2020

Flexible Methods for Accounting for Distributional Misspecification of the Response-Adaptive Randomization Ratio in Two-Arm Clinical Trials with Continuous or Survival Outcomes

Victoria C. Garcia
Virginia Commonwealth University

FLEXIBLE METHODS FOR ACCOUNTING FOR DISTRIBUTIONAL MISSPECIFICATION OF THE OPTIMAL RESPONSE-ADAPTIVE RANDOMIZATION RATIO IN TWO-ARM CLINICAL TRIALS WITH CONTINUOUS OR SURVIVAL OUTCOMES

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

by

VICTORIA C. GARCIA

M.P.H. Epidemiology, Eastern Virginia Medical School - 2014

B.S. Psychology, Virginia Commonwealth University - 2010

Chair: Adam P. Sima, Ph.D.

Associate Professor, Department of Biostatistics

Virginia Commonwealth University

Richmond, Virginia

July 2020

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Equal versus Response-Adaptive Randomization

Two considerations are paramount when designing a clinical trial. First is to ensure that the trial is sufficiently powerful to detect some clinically-meaningful difference between treatment groups, should one exist; while second is ensuring that the trial adheres to the ethical principles of research with human subjects set forth by the Belmont Report.[1,2] However, consider a phase III randomized clinical trial (RCT) where evidence exists suggesting that a novel treatment is superior to a current method of treatment (i.e., an active-control). In such a scenario, the power-versus-ethics dynamic of the RCT may be at-odds, a situation that will be discussed at length throughout the present work.

Equal randomization (ER) occurs when an equal number of trial participants are randomized to each study arm. In addition to being the most implement and conceptually-straightforward method of randomization in RCTs, ER is also the most powerful method of randomization when the responses obtained from each arm of the RCT vary equally. Therefore, implementing ER may save both time and money throughout the duration of the trial because a clinically-meaningful difference can be detected using response information obtained from fewer participants than would be required under any other type of randomization.[1] Equal randomization may be unethical if a

treatment difference exists despite its power and ease of implementation. Returning to the context of the phase III RCT where a novel treatment has shown evidence of outperforming a standard-of-care treatment, by definition, half of all trial participants will be randomized to the inferior active-control group.

In this instance, two ethical principles of research with human subjects are being violated. First is the principle of beneficence, or the "do no harm" principle which seeks to maximize the benefit and minimize the risk to human subjects.[2] Beneficence dictates that human subjects should undergo the most superior treatment available for their ailment. To allocate any subject to an inferior treatment could be considered harmful in that it is not acting in the subject's best interest: the subject is not benefiting from the best treatment available and may be experiencing a greater risk of experiencing poorer health outcomes as a result. The second principle of research with human subjects that may be violated is the justice principle which calls for the equitable treatment of and access to resources for every human subject.[2] Using similar logic, to deny any subject access to a superior treatment may be considered unjust treatment in that available resources that could benefit the subject are being withheld.

A more ethical randomization mechanism is one that, in the presence of a treatment difference, apportions subjects to competing treatment arms in such a way that power is maintained and the greatest number of subjects possible receives the most beneficial treatment.[3] Such a randomization method is called response-adaptive randomization (RAR). Response-adaptive randomization is a sequentially-updating process in which the treatment responses from participants who have completed the trial are used to influence randomization for incoming subjects such that randomization tends toward the better-performing treatment group, should one exist.[3–7] Existing RAR designs can be grouped into two broad categories: non-optimal (§1.2) and optimal designs (§1.3). Optimal RAR mechanisms strike a balance between randomization according to ethical and statistical pow-

er, while non-optimal RAR designs target ethical randomization. While optimal RAR is the focus of the present work, a brief description of non-optimal designs are provided in what follows.

1.2 Non-optimal Response-Adaptive Randomization Designs

There are a number of non-optimal RAR designs. Some randomize subjects according to a distributional link function (described in Appendix A.1), while others are adapted from methods originally purposed for binary response data or are based on nonparametric score functions. Examples of the former include the Doubly-adaptive Biased Coin Design [1,3,8–10] (DBCD), the Continuous Drop-the-Loser (CDL) design [11,12], and the Randomized Play-the-Winner (RPW) design [13]. A brief overview of these methods is available in Appendix A.2, and the DBCD as it relates to RAR is discussed later in this chapter.

Nonparametric scores can also be used to construct RAR designs for continuous and time-to-event response data. Rosenberger [14] developed nonparametric RAR designs for continuous data using treatment effect mapping (TEM), while Rosenberger and Seshaiyer [15] and Hallstrom et al. [16] used TEM for time-to-event response data. Bandyopadhyay and Biswas [17] developed the Wilcoxon-Mann-Whitney adaptive design (WAD) using components from both the Wilcoxon-Mann-Whitney (WMW) test [17–19] and the Friedman-Wei (FW) urn design [20–22]. Treatment effect mapping designs and WAD are described in more detail in Appendices A.3 and A.4, respectively.

1.3 Optimal Response-Adaptive Randomization Designs for Continuous Response Data

Non-optimal RAR designs target ethical randomization. Trials focused purely on the ethical aspect could allocate nearly all subjects to the superior treatment without regard for the necessary accrual of subjects in each treatment group to maintain an adequate level of power to detect a treatment difference. Thus, the determination of treatment superiority would be impossible. Optimal RAR,

on the other hand, optimizes an objective function based upon an importance criterion subject to a variance constraint.[4–7] An importance criterion is the primary ethical concern of the trial, such as minimizing the expected number of treatment failures, the total expected sample size, the total expected cost, or the expected number of subjects allocated to the inferior treatment. This importance criterion is translated into an objective function governed by the treatment-specific sample sizes and the moments of some assumed response distribution. A variance constraint is imposed upon the ethical concern represented by the objective function such that a larger number of subjects will receive the superior treatment when a treatment difference exists while also ensuring that the number of subjects necessary to maintain power will undergo the inferior treatment. If no treatment difference exists, the variance constraint ensures ER is maintained. The variance constraint is established by setting-equal a constant and the asymptotic variance of the estimated treatment difference, where the variance constraint is focused on the variance related to the statistic used to test the study hypothesis. Optimization occurs when the objective function is subject to the variance constraint, producing an optimization problem which returns the powerful and ethical optimal randomization ratio. Optimization details are provided in Appendix A.5 and will be referenced throughout the present work.

