JP Richert L Anglaret X Malvy D; COVERAGE study group Duvignaud A, Lhomme E. Home treatment of older people with symptomatic sars-cov-2 infection (covid-19): A structured summary of a study protocol for a multi-arm multi-stage (mams) randomized trial to evaluate the efficacy and tolerability of several experimental treatments to reduce the risk of hospitalisation or death in outpatients aged 65 years or older (coverage trial). *Trials*, 21(846), 2020.

[6] G Anderson, M LeBlanc, P Y Liu, and J Crowley. *Use of Covariates in Randomization and Analysis of Clinical Trials*. Handbook of Statistics in Clinical Oncology. CRC Press, third edition, 2011.

[7] J.B. Vermorken, M.K.B. Parmar, M.F. Brady, E.A. Eisenhauer, T. Hogberg, R.F. Ozols, J. Rochon, G.J.S. Rustin, S. Sagae, and R.H.M. Verheijen. Clinical trials in ovarian

carcinoma: study methodology. *Annals of Oncology*, 16:viii20 – viii29, 2005. 3rd International Ovarian Cancer Consensus of the GCIG, September, 2005: Germany.

[8] S. L. George. Reducing patient eligibility criteria in cancer clinical trials. *Journal of Clinical Oncology*, 14(4):1364–1370, 1996.

[9] Stuart J. Pocock and Richard Simon. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 31(1):103, 1975.

[10] Sarah Brown, Helen Thorpe, Kim Hawkins, and Julia Brown. Minimization—reducing predictability for multi-centre trials whilst retaining balance within centre. *Statistics in Medicine*, 24(24):3715–3727, 2005.

[11] Annette E. Hay, Sarit Assouline, Roland B. Walter, Richard F. Little, Anna Moseley, Sperling M Gail, Annie Im, James M. Foran, Jerald P. Radich, Min Fang, and et al. Accrual barriers and detection of early toxicity signal in older less-fit patients treated with azacitidine and nivolumab for newly diagnosed acute myeloid leukemia (aml) or high-risk myelodysplastic syndrome (mds) in the swog 1612 platform randomized phase ii/iii clinical trial. *Blood*, 134(Supplement 1):3905–3905, 2019.

[12] John W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society: Series B (Methodological)*, 11(1):15–44, 1949.

[13] Joseph Berkson and Robert P Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952.

[14] M. Othus, B. Barlogie, M. L. Leblanc, and J. J. Crowley. Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, 18(14):3731–3736, Jun 2012.

[15] Megan Othus, Aasthaa Bansal, Lisel Koepl, Samuel Wagner, and Scott Ramsey. Accounting for cured patients in cost-effectiveness analysis. *Value in Health*, 20(4):705–709, 2017.

[16] Maïlis Amico and Ingrid Van Keilegom. Cure models in survival analysis. *Annual Review of Statistics and Its Application*, 5(1):311–342, 2018.

[17] Mikkael A. Sekeres, Megan Othus, Alan F. List, Olatoyosi Odenike, Richard M. Stone, Steven D. Gore, Mark R. Litzow, Rena Buckstein, Min Fang, Diane Roulston, and et al. Randomized phase ii study of azacitidine alone or in combination with lenalidomide or with vorinostat in higher-risk myelodysplastic syndromes and chronic myelomonocytic

leukemia: North american intergroup study swog s1117. *Journal of Clinical Oncology*, 35(24):2745–2753, 2017.

[18] J.E. Anderson. Bone marrow transplantation for myelodysplasia. *Blood Reviews*, 14(2):63 – 77, 2000.

[19] R. A. Maller and S. Zhou. Testing for sufficient follow-up and outliers in survival data. *Journal of the American Statistical Association*, 89(428):1499–1506, 1994.

[20] Ross A. Maller and Xian Zhou. *Survival analysis with long-term survivors*. Wiley, 1996.

[21] Pao-Sheng Shen. Testing for sufficient follow-up in survival data. *Statistics & Probability Letters*, 49(4):313–322, 2000.

[22] Binbing Yu, Ram C. Tiwari, Kathleen A. Cronin, and Eric J. Feuer. Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in Medicine*, 23(11):1733–1747, 2004.

[23] A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2007.

[24] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[25] Stephen H. Petersdorf, Kenneth J. Kopecky, Marilyn Slovak, Cheryl Willman, Thomas Nevill, Joseph Brandwein, Richard A. Larson, Harry P. Erba, Patrick J. Stiff, Robert K. Stuart, and et al. A phase 3 study of gemtuzumab ozogamicin during induction and postconsolidation therapy in younger patients with acute myeloid leukemia. *Blood*, 121(24):4854–4860, 2013.

[26] Megan Othus, Aasthaa Bansal, Harry Erba, and Scott Ramsey. Bias in mean survival from fitting cure models with limited follow-up. *Value in Health*, 23(8):1034–1039, 2020.

[27] Elizabeth O Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric J Topol, and Nicholas J Schork. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Personalized Medicine*, 8(2):161–173, 2011.

[28] Charles H. Evans and Suzanne T. Ildstad. *Introduction*. National Academy Press, 2001.

[29] Patrick Bodilly Kane, Merlin Bittlinger, and Jonathan Kimmelman. Individualized therapy trials: Navigating patient care, research goals and ethics. *Nature Medicine*, 27(10):1679–1686, 2021.

[30] Ainhoa Madariaga, Lawrence Kasherman, Katherine Karakasis, Pamela Degendorfer, Ann M. Heesters, Wei Xu, Shahid Husain, and Amit M. Oza. Optimizing clinical research procedures in public health emergencies. *Medicinal Research Reviews*, 41(2):725–738, 2020.

[31] DEcIDE Methods Center N of 1 Guidance Panel, Christopher H Schmid, and Naihua Duan. page 33–49. Agency for Healthcare Research and Quality, 2014.

[32] Deborah R. Zucker, Robin Ruthazer, and Christopher H. Schmid. Individual (n-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. *Journal of Clinical Epidemiology*, 63(12):1312–1323, 2010.

[33] Salima Punja, Cecilia Bukutu, Larissa Shamseer, Margaret Sampson, Lisa Hartling, Liana Urichuk, and Sunita Vohra. N-of-1 trials are a tapestry of heterogeneity. *Journal of Clinical Epidemiology*, 76:47–56, 2016.

[34] Center for Drug Evaluation and Research. Ind submissions for individualized antisense oligonucleotide drug products: Administrative and procedural recommendations guidance for sponsor-investigators. Draft guidance, U.S. Department of Health and Human Services Food and Drug Administration, Silver Spring, MD, January 2021.

[35] L March, L Irwig, J Schwarz, J Simpson, C Chock, and P Brooks. N of 1 trials comparing a non-steroidal anti-inflammatory drug with paracetamol in osteoarthritis. *BMJ*, 309(6961):1041–1044, 1994.

[36] Naihua Duan, Richard L. Kravitz, and Christopher H. Schmid. Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *Journal of Clinical Epidemiology*, 66(8), 2013.

[37] Steven E. Arnold and Rebecca A. Betensky. Multicrossover randomized controlled trial designs in alzheimer disease. *Annals of Neurology*, 84(2):168–175, 2018.

[38] Abigail B. Shoben and Scott S. Emerson. Violations of the independent increment assumption when using generalized estimating equation in longitudinal group sequential trials. *Statistics in Medicine*, 33(29):5041–5056, 2014.

[39] Abigail B. Shoben, Kyle D. Rudser, and Scott S. Emerson. More data, less information? potential for nonmonotonic information growth using gee. *Journal of Biopharmaceutical Statistics*, 27(1):135–147, Jun 2016.

[40] Lawrence Samuel Friedman, Curt D. Furberg, and David L. DeMets. *Fundamentals of Clinical Trials*. Springer, 4 edition, 2010.

[41] S. Vohra, L. Shamseer, M. Sampson, C. Bukutu, C. H. Schmid, R. Tate, J. Nikles, D. R. Zucker, R. Kravitz, G. Guyatt, and et al. Consort extension for reporting n-of-1 trials (cent) 2015 statement. *BMJ*, 350(may14 17), May 2015.

[42] Antony J Porcino, Larissa Shamseer, An-Wen Chan, Richard L Kravitz, Aaron Orkin, Salima Punja, Philippe Ravaud, Christopher H Schmid, and Sunita Vohra. Spirit extension and elaboration for n-of-1 trials: Spent 2019 checklist. *BMJ*, page m122, 2020.

[43] G Guyatt, D Sackett, J Adachi, R Roberts, J Chong, D Rosenbloom, and J Keller. A clinician's guide for conducting randomized trials in individual patients. *Canadian Medical Association Journal*, 139(6), Sep 1988.

[44] Nicole B. Gabler, Naihua Duan, Sunita Vohra, and Richard L. Kravitz. N-of-1 trials in the medical literature. *Medical Care*, 49(8):761–768, 2011.

[45] Artur Araujo, Steven Julious, and Stephen Senn. Understanding variation in sets of n-of-1 trials. *Plos One*, 11(12), Jan 2016.

[46] Jae Won Lee and David L. Demets. Sequential comparison of changes with repeated measurements data. *Journal of the American Statistical Association*, 86(415):757–762, Sep 1991.

[47] Susan Todd, John Whitehead, and Karen M Facey. Point and interval estimation following a sequential clinical trial. *Biometrika*, 83(2):453–461, 1996.

[48] Xiaoyin (Frank) Fan, David L. DeMets, and K. K. Lan. Conditional bias of point estimates following a group sequential test. *Journal of Biopharmaceutical Statistics*, 14(2):505–530, 2004.

[49] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials*, 45:139–145, 2015.

[50] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. Fixed-effect versus random-effects models. *Introduction to Meta-Analysis*, page 77–86, 2010.

[51] JOHN WHITEHEAD. On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, 73(3):573–581, 1986.

[52] Scott S. Emerson and Thomas R. Fleming. Parameter estimation following group sequential hypothesis testing. *Biometrika*, 77(4):875–892, 1990.

[53] A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2007. 41-84.

[54] Susan Smith Ellenberg, Thomas R. Fleming, and David L. DeMets. page 122–134. Wiley, 2002.

[55] Peter C. O'Brien and Thomas R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549, 1979.

[56] S. J. POCOCK. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

[57] Brennan C. Kahan and Michael O. Harhay. Many multicenter trials had few events per center, requiring analysis via random-effects models or gees. *Journal of Clinical Epidemiology*, 68(12):1504–1511, 2015.

[58] Christopher Jennison and Bruce W. Turnbull. Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association*, 92(440):1330–1341, 1997.

[59] Zhengqing Li and David L. DeMets. On the bias of estimation of a brownian motion drift following group sequential tests. *Statistica Sinica*, 9(4):923–937, 1999.

Appendix A

APPENDIX FOR CHAPTER 1

A.1 Dynamic Balancing with Differing Experimental Arm Eligibility

The following is one possible schema for randomizing a newly enrolled participant to a treatment arm when balancing on L categorical stratification factors using a Pocock-Simon method for dynamic balancing. We assume 1:1 randomization in the following, but modification of the weighting scheme below easily allows for extension to other randomization ratios.

If the participant enrolls on the trial with no other previously randomized participants having the same combination of stratification variables, then the participant is randomly assigned to an eligible arm with equal weight given to all the eligible arms. Otherwise, we proceed with the following scheme:

1. For each eligible experimental arm, $E \in \mathbf{E} = \{\text{eligible experimental arms}\}$, and for each stratification factor $l = 1, \ldots, L$, tally the number of participants previously randomized to that arm $T_l(E)$ and randomized to the standard of care arm but eligible for that experimental arm $T_l(C_E)$. Only include participants with the same level of stratification factor l as the new participant.

2. Calculate the imbalance caused by adding the participant to each eligible study arm:

   (a) Choose one $Y \in \mathbf{S}$, where $\mathbf{S} = \{\text{eligible experimental arms and standard of care arm, } C\}$

   (b) Increment the tally in arm Y with $T_l(Y)^* = T_l(Y) + 1$ for all $l = 1, \ldots, L$ but keep the tallies the same for all other arms $Z \in \mathbf{S}$ with $Z \neq Y$: $T_l(Z)^* = T_l(Z)$ for $l = 1, \ldots, L$.

(Note: if $Y = C$, then this implies $T_l(C_E)^* = T_l(C_E) + 1$ for all $E \in \mathbf{E}$.)

(c) Compute the pairwise differences $|T_l(E)^* - T_l(C_E)^*|$ for all eligible experimental arms $E \in \mathbf{E}$

(d) The imbalance score for adding the participant to arm Y is

$$\sum_{l=1}^{L} \max_{\mathbf{E}} |T_l(E)^* - T_l(C_E)^*|$$

(e) Repeat this process for all arms in $\mathbf{S}$

3. Order each of the resulting $s = |\mathbf{S}|$ imbalance scores in order from largest (most imbalanced) to smallest (least imbalanced).

4. Then one choice of randomization weights is as follows: for $i = 1, \ldots, s$ the index in decreasing order of imbalance scores above, assign $p_i = \frac{i}{s(s+1)/2}$.

   That means the most imbalanced arm has the smallest randomization weight $p_1 = \frac{1}{s(s+1)/2}$ and the least imbalanced arm has the highest weight $p_s = \frac{s}{s(s+1)/2}$. This is one possible algorithm that ensures a strictly increasing sequence of weights from most to least imbalance.

   If an m-way tie ($m \in \{2, \ldots, s\}$) of imbalance scores exists at $i \in \{1, \ldots, s\}$, equally split the assigned weight $p_i + \cdots + p_{i+m-1}$ across the tied arms. (In other words, equally split the weight the arms would have had if they could have been ordered.)

