Author Manuscript

A Bayesian Group Sequential Small n Sequential Multiple Assignment Randomized Trial

Yan-Cheng Chao[1], Thomas M. Braun[1], Roy N. Tamura[2], and Kelley M. Kidwell[1]

[1]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, U.S.A.
[2]Health Informatics Institute, University of South Florida, Tampa, FL 33620, U.S.A.

Abstract

A small n, sequential, multiple assignment, randomized trial (snSMART) is a small sample multi-stage design where participants may be re-randomized to treatment based on intermediate endpoints. This design is motivated by ARAMIS (NCT02939573), an ongoing snSMART focusing on the evaluation of three drugs for isolated skin vasculitis. By formulating an interim decision rule for removing one of the treatments, we use a Bayesian model and the resulting posterior distributions to provide sufficient evidence that one treatment is inferior to the other treatments before enrolling more participants. By doing so, we can remove the worst performing treatment at an interim analysis and prevent the subsequent participants from receiving the removed treatment. Based on simulation results, we have evidence that the treatment response rates can still be unbiasedly and efficiently estimated in our new design, especially for the treatments with higher response rates. In addition, by adjusting the decision rule criteria for the posterior probabilities, we can control the probability of incorrectly removing an effective treatment.

Keywords: Interim analysis, adaptive design, rare disease, clinical trial, alpha-spending

1 Introduction

As an alternative to a traditional trial design, a small sample (n), sequential, multiple assignment, randomized trial (snSMART) can be used for efficient estimation of treatment effects in rare diseases (Tamura *et al.*, 2016). An snSMART is a multi-stage design where participants can be re-randomized at an interim timepoint based on their responses to initial treatment. A Randomized Multicenter Study for Isolated Skin Vasculitis (ARAMIS) is an ongoing snSMART of 90 participants designed to compare the effects of three active treatments for skin vasculitis (NCT02939573), and the motivating design for our proposed methods.

In contrast, a traditional sequential, multiple assignment, randomized trial (SMART), first proposed by Lavori and Dawson (2000) and Murphy (2005), is a multi-stage design used to evaluate the effects of tailored intervention sequences for treating disease, or dynamic treatment regimens (Murphy, 2003, 2005), with a relatively large number of participants. Thus, although an snSMART may seem similar to a traditional SMART, the two designs differ significantly in both their objective and assumed sample size.

Like traditional clinical trials, investigators may prefer a design that allows for the potential to remove an inferior treatment arm at an interim point during the trial. Adapting the snSMART design to allow for removing a treatment arm may also be favorable to participants because they are expected to receive a more effective treatment if the worst treatment is removed during the trial. Currently, no formal group sequential methods exist for an snSMART design, although many such methods exist for more traditional designs.

Frequentist interim analysis methods for clinical trials have been proposed by Stallard and Todd (2003), Stallard and Friede (2008), and Magirr *et al.* (2012). However, those methods assume that the study has a control arm, and any treatment that is not superior to the control is removed.

2

However, in our motivating snSMART design, there is no control arm, but rather three active treatment arms. Shih and Lavori (2013) did propose an alternative method in which they determine the current observed best treatment at each interim analysis, and all treatments shown to be inferior to the current best treatment are removed.

Bayesian approaches also exist for group sequential designs. Rosner and Berry (1995) focused on the posterior distribution of the difference in the treatment response rates to determine superiority at each interim analysis. However, they artificially divided their four treatments into two groups and performed two within-pair comparisons and one between-pair comparison, which is a limitation for application to a more general scenario of comparing multiple treatments. Yin *et al.* (2012) used the posterior predictive probability of treatment difference to decide early stopping boundaries in their Bayesian group sequential design. However, similar to many of the frequentist methods, Yin *et al.* (2012) also selected one treatment as the standard to which all other treatments were compared. Zhu *et al.* (2017) and Shi and Yin (2019) developed methods to control the overall Type I error rate in their Bayesian group sequential test, but only in the scenario of two treatment arms.

In our current work, we propose a Bayesian group sequential design that allows for removal of a worst performing treatment in an snSMART. Similar to a conventional group sequential design, before the start of an snSMART, we specify the number of interim analyses (looks) and the criteria for removing an arm at each interim analysis so that we control the overall probability of removing an arm under the scenario when three treatments have the same response rate. We describe our method in Section 2 and demonstrate the results of our approach via simulation in Section 3. We close with a discussion in Section 4.

3

2 Design

2.1 Standard snSMART design

2.1.1 General setup

The two-stage design of our motivating trial ARAMIS is shown in Figure 1(a); the original design had no interim analyses. In stage 1, participants are randomized equally to one of the three active treatments and then followed for six months, during which response to treatment may occur. In stage 2, stage 1 responders continue with the same treatment, while non-responders are re-randomized to one of the other two treatments that they did not initially receive. Participants are then followed for an additional six months for the occurrence of response to treatment. The length of stage 1 is the same as the length of stage 2, and stage 2 begins immediately after stage 1 ends. We emphasize that the term "stage" refers to the fixed period of time from a participant's receipt of a treatment to the end of their follow-up for response to that treatment.

2.1.2 Bayesian Joint Stage Model (BJSM) for an snSMART

Wei *et al.* (2018) developed a Bayesian Joint Stage Model (BJSM) to estimate the response rates of three treatments in an snSMART with binary outcomes. We briefly present the BJSM here because it is used in both the decision rule mentioned in Section 2.2.2 and the estimation of response rates at the end of a trial. For participant $i = 1, 2, \ldots, N$, where N is the number of participants, treatment $j = A, B, C$, and stage $k = 1, 2$, we let Y_{ik}^j be an indicator of response for participant i receiving treatment j in stage k. The stage 1 response rate to treatment j is denoted by π_j.

We then let $\beta_{1j}\pi_j$ denote the stage 2 response rate of the stage 1 responders to treatment j, with the assumption that $\beta_{1j} > 1$, so that if a participant responds in stage 1, they are at least as likely to respond again to the same treatment in stage 2. For stage 1 non-responders to treatment j, we let

4

$\beta_{0j}\pi_{j'}$ denote the reponse rate to treatment j' in stage 2, with the assumption that $\beta_{0j} < 1$, i.e. stage 1 non-responders are less likely to respond to either of the two other treatments in stage 2. Wei *et al.* (2018) referred to β_{1j} and β_{0j} as linkage parameters because they link the stage 1 response rates to the stage 2 response rates.

The Bayesian Joint Stage Model (BJSM) estimates the response rates of three treatments as follows:

$$Y_{i1}^{j}|\pi_j \sim \text{Bernoulli}(\pi_j) \tag{1}$$

$$Y_{i2}^{j'}|Y_{i1}^{j}, \pi_j, \pi_{j'}, \beta_{1j}, \beta_{0j} \sim \text{Bernoulli}((\beta_{1j}\pi_j)^{Y_{i1}^{j}}(\beta_{0j}\pi_{j'})^{1-Y_{i1}^{j}}) \tag{2}$$

$$\pi_j \sim \text{Beta}(\theta_1, \delta_1) \tag{3}$$

$$\beta_{0j} \sim \text{Beta}(\theta_2, \delta_2) \tag{4}$$

$$\beta_{1j} \sim \text{Pareto}(1, c) \tag{5}$$

Beta priors are used for π_j and β_{0j} because we assume that they range from 0 to 1, while Pareto$(1, c)$ is used for β_{1j} because it requires $\beta_{1j} > 1$. For more details about the specification of hyperparameters, see Wei *et al.* (2018). The response rate for each treatment is estimated from the posterior distribution of π_j using Markov Chain Monte Carlo (MCMC).

2.2 Group Sequential snSMART

2.2.1 General Setup

In stage 1, randomization will assign equal numbers of participants to each treatment; in contrast, the number of participants assigned to each treatment in stage 2 will depend upon the proportion of responders in stage 1. Thus, even without interim analyses, more participants are expected to

5

receive the better treatments in an snSMART. We now wish to determine if we can further increase the number of participants assigned to the better treatments if we allow for the removal of an inferior arm.

In a group sequential snSMART, treatment effects are estimated and compared at each interim analysis (or look) $l = 1, 2, \ldots, L$, where L is the maximum number of interim analyses performed during a trial. Here we will assume that $L = 2$ so that there are at most two looks in the snSMART. If an interim analysis suggests that one treatment is inferior to the others, then the treatment is removed and subsequent participants entering the trial no longer receive the removed treatment. If none of the treatments is considered inferior after look L, all three treatments are kept to the end of the trial. We note that "stage" and "look" are two different concepts in our group sequential snSMART design. Stage refers to a period of time specific to when each *participant* is followed for a response, while "look" refers to a period of time specific to the *entire study* when the accrued data are analyzed in an interim analysis.