Jennison and Turnbill[23] and Biswas and Mandal[5] developed early iterations of optimal RAR designs; each of these are described with greater detail in Appendices A.6 and A.7, respectively. From these, Zhang and Rosenberger[6] developed an optimal RAR design. In this design, assume two groups, a treatment T and an active control C, where groups are denoted by g, i.e., $g \in \{T, C\}$, and group-specific responses have a mean μ_g and variance σ_g^2. As well, assume σ_T^2 and σ_C^2 are unknown and smaller responses are more desirable. Then, this design minimizes the total expected

response using the following objective function:

$$\begin{aligned}
&\underset{n_T/n_C}{min}\left\{E\left[\sum_{g=T}^{C}\sum_{i=1}^{n_g}Y_{gi}\right]\right\} \\
&= \underset{n_T/n_C}{min}\left\{E\left[\sum_{i=1}^{n_T}Y_{Ti}+\sum_{i=1}^{n_C}Y_{Ci}\right]\right\} \\
&= \underset{n_T/n_C}{min}\left\{n_T\ \mu_T + n_C\ \mu_C\right\}, \qquad (1.1)
\end{aligned}$$

where $E[\cdot]$ is the expectation function. Solving this according to the asymptotic variance of the estimated treatment difference,

$$Var(\hat{\mu}_C - \hat{\mu}_T) = \frac{\sigma_C^2}{n_C} + \frac{\sigma_T^2}{n_T} = V, \text{ for some constant } V, \qquad (1.2)$$

where n_g is the number of subjects allocated to treatment g and setting V equal to one, as suggested by Biswas and Bhattacharya[7], gives the following optimal RAR ratio to T:

$$\rho_{ZR} = \frac{\sigma_T\sqrt{\mu_C}}{\sigma_T\sqrt{\mu_C} + \sigma_C\sqrt{\mu_T}}. \qquad (1.3)$$

A noted criticism of this design is that it degenerates if either of the group-specific response mean estimates, $\hat{\mu}_T$ or $\hat{\mu}_C$, are negative.[3,6] Otherwise, this design can be applied whenever response values are known to be positive. One proposed solution for negative responses is to add a sufficiently large positive constant, $y_{pos} \in \mathcal{R}^+$, such that all responses are positive-valued.[3,6] Then, Zhang and Rosenberger[6] propose the following probability of randomization to treatment T:

$$\rho_{ZR;y_{pos}} = \frac{\sigma_T\sqrt{\mu_C + y_{pos}}}{\sigma_T\sqrt{\mu_C + y_{pos}} + \sigma_C\sqrt{\mu_T + y_{pos}}}. \qquad (1.4)$$

Atkinson and Biswas[3] caution, however, that randomization will tend toward Neyman-Pearson randomization,

$$\rho_{NP} = \frac{\sigma_T}{\sigma_T + \sigma_C}, \qquad (1.5)$$

as y_{pos} tends toward infinity. Neyman-Pearson randomization is optimal, but not ethical. Therefore,

for responses taking negative values, Atkinson and Biswas [3] suggested assuming a distribution left-truncated at zero. Such a distribution is defined as $P(X = x) = f(x)$ for $x \geq 0$, 0 otherwise, where $f(x)$ is the density function of a random variable X.

Nevertheless, the Zhang and Rosenberger [6] randomization rule can be applied whenever response values are known to be positive. One such example is when responses are assumed to be exponentially-distributed, i.e., $Y_g \sim Exp(\theta_g)$, where the expected group-specific response is given by $E(Y_g) = \theta_g$. The objective function is:

$$\min_{n_T/n_C} \Big\{ n_T \, \theta_T + n_C \, \theta_C \Big\}, \tag{1.6}$$

and, since $Var(Y_g) = \theta_g^2$, the variance constraint of (1.2) takes the form:

$$Var(\hat{\theta}_C - \hat{\theta}_T) = \frac{\theta_C^2}{n_C} + \frac{\theta_T^2}{n_T} = V, \text{ for some constant } V. \tag{1.7}$$

Per this optimal RAR framework, solving (1.6) subject to (1.7) gives the following optimal RAR ratio to T for exponential responses:

$$\rho_{ZR;E} = \frac{\sqrt{\theta_T}}{\sqrt{\theta_T} + \sqrt{\theta_C}}. \tag{1.8}$$

Similar results can be obtained when response values followed gamma distributions, i.e., $Y_g \sim Gam(\alpha_g, \beta_g)$, as well. The expected group-specific response is given by $E(Y_g) = \alpha_g \beta_g$; therefore, the objective function to minimize the total expected response is given by:

$$\min_{n_T/n_C} \Big\{ n_T \, \alpha_T \beta_T + n_C \, \alpha_C \beta_C \Big\}; \tag{1.9}$$

and, since $Var(Y_g) = \alpha_g \beta_g^2$, the variance constraint is given by:

$$Var(\hat{\alpha}_C \hat{\beta}_C - \hat{\alpha}_T \hat{\beta}_T) = \frac{\alpha_C \beta_C^2}{n_C} + \frac{\alpha_T \beta_T^2}{n_T} = V, \text{ for some constant } V. \tag{1.10}$$

The solution to (1.9) subject to (1.10) yields a similar optimal randomization probability to T given

in (1.8):

$$\rho_{ZR;G} = \frac{\sqrt{\alpha_T}}{\sqrt{\alpha_T} + \sqrt{\alpha_C}}. \tag{1.11}$$

1.3.1 Tuning Based on Treatment Effect

Biswas and Bhattacharya[7] capitalized on the applicability of the Zhang and Rosenberger[6] design to any positive real-valued continuous response distribution in order to create an optimal RAR framework.[4,6,7] Introduced into this design are:

1. Ψ, a general representation of the function of the moments comprising the objective function,
2. $\rho_0 \in (0, 1)$, a prespecified target randomization probability for a given effect size,
3. $\tau \geq 0$, a tuning parameter used to achieve ρ_0 when a desired treatment difference is observed,
4. Additional constraints based on $n_g/(n_T + n_C) \geq b$ intended to bound the randomization ratio between b and $1 - b$, where $b \in [0, 1/2)$ and small.