A.2 Comparing Methods for Accommodating New Experimental Arms

We compare two approaches to accommodate newly added experimental arms in a platform trials with differing experimental arm eligibility. In the first method, which we will call "split data," we restart the stratified randomization procedure anew upon the addition of the new arm: we do not use any participants randomized prior to addition of the new arm in future randomization calculations. The second method, "combine data," allows participants

in continuing arms to be used in randomization calculations, but calculations involving the new arm only use participants enrolled after the addition of the new arm. This is more difficult to implement (as it requires the randomization system to use different subsets of participants), but it has the advantage of using more of the existing data.

Both methods provide appropriate balanced randomization at the prespecified allocation ratio with sufficient numbers of participants, but the allocation ratio may deviate from the specified ratio with random noise in a small window after the addition of the new arm. When the allocation ratio differs from the desired ratio, this can affect properties of hypothesis testing - this can be particularly important for sequential monitoring in clinical trials, which may rely on small tail probabilities early in a trial. As such, even local disruptions to the allocation ratio could have important effects in practice.

To assess the differences in disruption to the allocation ratio, we evaluate the deviation from desired allocation ratio (measured as the mean squared difference from the specified allocation ratio) and the variance of the observed allocation ratio for a fixed number of participants before and after the addition of a new experimental arm. Both the deviation and the variance should decrease to 0 with sufficient numbers of participants, so an increase in either quantity after the addition of a new arm indicates a disruption.

We simulate platform trials that begin with two experimental arms, have 2-3 stratification factors to balance on and add 1-2 additional experimental arms. We vary the addition of a third experimental arm at 100-300 pooled experimental arm participants, and, if we add a fourth experimental arm, it is added at 350 pooled experimental arm participants. We use the dynamic balancing scheme to assign study arms at a 1:1 allocation and close each experimental arm after it contains 200 participants.

In Tables A.1-A.2, we tabulate the average deviations and variances across 10,000 trials for platform trials with 3 or 4 total experimental arms and with 2 stratification factors. We perform these computations with a window size of 25 participants. We display the results for the allocation ratio from one of the two initial experimental arms, but results are similar for both initial arms, and conclusions are also similar when using 3 stratification factors rather

Table A.1: Mean Deviances and Variances (in 10^{-4}) of Allocation Ratio Across 10,000 Simulated Trials with One Added Experimental Arm (Arm 3), Two Stratification Factors and a Window Size of 25

		Deviation Pre-Arm 3	Variance Pre-Arm 3	Deviation Post-Arm 3	Variance Post-Arm 3
Add Arm 3 at 100	Combine	17.8	11.76	9.89	6.69
	Split	17.33	11.85	21.63	4.99
Add Arm 3 at 200	Combine	4.09	2.60	2.64	1.79
	Split	4.16	2.62	6.33	1.40
Add Arm 3 at 300	Combine	1.70	1.11	1.24	0.82
	Split	1.76	1.13	2.91	0.63

than 2.

Before the addition of any new experimental arm, the two approaches have similar deviation and variance (as expected, because the methods are identical before a new arm). The split data approach has consistently larger deviations than the combine data approach directly after the addition of a new experimental arm independent of the timing of the third experimental arm. However, with this window size of 25, we do observe that the split data approach has a smaller variance after the addition of a new arm.

With this window size, we see the split data approach suffers from the disruption more the later the third experimental arm is added: when added after at least 200 pooled experimental arm subjects, the deviation from the desired allocation ratio after the addition of a fourth experimental arm remains higher than the deviation before adding the third arm.

Notably, these represent small differences on the order of 10^{-4} or smaller and may not practically impact trials depending on the timing of interim analyses. We advocate for the combine data approach based on the deviations from the desired allocation ratio in this simulation study, but we acknowledge that the ease of implementation of the split data approach may be more appealing for those implementing these randomization systems.

Table A.2: Mean Deviations and Variances (in 10^{-4}) of Allocation Ratio Across 10,000 Simulated Trials with Two Added Experimental Arms (Arms 3 and 4), Two Stratification Factors and a Window Size of 25

		D.† Pre-Arm 3	V.‡ Pre-Arm 3	D. Post-Arm 3	V. Post-Arm 3	D. Post-Arm 4	V. Post-Arm 4
Add Arm 3 at 100	Combine	17.2	11.7	9.14	6.66	1.12	0.87
	Split	17.9	11.8	22.2	5.03	5.67	0.69
Add Arm 3 at 200	Combine	4.07	2.61	2.69	1.81	0.87	0.69
	Split	4.17	2.63	6.35	1.38	4.67	0.55
Add Arm 3 at 300	Combine	1.83	1.12	1.25	0.83	0.73	0.56
	Split	1.80	1.14	2.94	0.64	3.71	0.45

† Deviation, ‡ Variance

A.3 Characterizing Platform Trial Efficiency with Differing Arm Eligibility

Platform trials typically specify a maximum sample size for each experimental arm, but the size of the standard of care arm depends on the amount of common eligibility shared between experimental arms and on the allocation ratio. Even under equal allocation, when subjects are not all eligible to all (K) experimental arms, the probability of a given subject being randomized to the standard of care arm increases from $\frac{1}{K+1}$, the probability when all subjects are eligible to all experimental arms. We describe the relative efficiency of a platform trial with varying eligibility criteria across arms by comparing the probability of randomization to the standard of care against that probability in a trial with non-varying eligibility.

Platform trial efficiency arises via the size of the common standard of care arm, so efficiency can also be assessed by evaluating relative sizes of the standard of care arm. This evaluation requires specification of the planned sample size for each experimental arm in addition to the specification of arm eligibility probabilities, and we reach similar conclusions using the relative sample sizes as the relative probabilities.

In this section, we derive a general expression for the probability of randomizing to the standard of care arm for arbitrary arm eligibility and number of experimental arms. We

assume this calculation is done under the simplified scenario that all experimental arms are started and ending at the same time - a more general framework allowing for adding and dropping arms is possible, if needed.

With K experimental arms, there are $2^K - 1$ possible combinations (sets) of eligibility to those arms based on the power set and the requirement that subjects be eligible for at least one experimental arm. Consider organizing the eligibility combinations by their "order," i.e., the total number of eligible arms in a given combination (the cardinality of a given combination set).

Let $F_i^{(j)}$ be the ith combination of eligible arms of the jth order with $j = 1, \ldots, K$ and $i = 1, \ldots, \binom{K}{j}$.

For example, for $K = 3$ experimental arms:

$$
\begin{aligned}
&F_1^{(1)} = \{E_1\}, \quad F_2^{(1)} = \{E_2\}, \quad F_3^{(1)} = \{E_3\} \\
&F_1^{(2)} = \{E_1, E_2\}, \quad F_2^{(2)} = \{E_1, E_3\}, \quad F_3^{(2)} = \{E_2, E_3\} \\
&F_1^{(3)} = \{E_1, E_2, E_3\}
\end{aligned}
$$

For a new participant, define the random variable $B_i^{(j)}$ to indicate whether $F_i^{(j)}$ contains exactly all experimental arms to which the new participant is eligible and no more (in other words, an indicator that $F_i^{(j)}$ is the largest set by cardinality in the power set of the new participant's eligible experimental arms). Using the example above, if a new participant is eligible to E_1 and E_3 but not E_2, then we observe $B_2^{(2)} = 1$ and $B_i^{(j)} = 0$ for all $i \neq 2$ and $j \neq 2$.

Let $q_i^{(j)} = P(B_i^{(j)} = 1)$, with $\sum_{j=1}^{K} \sum_{i=1}^{\binom{K}{j}} q_i^{(j)} = 1$. The collection of $B_i^{(j)}$ over all i, j form a random vector drawn from a multinomial distribution with 2^{K-1} categories, exactly one trial and category probabilities given by $q_i^{(j)}$. We assume that new participants are drawn independently and identically from this distribution.

For a given new participant under iid sampling, by the law of total probability,

$$\begin{aligned} P(\text{randomize to } C) &= \sum_{j=1}^{K} \sum_{i=1}^{\binom{K}{j}} P(\text{randomize to } C \cap B_i^{(j)} = 1) \\ &= \sum_{j=1}^{K} \sum_{i=1}^{\binom{K}{j}} P(\text{randomize to } C | B_i^{(j)} = 1) P(B_i^{(j)} = 1). \end{aligned}$$

When balance is achieved, assuming a planned 1:1 allocation (this can be extended to other ratios), for each $j = 1, \ldots, K$ we have $P(\text{randomize to } C | B_i^{(j)} = 1) = \frac{1}{j+1}$ (every eligible study arm is equally likely). Then we can write

$$P(\text{randomize to } C) = \sum_{j=1}^{K} \frac{1}{j+1} \sum_{i=1}^{\binom{K}{j}} q_i^{(j)}.$$

A multi-arm trial with non-varying eligibility would require $q_1^{(K)} = 1$ (thus, implicitly, $q_i^{(j)} = 0$ for all other combinations $j = 1, \ldots, K-1$ and $i = 1, \ldots, \binom{K}{j}$). This results in $P(\text{randomize to } C) = \frac{1}{K+1}$. As we deviate from $q_1^{(K)} = 1$, the platform trial becomes less efficient, as the proportion randomized to the standard of care arm increases from $\frac{1}{K+1}$.

Appendix B

APPENDIX FOR CHAPTER 2

B.1 Estimation and Inference of Mixture Cure Model Parameters Improve with Longer Follow-Up

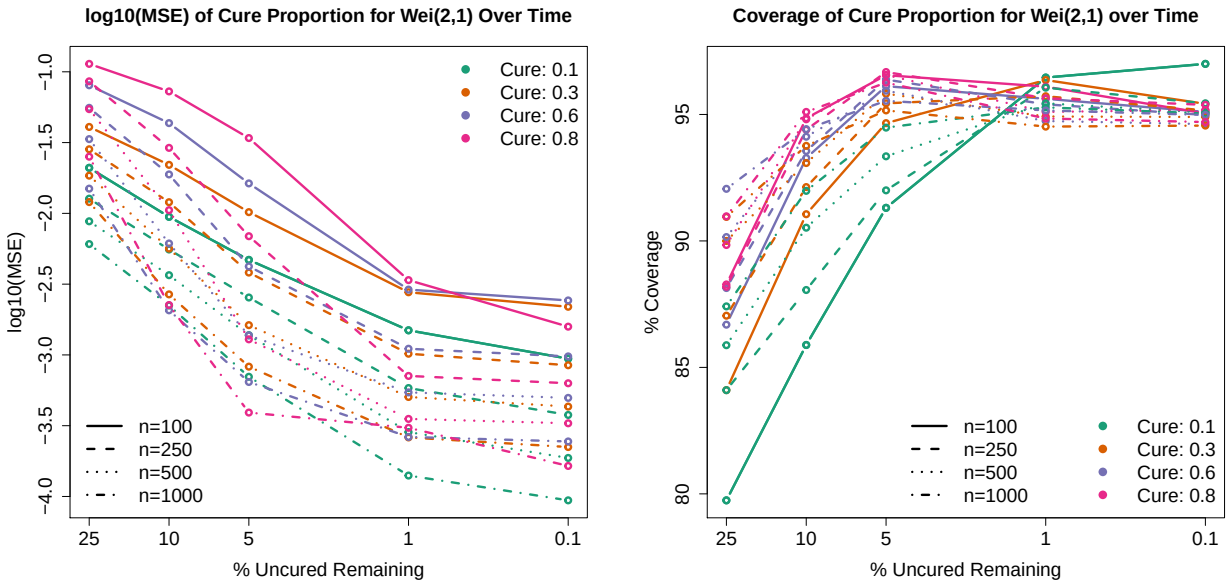
This section expands on the results of Yu et al. [22] and assesses the hypothesis that estimation and inference need not be limited to the setting in which we believe no uncured subjects remain at the end of the study, or $\tau_{F_0} < \tau_G$. We perform simulations to study the mean-squared error and coverage of 95% confidence intervals when fitting mixture cure models as follow-up time increases. (Note that here emphasis lies in the patterns across follow-up times.) These results are drawn from the simulation study described in Section 2.3.

We fit correctly-specified mixture cure models via maximum-likelihood estimation and compute the mean-squared error and coverage of Wald-based 95% confidence intervals for each model parameter estimate.

Researchers often focus on estimation and inference on the cure proportion, so we provide results of this parameter from a Weibull(2,1) mixture cure generating distribution in this section. Figures B.1a and B.1b demonstrate clear improvements in estimation and inference over time and with increased sample size. (We do notice that at small sample sizes of $n = 100$, confidence intervals may be larger than needed with conservative coverage.)

We emphasize that a similar trend of improvement over time remains across other distribution settings and across other model parameters. The shape and magnitude of improvement may differ, but mean-squared error decreases and confidence interval coverage approaches the nominal rate.

In addition to supporting the claim that researchers can compute statistics with nominal



(a) The mean-squared error for the cure fraction estimate in Weibull(2,1) mixture cure distributions decreases with longer follow-up times and with sample size.

(b) The coverage of confidence intervals for the cure fraction estimate in Weibull(2,1) mixture cure distributions approach nominal levels with longer follow-up times and with sample size.

Figure B.1: Results to Assess the Properties of Cure Fraction Estimation for Weibull(2,1) Mixture Cure Distributions with Longer Follow-Up Time.

coverage and low mean-squared error without $\tau_{F_0} < \tau_G$ holding, figures such as these can also quantify the length of time most helpful in a trade-off between follow-up time and statistical validity (as measured by mean-squared error and coverage). For the Weibull model, it seems that mean-squared error and coverage improve dramatically from 25% uncured remaining to 5% uncured remaining. However, with fewer than 1% uncured remaining, additional follow-up seems to show diminishing gains. As such, we might reasonably consider 1% or lower uncured remaining to be strong evidence for sufficient follow-up, and we might be skeptical of the follow-up when many more than 5% of uncured remain in the study.