If an interim analysis suggests removal of a treatment, the trial continues such that stage 1 non-responders to that inferior treatment are randomized equally to the two non-inferior treatments, while stage 1 non-responders to each of the non-inferior treatments are deterministically switched to the non-inferior treatment they had not received. In addition, stage 1 responders continue to receive the same treatment in stage 2 regardless of whether or not the treatment has been removed. An example of a two-stage snSMART design after treatment A is removed at look l is demonstrated in Figure 1(b).

In order to better describe the process of the trial, we demonstrate an example of a group sequential snSMART with two interim analyses, in Figure 2. Here we assume that three participants are enrolled in the trial every month, and recruitment continues for thirty months. The interim

6

analyses are planned after the 30th and 60th patients have completed stage 1. When the stage 1 outcome from the 30th participant is collected (marked by the first dashed box at month 16 in Figure 2), the first look occurs and response rates are estimated using the BJSM, and consideration of removing a treatment is based on the decision rule presented in Figure 3, the details for which are found in Section 2.2.2. We note that the stage 2 outcomes from some early participants are available for model fitting when the interim analysis is conducted, but not all participants will have stage 2 outcomes.

If a treatment is removed at the first look, the second look would not occur. If no arm is removed at the first look, the second look would occur when the stage 1 outcome from the 60th participant is collected (marked by the second dashed box at month 26 in Figure 2). At this point, whether an arm is removed depends on the result from the BJSM and the decision rule, but no more looks would be conducted until the final data analysis at the end of the trial. After the trial ends, we apply the BJSM to estimate the response rates of the three treatments using the stage 1 and stage 2 response indicators from all participants. Note that if the trial had been designed with only one look, that look could be conducted when the stage 1 outcome from the 45th participant was collected.

2.2.2 Bayesian Decision Rules

To consider the removal of a treatment arm, we introduce a two-step decision rule based on the posterior distributions of the response rates at each interim look l. The sample size for each look l is N_l, which is a cumulative number of all the accrued participants until look l, and the total sample size for the snSMART is denoted by N_T. In our design, an equal number of participants is accrued between looks, i.e., $N_l - N_{l-1} = N_T/(L+1)$. At each look, the BJSM is able to produce posterior draws of the response rates of all treatments even though stage 2 outcomes may be missing from

7

some participants. In this case, the participants that provide Y_{i2}^j are a subset of the participants that provide Y_{i1}^j.

We let $P_{j,l} = P_l(\pi_j > \pi_{j'}$ for all $j' \neq j|\text{Data}_l)$ denote the interim posterior probability that treatment j has the *greatest* response rate given the data up to the look l, and the posterior probability $Q_{j,l} = P_l(\pi_j < \pi_{j'}$ for all $j' \neq j|\text{Data}_l)$ denote the interim posterior probability that treatment j has the *smallest* response rate given the data up to the look l, where Data_l are all available Y_{i1}^j and Y_{i2}^j for all $j = A, B, C$ at look l. The first step of the decision rule is based on $P_{j,l}$ and the second step is based upon $Q_{j,l}$, conditional upon the value of $P_{j,l}$. A visual presentation of the detailed two-step decision rule is shown in the Figure 3.

Specific steps are:

1. For each treatment $j = A, B, C$, compute $P_{j,l}$ and compare to the pre-specified cutoff τ_l.

2. (a) If $P_{j,l} > \tau_l$ for *any* of the $j = A, B, C$, then compute $Q_{j',l}$ for treatments $j' \neq j$ and remove the treatment with higher $Q_{j',l}$.

   (b) If $P_{j,l} \leq \tau_l$ for *all* $j = A, B, C$, then compute $Q_{j,l}$ for all j and compare the posterior probability $Q_{j,l}$ with the pre-specified cutoff ψ_l. If $Q_{j,l} > \psi_l$ for any of the $j = A, B, C$, then remove treatment j. Otherwise, keep all three treatments.

Our two-step approach is quite intuitive. If enough evidence shows that one treatment is best (Step 2(a)), then one of the two inferior treatments should be removed. Similarly, if no single best treatment is identified, but there is enough evidence that one treatment is worst (Step 2(b)), then the worst treatment should be removed. Since we want to guarantee that at least two treatments remain until the end of the trial, at most one treatment can be removed at an interim analysis, after which, no more interim analyses would be conducted. Thus, when we refer to a design with L

8

looks in the following sections, we mean that at most L looks may take place. If a treatment arm is removed at an early look, the total number of looks may be smaller than L.

The thresholds τ_l and ψ_l used in Steps 1 and 2 can be selected by a user through a grid search as follows. First, consider a "null" setting in which all three treatments have the same response rate ($\pi_A = \pi_B = \pi_C$). If we let α_l denote the probability of incorrectly removing an arm from the trial at look l, the overall probability of making such an incorrect decision during the trial is equal to $\alpha = \sum_{l=1}^{L} \alpha_l$. Thus, for a pre-defined value of α, we recommend assigning the same values to each τ_l and to each ψ_l in a range from 0.98 to 0.80 with a step size of 0.02. Simulations are then run with these pre-assigned τ_l and ψ_l under the "null" scenario and the resulting value of α is recorded to obtain an approximate range of values assigned to τ_l and ψ_l that all result in our pre-specified α. We can then apply these values to new "non-null" settings in which all three treatments do not have the same response rates to assess the probability that an inferior arm is now correctly dropped.

Without loss of generality, we assume that $\pi_A \leq \pi_B \leq \pi_C$. There are four possible scenarios for the values of these response rates. We describe how our two-step decision rule works in each of these scenarios.

(1) $\pi_A = \pi_B = \pi_C$: $P_{j,l} > \tau_l$ is unlikely to be true for $j = A, B, C$, meaning that none of the arms is superior, then $Q_{j,l} > \psi_l$ is also unlikely to be true. The rule results in keeping all three arms.

(2) $\pi_A < \pi_B = \pi_C$: $P_{j,l} > \tau_l$ is unlikely to be true because $P_{B,l}$ and $P_{C,l}$ should be close, but $Q_{A,l} > \psi_l$ is likely to be true. The rule results in removing arm A.

(3) $\pi_A = \pi_B < \pi_C$: $P_{C,l} > \tau_l$ is likely to be true. The rule results in removing either arm A or arm B with nearly identical probabilities.

9

(4) $\pi_A < \pi_B < \pi_C$: $P_{C,l} > \tau_l$ is likely to be true. The rule results in removing arm A more often than arm B because $Q_{A,l} > Q_{B,l}$ is more likely to be true.

Although our decision rule is comprised of two steps, we could modify the rule to only have one step based solely on each $Q_{j,l}$. Specifically, if any of the $Q_{j,l}$ exceeds the pre-specified ψ_l, treatment j should be removed. Thus, in the one-step rule, we only consider inferiority of a treatment, whereas in the two-step rule we also consider superiority of a treatment. We investigate the operating characteristics of group sequential snSMARTs with both one-step and two-step decision rules in Section 3.2.

2.2.3 Estimation of treatment effects under the decision rule

In an snSMART without interim analyses, response rates are estimated by pooling the first and second stage outcomes using the BJSM. We will show that due to the sequential randomization, each response rate obtained from the BJSM is an unbiased estimate of the true treatment response rate. In our group sequential snSMART, it is possible that stage 2 randomization is not conducted for some first stage responders because one treatment arm is removed. We now justify that an unbiased estimate of the response rate can be obtained even when the second stage treatment allocation is deterministic for some non-responders.

To distinguish from the observed first and second stage outcomes Y_1^j and $Y_2^{j'}$ (subscript i is omitted here for simplicity), respectively, we denote the counterfactual outcomes for first stage treatment j and second stage treatment j' by $Y_1(j)$ and $Y_2(j,j')$. We also denote the first and second stage treatment assignments by J_1 and J_2. Under the consistency assumption, the individual with observed treatment $J_1 = j$ or $(J_1, J_2) = (j, j')$ has the observed outcomes Y_1^j and $Y_2^{j'}$ equal to his counterfactual outcomes $Y_1(j)$ and $Y_2(j,j')$. In addition, randomization guarantees that the assignment of treatment is independent of the counterfactual outcomes, or $J_1 \perp Y_1(j)$, $J_1 \perp$

10

$Y_2(j, j')$ and $J_2 \perp Y_2(j, j')$. For the first stage outcomes, under the consistency assumption and randomization:

$$P(Y_1^j = 1 | J_1 = j) = P(Y_1(j) = 1 | J_1 = j) \quad \text{(consistency)}$$

$$= P(Y_1(j) = 1) \quad \text{(first stage randomization)}$$

$$= \pi_j$$

The observed response rate of participants who did not respond to j in the first stage and receive j' in the second stage can be expressed by $P(Y_2^{j'} = 1 | J_1 = j, Y_1^j = 0, J_2 = j')$. Thus, under the consistency assumption and randomization:

$$P(Y_2^{j'} = 1 | J_1 = j, Y_1^j = 0, J_2 = j') = P(Y_2(j, j') = 1 | J_1 = j, Y_1^j = 0, J_2 = j') \quad \text{(consistency)}$$

$$= P(Y_2(j, j') = 1 | J_1 = j, Y_1^j = 0) \quad \text{(second stage randomization)}$$

$$= P(Y_2(j, j') = 1 | J_1 = j, Y_1(j) = 0) \quad \text{(consistency)}$$

$$= P(Y_2(j, j') = 1 | Y_1(j) = 0) \quad \text{(first stage randomization)}$$

$$= \beta_{0j} \pi_{j'}$$

The relationship of observed and true second stage response rates for first stage responders to treatment j can be derived using a similar approach. Thus, valid inference can be made for π_j with the observed response rates from both stages using the BJSM in an snSMART without interim analysis, meaning that the estimated response rates from BJSM are unbiased.