Biswas and Bhattacharya[4] applied these modifications to the Zhang and Rosenberger design. Representing the moments of the objective function with $\Psi(\cdot)$, the objective function of (1.1) is given by:

$$\underset{n_T/n_C}{min} \left\{ n_T \ \Psi(\mu_T) + n_C \ \Psi(\mu_C) \right\}. \tag{1.12}$$

Subjecting this to the variance constraint of (1.2) yields the optimal RAR ratio to T given by:

$$\rho = \frac{\sigma_T \sqrt{\Psi(\mu_C)}}{\sigma_T \sqrt{\Psi(\mu_C)} + \sigma_C \sqrt{\Psi(\mu_T)}}, \tag{1.13}$$

The choice of $\Psi(\cdot)$ is objective-specific based upon the desired importance criterion.[7] Where the goal of Zhang and Rosenberger's design[6] was to minimize the total expected mean response, Biswas and Bhattacharya[4] suggested using $\Psi(x) = x^\tau$. The desired randomization ratio, ρ_0, is defined

according to μ_{g0} and σ_{g0}, the group-specific moments necessary to observe a desired treatment effect. Therefore,

$$\rho_0 = \frac{\sigma_{T0}\ \Psi(\mu_{C0})}{\sigma_{T0}\ \Psi(\mu_{C0}) + \sigma_{C0}\ \Psi(\mu_{T0})} \tag{1.14}$$

$$= \frac{\sigma_{T0}\ \mu_{C0}^{\tau/2}}{\sigma_{T0}\ \mu_{C0}^{\tau/2} + \sigma_{C0}\ \mu_{T0}^{\tau/2}}. \tag{1.15}$$

It follows that:

$$\tau = 2\ \frac{log\left(\frac{1-\rho_0}{\rho_0}\right)\ +\ log\left(\frac{\sigma_{T0}}{\sigma_{C0}}\right)}{log(\mu_{C0}/\mu_{T0})}. \tag{1.16}$$

This value, τ, permits the integration and attainment of a predetermined target randomization proportion within the randomization procedure itself. This ensures that the randomization proportion occurs at approximately the same rate for a given treatment difference regardless of the scale of the outcome measure.

Subjecting (1.12) to (1.2) using the Biswas and Bhattacharya's[4] $\Psi(x) = x^{\tau}$ generalization of Zhang and Rosenberger's[6] framework yields the optimal RAR ratio given by:

$$\rho' = \frac{\sigma_T\ \mu_C^{\tau/2}}{\sigma_T\ \mu_C^{\tau/2} + \sigma_C\ \mu_T^{\tau/2}}, \tag{1.17}$$

and bounded as follows:

$$\rho = \begin{cases} b, & \text{if } \{\Psi(\mu_T) > 0 \text{ and } \Psi_(\mu_C) > 0 \text{ and } \rho' < b\} \\ \rho, & \text{if } \{\Psi(\mu_T) > 0 \text{ and } \Psi(\mu_C) > 0 \text{ and } b \leq \rho' \leq 1-b\} \\ 1-b, & \text{if } \{\Psi(\mu_T) > 0 \text{ and } \Psi(\mu_C) > 0 \text{ and } \rho' > 1-b\} \\ b, & \text{if } \{\Psi(\mu_T) > 0 \text{ and } \Psi(\mu_C) < 0\} \\ 1-b, & \text{if } \{\Psi(\mu_T) < 0 \text{ and } \Psi(\mu_C) > 0\} \\ 1-b, & \text{if } \{\Psi(\mu_T) < 0 \text{ and } \Psi(\mu_C) < 0 \text{ and } \frac{\sigma_T}{\sigma_C} < \sqrt{\frac{\Psi(\mu_T)}{\Psi(\mu_C)}}\} \\ b, & \text{if } \{\Psi(\mu_T) < 0 \text{ and } \Psi(\mu_C) < 0 \text{ and } \frac{\sigma_T}{\sigma_C} > \sqrt{\frac{\Psi(\mu_T)}{\Psi(\mu_C)}}\}, \end{cases}$$

where ρ represents the probability that a subject is randomized to treatment group T. Because the Biswas and Bhattacharya [7] optimal RAR design provides a general framework for the optimal RAR process, is bounded between b and $1-b$, and allows randomization to target a desired proportion which is achieved when a desired treatment difference is observed, this framework is the one used in the present work for continuous response data.

1.4 Motivation: The Problem of Misspecification

For most types of statistical analyses, data are typically assessed for some distributional assumption prior to statistical analysis as a safeguard against erroneous statistical conclusions. This is not the case for response-adaptive RCTs. Recalling Section 1.1, when planning a RCT, an investigator must know the number of subjects that will be enrolled into the trial in order to detect a clinically-meaningful difference between treatment groups. Typically, in the context of RCTs with continuous response data, this value is determined by assuming subject responses will follow a normal distribution. Section 1.3 described how optimal RAR is dictated by the moments of the assumed response distribution. Therefore, for RCTs employing optimal RAR, it is ideal that the

observed response data match the assumed distribution. If the observed responses do not follow the intended distribution, the randomization mechanism may be misspecified. Consequently, participant randomization may differ from what was intended, possibly producing study characteristics (i.e., ethics and power) dissimilar from those originally anticipated. These effects may be more profound in smaller samples and scenarios where the intended response distribution is symmetric yet observed response data are skewed or vice versa. Though means tend to be robust against distributional misspecification, treatment group variance estimation is heavily contingent upon the correct specification of the assumed distribution.[24]

1.4.1 The Biasedness Dilemma

Consider the randomization ratio of (1.17). Assume that ρ' is already bounded and is, therefore, ρ. As discussed in Section 1.3, (1.17) is a function of some unknown group-specific population parameters μ_g and σ_g^2. However, pursuant to the optimal RAR process,[4–7] randomization is dictated by the observed response information obtained from subjects who have already completed the trial. Therefore, the observed optimal RAR ratio for continuous response data is calculated using the estimated treatment-specific mean and variance of the response data obtained from subjects $\{1, 2, \cdots, i\}$, denoted $\hat{\mu}_{gi}$ and $\hat{\sigma}_{gi}^2$, respectively. Therefore, the RAR ratio for participant $i+1$ to T is based on the sample statistics, and is denoted by:

$$\hat{\rho}_{T;i+1} = \frac{\hat{\sigma}_{Ti}\ \hat{\mu}_{Ci}^{\tau/2}}{\hat{\sigma}_{Ti}\ \hat{\mu}_{Ci}^{\tau/2} + \hat{\sigma}_{Ci}\ \hat{\mu}_{Ti}^{\tau/2}}. \tag{1.18}$$

By Jensen's Inequality,[24] replacing the distributional parameters with sample estimates could result in a biased estimate of the randomization ratio. Despite this, plug-in estimators, specifically maximum likelihood estimators (MLEs), are convention and, therefore, are used in the present work.