We believe that these results may support a shift for the discussion of "sufficient" follow-up time being a single number for a given study to being a range of times depending on a researcher's decision on such a trade-off.

B.2 Regularity Conditions for Asymptotic Properties

Here, we list the regularity conditions necessary for Theorems 1 and 2. We invoke van der Vaart [23] Theorems 5.41 and 5.42.

Theorem 5.41 establishes asymptotic normality of maximum likelihood estimators (in the view of maximum likelihood estimation as Z-estimation with the score equations) and 5.42 ensures consistent roots exist in this setting. Define $\psi_\eta(x) \equiv \dot{l}_n(\eta) = \frac{d}{\eta_i} l(\eta; y, \delta)$ and $m_\eta(x) \equiv l(\eta; y, \delta)$ (with $X = (Y, \Delta)$).

The regularity conditions needed for consistency and asymptotic normality of maximum likelihood estimation are as follows.

R1. The parameter of interest η_0 is a local maximum of $\mathbb{E}m_\eta(Y, \Delta)$ and $0 = \mathbb{E}\psi_{\eta_0}(Y, \Delta)$.

R2. The score equations $\psi_\eta(y, \delta)$ are twice continuously differentiable for every (y, δ).

R3. The second moment $\mathbb{E}||\psi_{\eta_0}(Y, \Delta)||^2 < \infty$.

R4. The matrix $\mathbb{E}\dot{\psi}_{\eta_0}(Y, \Delta)$ exists and is nonsingular.

R5. The second-order derivatives of the score equations are dominated by a fixed integrable function $\ddot{\psi}(y, \delta)$ for every η in a neighborhood of η_0.

R6. $S(t)$ is uniformly bounded away from 0 when $t \in [0, \tau]$.

By van der Vaart [23], these conditions are sufficient in the setting of maximum likelihood estimation from common parametric families but are likely stricter than necessary. The Exponential, Gamma, Weibull and Log-Logistic families used with Uniform censoring in this report, as popular choices in applied survival analysis, clearly meet these conditions due to their smoothness. The final condition should hold in real-world conditions with finite follow-up.

In addition, for consistency and asymptotic normality of $\hat{r}_n$, we do require $\eta \mapsto r(\eta)$ be continuous and differentiable at η_0. But the listed conditions are frequently sufficient for this because $r(\eta)$ is a smooth transformation of η in these parametric models.

These theorems do not preclude the existence of other, inconsistent roots. We have not experienced this in practice, but if needed to address multiple roots, van der Vaart [23] suggests considering a preliminary consistent estimate and then choose the closest root. In particular, we recommend the root closest to $\tilde{\pi}_n = \tilde{S}_{KM,n}(\tau)$ (where $\tilde{S}_{KM,n}(\tau)$ is the estimate of survival probability at the end of the study by Kaplan-Meier) because this is a simple, consistent non-parametric estimator of π_0. [20]

B.3 Extended Simulation Results

In this paper, we present the results from the Weibull(2,1) mixture cure distribution, but patterns are similar across distributions. We report $\gamma_\pi(u)$ in Table B.1 for each procedure (RECeUS-AIC, $\hat{\alpha}_n$ and $\tilde{\alpha}_n$) across the varying $\pi \in [0, 0.8]$ and $u \in \{0.25, 0.1, 0.05, 0.01, 0.001\}$. In addition to $\gamma_\pi(u)$, the RECeUS procedure also includes the true ratio r for additional context: if r is small, then we expect $\gamma_\pi(u)$ to be large and the opposite if r is large. We summarize the key results in the main text.

Table B.1: Rates for Concluding Sufficient Follow-Up ($\gamma_\pi(u)$) by Procedure, Percentage of Uncured Remaining, Sample Size and Cure Fraction in 5000 Simulations

		RECeUS-AIC: $\hat{r}_n < 0.05$ and $\hat{\pi}_n > 0.025$					$\hat{\alpha}_n < 0.05^*$					$\tilde{\alpha}_n < 0.05^*$				
		% Uncured Remaining†					% Uncured Remaining†					% Uncured Remaining†				
Cure Fraction	Sample Size	25	10	5	1	0.1	25	10	5	1	0.1	25	10	5	1	0.1
0%	100	0.122	0.084	0.055	0.029	0.002	0.491	0.473	0.413	0.295	0.133	0.043	0.044	0.040	0.033	0.018
	250	0.067	0.037	0.026	0.012	0.000	0.526	0.505	0.467	0.344	0.193	0.029	0.024	0.032	0.028	0.018
	500	0.038	0.017	0.005	0.001	0.000	0.543	0.517	0.522	0.380	0.227	0.016	0.018	0.021	0.021	0.017
	1000	0.012	0.002	0.000	0.000	0.000	0.540	0.538	0.521	0.431	0.279	0.010	0.011	0.015	0.017	0.018
	True Ratio	1.0000	1.0000	1.0000	1.0000	1.0000										
10%	100	0.132	0.101	0.107	0.322	0.811	0.523	0.542	0.571	0.684	0.810	0.049	0.052	0.064	0.110	0.220
	250	0.070	0.062	0.066	0.373	0.928	0.552	0.554	0.581	0.683	0.818	0.026	0.034	0.041	0.082	0.193
	500	0.047	0.035	0.039	0.387	0.969	0.535	0.552	0.585	0.671	0.809	0.019	0.024	0.025	0.060	0.164
	1000	0.019	0.005	0.005	0.339	0.998	0.543	0.570	0.589	0.649	0.782	0.013	0.014	0.015	0.042	0.116
	True Ratio	0.7262	0.5409	0.3292	0.0599	0.0052										
30%	100	0.149	0.146	0.215	0.589	0.922	0.549	0.583	0.631	0.750	0.851	0.064	0.060	0.083	0.145	0.250
	250	0.086	0.078	0.149	0.699	0.986	0.572	0.583	0.621	0.732	0.838	0.034	0.040	0.053	0.114	0.220
	500	0.064	0.071	0.098	0.807	0.999	0.573	0.589	0.594	0.699	0.822	0.024	0.027	0.035	0.077	0.182
	1000	0.044	0.038	0.067	0.920	1.000	0.570	0.599	0.603	0.680	0.811	0.017	0.017	0.020	0.050	0.140
	True Ratio	0.4692	0.2820	0.1406	0.0208	0.0017										
60%	100	0.193	0.243	0.325	0.650	0.907	0.589	0.629	0.665	0.778	0.856	0.090	0.090	0.110	0.177	0.272
	250	0.135	0.169	0.277	0.765	0.965	0.590	0.605	0.645	0.747	0.849	0.053	0.056	0.071	0.141	0.234
	500	0.091	0.120	0.255	0.866	0.991	0.585	0.602	0.637	0.725	0.837	0.035	0.039	0.053	0.107	0.216
	1000	0.071	0.092	0.222	0.923	0.999	0.571	0.589	0.613	0.702	0.821	0.024	0.023	0.035	0.071	0.180
	True Ratio	0.3065	0.1641	0.0756	0.0105	0.0009										
80%	100	0.206	0.253	0.339	0.584	0.820	0.613	0.642	0.688	0.776	0.874	0.111	0.121	0.142	0.199	0.315
	250	0.204	0.260	0.383	0.727	0.957	0.606	0.626	0.690	0.777	0.862	0.080	0.096	0.113	0.176	0.271
	500	0.142	0.210	0.364	0.826	0.982	0.585	0.635	0.659	0.766	0.853	0.055	0.067	0.085	0.143	0.234
	1000	0.102	0.153	0.347	0.901	0.996	0.594	0.614	0.633	0.733	0.846	0.041	0.046	0.048	0.104	0.218
	True Ratio	0.2490	0.1284	0.0578	0.0079	0.0006										

* The $\hat{\alpha}_n$ and $\tilde{\alpha}_n$ procedures also use as screening that the last observation time must be a censored observation.

† In this setting, the exact follow-up times generating the data were 1.25, 1.5, 1.75, 2.25 and 2.75. These are the follow-up times corresponding to the percentiles rounded to the nearest quarter in order to more closely represent clinical trial data.

We can also compare RECeUS-AIC to RECeUS fit with the correctly-specified model. We include this comparison for the Weibull(2,1) mixture-cure distribution in Table B.2. As the results between RECeUS-AIC and RECeUS-Weibull differ, we can immediately conclude that AIC does not always correctly select the Weibull model under these settings. Correct model specification does have improved behavior, but the RECeUS-AIC procedure has acceptable properties and model selection reduces sensitivity to model misspecification (over choosing a single model a priori).

Table B.2: Rates for Concluding Sufficient Follow-Up ($\gamma_\pi(u)$) in RECeUS-Weibull and RECeUS-AIC by Percentage of Uncured Remaining, Sample Size and Cure Fraction in 5000 Simulations

		RECeUS-Weibull*					RECeUS-AIC				
		% Uncured Remaining†					% Uncured Remaining†				
Cure Fraction	Sample Size	25%	10%	5%	1%	0.1%	25%	10%	5%	1%	0.1%
0%	100	0.024	0.013	0.019	0.036	0.003	0.122	0.084	0.055	0.029	0.002
	250	0.001	0.000	0.001	0.003	0.000	0.067	0.037	0.026	0.012	0.000
	500	0.000	0.000	0.000	0.000	0.000	0.038	0.017	0.005	0.001	0.000
	1000	0.000	0.000	0.000	0.000	0.000	0.012	0.002	0.000	0.000	0.000
10%	100	0.039	0.042	0.075	0.456	0.964	0.132	0.101	0.107	0.322	0.811
	250	0.002	0.003	0.010	0.424	0.998	0.070	0.062	0.066	0.373	0.928
	500	0.000	0.000	0.000	0.386	1.000	0.047	0.035	0.039	0.387	0.969
	1000	0.000	0.000	0.000	0.320	1.000	0.019	0.005	0.005	0.339	0.998
30%	100	0.102	0.126	0.242	0.790	0.996	0.149	0.146	0.215	0.589	0.922
	250	0.020	0.028	0.114	0.880	1.000	0.086	0.078	0.149	0.699	0.986
	500	0.001	0.003	0.036	0.951	1.000	0.064	0.071	0.098	0.807	0.999
	1000	0.000	0.000	0.006	0.988	1.000	0.044	0.038	0.067	0.920	1.000
60%	100	0.236	0.302	0.453	0.876	0.996	0.193	0.243	0.325	0.650	0.907
	250	0.098	0.159	0.362	0.953	1.000	0.135	0.169	0.277	0.765	0.965
	500	0.027	0.073	0.305	0.990	1.000	0.091	0.120	0.255	0.866	0.991
	1000	0.003	0.022	0.223	0.999	1.000	0.071	0.092	0.222	0.923	0.999
80%	100	0.412	0.459	0.580	0.867	0.988	0.206	0.253	0.339	0.584	0.820
	250	0.243	0.337	0.538	0.937	1.000	0.204	0.260	0.383	0.727	0.957
	500	0.125	0.228	0.504	0.985	1.000	0.142	0.210	0.364	0.826	0.982
	1000	0.038	0.127	0.454	0.998	1.000	0.102	0.153	0.347	0.901	0.996

* The RECeUS-Weibull procedure employs the correctly-specified Weibull model for estimation.

† In this setting, the exact follow-up times generating the data were 1.25, 1.5, 1.75, 2.25 and 2.75. These are the follow-up times corresponding to the percentiles rounded to the nearest quarter in order to more closely represent clinical trial data.

Appendix C

APPENDIX FOR CHAPTER 3

C.1 Verifying Monotonic Information Growth

Shoben et al. [39] discuss conditions for nonmonotonic information growth and its impact on sequential monitoring. They note that highly correlated data - as we might expect from N-of-1 trials - can lead to nonmonotonic information growth. In this section, we verify that the proposed approach of Section 3.2.2 does not suffer from nonmonotonic growth.

First, we introduce matrix notation for the analysis model for analyzing one N-of-1 trial. Let Y_{ij} be the outcome measurement in treatment block $i = 1, \ldots, B$ and treatment period $j = 1, \ldots, J$, and let Z_{ij} be a binary indicator for the treatment variable. Recall that $Z_{ij} = 1$ exactly $J/2$ times per block and 0 otherwise by the assumption of balanced treatment blocks. Denote X_i as the design matrix with the form

$$X_i = \begin{bmatrix} 1 & Z_{i1} \\ \vdots & \vdots \\ 1 & Z_{iJ} \end{bmatrix}.$$

The analysis model can then be written in matrix form as

$$\mathbb{E}[Y_i|\gamma_i, \beta] = \gamma_i 1_{J\times 1} + X_i\beta,$$

Where γ_i is the (scalar) random intercept by treatment block and the parameter vector $\beta = (\alpha, \theta)^T$ contains α the fixed intercept and θ the true ATE for the subject of the N-of-1 trial. The matrix $1_{J\times 1}$ is a matrix of dimension $J \times 1$ with all elements equal to 1. The matrix Y_i is the vector $(Y_{i1}, \ldots, Y_{iJ})^T$.

Let σ^2 be the error variance, and we assume $\gamma_i \sim N(0, g^2)$ with the errors being independent of the random intercepts. Thus, we can write the covariance matrix for Y_i as

$$\Sigma \equiv Var(Y_i) = \sigma^2 I_{J\times J} + g^2 1_{J\times J},$$

Where $I_{J\times J}$ indicates the identity matrix of dimension $J \times J$.