In a group sequential snSMART, if arm A is removed after an interim analysis, the subsequent participants are not randomized to A, and the non-responders to B (or C) in the first stage are assigned C (or B) in the second stage deterministically (Figure 1(b)). The failure to conduct second stage randomization may undermine the above derivation such that $P(Y_2(B, C) = 1 | J_1 = B, Y_1^B = 0, J_2 = C) \neq P(Y_2(B, C) = 1 | J_1 = B, Y_1^B = 0)$. However, in this specific case, we see

11

that the condition "$J_2 = C$" is equivalent to the condition "$J_1 = B$ and $Y_1^B = 0$", and this idea can be generalized to situations where other second stage response rates are of interest. Thus, $P(Y_2(B,C) = 1|J_1 = B, Y_1^B = 0, J_2 = C) = P(Y_2(B,C) = 1|J_1 = B, Y_1^B = 0)$ is valid for group sequential snSMART even if the second stage randomization does not occur for some first stage non-responders, leading to the conclusion that the second stage response rate of C obtained from the observed outcomes, $P(Y_2^C = 1|J_1 = B, Y_1^B = 0, J_2 = C)$ is still an unbiased estimate of the true second stage response rate, $\beta_{0B}\pi_C$.

3   Simulation

3.1   Data generation

We conducted simulation studies to examine the impact of interim analyses in an snSMART in four specific scenarios: (1) $\pi_A = \pi_B = \pi_C = 0.25$; (2) $\pi_A = 0.25, \pi_B = \pi_C = 0.5$; (3) $\pi_A = \pi_B = 0.25, \pi_C = 0.5$; (4) $\pi_A = 0.25, \pi_B = 0.45, \pi_C = 0.65$. For analysis with the BJSM, we let $\beta_{1A} = \beta_{1B} = \beta_{1C} = 1.5$ and $\beta_{0A} = \beta_{0B} = \beta_{0C} = 0.8$. The prior distributions for π_j, β_{1j} and β_{0j} are Beta$(0.4, 1.6)$, Pareto$(1, 3)$ and Beta$(1.6, 0.4)$, respectively, which have respective prior means of 0.2, 1.5, and 0.8. The hyperparameters of the prior distributions were chosen based on the prior knowledge of the stage 1 and stage 2 treatment effects motivated by ARAMIS.

We examined a group sequential snSMART that uses a maximum of 1 look, one that uses a maximum of 2 looks, as well as a traditional snSMART with no interim analyses. The interim analyses will be based on both the one-step and two-step decision rules described in Section 2.2.2. We also examine accrual rates of 2, 3, and 5 participants per month. In all trials, the number of participants was $N_T = 90$ and values for τ_l and ψ_l in the decision rule were chosen such that the probability of dropping a treatment in scenario 1 is close to a pre-specified value of $\alpha = 0.1$.

12

3.2   Simulation results

Table 1 presents a summary of the simulations for all four scenarios when three participants accrue each month. In this table we wish to see how operating characteristics first change as a function of the decision rule, and then how they change as a function of the number of interim analyses.

By comparing the top two rows of Table 1 to the middle two rows, we find that the probability of correctly removing an arm in scenario 2 is relatively unaffected whether one step or two steps are used in the decision rule. However, in scenarios 3 and 4, we see that the two-step rule performs better than the one-step rule, with an increase of 20-30 percentage points in the probability of removing a treatment arm. We note that this observed difference in probability of correctly removing a treatment arm increases as N_T increases (data not shown). Thus, a two-step rule is preferred to a one-step rule.

Next, we compare the middle two rows of Table 1 to the bottom two rows to assess the impact of moving from one interim analysis to two interim analyses. In all of scenarios 2, 3, and 4, we see that the probability of correctly removing a treatment arm increases when two interim analyses are performed relative to one interim analysis. When $N_T = 300$ (data not shown), the benefit of two interim analyses is no longer apparent, mostly because with such a large sample size, the probability of correctly removing a treatment arm with one look already reaches 0.95.

In Figure 4, we assess how interim analyses impact the number of stage 2 participants assigned to the best treatment in a group sequential snSMART. The height of each bar represents the ratio of the number of participants assigned to each treatment relative to the number of participants that would occur in an snSMART without interim analyses. In scenario 1, we see bar heights close to 1.0, indicating that interim analyses have little impact on patient allocation, relative to no interim analyses, because all three response rates are equal.

13

In scenarios 2, 3 and 4, we see bars with heights greater than 1.0 corresponding to treatments with the highest response rate and bars with heights less than 1.0 for treatments with the lowest response rate. This indicates that including interim analyses leads to assigning more participants to the better performing treatments compared to the snSMART without interim analyses. Furthermore, the ratio for the best treatment is highest when the two-step decision rule is used with two interim analyses, which agrees with the pattern of probabilities of correctly removing a treatment arm shown in Table 1. We obtained a similar pattern if we focused on the stage 1 participant counts (data not shown). Thus, with regard to participant assignment, a two-step decision rule with two interim analyses is preferred for all scenarios for $N_T = 90$.

In Table 2, we assess how interim analyses impact the numbers of responders to each treatment in each scenario. In scenario 1, since all response rates are equal, there are almost equal numbers of participants responding to each treatment. However, in scenarios 2, 3 and 4, we see that incorporating interim analyses leads to more responders to the treatments with higher response rates. Most importantly, when the response rates of three treatments are not equal, a group sequential design has more responders than that of a design without interim analyses. Together with the result in Figure 4, we conclude that group sequential snSMARTs allocate more participants to the better treatment, and more participants can benefit from their assigned treatment.

In Figure 5, we assess the impact of interim analyses on the bias and root mean-squared error (rMSE) of the response rates using the BJSM. We focus solely on a design with two interim analyses that use the two-step decision rule, as that design was seen to be best in terms of patient assignment. In general, the interim analysis does appear to lead to a slightly higher bias, but the overall biases still remain small compared to the true response rates. We note that the bias corresponding to the worst treatment can be higher than the bias of the other treatments, which is expected because fewer participants are assigned to the worst treatment. As with bias, rMSE is im-

14

pacted to a small degree when interim analyses are incoporated in the design. Although there is a little impact on the rMSE of the best treatment, the efficiency corresponding to the worst treatment is compromised in the group sequential snSMART, again because fewer participants are assigned to this treatment when interim analyses are used. Furthermore, the conditional bias using only the simulations where a treatment arm was removed increased slightly in the scenarios where (1) P_{drop} was small or (2) the response rates of a treatment was small (results not shown). This increase is expected because these biases were calculated using the results from fewer simulations and/or fewer participants assigned to a treatment. When neither of the above conditions was true, the conditional bias was almost as small as the marginal bias shown in Figure 5.

In Table 3, we examine how the probability of correctly removing a treatment is impacted by the accrual rate, as faster (slower) accrual implies a higher (lower) proportion of participants who have not completed stage 2 by the time of the interim analysis. The top two rows of Table 3 summarize when accrual is faster (5 participants/month), the middle two rows are the original accrual (3 participants/month), and the bottom two rows correspond to slower accrual (2 participants/month).

In scenarios 2, 3, and 4, we see generally as the accrual rate increases, there is a decrease in the probability of correctly removing a treatment arm, which is likely due to the increasing proportion of missing stage 2 outcomes. Correspondingly, when the accrual rate is slower, more stage 2 outcomes from participants can be collected for model fitting and there is an increase in the probability of correctly removing a treatment arm. Nonetheless, although the slower accrual rate leads to a slightly higher probability of correctly removing a treatment arm, the slower accrual rate also leads to a longer trial. Certainly the accrual rate will vary with the rarity of the disease and the number of sites that recruit participants, but overall, we expect that realistic rates of accrual will only slightly affect the probabilities of correctly removing a treatment arm.

15

4 Discussion

We provide a framework for incorporating interim analyses into an snSMART to potentially remove one of three treatment arms. With the proposed two-step Bayesian decision rule, a group sequential snSMART with two interim analyses may be more appealing to both those designing the trial and those participating in the trial. In a group sequential snSMART, fewer participants are expected to receive the worst treatment and the estimation of the response rate of the best treatment is not compromised relative to an snSMART without interim analyses. Similar to traditional group sequential designs, we can control the overall probability of removing an arm under a "null" scenario when three response rates are equal by using simulations to determine the values used for the cutoff values in the decision rule.