1.4.2 Objective and Outline of Present Work

There exists a paucity of information regarding the trial ramifications that may occur if the distributional assumption of the participant response data is not met. That is, existing literature falls short of addressing the behavior of the derived methodologies in the presence of distributional misspecification of the observed response data, a conceivable possibility given the assumed distribution is chosen prior to data collection.[4–7,23] This is an important consideration given randomization of incoming participants is contingent upon the response data obtained to that point, and the estimation of the mean and variance used to estimate the RAR ratio is effected by the specification of the response data. Traditionally, the DBCD is employed to account for unintended trial characteristics resulting from a small sample size, e.g., large variance.[8–10] This procedure operates as a function of the optimal RAR ratio, and therefore is not, in itself, optimal. The DBCD dampens the effect of the power-versus-ethics dynamic of the optimal RAR framework, no longer randomizing subjects based on the observed effect size. Of interest in the present work, however, is to study the effects of misspecification on optimal RAR. Therefore, the DBCD was not used.

The present work sought to identify flexible methodologies that produced an estimator of the optimal RAR ratio that was robust against departures from a prespecified parametric response distribution. To this end, nonparametric methods were first used to estimate the empirical distribution of the observed response data in order to obtain distribution-free estimates of the group-specific sample moments (Chapter 2). These were, in turn, used to develop an estimator of the RAR ratio that is robust to distributional misspecification. A step further was the utilization of the weighted-average approach, a method that allowed the randomization ratio to adjust automatically according to the distributional fit of a set of parametric candidate response distributions, regardless of the a priori distributional assumption (Chapter 3). Chapters 2 and 3 detail these approaches in the context of continuous response data. Chapter 4 centers about right-censored survival out-

comes, evaluating the misspecification of the conventional Zhang and Rosenberger[25] RAR ratio and developing a RAR ratio using the cumulative hazard function. The final chapter (Chapter 5) discusses implications, limitations, and future work.

Chapter 2

Robust Estimation of Continuous Moments Using the Empirical Distribution Function

2.1 Distribution-independent Estimation of Moments

The goal of the present chapter is to obtain distribution-independent estimates of the group-specific means and SDs of the observed response data. These estimates, in turn, are used to produce an estimate of the RAR ratio that is robust against distributional misspecification. Distribution-independent moment estimates can be obtained using the cumulative distribution function [26–29] (CDF), $F(y)$, as follows. For a random variable Y,

$$E(Y) = \int_0^\infty (1 - F(y))dy, \tag{2.1}$$

$$E(Y^2) = 2\int_0^\infty y(1 - F(y))dy, \tag{2.2}$$

$$\text{Then, } Var(Y) = E(Y^2) - E(Y)^2, \tag{2.3}$$

$$\text{and, } SD(Y) = Var(Y)^{1/2}, \tag{2.4}$$

where (2.1) and (2.2) are derived in Muldowney et al. [29] The CDF lacks an easily-estimable closed-form solution for observed response data that are not assumed to follow a parametric distribution. [26–28] Therefore, obtaining distribution-independent moment estimates requires an empirical

estimate of the CDF (eCDF).[26–29] The eCDF can be estimated by sorting the observed response data into ascending sequence and setting the mass of each observed response value at a height equal to the number of responses less than or equal to that response value, divided by the total number of observed responses. While this is a simple and effective approach for obtaining the underlying distribution function, the resulting data-generating function is a step function modeling a discrete random variable. Therefore, because the focus of this work is on continuous response data, the relationship between the survival function[26–28] (SF), $S(y)$, and the CDF,

$$F(y) = 1 - S(y), \tag{2.5}$$

was exploited. With this, continuous fits of the empirical distribution function, denoted $\hat{F}(y)$, could be obtained such that distribution-independent estimates of the means and SDs could be estimated from the observed response data directly. A number of methodologies have been developed to estimate the SF,[30–32] but piecewise-linear estimation of the Kaplan-Meier (KM) estimate of the SF (PLKM) by Kaczynski et al.[33] and Hazard Estimation with Flexible Tails (HEFT) by Kooperberg et al.[34] were utilized in the present work. The PLKM and HEFT methods were selected for their simplicity and complexity, respectively, and are detailed in Sections 2.1.1 and 2.1.2, respectively. Once these empirical SF modeling approaches yielded $\hat{F}(y)$, group-specific distribution-independent estimates of the mean and SD, $\widehat{E(Y_g)}$ and $\widehat{SD(Y_g)}$, respectively, were obtained as follows:

$$\widehat{E(Y_g)} = \int_0^\infty (1 - \hat{F}(y_g))dy_g, \tag{2.6}$$

$$\widehat{E(Y_g^2)} = 2\int_0^\infty y_g(1 - \hat{F}(y_g))dy_g, \tag{2.7}$$

$$\text{Then, } \widehat{Var(Y_g)} = \widehat{E(Y_g^2)} - \widehat{E(Y_g)}^2, \tag{2.8}$$

$$\text{and, } \widehat{SD(Y_g)} = \widehat{Var(Y_g)}^{1/2}. \tag{2.9}$$

Plugging these empirical moment estimates into the equation for the RAR ratio given by (1.18) produces a robust estimator of the RAR ratio.