If the study ends at treatment block m, then the estimate of β is (by the assumptions for independent treatment blocks)

$$\hat{\beta}^{(m)} = \left(\sum_{i=1}^{m} X_i^T \Sigma^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{m} X_i^T \Sigma^{-1} Y_i\right)$$

with variance

$$Var(\hat{\beta}^{(m)}) = \left(\sum_{i=1}^{m} X_i^T \Sigma^{-1} X_i\right)^{-1}.$$

The approximate information, assuming large sample theory, is the inverse of the variance.

$$I(\hat{\beta}^{(m)}) = \sum_{i=1}^{m} X_i^T \Sigma^{-1} X_i$$

We can focus on the information contribution from an arbitrary treatment block i,

$$I^{(i)} = X_i^T \Sigma^{-1} X_i = X_i^T (\sigma^2 I_{J\times J} + g^2 1_{J\times J})^{-1} X_i.$$

By the Sherman-Morrison formula we can write

$$(\sigma^2 I_{J\times J} + g^2 1_{J\times J})^{-1} = \frac{1}{\sigma^2} I_{J\times J} - \frac{g^2}{\sigma^2(\sigma^2 + Jg^2)} 1_{J\times J}.$$

Then we have

$$\begin{aligned} I^{(i)} &= \frac{1}{\sigma^2} X_i^T X_i - \frac{g^2}{\sigma^2(\sigma^2 + Jg^2)} X_i^T 1_{J\times J} X_i \\ &= \frac{1}{\sigma^2} \begin{bmatrix} J & \sum_{j=1}^{J} Z_{ij} \\ \sum_{j=1}^{J} Z_{ij} & \sum_{j=1}^{J} Z_{ij}^2 \end{bmatrix} - \frac{g^2}{\sigma^2(\sigma^2 + Jg^2)} \begin{bmatrix} J^2 & J\sum_{j=1}^{J} Z_{ij} \\ J\sum_{j=1}^{J} Z_{ij} & (\sum_{j=1}^{J} Z_{ij})^2 \end{bmatrix}. \end{aligned}$$

Let $T_i = \sum_{j=1}^{J} Z_{ij}$. Because Z_{ij} is a binary variable, it is always true that $Z_{ij} = Z_{ij}^2$, so $T_i = \sum_{j=1}^{J} Z_{ij}^2$ as well. Then we can write

$$I^{(i)} = \frac{1}{\sigma^2 + Jg^2} \begin{bmatrix} J & T_i \\ T_i & \frac{(\sigma^2+Jg^2)T_i - g^2T_i^2}{\sigma^2} \end{bmatrix}.$$

The information contribution for the $\hat{\theta}$ component of $\hat{\beta}^{(m)}$ for treatment block i is then

$$\begin{aligned} I(\hat{\theta})^{(i)} &= \frac{1}{\sigma^2 + Jg^2} \frac{(\sigma^2 + Jg^2)T_i - g^2T_i^2}{\sigma^2} - \left(\frac{T_i}{\sigma^2 + Jg^2}\right)\left(\frac{\sigma^2 + Jg^2}{J}\right)\left(\frac{T_i}{\sigma^2 + Jg^2}\right) \\ &= \frac{1}{\sigma^2 + Jg^2}\left[\frac{(\sigma^2 + Jg^2)T_i - g^2T_i^2}{\sigma^2} - T_i^2/J\right] \\ &= \frac{1}{\sigma^2}(T_i - T_i^2/J) \end{aligned}$$

When we assume balanced treatment block, in fact, $T_i = \frac{J}{2}$ and $T_i^2 = \frac{J^2}{4}$ for all i. So

$$I(\hat{\theta})^{(i)} = \frac{1}{\sigma^2}\left(\frac{J}{2} - \frac{J}{4}\right) = \frac{J/2}{\sigma^2}$$

Thus the information contribution from each block $I(\hat{\theta})^{(i)} > 0$, so information growth is monotonic. Further, because $I(\hat{\theta})^{(i)}$ is independent of the index i, we can see that the information increases linearly with the number of treatment blocks with rate $\frac{J/2}{\sigma^2}$.

For unbalanced treatment blocks, as long as both treatments appear at least once within each block (i.e., $0 < T_i < J$ for all i), the derivation would be similar but the rate may not be independent of index i.

Throughout this derivation, we assume the number of periods per block is a fixed number J - the general setting with number of periods possibly differing by block may not necessarily have monotonic information growth. This could occur, for example, due to missing data.

C.2 Bias-Adjusted Point Estimators

Many authors have documented the bias of estimates following a group sequential trial. Intuitively, this arises because trials will more frequently end because of an overestimate

of the effect size, while estimates closer to the null do not frequently lead to stopping. Researchers have proposed several methods for reducing or eliminating this bias. We note that some authors contend that this bias is not a problem in larger group-sequential trials, but in the N-of-1 setting this adjustment seems critical.

In a series of N-of-1 trials, we anticipate individual trials may differ in their monitoring, so the bias from each trial estimate may differ.

We study two types of estimators for de-biasing the effect size estimate from the linear mixed model in addition to the naive estimator of the ATE $\hat{\theta}$. Let $\hat{\vartheta}$ be the observed naive point estimate of the ATE, $M \in \{2, \ldots, B\}$ be the random variable for the treatment block at which the study stopped and m be its observed value.

1. Bias-Adjusted Mean (BAM), the parameter value $\tilde{\theta}^{\text{BAM}}$ that solves

$$\mathbb{E}[\hat{\theta}; \theta = \tilde{\theta}^{\text{BAM}}] = \hat{\vartheta}.$$

2. Median Unbiased Estimate (MUE), the parameter value $\tilde{\theta}^{\text{MUE}}$ that solves

$$P((M, \hat{\theta}) >_o (m, \hat{\vartheta}); \theta = \tilde{\theta}^{\text{MUE}}) = \frac{1}{2}.$$

   This estimator requires specifying an ordering for $>_o$, and researchers have previously used orderings based on the analysis stage (AT), the sample mean (SM) and the likelihood ratio statistic (LR).

The density of the naive estimator is considered to be the sequential density derived by Armitage, McPherson and Rowe. [52] This appeals to the asymptotic normality of the naive estimator. The probability of the MUE is taken with respect to this density, and the expectation of the BAM is taken with respect to this density and averaging over the random variable of treatment block.

By construction, we observe that each of these estimators is a member of the class of Z-estimators. To verify the asymptotic normality of these estimators, we employ van der

Vaart theoerm 5.41. [53] Because the estimating equations are smooth functions of the Gaussian density, they satisfy the differentiability conditions needed. As such, we employ the expressions for the asymptotic standard error for the estimates of standard error.

C.3 Additional Simulation Results

C.3.1 Bias and Mean-Squared Error for One Sequentially-Monitored N-of-1 Trial

In Tables C.1-C.2, we present the full bias and mean-squared error results with 2 looks at the data.

C.3.2 Bias for A Series of Sequentially-Monitored N-of-1 Trials

Figure C.1 displays the bias of the combined point estimator at different effect sizes, number of blocks and periods per block with all constituent trials using an OBF boundary shape and 2 looks at the data. The difference in bias between 5 trials and 10 trials within a series does not appear to be importantly different when we summarize 100-200 simulated series. It is possible that, with a larger number of simulations, we may be able to detect patterns in the bias across numbers of trials within a series.

C.3.3 Bias and Mean-Squared Error for A Series of Sequentially-Monitored N-of-1 Trials, Preliminary Results Allowing for Boundary Shapes to Vary within the Series

In this section, we present preliminary results when allowing for constituent trials in a series to have either a Pocock boundary shape or an OBF shape (for either shape, there remain 2 looks at the data). We vary the number from 0% to 100% OBF constituent trials (i.e., series comprised of only Pocock trials to only OBF trials) with 5 or 10 N-of-1 trials in a series. We report across similar settings as in Section C.3.2 and continue to summarize across 200 series with 5 constituent trials or 100 series with 10 constituent trials.

Figures C.2 and C.3 display the preliminary results. We see series with more OBF constituent trials have consistently smaller MSE across non-zero effect sizes, but the difference

Table C.1: Bias across 1,000 N-of-1 trials of candidate point estimators with 2 looks

Blocks	Periods	Effect	MUE.SM	MUE.AT	MUE.LR	BAM	Naïve
		0	0.01	0.01	0.01	0.01	0.01
		0.5	0.04	0.04	0.04	0.02	0.05
		0.75	0.05	0.06	0.05	0.03	0.07
		0.9	0.06	0.06	0.06	0.03	0.08
	2	1	0.06	0.06	0.06	0.03	0.08
		1.1	0.06	0.06	0.06	0.03	0.09
		1.25	0.05	0.05	0.05	0.03	0.09
		1.5	0.05	0.05	0.05	0.03	0.09
		3	0.02	0.02	0.02	0.01	0.02
13		0	0.01	0.01	0.01	0.01	0.01
		0.5	0.03	0.04	0.03	0.02	0.05
		0.75	0.04	0.05	0.04	0.03	0.07
		0.9	0.04	0.04	0.04	0.03	0.07
	6	1	0.04	0.03	0.04	0.02	0.06
		1.1	0.03	0.03	0.03	0.02	0.05
		1.25	0.02	0.02	0.02	0.01	0.04
		1.5	0.02	0.02	0.02	0.01	0.02
		3	0.02	0.02	0.02	0.02	0.02
		0	-0.01	-0.01	-0.01	-0.01	-0.01
		0.5	0.00	0.01	0.00	-0.01	0.01
		0.75	0.02	0.02	0.02	0.00	0.04
		0.9	0.02	0.03	0.02	0.00	0.05
	2	1	0.03	0.03	0.03	0.01	0.06
		1.1	0.03	0.02	0.03	0.00	0.06
		1.25	0.02	0.02	0.02	0.00	0.06
		1.5	0.01	0.01	0.01	-0.01	0.03
		3	-0.01	-0.01	-0.01	-0.01	-0.01
26		0	0.00	0.00	0.00	0.00	0.00
		0.5	0.02	0.03	0.02	0.01	0.04
		0.75	0.02	0.02	0.02	0.01	0.04
		0.9	0.01	0.01	0.01	0.00	0.02
	6	1	0.00	0.00	0.00	-0.01	0.01
		1.1	0.00	0.00	0.00	-0.01	0.00
		1.25	0.00	0.00	0.00	-0.01	0.00
		1.5	0.00	0.00	0.00	0.00	0.00
		3	0.00	0.00	0.00	0.00	0.00

Table C.2: Mean-squared error across 1,000 N-of-1 trials of candidate point estimators with 2 looks

Blocks	Periods	Effect	MUE.SM	MUE.AT	MUE.LR	BAM	Naïve
13	2	0	0.17	0.17	0.17	0.16	0.17
		0.5	0.18	0.18	0.18	0.17	0.19
		0.75	0.19	0.20	0.19	0.18	0.20
		0.9	0.19	0.20	0.19	0.19	0.21
		1	0.20	0.20	0.20	0.19	0.21
		1.1	0.20	0.21	0.20	0.20	0.21
		1.25	0.21	0.21	0.21	0.20	0.21
		1.5	0.21	0.22	0.21	0.21	0.21
		3	0.27	0.27	0.27	0.28	0.27
	6	0	0.06	0.06	0.06	0.05	0.06
		0.5	0.07	0.07	0.07	0.07	0.08
		0.75	0.07	0.08	0.07	0.07	0.08
		0.9	0.08	0.08	0.08	0.08	0.08
		1	0.08	0.08	0.08	0.08	0.07
		1.1	0.08	0.08	0.08	0.08	0.08
		1.25	0.09	0.09	0.09	0.09	0.08
		1.5	0.09	0.09	0.10	0.10	0.09
		3	0.10	0.10	0.10	0.10	0.10
26	2	0	0.08	0.08	0.08	0.08	0.08
		0.5	0.09	0.09	0.09	0.08	0.09
		0.75	0.09	0.10	0.09	0.09	0.10
		0.9	0.10	0.10	0.10	0.10	0.11
		1	0.10	0.10	0.10	0.10	0.11
		1.1	0.11	0.11	0.11	0.10	0.11
		1.25	0.11	0.11	0.11	0.11	0.11
		1.5	0.12	0.13	0.12	0.12	0.11
		3	0.15	0.15	0.15	0.15	0.15
	6	0	0.03	0.03	0.03	0.03	0.03
		0.5	0.04	0.04	0.04	0.04	0.04
		0.75	0.04	0.04	0.04	0.04	0.04
		0.9	0.04	0.05	0.04	0.05	0.04
		1	0.05	0.05	0.05	0.05	0.04
		1.1	0.05	0.05	0.05	0.05	0.05
		1.25	0.05	0.05	0.05	0.05	0.05
		1.5	0.05	0.05	0.05	0.05	0.05
		3	0.05	0.05	0.05	0.05	0.05

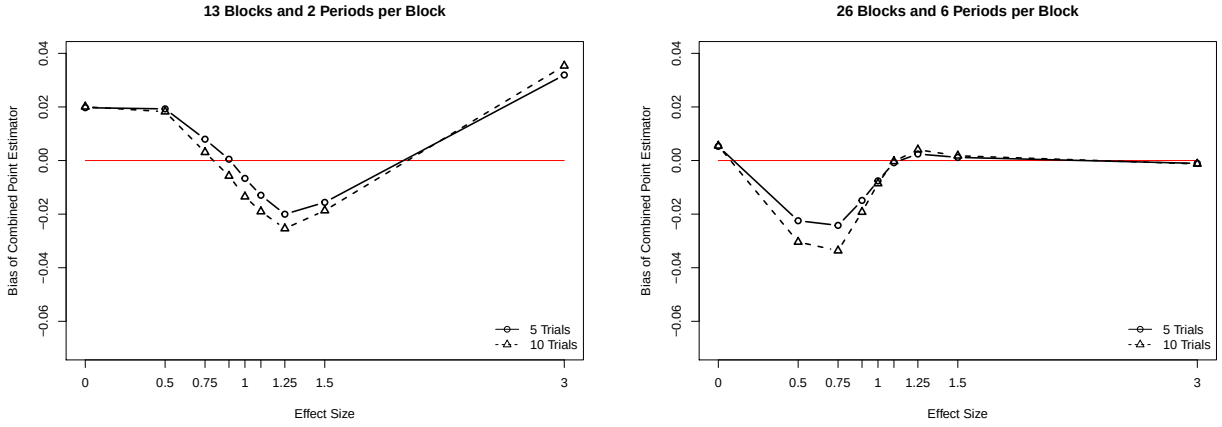


Figure C.1: Bias for the combined point estimator from a series of 5 or 10 sequentially-monitored N-of-1 trials with varying numbers of blocks and periods per block and an OBF boundary shape with 2 looks at the data. Bias of 0 indicated in red.

between the shapes decreases with more constituent trials, treatment blocks and periods per block. We do not see a consistent pattern for bias. Similar to above, a larger number of simulations may provide the precision to detect other patterns.