Our group sequential snSMART design can be used more flexibly in real practice. First, the proposed decision rule can be extended if there are interactions between stage 1 and 2 treatments that vary depending upon which treatments are used. Second, we assumed that interim analyses were performed when stage 1 outcomes were collected from a fixed number of participants at equal intervals. Instead, we can easily adjust the design to accommodate interim analyses at any interval of time. Third, the prior distributions of the response rates and linkage parameters can also be changed to reflect prior beliefs in the treatment response rates and linkage parameters. We assumed a Pareto distribution for the linkage parameters β_1 because we believed that responders were more likely to respond again in stage 2 had they already responded in the stage 1. However, we can change this prior distribution to a gamma or lognormal distribution, which ranges from 0 to infinity, under different assumptions for the responders. Similarly, the other prior distributions and their hyperparameters could differ given the specific trial setting. Based upon other simulations (results not shown), even if the prior distributions are centered away from the true parameter values,

16

estimation of the response rates shows little bias.

We note that in a traditional group sequential design, the number of interim analyses is often decided by many factors, including the total sample size, the power under the expected treatment effect difference, the effort to carry out interim analyses (Jennison and Turnbull, 1999). Practioners can decide an appropriate number of interim analyses through simulation studies after the total sample size, power under expected treatment effect difference, accrual rate and maximum number of interim analyses are pre-specified in group sequential snSMART designs. In small sample scenarios, such as 90 participants in our simulations, we do not recommend more than two interim analyses. A greater number of interim analyses will not substantially enhance the probability of correctly removing an arm because insufficient information will be available for decision making at the earlier interim analyses. Furthermore, if one wants to remove an arm more quickly when some early evidence of strong inferiority can be identified, then earlier interim analysis would be desired. On the contrary, if one wants to be more conservative about making a decision to remove an arm, a late interim analysis would be preferred.

Choosing the specific values of response rates under scenario 1 is arbitrary as long as the three response rates are equal. In our simulations we chose 0.25 as the "null" response rates for all three treatments because this response rate was considered ineffective across treatments for our setting. Although different response rates for scenario 1 might change the chosen threshold values τ_l and ψ_l, we have found that the small difference in threshold values does not greatly change the operating characteristics of the group sequential snSMART in scenarios 2, 3 and 4 (data not shown). In addition, we investigated simulation studies with different true "null" response rates, where the threshold values were chosen assuming null response rates of 0.25, but true null response rates were 0.35 or 0.45. For both "null" values of 0.35 and 0.45, we found $\alpha = 0.09$, which was very close to the nominal value of 0.10.

17

The posterior probabilities $Q_{j',l}$ of the two-step decision rule can be equal in extremely rare cases because these two probabilities were computed using the posterior draws from MCMC. For example, in scenario 3, where treatments A and B have the same response rate that is smaller than that of C, it is possible, though very unlikely, that $Q_{A,l}$ and $Q_{B,l}$ are equal at the second step of the decision rule. As a solution, one could randomly remove one of the two treatments or instead decide not to remove either arm and wait for a later look to make a decision.

Our group sequential snSMART is preferred for rare disease trials or trials where the accrual rate is relatively slow. If patient accrual is much faster than the timing of outcome measurements, most treatment allocations will be completed before interim analyses can be performed. In this case, the removal of a treatment arm will have a very limited effect in allocating patients to potentially better treatments.

Our two-step decision rule is currently only applicable to a three arm trial, where there is a single best or worst treatment if three treatments do not have the same response rate. Thus, future work includes the development of a more general decision rule that can be applied to an snSMART with more than three arms. Moreover, if many arms are compared at the same time, we would like to develop a decision rule that can remove more than one arm.

5 Acknowledgment

18

References

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

Lavori, P. W. and Dawson, R. (2000). A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **163**(1), 29–38.

Magirr, D., Jaki, T., and Whitehead, J. (2012). A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika*, **99**(2), 494–501.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(2), 331–355.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, **24**(10), 1455–1481.

Rosner, G. L. and Berry, D. A. (1995). A Bayesian group sequential design for a multiple arm randomized clinical trial. *Statistics in Medicine*, **14**(4), 381–394.

Shi, H. and Yin, G. (2019). Control of Type I error rates in Bayesian sequential designs. *Bayesian Analysis*, **14**(2), 399–425.

Shih, M.-C. and Lavori, P. W. (2013). Sequential methods for comparative effectiveness experiments: Point of care clinical trials. *Statistica Sinica*, **23**(4), 1775–1791.

Stallard, N. and Friede, T. (2008). A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine*, **27**(29), 6209–6227.

19

Stallard, N. and Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine*, **22**(5), 689–703.

Tamura, R. N., Krischer, J. P., Pagnoux, C., Micheletti, R., Grayson, P. C., Chen, Y.-F., and Merkel, P. A. (2016). A small n sequential multiple assignment randomized trial design for use in rare disease research. *Contemporary Clinical Trials*, **46**, 48–51.

Wei, B., Braun, T. M., Tamura, R. N., and Kidwell, K. M. (2018). A Bayesian analysis of small n sequential multiple assignment randomized trials (snSMARTs). *Statistics in Medicine*, **37**, 3723–3732.

Yin, G., Chen, N., and Jack Lee, J. (2012). Phase II trial design with Bayesian adaptive randomization and predictive probability. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**(2), 219–235.

Zhu, H., Yu, Q., and Mercante, D. E. (2017). A Bayesian sequential design with binary outcome. *Pharmaceutical Statistics*, **16**(3), 192–200.

20

Looks	Steps	τ_l	ψ_l	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
				P_{drop}	$1-P_b$	P_w	P_{drop}	$1-P_b$	P_w	P_{drop}	$1-P_b$	P_w	P_{drop}	$1-P_b$	P_w
							$N_T = 90$								
1	1	NA	0.90	0.10	NA	NA	0.52	1.00	1.00	0.21	1.00	1.00	0.56	1.00	0.99
		NA	0.89	0.10	NA	NA	0.55	1.00	1.00	0.22	1.00	1.00	0.58	1.00	0.99
1	2	0.91	0.95	0.10	NA	NA	0.46	0.96	0.96	0.54	1.00	1.00	0.78	1.00	0.93
		0.95	0.91	0.10	NA	NA	0.55	0.97	0.97	0.47	1.00	1.00	0.78	1.00	0.94
1	2	0.96, 0.96	0.96, 0.96	0.10	NA	NA	0.57	0.97	0.97	0.55	0.99	0.99	0.80	1.00	0.91
		0.96, 0.95	0.96, 0.95	0.10	NA	NA	0.61	0.97	0.97	0.60	1.00	1.00	0.84	1.00	0.91

Table 1: The proportion of runs that drop an arm (P_{drop}), the proportion of not dropping the best treatment if an arm is dropped ($1-P_b$), and the proportion of dropping the worst treatment if an arm is dropped (P_w) for all four scenarios listed in Section 3.2 with different type of dropping rule (one-step or two-step), different number of interim analyses (one look or two looks) and dropping threshold. Accrual rate is 3 people/ month and accrual time is 30 months for a total of 90 participants. For each case, 1000 runs are conducted.

Looks	Steps	τ_l	ψ_l	Treatment	Mean number of treatment responders in stage 2			
					Scenario 1	Scenario 2	Scenario 3	Scenario 4
					$N_T = 90$			
NA	NA	NA	NA	A	7.52	7.53	7.52	7.53
				B	7.41	14.89	7.42	13.39
				C	7.50	15.02	15.01	19.51
				Total	22.43	37.43	29.95	40.43
1	1	NA	0.89	A	7.29	4.48	6.26	4.17
				B	7.25	20.35	6.46	15.96
				C	7.24	20.28	21.10	31.86
				Total	21.78	45.11	33.81	51.99
2	1	0.95	0.91	A	7.27	4.55	6.07	3.83
				B	7.24	20.31	6.26	16.00
				C	7.28	20.14	22.08	32.84
				Total	21.79	45.01	34.40	52.67
2	2	0.96, 0.95	0.96, 0.95	A	7.27	4.50	5.86	3.78
				B	7.27	20.40	6.30	15.86
				C	7.34	19.96	22.20	33.21
				Total	21.88	44.85	34.36	52.85

Table 2: The average numbers of responders to the treatments in the second stage of a standard snSMART (snSMART without interim analyses) or a group sequential snSMART with the given type of rule (one-step or two-step), for a given number of interim analyses (one look or two looks) under all four scenarios listed in Section 3.2. The mean numbers of responders to each treatment and all treatments are listed for each design under each scenario.