2.1.1 Piecewise-linear Estimation of the Kaplan-Meier Estimate of the Survival Function

Mentioned previously, methods of estimating the SF were exploited so that continuous distribution functions could be obtained from the observed response data. The first method selected for this task relied upon one of the most common and straightforward estimators of the SF: the KM estimator, denoted $S_{KM}(y)$.[31–33] For survival outcomes, $S_{KM}(y)$ may be interpreted as the probability of surviving until immediately prior to a given time y.[i] In the absence of censoring, however, $S_{KM}(y)$ is simply the complement of the previously-detailed eCDF.[32]

2.1.1.1 Fitting PLKM

Letting j index a set of K observed subject responses, denoted $y_{(j)}$ $(j = 1, 2, \cdots, K)$, ordered such that $y_{(1)} < y_{(2)} < \cdots < y_{(K)}$, where $y_{(j)}$ denotes the j^{th} ordered value in the set of K observed responses, the KM probability of survival corresponding to each observed response value are sequenced in descending order, $S_{KM}(y_{(1)}) > S_{KM}(y_{(2)}) > \cdots > S_{KM}(y_{(K)})$, generating a decreasing step function. In order to obtain a continuous estimate of $S_{KM}(y)$, information between each observed response value (i.e., each downward step) was interpolated using a series of piecewise-linear functions.[33] To this end, knot points were placed at each observed response value at heights given by $h_{(j)} = \frac{K-j}{K-1}$, such that $h_{(1)} = 1 > h_{(2)} > \cdots > h_{(K-1)} > h_{(K)} = 0$, and were connected using line segments with slope $\frac{h_{(j)}-h_{(j+1)}}{y_{(j)}-y_{(j+1)}}$ for $j = 1, 2, \cdots, K-1$. This concatenation of line segments produced $\hat{S}^P(y)$, a continuous piecewise-linear approximation of the discrete KM estimate of the SF.

[i] y, versus more conventional t notation, used to represent survival time (1) to ensure consistent representation of observed response data, and (2) because present methods, though typically discussed in a survival context, are centered about continuous response data in this chapter.

2.1.1.2 Moment Estimation Using PLKM Fit

Using $\hat{S}^P(y)$, distribution-independent moments were estimated using numerical integration by the trapezoidal rule.[35] Specifically, $\widehat{E(Y)}^P$ was found by summing the area of the trapezoids bounded on the right and the left by $y_{(j+1)}$ and $y_{(j)}$, respectively, and above and below by $\hat{S}^P(y)$ and 0, respectively. Overlooking group affiliation[ii] for the following derivations, (2.6) gives:

$$\widehat{E(Y)}^P = \int_{y_{(1)}}^{y_{(K)}} \hat{S}^P(y)dy \tag{2.10}$$

$$\approx \frac{1}{2}\sum_{j=1}^{K-1}\Big(h_{(j)} + h_{(j+1)}\Big)\Delta y_j, \tag{2.11}$$

where $\Delta y_j = (y_{(j+1)} - y_{(j)})$. Next, to obtain $\widehat{E(Y^2)}^P$, $\hat{S}_2^P(y)$, a piecewise-linear curve representing the height of each knot point multiplied by the observed response value corresponding to the location of that knot point, i.e.,

$$\hat{S}_2^P(y) = h_{(j)} \times y_{(j)}, \tag{2.12}$$

was calculated. With this, $\widehat{E(Y^2)}^P$ was calculated by summing the same trapezoids used to calculate $\widehat{E(Y)}^P$, except bounded above by $\hat{S}_2^P(y)$ instead of $\hat{S}^P(y)$. Building upon (2.7),

$$\widehat{E(Y^2)}^P = 2\int_{y_{(1)}}^{y_{(K)}} y\hat{S}_2^P(y)dy \tag{2.13}$$

$$\approx 2\sum_{j=1}^{K-1}\Big(h_{(j)} \times y_{(j)} + h_{(j+1)} \times y_{(j+1)}\Big)\Delta y_j. \tag{2.14}$$

The distribution-independent estimate of the variance follows directly from (2.8) and $\widehat{SD(Y)}^P$ from (2.9). With this, defining $\hat{\rho}_{T;i+1}$ of (1.18), a function of $\{\hat{\mu}_{gi},\ \hat{\sigma}_{gi}\}$, as $\hat{\rho}^P_{T;i+1}$, a function of $\left\{\widehat{E(Y_{gi})}^P,\ \widehat{SD(Y_{gi})}^P\right\}$ gives the robust RAR estimator obtained using the PLKM empirical estimation method.

[ii] Group affiliation dropped for remainder of PLKM moment derivations for ease of notation and without loss of generality.

2.1.2 Hazard Estimation with Flexible Tails

The second method for obtaining a continuous distribution function from the observed response data was HEFT estimation, a more sophisticated and procedurally-complex estimation method than PLKM.[34,36] Hansen et al. states that, in survival analysis, the hazard function, $\lambda(y)dy$, "is often of-interest since it can be interpreted as the probability that someone dies in the next time-interval of infinitesimal length dy, given he is alive at time y." [36] Attempting to model the hazard function directly can be problematic, however.[31,32,34,36,37] Therefore, Kooperberg et al.[34,37] and Hansen et al.[36] suggested modeling the logarithm of the unknown hazard function (i.e., log-hazard function) instead. Compared to the hazard, density, or distribution function, positivity constraints are not a concern when modeling the log-hazard function, and such an approach lends itself well to Cox proportional hazards modeling. Described in the following section, the HEFT estimation method uses cubic splines to model the log-hazard function in a way that allows for and captures a wide range of tail behavior.

Polynomial splines are piecewise polynomials of some degree d, and knots mark the breakpoints from one polynomial to the next.[34,36] Splines satisfy smoothness constraints that describe how the different pieces are to be joined, typically in terms of s, the number of continuous derivatives exhibited by the piecewise polynomials. Specifically, cubic splines are piecewise cubic functions having two continuous derivatives, allowing jumps in the third derivative at the knots. Given a degree d and a knot vector $\boldsymbol{y} = (y_1, y_2, \cdots, y_K)'$, the collection of polynomial splines having s continuous derivatives forms a linear space. The collection of linear splines with knot sequence $\boldsymbol{y}$ is spanned by the functions:

$$1, x, (x - y_1)_+, \cdots, (x - y_K)_+, \tag{2.15}$$

where $(\cdot)_+ = max(\cdot, 0)$. This is the truncated power basis. In the case of the cubic spline, where

$d = 3$ and $s = 2$, the basis, $\boldsymbol{B}(\boldsymbol{y})$, is spanned by:

$$1, x, x^2, x^3, (x - y_1)^3_+, \cdots, (x - y_K)^3_+. \tag{2.16}$$

Technical details regarding polynomial spline modeling are available in Chapter 2 of Hansen et al. [36]