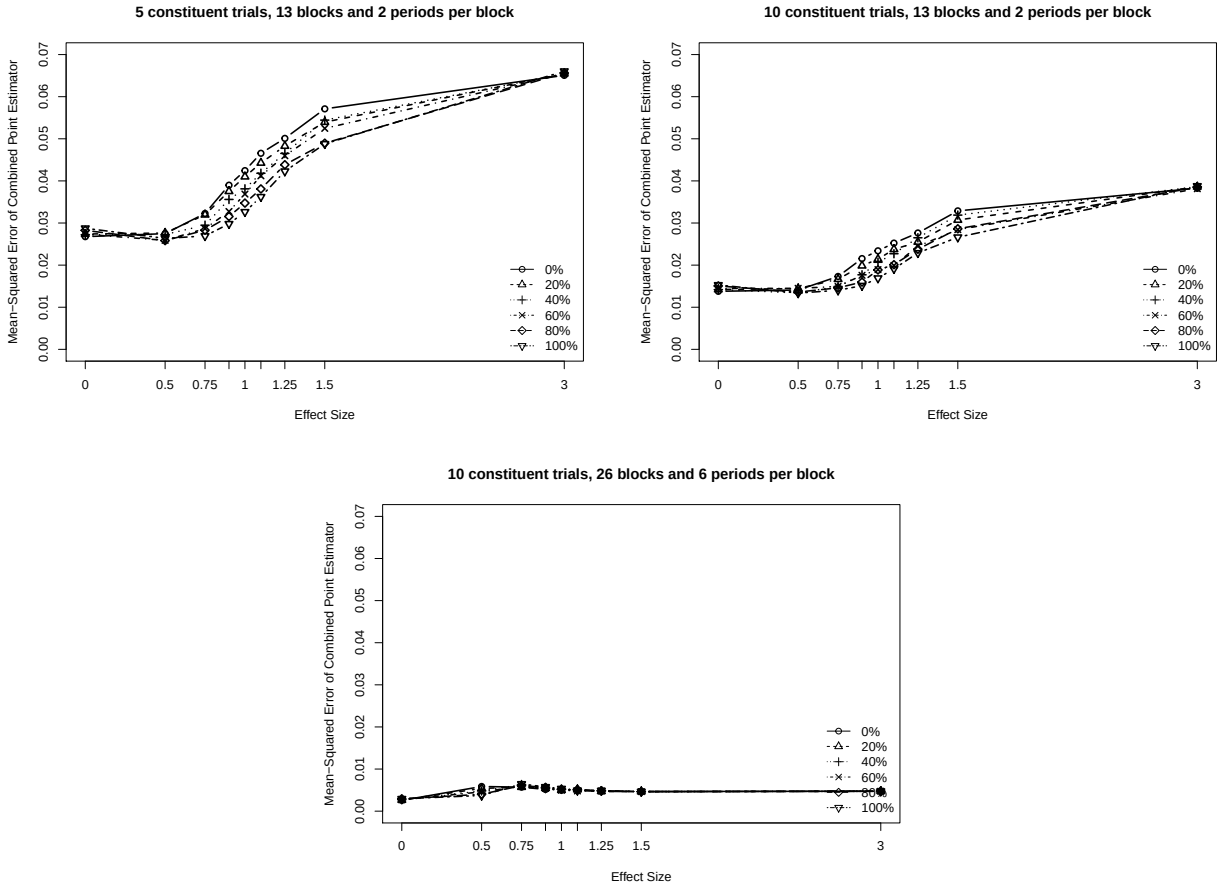


Figure C.2: Mean-squared error for the combined point estimator from a series of 5 or 10 sequentially-monitored N-of-1 trials with varying numbers of blocks and periods per block and varying OBF boundary shape.

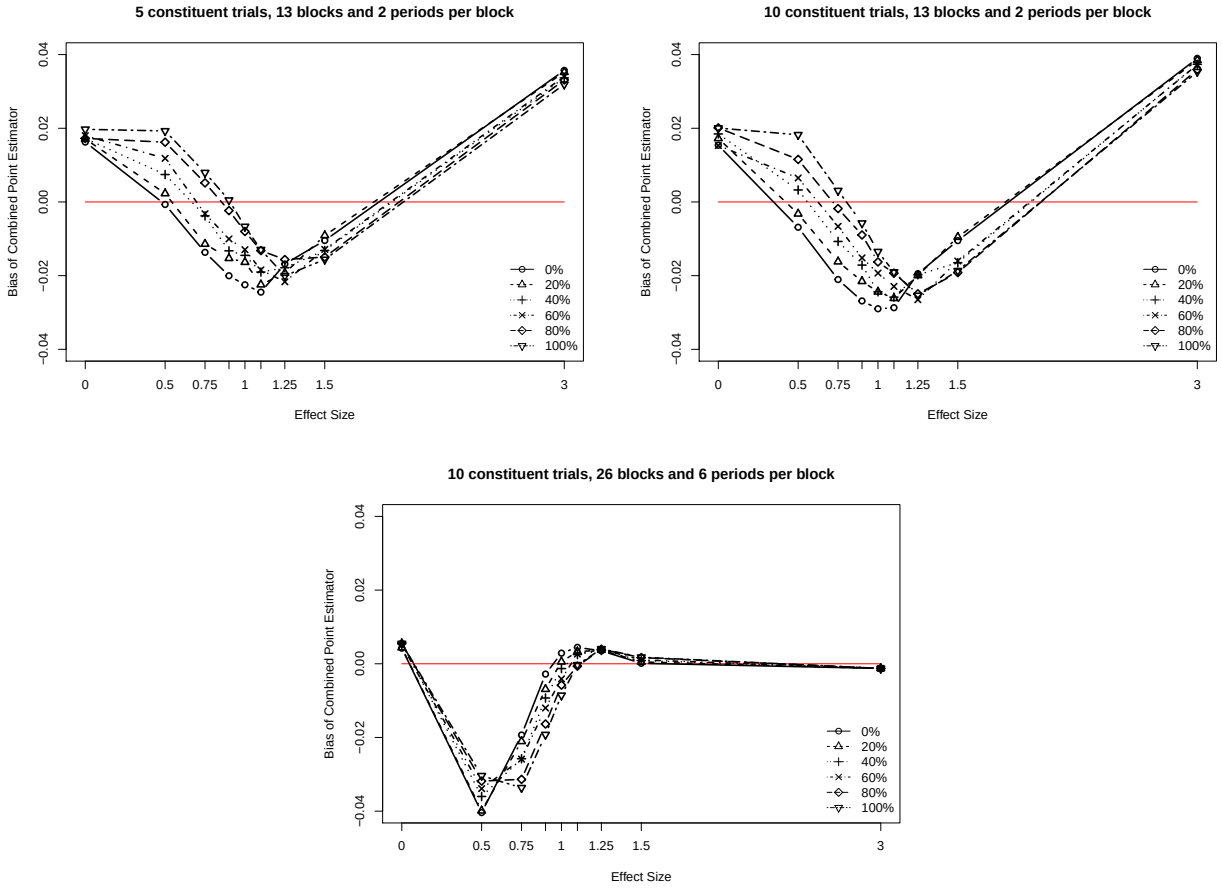


Figure C.3: Bias for the combined point estimator from a series of 5 or 10 sequentially-monitored N-of-1 trials with varying numbers of blocks and periods per block and varying OBF boundary shape. Bias of 0 indicated in red.

Appendix D

SOFTWARE AND SOFTWARE-RELATED QUALITY CONTROL OF THIS DISSERTATION

Code for reproducing results from this dissertation will be available at https://github.com/srselukar. This section describes R packages used in the code and procedures for validating the code base for each project.

In general, the code is written as nested functions, and each function is verified independently. For functions from packages available on CRAN, no further verification has been done. For user-created functions, I verify by testing multiple iterations of the input and checking the output - if the input is simulated data, then independent simulation runs under different settings were used.

D.1 Stratified Randomization for Platform Trials with Differing Experimental Arm Eligibility

1. Simulating treatment assignments for platform trials: code generates subject data with eligibility information and sequentially assigns each subject according to the proposed methodology then returns the dataset with treatment assignments and eligibility information