22

Rate × Month	τ_l	ψ_l	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
			P_{drop}	$1-P_b$	P_w	P_{drop}	$1-P_b$	P_w	P_{drop}	$1-P_b$	P_w	P_{drop}	$1-P_b$	P_w
5 × 18	0.95, 0.96	0.95, 0.96	0.10	NA	NA	0.51	0.96	0.96	0.48	1.00	1.00	0.74	1.00	0.90
	0.96, 0.95	0.96, 0.95	0.10	NA	NA	0.54	0.97	0.97	0.53	1.00	1.00	0.77	1.00	0.90
3 × 30	0.96, 0.96	0.96, 0.96	0.10	NA	NA	0.57	0.97	0.97	0.55	0.99	0.99	0.80	1.00	0.91
	0.96, 0.95	0.96, 0.95	0.10	NA	NA	0.61	0.97	0.97	0.60	1.00	1.00	0.84	1.00	0.91
2 × 45	0.95, 0.96	0.95, 0.96	0.10	NA	NA	0.59	0.98	0.98	0.59	1.00	1.00	0.81	1.00	0.92
	0.96, 0.95	0.96, 0.95	0.10	NA	NA	0.62	0.98	0.98	0.63	1.00	1.00	0.85	1.00	0.92

Table 3: The proportion of runs that drop an arm (P_{drop}), the proportion of not the best treatment if an arm is dropped ($1-P_b$), and the proportion of dropping the worst treatment if an arm is dropped (P_w) for all four scenarios listed in Section 3.2 with different accrual rates and times, but same total sample sizes ($N_T = 90$) and same two-step rule and number of interim analyses (two looks). For each case, 1000 runs are conducted.
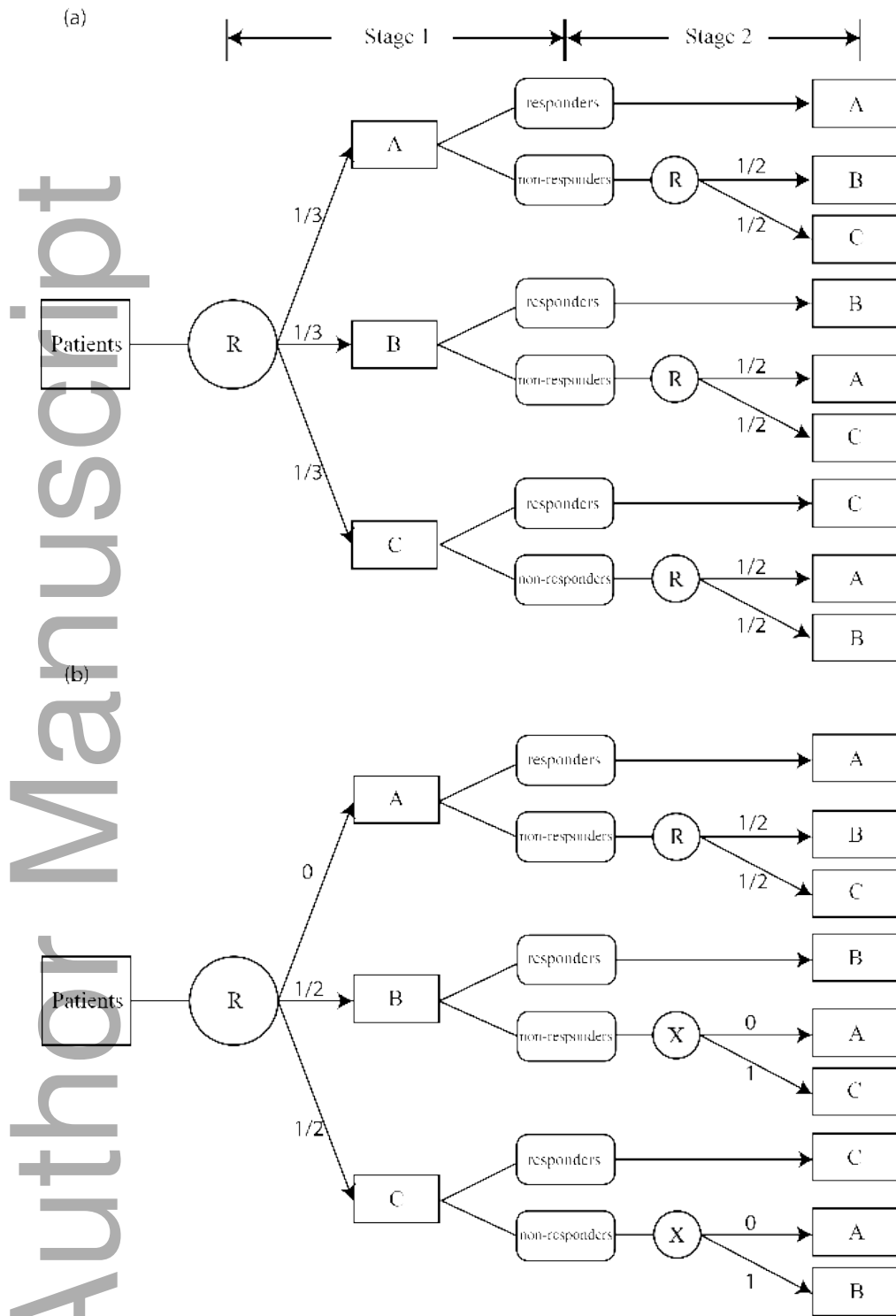
Figure 1: (a) A group sequential small n sequential multiple assignment randomized trial (snSMART) design before an arm being removed, which is also an snSMART design without interim analysis. (b) A group sequential snSMART design after treatment *A* is removed. The numbers around the arrows indicate the probabilities that a participant is assigned to the treatment. R represents randomization to the following treatments. X represents deterministic assignment to the following treatment.
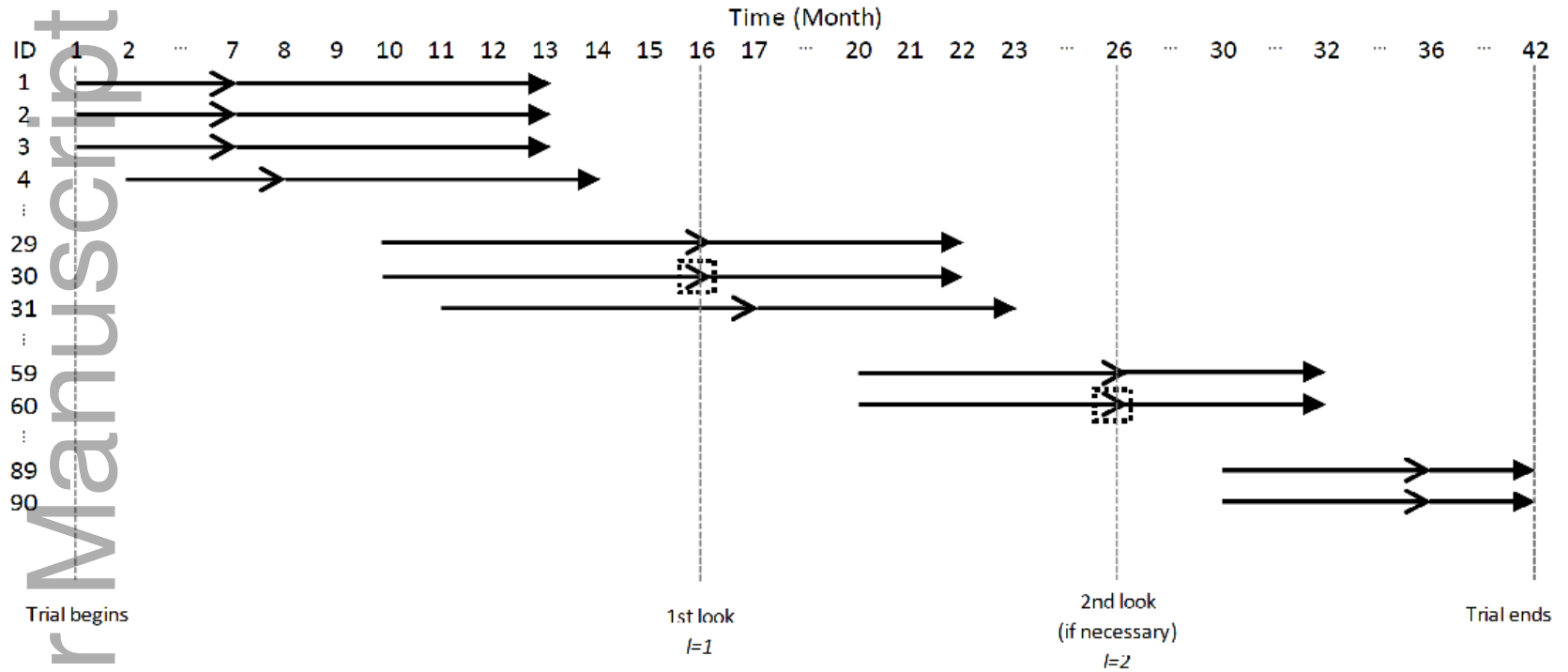
24