2.1.2.1 Fitting HEFT

Let f be a positive density function on $(0, \infty)$, $\varphi = log(f)$ be the log-density function, $\lambda = f/(1-F)$ be the hazard function, $\alpha = log(\lambda)$ be the log-hazard function, and, $Q = F^{-1}$ be the quantile function such that $Q(F(y)) = y$ for $y > 0$ and $F(Q(p)) = p$ for $0 < p < 1$. Then, for $y \geq 0$,

$$1 - F(y) = exp\bigg(- \int_0^y \lambda(u)du \bigg) \tag{2.17}$$

$$= exp\bigg(- \int_0^y exp(\alpha(u))du \bigg). \tag{2.18}$$

And, since (i) $F(y) < 1$ for $0 < y < \infty$ and (ii) $\lim_{y \to \infty} F(y) = 1$, it was concluded that

$$\int_0^y exp(\alpha(u))du < \infty \text{ for } 0 < y < \infty, \text{ and} \tag{2.19}$$

$$\int_0^\infty exp(\alpha(y))dy = \infty. \tag{2.20}$$

Noting $\lambda = exp(\alpha)$, for $y > 0$,

$$f(y) = exp(\alpha(y))\ exp\bigg(- \int_0^y exp(\alpha(u))du \bigg), \text{ and} \tag{2.21}$$

$$\varphi(y) = \alpha(y) - \int_0^y exp(\alpha(u))du. \tag{2.22}$$

Now, in an arrangement similar to that of the PLKM method, consider $K \geq 3$ observed responses given by $y_{(j)}$ $(j = 1, 2, \cdots, K)$ such that $0 < y_{(1)} < y_{(2)} < \cdots < y_{(K)} < \infty$, and let $p = K - 2$. As well, let $\mathbb{G}$ be a p-dimensional space of twice-continuously differentiable functions s on $[0, \infty)$ such that s is constant on $[0, y_{(1)}]$ and $[y_{(K)}, \infty)$, and the restriction of s to each of the intervals

$[y_{(1)}, y_{(2)}], \cdots, [y_{(K-1)}, y_{(K)}]$ is a cubic polynomial. With this, the functions in $\mathbb{G}$ are cubic splines having simple knots at $y_{(1)}, y_{(2)}, \cdots, y_{(K)}$. Now, let $B_1(y), B_2(y), \cdots, B_p(y)$ be a basis of this space such that $B_p(y) = 1$ on $[0, \infty)$ and $B_1(y), B_2(y), \cdots, B_{p-1}(y) = 0$ on $[y_{(K)}, \infty)$. Take ε to be the 75^{th} percentile of the observed response data, and set

$$B_{-1}(y) = log\left(\frac{y}{y+\varepsilon}\right) \text{ and } B_0(y) = log(y+\varepsilon), \text{ for } y > 0. \tag{2.23}$$

Kooperberg et al.[34] and Hansen et al.[36] provide details motivating the inclusion of these additional log terms. Then, $B_{-1}(y), B_0(y), B_1(y), \cdots, B_p(y)$ is a basis of the linear space spanned by $\mathbb{G} \cup \{B_{-1}, B_0\}$ such that:

$$\alpha(\cdot|\boldsymbol{\theta}) = \theta_{-1}B_{-1}(y) + \theta_0 B_0(y) + \theta_1 B_1(y) + \cdots + \theta_p B_p(y), \tag{2.24}$$

$$\boldsymbol{\theta} = (\theta_{-1}, \theta_0, \theta_1, \cdots, \theta_p)' \in \mathbb{R}^K, \text{ and} \tag{2.25}$$

$$\boldsymbol{\Theta} = \left\{\boldsymbol{\theta} \in \mathbb{R}^K : \int_0^y exp(\alpha(u|\boldsymbol{\theta}))du < \infty \text{ for } 0 < y < \infty \text{ and } \int_0^\infty exp(\alpha(y|\boldsymbol{\theta}))dt = \infty\right\} \tag{2.26}$$

$$= \left\{\boldsymbol{\theta} = (\theta_{-1}, \theta_0, \theta_1, \cdots, \theta_p)' \in \mathbb{R}^K : \theta_{-1} > -1 \text{ and } \theta_0 \geq -1\right\}. \tag{2.27}$$

With this, $\alpha(\cdot|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, is the log-hazard function fit to the observed response data using cubic spline modeling per the HEFT method. Additionally, given basis coefficients $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the corresponding hazard, density, log-density, and survival functions are, respectively, given by $\lambda(\cdot|\boldsymbol{\theta}) = exp(\alpha(\cdot|\boldsymbol{\theta}))$, and, for $y > 0$,

$$f(y|\boldsymbol{\theta}) = exp(\alpha(t|\boldsymbol{\theta}))\, exp\left(-\int_0^y exp(\alpha(u|\boldsymbol{\theta}))du\right), \tag{2.28}$$

$$\varphi(y|\boldsymbol{\theta}) = \alpha(t|\boldsymbol{\theta}) - \int_0^y exp(\alpha(u|\boldsymbol{\theta}))du, \text{ and} \tag{2.29}$$

$$1 - F(y|\boldsymbol{\theta}) = exp\left(-\int_0^y exp(\alpha(u|\boldsymbol{\theta}))du\right). \tag{2.30}$$