    - Only base R functions are used
    - No additional verification beyond the default

```
## Output:
# Data frame ordered by enrollment containing
# 1. strat factors
```

```
# 2. eligibility at enrollment for each experimental arm
# 3. assigned arm

## Inputs:
# 1. Time (# subjects) when arm added: (assuming 200 per arm) 100-300 total
# subjects in experimental arms
# 2. # initial experimental arms: fixed at 2
# 3. Probability of each eligibility combination: fixed at all equal
# 4. Planned sample size for arms: fixed at 200
# 5. Binary stratification factors: varies from 2-3
#       a. Factor 1: p=0.5
#       b. Factor 2: p=0.3
#       c. (Optional) Factor 3: p=0.25



### Simulate data
simDatGen <- function(probEligInit=NULL,probEligAdd=NULL,#  vectors of probabilitie
# for each possible combination (PROBABILITIES NEED TO MATCH THE SAMPLING FRAME)
probStrat1=0.5,probStrat2=0.3,probStrat3=NA, # probabilities for the success
# of each strat factor; by default, only 2 factors
numSubjInit,numSubjAdd, # number of subjects before arm added and
# number of subjects after
numSims){

numExp <- 2 # fixed at 2 experimental arms for now

if (is.null(probEligInit)) probEligInit <- rep(1/(2**numExp-1),2**numExp-1)
```

```
# if null, equal probabilities
if (is.null(probEligAdd)) probEligAdd <- rep(1/(2**(numExp+1)-1),2**(numExp+1)-1)

numSubj <- numSubjInit+numSubjAdd

### Simulate Eligibility
eligInit <- matrix(unlist(
sample(list(c(1,0,NA),
c(0,1,NA),
c(1,1,NA)),
numSubjInit * numSims,
replace=TRUE,
prob=probEligInit
)
),ncol=3,byrow=TRUE)

eligAdd <- matrix(unlist(
sample(list(c(1,0,0),
c(0,1,0),
c(1,1,0),
c(1,0,1),
c(0,1,1),
c(1,1,1),
c(0,0,1)),
numSubjAdd * numSims,
replace=TRUE,
prob=probEligAdd
)
```

```
),ncol=3,byrow=TRUE)

### Simulate Binary Stratification Factors
strat1 <- sample(c("strat1_F","strat1_S"),numSubj * numSims,
prob=c(1-probStrat1,probStrat1),replace=TRUE)
strat2 <- sample(c("strat2_F","strat2_S"),numSubj * numSims,
prob=c(1-probStrat2,probStrat2),replace=TRUE)
if (!is.na(probStrat3)) strat3 <- sample(c("strat3_F","strat3_S"),numSubj * numSims
prob=c(1-probStrat3,probStrat3),replace=TRUE) else strat3 <- NULL


### Create output datasets for initial subjects and subjects after adding arm;
## assign everyone NA for treatment arm
initData <- cbind(strat1[1:(numSubjInit * numSims)],
strat2[1:(numSubjInit * numSims)],
strat3[1:(numSubjInit * numSims)],
rep(NA,numSubjInit * numSims),
eligInit)

addData <- cbind(strat1[(numSubjInit*numSims+1):(numSubj*numSims)],
strat2[(numSubjInit*numSims+1):(numSubj*numSims)],
strat3[(numSubjInit*numSims+1):(numSubj*numSims)],
rep(NA,numSubjAdd * numSims),
eligAdd)

return(list(
initData,
addData
```

```
))
}

### Compute randomization tables given already randomized data
## Input:
# Dataset with columns for each strat factor,
# assigned study arm (or NA),
# eligibility (1/0) for each experimental arm (NA for arm not added)
# Number of strat factors
# Which participant to randomize (row index)

## Output:
# Table(s) of counts of subjects with same strat factor level
# One table for each pairwise comparison with control

funTab <- function(x,numFactors,rowIdx){ # assumes that x is a data frame
# with columns 1:numFactors are strat factors and
# there exists "arm" a variable for treatment arm
elig <- as.numeric(x[rowIdx,(numFactors+1+1):ncol(x)]) # grab the
# experimental arm eligibilities
if (sum(elig,na.rm=TRUE)==1){ # if not eligible to multiple experimental arms
tabOut <- matrix(NA,ncol=2,nrow=numFactors) # table for exactly one arm or control
for (i in 1:numFactors) tabOut[i,] <- table( # count the number of subjects
# randomized to control or the
# eligible exp arm of the same strat level
factor(x[
x[,i]==x[rowIdx,i] & # other subjects with the same strat level
x[,(numFactors+1)+which(elig==1) # pick out the eligible experimental arm
```

```
]==1, # so that they have the same (or more) eligibility
(numFactors+1) # pick out the arm variable (because we are counting
# how many in the same arm already)
],
levels=c(0,which(elig==1)))
)
return(tabOut)
} else {
tabOut <- vector("list",sum(elig,na.rm=TRUE)) # a table for each arm and
# associated control
for (j in 1:sum(elig,na.rm=TRUE)){
tabOut[[j]] <- matrix(NA,ncol=2,nrow=numFactors)
for (i in 1:numFactors) tabOut[[j]][i,] <- table( # count the number of subjects
# randomized to control or the
# eligible exp arm of the same strat level
factor(x[
x[,i]==x[rowIdx,i] & # other subjects with the same strat level
x[,(numFactors+1)+which(elig==1)[j] # pick out the eligible experimental arm
]==1, # so that they have the same (or more) eligibility
(numFactors+1) # pick out the arm variable
],
levels=c(0,which(elig==1)[j]))
) # creates a table for each experimental arm j=1,...,sum(elig,na.rm=TRUE)
}
return(tabOut)
}
}
```

```
### One simulation run:
## Input:
# One data set of pre-arm add data and one data set of post-arm add data
# When to add arm (# of subjects)

## Output:
# Data frame ordered by enrollment containing
# 1. strat factors
# 2. eligibility at enrollment for each experimental arm
# 3. assigned arm

## Outline:
# 1. Loop over all subjects in the pre-arm add data with
# stratified randomization method
# 2. Switch to post-arm add data;
# break loop and output data when 200 in every arm

assignArms <- function(datInInit,datInAdd, # one dataset each for pre- and
# post-add
numFactors, # number of stratification factors
numArmsInit=2, # number of experimental arms before adding
pBalance=0.67, # weight given to the least imbalanced arm for randomization
combData, # an indicator variable to decide whether to combine pre-add data
# into calculations after adding arm
numAdd, # number of pooled experimental arm subjects
# when we add the new experimental arm
armSize=200){ # the number needed in each experimental arm to close the arm
```

```
### start with initial data
# loop over each subject in pre-add data and assign an arm
# count number randomized to each experimental arm
# end loop when numAdd subjects randomized (pooled) to experimental arms

dat <- datInInit

# dat[,numFactors+1] will store the arm variable

for (j in 1:nrow(dat)){ # plan to run through all of the subjects in the input
# initial data: the input data will be larger than needed
# to hit the desired number of experimental arm subjects
elig.j <- c(0,which(dat[j,(numFactors+1+1):(numFactors+1+numArmsInit)]==1))
# create a vector of the control arm (0) and experimental arms to
# which subject is eligible
if (j==1){ # need to deal with first subject differently (randomly assign
# based on eligibility)
dat[,numFactors+1][j] <- sample(elig.j,1)
} else {
N <- length(elig.j) # number of arms to which subject is eligible
G <- rep(NA,N) # will hold the balance for each arm possible
pUnbalance <- (1-pBalance)/(N-1) # the weights for the more imbalanced study arms
p <- rep(pUnbalance,N) # hold the probabilities of allocation,
# initially assign least favorable weight

tab.j <- funTab(dat,numFactors,j) # get the pairwise tabulations for
# the subject under consideration
```

```
if (!is.list(tab.j)){ # if tab.j is not a list,
# then it is only one table because subject is only eligible to 1 experimental arm
for (armIdx in 1:N){ # N=2 based on the if statement
tab.j.arm <- tab.j
tab.j.arm[,armIdx] <- tab.j.arm[,armIdx]+1 # for each eligible study arm,
# we will add one to its count to check the imbalance for adding to that arm
G[armIdx] <- sum(max(dist(tab.j.arm[1,],"manhattan")),
max(dist(tab.j.arm[2,],"manhattan")),
ifelse(numFactors==3,max(dist(tab.j.arm[1,],"manhattan")),0) # also add over
# the 3rd stratification factor if it exists
)
}
} else { # if tab.j is a list, then there are multiple tables
# because of multiple experimental arm eligibilities
Gtab1 <- Gtab2 <- matrix(nrow=N,ncol=N-1) # store the imbalance for each
# stratification factor for adding to each arm, over all pairwise comparisons (N-1)
if (numFactors==3) Gtab3 <- matrix(nrow=N,ncol=N-1) # if a 3rd factor is included
for (armIdx in 0:(N-1)){ # for each eligible study arm, we will add one to
# its count to check the imbalance for adding to that arm
for (tabIdx in 1:(N-1)){ # calculate the imbalance for each table (number of
#  pairwise tables is N-1 > 1 in this else statement)
tab.j.arm <- tab.j[[tabIdx]] # grab one pairwise table

if (armIdx==0) tab.j.arm[,1] <- tab.j.arm[,1]+1 # the study arm being added is
# the contorl arm: add one to the column in every pairwise table
if (tabIdx==armIdx) tab.j.arm[,2] <- tab.j.arm[,2]+1  # the study arm being added
# is the experimental arm  of the current pairwise table: add one
```

```
Gtab1[armIdx+1,tabIdx] <- dist(tab.j.arm[1,],"manhattan")
Gtab2[armIdx+1,tabIdx] <- dist(tab.j.arm[2,],"manhattan")
if (numFactors==3) Gtab3[armIdx+1,tabIdx] <- dist(tab.j.arm[3,],"manhattan")
}
G[armIdx+1] <- sum( # sum across stratification factors for the imbalance due to
# adding to the given study arm
max(Gtab1[armIdx+1,]), # imbalance for strat factor 1 and given study arm
max(Gtab2[armIdx+1,]), # strat factor 2
ifelse(numFactors==3,max(Gtab3[armIdx+1,]),0) # optional strat 3
)
}
}

dup.j <- duplicated(G) | duplicated(G,fromLast = TRUE) # gets all indices of
# the duplicated arm imbalances
if (!any(dup.j)){ # no duplications
p[which.min(G)] <- pBalance # give the most balanced arm a
# higher probability of allocation
dat[,numFactors+1][j] <- sample(elig.j,1,prob=p)
} else if (length(dup.j)==2){ # only eligible for ctrl + one treatment and
# both tied, so give equal weight in randomizing
dat[,numFactors+1][j] <- sample(elig.j,1)
} else { # eligible for multiple arms,
# need to consider whether the duplicates are above or below the
# smallest imbalance score
if (min(G[dup.j])==min(G)){ # the smallest imbalance is duplicated
pTie <- (pBalance+(sum(G==min(G))-1)*pUnbalance)
p[G==min(G)] <- pTie/sum(G==min(G))
```

```
p[G!=min(G)] <- (1-pTie)/sum(G!=min(G))
dat[,numFactors+1][j] <- sample(elig.j,1,prob=p)
} else { # the smallest imbalance is not duplicated, so ignore duplicates
p[which.min(G)] <- pBalance # give the most balanced arm a
# higher probability of allocation
dat[,numFactors+1][j] <- sample(elig.j,1,prob=p)
}
}

if (sum(dat[,numFactors+1]==1 | dat[,numFactors+1]==2,na.rm=TRUE) > numAdd) {
# stop the loop once we hit the desired number of experimental arm subjects
initSubj <- j
break
}
}

}

datInit <- dat[1:initSubj,]
rm(dat) # clean it up just in case

### now deal with the subjects after adding the additional experimental arm
# repeat the process but with post-add data
# use combData variable to decide if funTab() function includes the
# pre-add data or not
# keep counter of total number randomized to each arm
# end loop when all arms have armSize subjects
```

```
if (combData) { # either combine the initial data (for those assigned) with the
# data with add or keep separate
dat <- rbind(datInit,datInAdd) # the combined data of the
# actually-assigned initial subjects + the potential data after adding arm
addIdx <- (initSubj+1):(initSubj+nrow(datInAdd)) # only loop over new,
# non-assigned subjects
preAddOffset <- rep(0,numArmsInit+1) # because we include the pre-add data,
# we do not need an offset
} else  {
dat <- datInAdd # only use additional data for stratified randomization procedure
addIdx <- 1:nrow(datInAdd)
preAddOffset <- table(factor(datInit[,numFactors+1],levels=1:(numArmsInit+1)))
# we want to include this offset to properly count the
# arm sizes in the following loop
}

for (j in addIdx){ # run through all of the subjects in the input add data:
# the input data will be larger than needed to
# hit the desired number of experimental arm subjects
elig.j <- c(0,which(dat[j,(numFactors+1+1):ncol(dat)]==1)) # create a vector of
# the control arm (0) and experimental arms to which subject is eligible

if (any(elig.j > 0)){ # if no experimental arm eligibility, skip this subject

if (j==1){ # need to deal with first subject differently (randomly
# assign based on eligibility)
dat[,numFactors+1][j] <- sample(elig.j,1)
} else {
```

```
N <- length(elig.j) # number of arms to which subject is eligible
G <- rep(NA,N) # will hold the balance for each arm possible
pUnbalance <- (1-pBalance)/(N-1)
# the weights for the more imbalanced study arms
p <- rep(pUnbalance,N) # hold the probabilities of allocation,
#  initially assign least favorable weight

tab.j <- funTab(dat,numFactors,j) # get the pairwise tabulations for
# the subject under consideration
if (!is.list(tab.j)){ # if tab.j is not a list,
# then it is only one table because subject is only eligible to 1 experimental arm
for (armIdx in 1:N){ # N=2 based on the if statement
tab.j.arm <- tab.j
tab.j.arm[,armIdx] <- tab.j.arm[,armIdx]+1 # for each eligible study arm,
# we will add one to its count to check the imbalance for adding to that arm
G[armIdx] <- sum(max(dist(tab.j.arm[1,],"manhattan")),
max(dist(tab.j.arm[2,],"manhattan")),
ifelse(numFactors==3,max(dist(tab.j.arm[1,],"manhattan")),0) # also add over the
# 3rd stratification factor if it exists
)
}
} else { # if tab.j is a list, then there are multiple tables because
# of multiple experimental arm eligibilities
Gtab1 <- Gtab2 <- matrix(nrow=N,ncol=N-1) # store the imbalance for each
# stratification factor for adding to each arm,
# over all pairwise comparisons (N-1)
if (numFactors==3) Gtab3 <- matrix(nrow=N,ncol=N-1) # if a 3rd factor is included
for (armIdx in 0:(N-1)){ # for each eligible study arm,
```

```
#  we will add one to its count to check the imbalance for adding to that arm
for (tabIdx in 1:(N-1)){ # calculate the imbalance for
# each table (number of pairwise tables is N-1 > 1 in this else statement)
tab.j.arm <- tab.j[[tabIdx]] # grab one pairwise table

if (armIdx==0) tab.j.arm[,1] <- tab.j.arm[,1]+1 # the study arm being added is
# the control arm: add one to the column in every pairwise table
if (tabIdx==armIdx) tab.j.arm[,2] <- tab.j.arm[,2]+1  # the study arm being added
# is the experimental arm  of the current pairwise table: add one

Gtab1[armIdx+1,tabIdx] <- dist(tab.j.arm[1,],"manhattan")
Gtab2[armIdx+1,tabIdx] <- dist(tab.j.arm[2,],"manhattan")
if (numFactors==3) Gtab3[armIdx+1,tabIdx] <- dist(tab.j.arm[3,],"manhattan")
}
G[armIdx+1] <- sum( # sum across stratification factors for
# the imbalance due to adding to the given study arm
max(Gtab1[armIdx+1,]), # imbalance for strat factor 1 and given study arm
max(Gtab2[armIdx+1,]), # strat factor 2
ifelse(numFactors==3,max(Gtab3[armIdx+1,]),0) # optional strat 3
)
}
}

dup.j <- duplicated(G) | duplicated(G,fromLast = TRUE) # gets all indices of
# the duplicated arm imbalances
if (!any(dup.j)){ # no duplications
p[which.min(G)] <- pBalance # give the most balanced arm a
#  higher probability of allocation
```

```
dat[,numFactors+1][j] <- sample(elig.j,1,prob=p)
} else if (length(dup.j)==2){ # only eligible for ctrl + one treatment and
# both tied, so give equal weight in randomizing
dat[,numFactors+1][j] <- sample(elig.j,1)
} else { # eligible for multiple arms, need to consider whether the duplicates are
# above or below the smallest imbalance score
if (min(G[dup.j])==min(G)){ # the smallest imbalance is duplicated
pTie <- (pBalance+(sum(G==min(G))-1)*pUnbalance)
p[G==min(G)] <- pTie/sum(G==min(G))
p[G!=min(G)] <- (1-pTie)/sum(G!=min(G))
dat[,numFactors+1][j] <- sample(elig.j,1,prob=p)
} else { # the smallest imbalance is not duplicated, so ignore duplicates
p[which.min(G)] <- pBalance # give the most balanced arm a
# higher probability of allocation
dat[,numFactors+1][j] <- sample(elig.j,1,prob=p)
}
}

expArmSizes <- preAddOffset + table(factor(dat[,numFactors+1],
levels=1:(numArmsInit+1))) # counter of the size of each experimental arm

if (any(expArmSizes >= armSize)){ # close experimental arms that are full
dat[(j+1):nrow(dat),(numFactors+1)+which(expArmSizes >= armSize)] <- NA # make all
# later subjects ineligible to full arms
}

if (all(expArmSizes >= armSize)) { # stop the loop once we hit the desired number
# of subjects in all experimental arms
```

```
break
}
}

}
}

# store the final assigned data
if (combData){
datFin <- dat[1:j,]
} else {
datFin <- rbind(datInit,dat[1:j,])
}
finSubj <- nrow(datFin)
outArmSizes <- table(factor(datFin[,numFactors+1],levels=0:(numArmsInit+1)))

return(list(datInit,datFin, # record each output data set
initSubj,finSubj, # record the time (subject number) at which each "stage" ended
outArmSizes # record the final arm sizes
)
)
}
```