Figure 2: An illustration of how a group sequential snSMART with at most two interim analyses (looks) proceeds. Each row is an ID of a participant, and each column is the number of months after the trial begins. The participants are enrolled in the study at the rate of 3 people per month. Enrollment takes 30 months in total. The outcomes in each stage can be obtained from participants six months after the treatment assignment and the second stage treatments are assigned to participants immediately after their first stage outcomes are obtained. ⟶ shows the time duration when the participants are in the first stage, and ⟶ shows the time duration when the participants are in the second stage. Two dashed boxes indicate the events when interim analyses are conducted. Although the arrows of some participants may be aligned at the same start and end points, it represents that they start and end in the same months, not necessarily the same days.
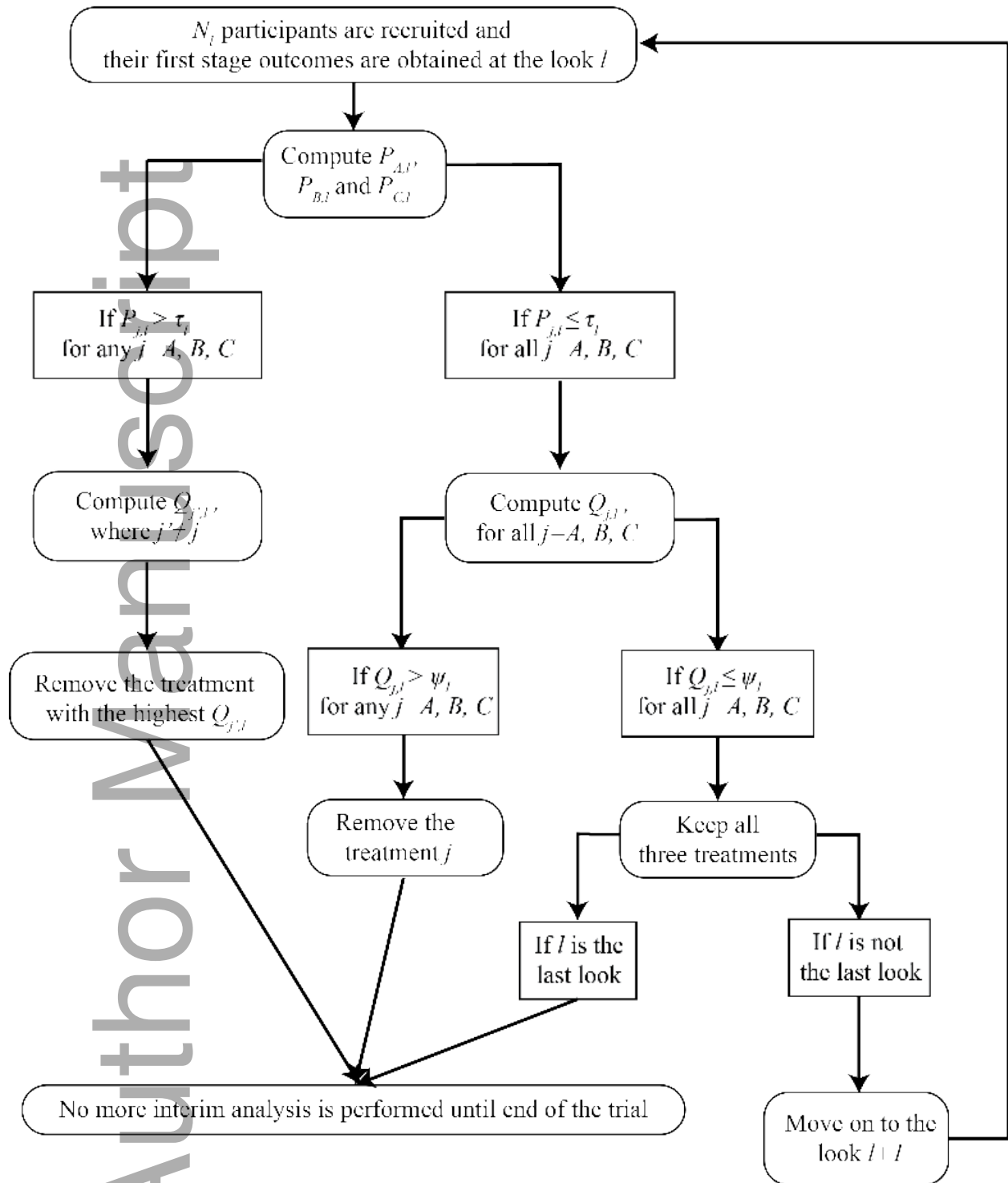
N_l participants are recruited and
their first stage outcomes are obtained at the look l

Compute $P_{A,l}$, $P_{B,l}$ and $P_{C,l}$

If $P_{j,l} > \tau_l$
for any $j \in A, B, C$

If $P_{j,l} \leq \tau_l$
for all $j \in A, B, C$

Compute $Q_{j',l}$,
where $j' \neq j$

Compute $Q_{j,l}$,
for all $j = A, B, C$

Remove the treatment
with the highest $Q_{j,l}$

If $Q_{j,l} > \psi_l$
for any $j \in A, B, C$

If $Q_{j,l} \leq \psi_l$
for all $j \in A, B, C$

Remove the
treatment j

Keep all
three treatments

If l is the
last look

If l is not
the last look

No more interim analysis is performed until end of the trial

Move on to the
look $l+1$

Figure 3: The detailed procedure of the proposed two-step Bayesian decision rule performed at an interim analysis l. If an one-step rule is applied, then the procedure starts from computing $Q_{j,l}$, $j = A, B, C$.