2.1.2.2 Moment Estimation Using HEFT Fit

The aforedescribed modeling of $\alpha(\cdot|\boldsymbol{\theta})$ and $1 - F(y|\boldsymbol{\theta})$ was performed in `R` using the `polspline` package, where $\alpha(\cdot|\boldsymbol{\theta})$ was obtained using the `polspline::heft` function and $F(y|\boldsymbol{\theta})$ was obtained using the `polspline::pheft` function.[38,39] Denoting $\hat{S}^H(y) = 1 - F(y|\boldsymbol{\theta})$, once more, numerical integration by the trapezoidal rule was used to obtain the distribution-independent mean, $\widehat{E(Y)}^H$, and SD, $\widehat{SD(Y)}^H$. That is, $\widehat{E(Y)}^H$, $\hat{S}_2^H(y)$, and $\widehat{E(Y^2)}^H$ were obtained using (2.11), (2.12), and (2.14), respectively, but with a single change. Where the PLKM method inherently produced $K-1$ trapezoids for numerical integration across the $j = 1, 2, \cdots, K$ ordered subject responses, the HEFT method produces a smooth fit of the SF. Therefore, the minimum and maximum observed response values, $y_{(1)}$ and $y_{(K)}$, respectively, were obtained, and $K = 1,000$ equally-spaced intervals were used to produce $K - 1 = 999$ trapezoids of equal width between $y_{(1)}$ and $y_{(K)}$. Letting the left-hand side of a given trapezoid be j and the right-hand side be $j + 1$, the top of the trapezoid was given by the line segment connecting $h_{(j)} = \hat{S}^H(y_{(j)})$ and $h_{(j+1)} = \hat{S}^H(y_{(j+1)})$ (or $h_{(j)} = \hat{S}_2^H(y_{(j)})$ and $h_{(j+1)} = \hat{S}_2^H(y_{(j+1)})$ for second moment estimation). With this, (2.11), (2.12), and (2.14) was applied, and $\widehat{E(Y_{gi})}^H$ and $\widehat{SD(Y_{gi})}^H$ were supplanted into (1.18) to produce the HEFT-derived robust RAR ratio estimator $\hat{\rho}_{T;i+1}^H$.

2.2 Methods Evaluation and Simulation Process

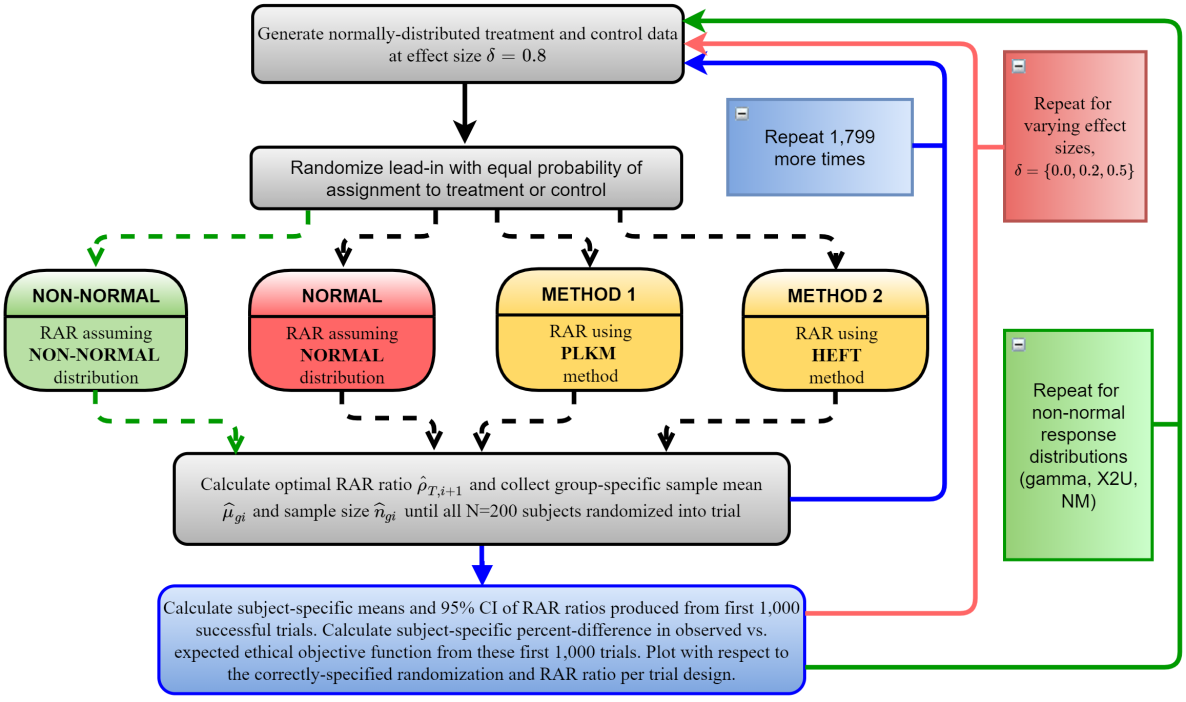
A simulation study was performed to determine how well the PLKM and HEFT methods accounted for the distributional misspecification of the RAR ratio. The goal of this simulation study was to assess the bias, precision, and ethicality of the RAR ratios obtained using each of the empirical methods of moment estimation. Bias and precision were evaluated to gauge how well the distribution-independent empirical estimation methods response-adaptively randomized subjects into a trial when observed responses followed a parametric continuous distribution. The means

and 95% confidence intervals (CI) of the RAR ratios comprised of moments estimated using either the PLKM or HEFT empirical estimation methods averaged over 1,000 simulated trials, denoted $\overline{\hat{\rho}}_{\mathrm{P}}$ and $\overline{\hat{\rho}}_{\mathrm{H}}$, respectively, were plotted with respect to both the means and 95% CIs of the RAR ratios estimated using the MLEs of the correctly-specified, true response distribution, denoted $\overline{\hat{\rho}}_{\mathrm{T}}$, and the RAR ratio intended by the trial design, denoted ρ_d. Means of the 1,000 simulated trials were plotted to measure bias, where smaller distances from the empirically-estimated $\overline{\hat{\rho}}_{\mathrm{P}}$ or $\overline{\hat{\rho}}_{\mathrm{H}}$ to $\overline{\hat{\rho}}_{\mathrm{T}}$ or ρ_d represented less bias, and, thereby, greater accuracy of the empirical methods. Ninety-five percent CIs about the mean RAR ratios were provided as a measure of the precision of these estimators.

Plots of the percent-difference between the observed and expected ethical objective function ($\%_{\text{O-E}}$) were used to evaluate whether or not the empirical estimators of the RAR ratio randomized participants into the trial in a manner consistent with the ethical objective of RAR. If RAR adhered to the intended ethical objective, the observed ethical objective function should have matched the ethical objective function intended by the trial design. Therefore, estimators of the RAR ratio with $\%_{\text{O-E}}$ closer to zero were considered better behaved and, thereby, better performing in this context. Details for the calculation of the $\%_{\text{O-E}}$ metric are provided later in this section. Finally, type I error, i.e., detecting a treatment difference when one does not exist, was evaluated using an independent samples t-test based on the observed responses from the first 100 participants enrolled into the simulated trials where no treatment effect was generated (i.e., $\delta = 0.0$, described subsequently).