2. Summarizing different approaches after addition of experimental arms: for each experimental arm, as each subject accrues, the code calculates the ratio of subjects assigned to the experimental arm compared to eligible subjects assigned to the standard of care

arm and summarizes how the ratio changes in the window before and after addition of new experimental arms

- Only base R functions are used
- In addition to the default verification of the simulation functions, the methods were checked to ensure balance was being achieved by assessing whether the ratios described above approach 1 as subjects accrue

D.2 RECeUS: Ratio Estimation of Censored Uncured Subjects, A Different Approach for Studying Sufficient Follow-Up in Studies of Long-Term Survivors

1. Simulating data with long-term survivors: code uses the inverse transform method to sample from a specified mixture-cure distribution and returns observed times and censoring

   - Uses R package: flexsurvcure (available on CRAN)
   - No additional verification beyond the default

```
### Simulate data
## Expect:
## n=sample size per simulation,
## sims=number of simulations,
## seed=seed for random draw
## dist=type of uncured distribution (exp, wei, llogis, g.gam),
## params=param vector that includes truCure as first element and
## the parameters of the uncured distribution after truCure,
## tauP=what percentile of uncured distribution is admin censoring

simData <- function(n=100,sims=1,seed=2019,params=c(0.3,1),dist="exp",tau=1.5,
```

```
accrualTime=NULL){
if (is.null(accrualTime)) accrualTime <- tau/2

truCure <- params[1]
truTheta <- params[-1]

set.seed(seed)
u <- runif(sims * n)

cens <- runif(sims * n,min=tau-accrualTime,max=tau) # akin to uniform accrual:
# note that min(T,tau-A)=min(T,C) where C=tau-A~Unif(tau-accrual,tau)

# Inverse transform sampling: X = q(u / (1-pi)) [equal in distribution;
# q is quantile function; u is uniform sample; pi is cure]
if (dist=="exp"){
truT <- ifelse(!is.nan(qexp(u/(1-truCure),rate=truTheta[1])),
qexp(u/(1-truCure),rate=truTheta[1]),Inf
)
}
if (dist=="wei"){
truT <- ifelse(!is.nan(qweibull(u/(1-truCure),shape=truTheta[1],
scale=truTheta[2])),
qweibull(u/(1-truCure),shape=truTheta[1],scale=truTheta[2]),Inf
)
}
if (dist=="llogis"){
truT <- ifelse(!is.nan(qllogis(u/(1-truCure),shape=truTheta[1],
scale=truTheta[2])),
```

```
qllogis(u/(1-truCure),shape=truTheta[1],scale=truTheta[2]),Inf
)
}
if (dist=="gam"){ # uses RATE parameterization in flexsurv!!
# use rate here to be consistent
truT <- ifelse(!is.nan(qgamma(u/(1-truCure),shape=truTheta[1],
rate=truTheta[2])),
qgamma(u/(1-truCure),shape=truTheta[1],rate=truTheta[2]),Inf
)
}
if (dist=="g.gam"){
truT <- ifelse(!is.nan(qgengamma(u/(1-truCure),mu=truTheta[1],
sigma=truTheta[2],Q=truTheta[3])),
qgengamma(u/(1-truCure),mu=truTheta[1],sigma=truTheta[2],Q=truTheta[3]),Inf
)
}

# Observed in sample
obsY <- pmin(cens,truT)
obsDelta <- truT <= cens

return(data.frame(Y=obsY,D=obsDelta))
}
```

2. Fitting models to simulated data: code fits class of parametric models and calculates sufficient follow-up statistics for the simulated data and returns all calculated estimates

    - Uses R package: flexsurvcure (available on CRAN)

- In addition to the default verification of the simulation functions, the information of the RECeUS $\hat{r}_n$ statistic was calculated numerically in R and compared to simulated results

```
### Analysis Functions
## Wrapper function for computing MLE
mleFun <- function(dat,dist="exp"){ # expects a df with columns Y,D
if (dist=="exp"){
tmp <- flexsurvcure(Surv(Y,D)~1,data=dat,dist="exp")
return(c(pi=tmp$res[1,1],rate=tmp$res[2,1],
piLB=tmp$res[1,2],piUB=tmp$res[1,3],
rateLB=tmp$res[2,2],rateUB=tmp$res[2,3],
AIC=tmp$AIC))
}
if (dist=="expUnc"){
tmp <- flexsurvreg(Surv(Y,D)~1,data=dat,dist="exp")
return(c(pi=0,rate=tmp$res[1,1],
piLB=NA,piUB=NA,
rateLB=tmp$res[1,2],rateUB=tmp$res[1,3],
AIC=tmp$AIC))
}
if (dist=="wei"){
tmp <- flexsurvcure(Surv(Y,D)~1,data=dat,dist="weibull")
return(c(pi=tmp$res[1,1],shape=tmp$res[2,1],scale=tmp$res[3,1],
piLB=tmp$res[1,2],piUB=tmp$res[1,3],
shapeLB=tmp$res[2,2],shapeUB=tmp$res[2,3],
scaleLB=tmp$res[3,2],scaleUB=tmp$res[3,3],
AIC=tmp$AIC))
```

```
}
if (dist=="weiUnc"){
tmp <- flexsurvreg(Surv(Y,D)~1,data=dat,dist="weibull")
return(c(pi=0,shape=tmp$res[1,1],scale=tmp$res[2,1],
piLB=NA,piUB=NA,
shapeLB=tmp$res[1,2],shapeUB=tmp$res[1,3],
scaleLB=tmp$res[2,2],scaleUB=tmp$res[2,3],
AIC=tmp$AIC))
}
if (dist=="llogis"){
tmp <- flexsurvcure(Surv(Y,D)~1,data=dat,dist="llogis")
return(c(pi=tmp$res[1,1],shape=tmp$res[2,1],scale=tmp$res[3,1],
piLB=tmp$res[1,2],piUB=tmp$res[1,3],
shapeLB=tmp$res[2,2],shapeUB=tmp$res[2,3],
scaleLB=tmp$res[3,2],scaleUB=tmp$res[3,3],
AIC=tmp$AIC))
}
if (dist=="llogisUnc"){
tmp <- flexsurvreg(Surv(Y,D)~1,data=dat,dist="llogis")
return(c(pi=0,shape=tmp$res[1,1],scale=tmp$res[2,1],
piLB=NA,piUB=NA,
shapeLB=tmp$res[1,2],shapeUB=tmp$res[1,3],
scaleLB=tmp$res[2,2],scaleUB=tmp$res[2,3],
AIC=tmp$AIC))
}
if (dist=="gam"){
tmp <- flexsurvcure(Surv(Y,D)~1,data=dat,dist="gamma")
return(c(pi=tmp$res[1,1],shape=tmp$res[2,1],rate=tmp$res[3,1],
```

```
piLB=tmp$res[1,2],piUB=tmp$res[1,3],
shapeLB=tmp$res[2,2],shapeUB=tmp$res[2,3],
rateLB=tmp$res[3,2],rateUB=tmp$res[3,3],
AIC=tmp$AIC))
}
if (dist=="gamUnc"){
tmp <- flexsurvreg(Surv(Y,D)~1,data=dat,dist="gamma")
return(c(pi=0,shape=tmp$res[1,1],rate=tmp$res[2,1],
piLB=NA,piUB=NA,
shapeLB=tmp$res[1,2],shapeUB=tmp$res[1,3],
rateLB=tmp$res[2,2],rateUB=tmp$res[2,3],
AIC=tmp$AIC))
}
if (dist=="g.gam"){
tmp <- flexsurvcure(Surv(Y,D)~1,data=dat,dist="gengamma")
return(c(pi=tmp$res[1,1],mu=tmp$res[2,1],sigma=tmp$res[3,1],shape=tmp$res[4,1],
piLB=tmp$res[1,2],piUB=tmp$res[1,3],
muLB=tmp$res[2,2],muUB=tmp$res[2,3],
sigmaLB=tmp$res[3,2],sigmaUB=tmp$res[3,3],
shapeLB=tmp$res[4,2],shapeUB=tmp$res[4,3],
AIC=tmp$AIC))
}
if (dist=="g.gamUnc"){
tmp <- flexsurvreg(Surv(Y,D)~1,data=dat,dist="gengamma")
return(c(pi=0,mu=tmp$res[1,1],sigma=tmp$res[2,1],shape=tmp$res[3,1],
piLB=NA,piUB=NA,
muLB=tmp$res[1,2],muUB=tmp$res[1,3],
sigmaLB=tmp$res[2,2],sigmaUB=tmp$res[2,3],
```

```
shapeLB=tmp$res[3,2],shapeUB=tmp$res[3,3],
AIC=tmp$AIC))
}
}

## M-Z test statistic Maller Zhou (1994)
mzTest <- function(dat){ # expects a df with columns Y,D
maxE <- max(dat$Y[dat$D==1]) # last event time
if (max(dat$Y) > maxE){
plat <- max(dat$Y) - maxE
numBefore <- sum(dat$D==1 & dat$Y > (maxE-plat)) # number of events that
# are plateau length before the last event
return(
(1-numBefore/nrow(dat))**nrow(dat) # alphahat
)
} else return(NA) # if the last observation is an event, then this test fails
}

## qn test statistic Maller Zhou (1996)
qnTest <- function(dat){ # expects a df with columns Y,D
maxE <- max(dat$Y[dat$D==1]) # last event time
maxAll <- max(dat$Y)
if (maxAll > maxE){
numPlat <- sum(dat$D==1 & dat$Y > (2*maxE-maxAll) & dat$Y <= maxE) # number of
# events that are between (2*maxE-maxAll) (exclusive) and maxE (inclusive)
return(
numPlat/nrow(dat) # qn
)
```

```
} else return(NA) # if the last observation is an event, then this test fails
}



shenTest <- function(dat){ # Shen 2000
maxE <- max(dat$Y[dat$D==1]) # last event time
maxAll <- max(dat$Y)
if (maxAll > maxE){
w <- (maxAll - maxE)/maxAll
tauG <- w*maxE+(1-w)*maxAll
numBefore <- sum(dat$D==1 &
(dat$Y >= tauG*maxE/maxAll & dat$Y <= maxE)
) # number of events that are "plateau" length before the last event (plateau here
#  is different than before)
return(
(1-numBefore/nrow(dat))**nrow(dat) # alphatilde
)
} else return(NA) # if the last observation is an event, then this test fails
}
```

D.3 A Framework for Sequential Monitoring of One N-of-1 Trial and Combining Results across a Series of Sequentially-Monitored N-of-1 Trials

1. Simulating data from N-of-1 trials: code generates each N-of-1 trial based on correctly-specified linear mixed-effects model then returns the naive estimate after every treatment block and its standard error

    - Uses R packages (all available on CRAN):
        - gtools - Computing possible combinations of treatment assignment sequences

  - mvtnorm - Generate data
  - lme4 - Analyze data at each treatment block
- No additional verification beyond the default

```
### Function to simulate data for one (n-of-1) trial
## 1. generate the treatment assignments
## 2. construct fixed design matrix and covariance matrix
## 3. simulate one trial

## Inputs:
# 1. numPeriods: number of periods per block
#   NOTE: currently, must be multiple of 2 because balanced
# treatment frequency within  blocks
# 2. numBlocks: number of independent blocks
# 3. fixIntercept: fixed effect under "control" (default to 0,
# do not plan on changing)
# 4. fixSlope: fixed treatment effect
# 5. randVar: Random effect variance (i.e., within-block covariance) (default to 1)
# 6. errVar: Error variance (default to 1, do not plan on changing)
# 7. inSeqs: matrix of possible sequences for periods within a block
#   each row is possible sequence (not simulation input:
# generated by simData function by numPeriods/numBlocks/timeBlock arguments)

## Output:
# Dataframe with names y (outcome vector), trt (treatment assignment) with
#  nrow()=numPeriods*numBlocks
simOneTrial <- function(numPeriods,
numBlocks,
```

```
fixIntercept,
fixSlope,
randVar,
errVar,
inSeqs){
fixParams <- c(fixIntercept,fixSlope)

### generate treatment assignments
blockSeqs <- sample(1:nrow(inSeqs),numBlocks,replace=TRUE) # indicates which
# of the sequences used for each of the blocks; assignments may be repeated
trtAsns <- c(t(inSeqs[blockSeqs,])) # converts the sequences above to a vector
# of the assignments for the full trial

### construct fixed design matrix and covariance matrix
fixDesignMat <- cbind(rep(1,numPeriods*numBlocks),trtAsns)

blockCovar <- errVar*
diag(numPeriods)+randVar*matrix(1,nrow=numPeriods,ncol=numPeriods)
trialCovar <- diag(numBlocks) %x% blockCovar

### simulate one trial
y <- rmvnorm(1,
mean=(fixDesignMat %*% fixParams),
sigma=trialCovar
)
return(cbind(block=rep(1:numBlocks,each=numPeriods),y=c(y),trt=trtAsns))
}
```

```
### Function to simulate data for all trials in a simulation set
# (calls simOneTrial)
## Inputs:
# 1. numTrials: number of trials to simulate (default to 1 for testing)
# 2. timeBlock: binary variable to indicate whether
#   sequences must be ABBA/BAAB-type to account for time confounding,
#   or if other sequences are also acceptable
#   (defaults to FALSE, not currently supported)
# 3+. arguments for simOneTrial

## Output:
# array with dim()=c(numPeriods*numBlocks,3,numTrials)
# one layer for each trial, each layer has 3 columns (y, trt and block,
# inherited from simOneTrial) and numPeriods*numBlocks rows
simData <- function(numTrials=1,
timeBlock=FALSE,
numPeriods=4,
numBlocks=3,
fixIntercept=0,
fixSlope=1,
randVar=1,
errVar=1){

if (numPeriods %% 2 != 0 ) stop("Unbalanced treatment frequency within blocks")

### generate possible treatment sequences for each block
tmpSeqs <- permutations(numPeriods,numPeriods,
c(rep(0,numPeriods/2),rep(1,numPeriods/2)),
```

```
set=FALSE) # calculates all possible sequences (with some duplicates)
posSeqs <- tmpSeqs[!duplicated(tmpSeqs),]

if (timeBlock){
stop("Not supported") # need to figure out how to select only the
# time confounding sequences from posSeqs
} else {
inSeqs <- posSeqs
}

### call simOneTrial
out <- replicate(numTrials,
simOneTrial(numPeriods,
numBlocks,
fixIntercept,
fixSlope,
randVar,
errVar,
inSeqs))

return(out)
}
```