26
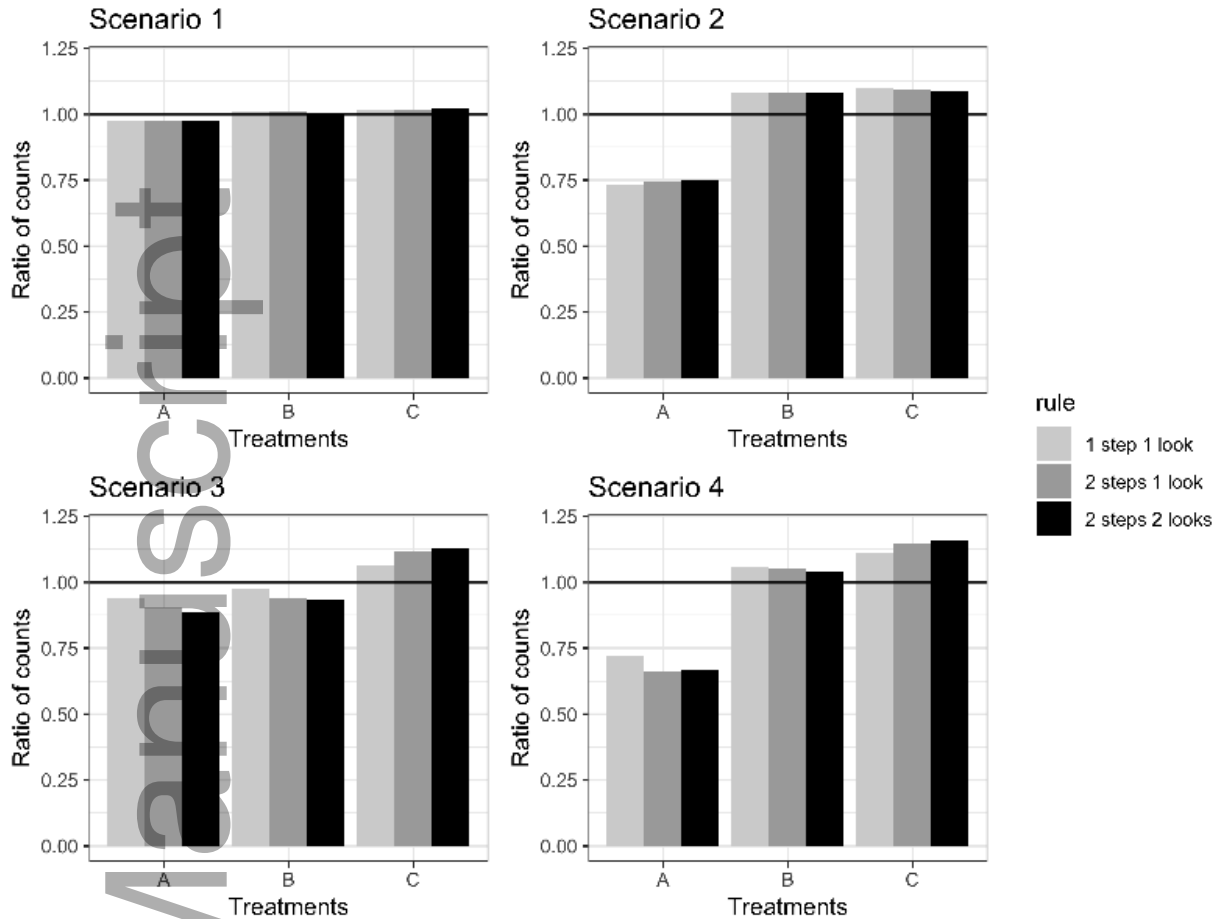
Figure 4: The ratio of the second stage participant count under a group sequential snSMART with the given rule (one-step or two-step) and number of maximum interim analyses (one look or two looks) to the second stage participant count under an snSMART without interim analyses. The four scenarios are listed in the Section 3.2. The total number of participants on trial $N_T = 90$.
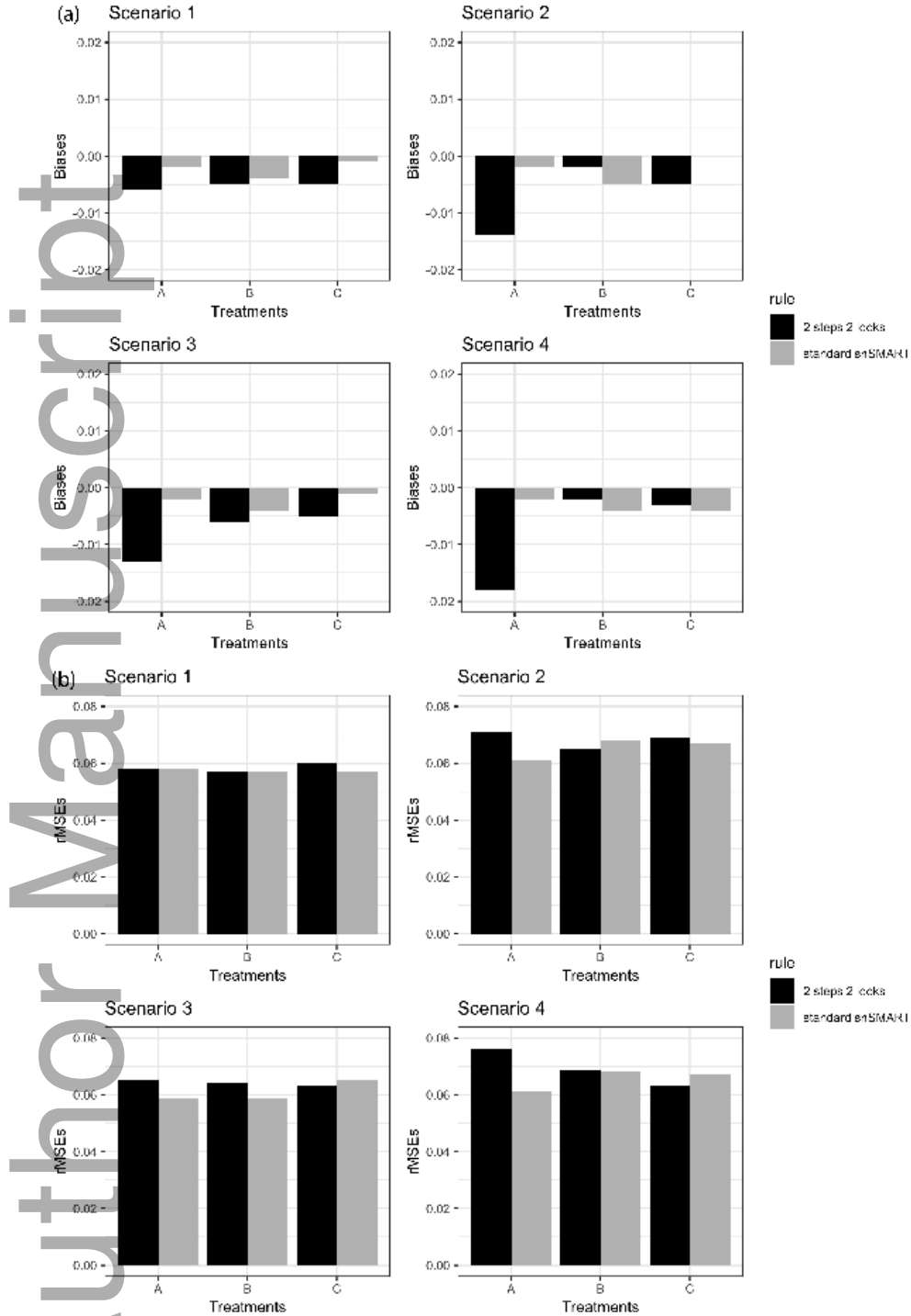
27