2. Summarizing type-1 error, power, early stopping and average number of blocks at stopping: code computes monitoring boundaries then returns decision and numbers of blocks at stopping

    - Uses R package: gsDesign (available on CRAN)

- No additional verification beyond the default

3. Summarizing bias and mean-squared error for candidate point estimators: code computes monitoring boundaries, candidate point estimates and estimates of standard error

   - Uses R package: RCTdesign (not available on CRAN)
   - In addition to the default verification of the simulation functions, the RCTdesign package functions were verified in the following ways:
     - Compared monitoring boundaries with those returned by gsDesign package
     - Compared the "lower-level" functions (e.g., sequential monitoring density, expectation of naive estimate) against user-written functions written in base R
     - Compared higher-level functions (e.g., computing point estimates) against user-written functions written with lower-level RCTdesign functions

```
### Function to create monitoring boundaries (requires RCTdesign)
# only 3 specific boundary shapes supported currently

# Inputs:
# 1. stopTimes: vector of stopping times (i.e., blocks after which
# data are analyzed)
#   stopTimes[1] >= 2 and stopTimes[length(stopTimes)] == last block of trial
# 2. totBlocks: total number of blocks (max "sample size")
# 3. boundType: type of boundary (supporting symmetric OBF, symmetric Pocock and
# asymmetric with OBF upper and Pocock lowerr)
# 4. alpha: one-sided alpha level (defaults to 0.025, not planning to change)
```

```
# Outputs:
# seqDesign object
makeBounds <- function(stopTimes,
boundType="symOBF",
alpha=0.025){

if (boundType=="symPoc"){
out <- seqDesign(nbr.analysis=length(stopTimes),sample.size=stopTimes,
arms=1,
design.family="Pocock",
test.type="two.sided")
} else if (boundType=="symOBF"){
out <- seqDesign(nbr.analysis=length(stopTimes),sample.size=stopTimes,
arms=1,
design.family="OBF",
test.type="two.sided")
} else if (boundType=="asym") { # asymmetric with OBF upper and Poc lower;
# NOTE: allows for conclusion of null at last analysis
# (if no crossing at last stage)
upBound <- seqDesign(nbr.analysis=length(stopTimes),sample.size=stopTimes,
arms=1,
design.family="OBF",
test.type="two.sided")$boundary[,4]
loBound <- seqDesign(nbr.analysis=length(stopTimes),sample.size=stopTimes,
arms=1,
design.family="Pocock",
test.type="two.sided")$boundary[,1]
```

```
out <- seqDesign(nbr.analysis=length(stopTimes),sample.size=stopTimes,
arms=1,
exact.constraint=cbind(loBound,0,0,upBound),
test.type="two.sided")
} else stop("Not supported")

return(out)
}

### Function to return bias-adjusted estimates from a naive estimate at stopping
# (requires RCTdesign)

# Inputs:
# 1. boundObj: the seqDesign object used for monitoring
# 2. stopIdx: index of stopping
# 3. est.Naive: naive estimate at stopping
# 4. se.Naive: naive estimate of standard error at stopping
# 5. h: increment size for numerical derivative

# Outputs: list of two vectors of length 6
# 1. vector of naive + bias-adjusted estimates
# 2. vector of corresponding estimated standard errors

calcEst <- function(boundObj,stopIdx,est.Naive,se.Naive,h){
newBound <- update.seqDesign(boundObj,
exact.constraint=seqBoundary(boundObj,scale="Z"),design.family="Z",
# otherwise, update may change the Z boundaries! (esp for asymmetric bounds)
```

```
sd=se.Naive*sqrt(boundObj$specification$sample.size[stopIdx])
# need to update the input standard deviation (note that seqDesign expects
# SD not SE)
)

### RCTdesign natively returns all bias-adjusted estimates except conditional BAM
infOut <- seqInference(newBound,observed=est.Naive,analysis.index = stopIdx,
inScale="X")
estMUE.SM <- infOut[,4]
estMUE.AT <- infOut[,5]
estMUE.LR <- infOut[,6]
est.BAM <- infOut[,7]

## calculate CBAM


est.CBAM <- tryCatch(
myCBAM(desIn=newBound,observed=est.Naive,analysis.index=stopIdx,
searchRange=est.Naive+c(-1,1)),
error=function(e) return(NA)
)
if (is.na(est.CBAM)) {
tryCatch(
myCBAM(desIn=newBound,observed=est.Naive,analysis.index=stopIdx,
searchRange=est.Naive+c(-3,3)),
error=function(e) return(NA)
)
}
```

```
## collect all estimates

estOut <- cbind(
estMUE.SM,estMUE.AT,estMUE.LR,
est.BAM,
est.CBAM,
est.Naive
)

### Use Z-estimator large-sample theory to produce estimated variance/SE

pSeqOut <- cbind(
(pSeq(newBound,observed=estOut[,1],theta=estOut[,1]-h,analysis.index=stopIdx)[,4]-
pSeq(newBound,observed=estOut[,1],theta=estOut[,1]+h,analysis.index=stopIdx)[,4])/
(2*h), # MUE.SM
(pSeq(newBound,observed=estOut[,2],theta=estOut[,2]-h,analysis.index=stopIdx)[5]-
pSeq(newBound,observed=estOut[,2],theta=estOut[,2]+h,analysis.index=stopIdx)[,5])/
(2*h), # MUE.AT
(pSeq(newBound,observed=estOut[,3],theta=estOut[,3]-h,analysis.index=stopIdx)[,6]-
pSeq(newBound,observed=estOut[,3],theta=estOut[,3]+h,analysis.index=stopIdx)[,6])/
(2*h) # MUE.LR
)

meanSeqOut <- (meanSeq(newBound,estOut[,4]+h)[2]-
meanSeq(newBound,estOut[,4]-h)[2])/(2*h)
condmeanSeqOut <- (condMeanSeq(newBound,stopIdx,estOut[,5]+h)-
condMeanSeq(newBound,stopIdx,estOut[,5]-h))/(2*h)
```

```
B <- cbind(
pSeqOut,
meanSeqOut[[1]],
condmeanSeqOut
)

A <- (cbind( # delta method
(pSeq(newBound,observed=estOut[,1]-h,theta=estOut[,1],analysis.index=stopIdx)[,4]-
pSeq(newBound,observed=estOut[,1]+h,theta=estOut[,1],analysis.index=stopIdx)[,4])/
(2*h), # MUE.SM
(pSeq(newBound,observed=estOut[,2]-h,theta=estOut[,2],analysis.index=stopIdx)[5]-
pSeq(newBound,observed=estOut[,2]+h,theta=estOut[,2],analysis.index=stopIdx)[,5])/
(2*h), # MUE.AT
(pSeq(newBound,observed=estOut[,3]-h,theta=estOut[,3],analysis.index=stopIdx)[,6]-
pSeq(newBound,observed=estOut[,3]+h,theta=estOut[,3],analysis.index=stopIdx)[,6])/
(2*h), # MUE.LR

1,
1 # no need for Delta method for BAM or CBAM because derivative is 1
)**2)*
((se.Naive**2)) # NOTE: naive SE takes into account sample size

seOut <- cbind(
sqrt( # SE instead of variance
(A/(B**2)) # large-sample approximation for variance of Z-estimators
# (NOTE: naive SE already took into account sample size)
),
```

```
se.Naive
)

return(list(
estOut,seOut
))

}

## Helper functions to calculate CBAM
# calculate the conditional expectation
condMeanSeq <- function(desIn,anaIdx,theta){

# calculate denominator of conditional likelihood
# (probability of stopping at look=anaIdx)
totLooks <- nrow(seqBoundary(desIn))
if (totLooks > 1) {
cumeProbs <- (seqEvaluate(dsn=desIn,theta=theta,pwr=NULL)[[3]][4+(1:anaIdx)])
# use RCTdesign to calculate the probabilities
if (anaIdx == 1) probM <- cumeProbs[1] else probM <- diff(cumeProbs)[anaIdx-1]
# need to use diff() to calculate the probability of each stage > first
} else probM <- 1 # automatically stops if only 1 look



integCondSeq(x=desIn,analysis.index=anaIdx,observed=Inf,theta=theta,task="e")[1,4]/
probM # need to use observed=Inf to indicate a full expectation,
# but only at stopping stage (not mean across all stages)
```

```
}

# find the root over conditional expectations
# (note: analogous to BAM but using conditional density)
myCBAM <- function(desIn,observed,analysis.index,searchRange=c(-3,8)) {
if ( # do not return estimate if should not have stopped
analysis.index != nrow(desIn$boundary) & # not the last look
!(observed < desIn$boundary[analysis.index,1] |
 observed > desIn$boundary[analysis.index,4])
 # did not cross the boundary at that look
) {
return(NA)
} else return(uniroot(function(t) as.numeric(condMeanSeq(desIn,analysis.index,t))-
observed,interval=searchRange)$root)
# return CBAM when stopped properly
}


### Function to use input monitoring boundary and simulation results to
# output bias-adjusted estimates with standard errors

## Inputs:
# 1. dataIn: simulation results - a matrix of the estimates and
# standard errors with
#   nrow() = number of simulations
#   ncol() = 2*(number of trial blocks-1) with
#   odd columns the estimates after each block (no analysis after block 1) and
```

```
#   even columns the (observed) standard error
# 2. stopTimes: vector of stopping times (i.e., blocks after which data would have
#   stopTimes[1] >= 2 and stopTimes[length(stopTimes)] == last block of trial
# 3. boundIn: seqDesign object
# 4. h: increment size for numerical derivative

## Outputs: list of 2 matrices (numSims rows and 6 columns, one for each estimator)
# 1. estimates
# 2. estimated standard errors

## Important: block 1 estimate/SE not included in simulation results -
# impacts the indexing

anaSims <- function(dataIn,stopTimes,boundIn,
h=1e-6){

numBlocks <- stopTimes[length(stopTimes)]

### First find the block the trial stops at
bounds <- matrix(
seqBoundary(boundIn,scale="Z")[,c(1,4)],
ncol=2)
stopFun <- function(x) min(which(x < bounds[,1] | x > bounds[,2]),nrow(bounds))
# return logical index of stop

if (length(stopTimes) > 1){
statIn <- dataIn[,2*(stopTimes-1)-1]/dataIn[,2*(stopTimes-1)]
# transform the simulation data to wald statistic for monitoring boundary
```

```
} else {
statIn <- matrix(
dataIn[,2*(stopTimes-1)-1]/dataIn[,2*(stopTimes-1)],
# transform the simulation data to wald statistic for monitoring boundary
ncol=length(stopTimes)
)
}



stopIdx <- apply(statIn,1, # look at each row's Wald statistics
FUN = stopFun # find the index of stopping for that trial
)

stopBlock <- stopTimes[stopIdx] # return the stopping block

### Find the corresponding naive estimate

est.Naive <-  dataIn[,2*(stopTimes-1)-1][cbind(1:nrow(dataIn),stopIdx)]
# this picks out the specific stopIdx column of each row to
# get the resultant naive estimates
se.Naive <- dataIn[,2*(stopTimes-1)][cbind(1:nrow(dataIn),stopIdx)]
# this repeats to return the naive SE

### Loop over each trial to return the desired estimates/standard errors
# Need to run this loop because each bias-adjusted estimate requires an
# update to the trial's SD
```

```
estOut <- seOut <- matrix(ncol=6,
nrow=nrow(dataIn))

for (i in 1:nrow(dataIn)){
tmp <- tryCatch(calcEst(boundIn,stopIdx[i],est.Naive[i],se.Naive[i],h),
error=function(e) return(NA))

if (length(tmp)==2){ # some minor error handling if a given design did not work
estOut[i,] <- tmp[[1]]
seOut[i,] <- tmp[[2]]
}

}

### Report out estimates with estimated standard errors

return(list(
estOut,
seOut
))
}
```

4. Summarizing bias and mean-squared error for combined point estimator: code groups the above-generated bias-adjusted means into series of 5 or 10 trials then returns the combined point estimate

    - Uses R package: metafor (available on CRAN)

- No additional verification beyond the default