Figure 5: (a) The bias of the estimated response rates under the four scenarios listed in the Section 3.2. (b) The root mean squared error (rMSE) of the estimated response rates under the same four scenarios. "2 steps 2 looks" means the group sequential snSMART design using the two-step decision rule with at most two looks, and the "standard snSMART" means the snSMART without interim analyses. The total number of participants on trial $N_T = 90$.
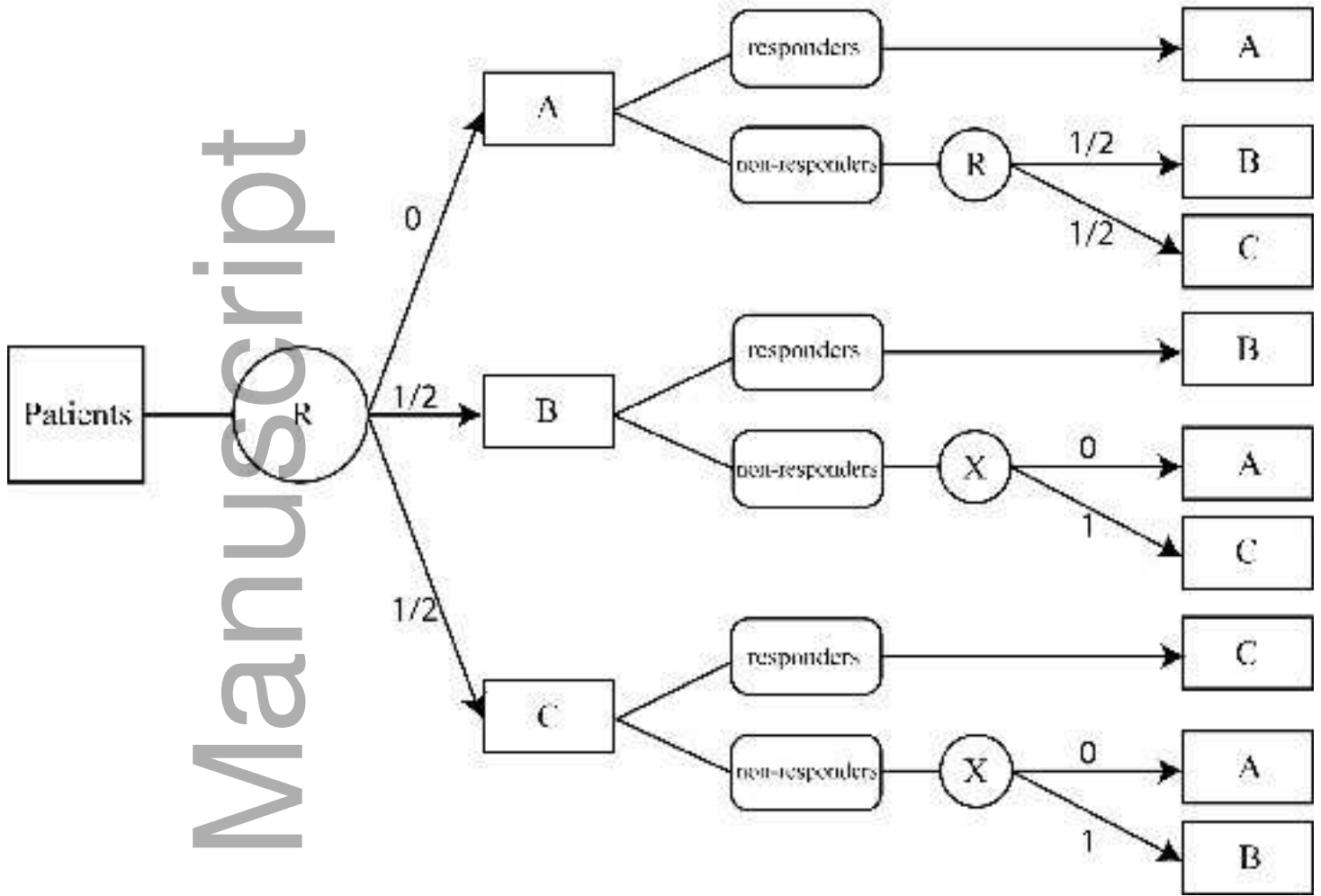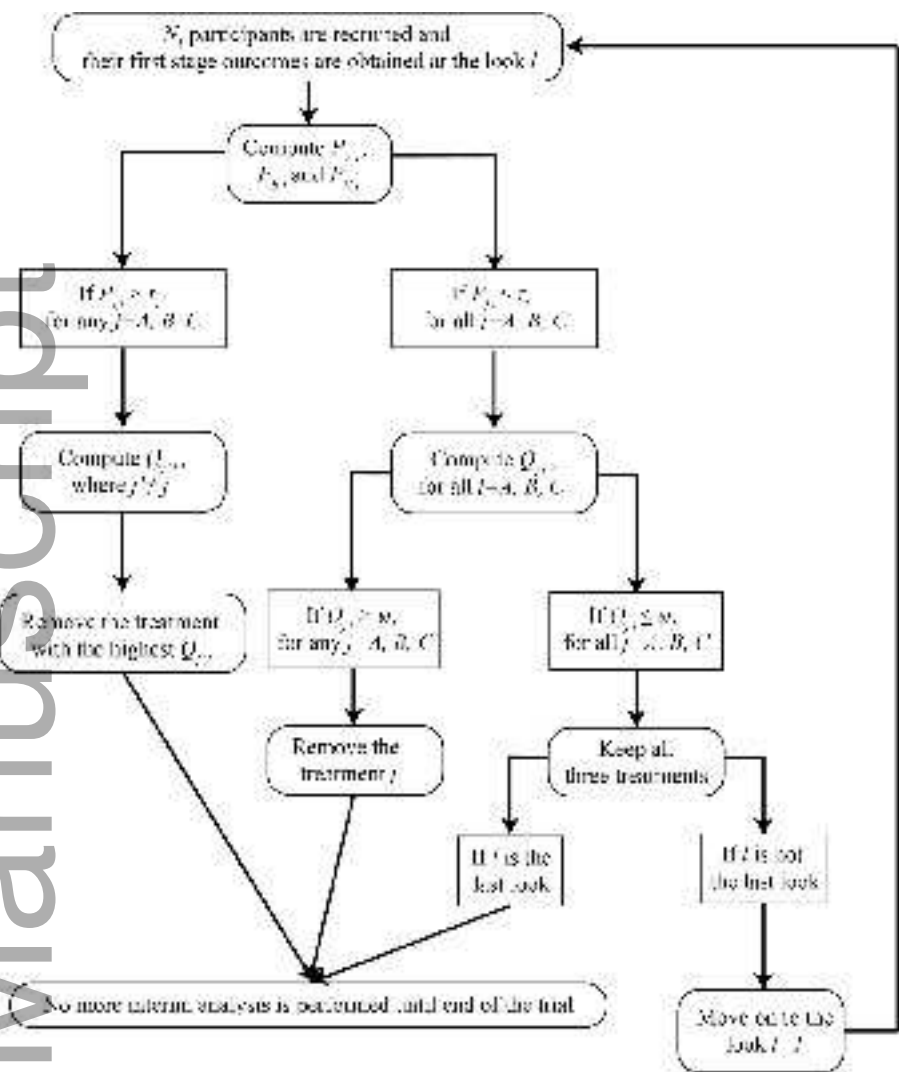
28
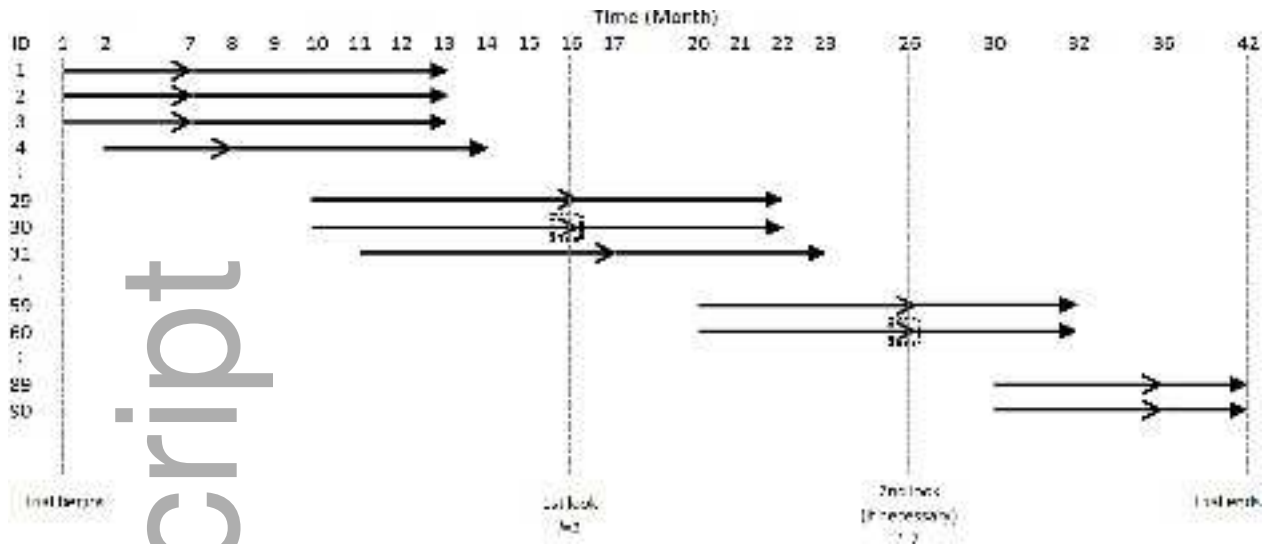
rssc_12406_fig3.png

(a)
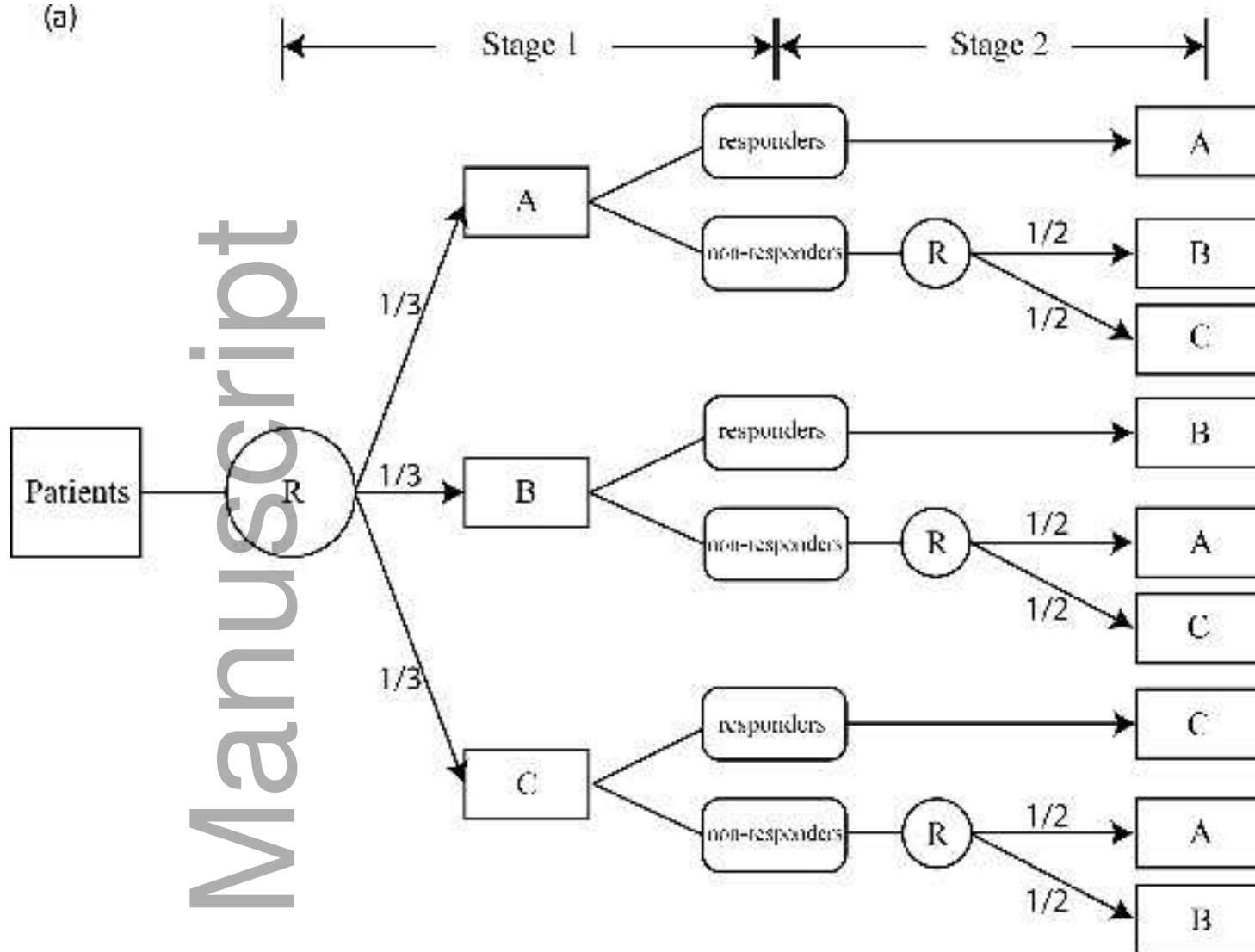
rssc_12406_figure1a.png

(b)

rssc_12406_figure1b.png

rssc_12406_figure2.png