Figure 2.1 provides a visual representation of the complete simulation process. In these simulations, group-specific participant response data were generated to follow four parametric response distributions: two exponential family distributions, the symmetrical normal distribution and the right-skewed gamma distribution, and two non-exponential family mixture distributions, the heavily right-skewed χ^2-Uniform (X2U) mixture distribution with a long, thick tail and a bimodal

Figure 2.1: Simulation process for evaluation of empirical estimation methods using parametric continuous response data.



mixture of two normal random variables, i.e., the Normal-Mixture (NM) distribution. Algorithms describing trial generation and participant randomization as well as the simulation of X2U and NM response data are provided in Appendices A.8-A.10, respectively. These distributions were selected to show the behavior of RAR in increasingly complex settings. Group-specific means and SDs were obtained by estimating the MLEs of the parameters that characterize each response distribution. Derivation of the normal, gamma, X2U, and NM distribution MLEs, along with their means and SDs, are provided in Appendices A.11-A.14, respectively. These MLE-derived means and SDs were supplanted into the RAR ratio given by (1.18) in Section 1.4.1 to produce distribution-specific estimators of the RAR ratio.

Four treatment differences, or effect sizes, denoted δ, were considered in this simulation process.

These treatment differences were calculated using Cohens D assuming a pooled SD:

$$\delta = \frac{\mu_{C0} - \mu_{T0}}{\sqrt{(1-\rho_0)\,\sigma_{C0}^2 \;+\; \rho_0\,\sigma_{T0}^2}}, \tag{2.31}$$

where ρ_0 was used as a proxy for the group-specific sample sizes.[40,41] With this, group-specific response data reflecting no ($\delta = 0.0$), small ($\delta = 0.2$), moderate ($\delta = 0.5$), and large ($\delta = 0.8$) treatment differences were generated.[40,41] The aforementioned evaluation of type I error was incorporated into the scenarios where no treatment difference existed ($\delta = 0.0$).

Table 2.1: Distribution-specific parameter values for simulation of treatment and control group response data by effect size.

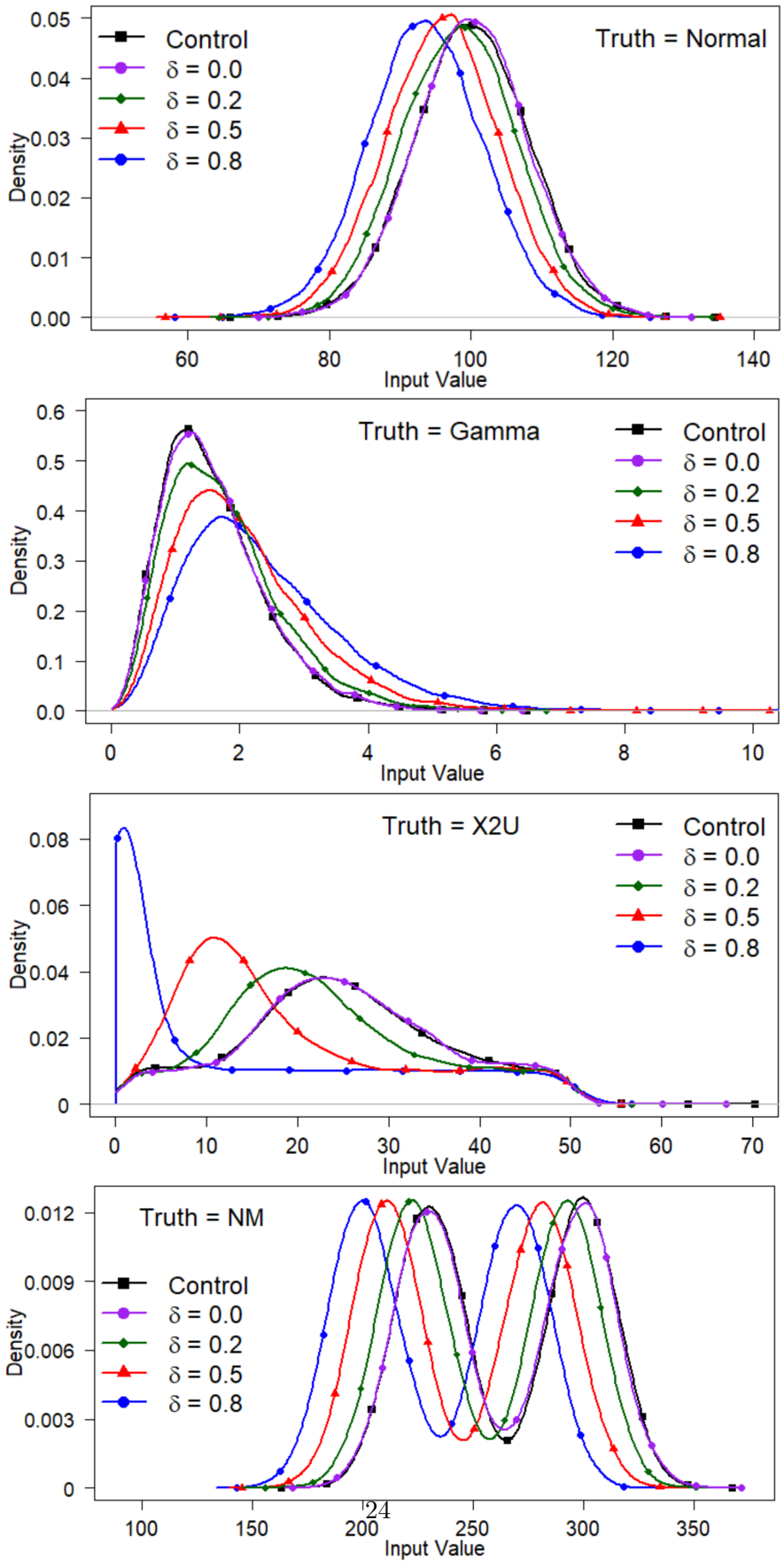
	Parameter values fixed across effect sizes		
Truth	Type	Value	τ for $\rho_0 = 0.75$
Normal	Scale/SD, σ_0	8.00	33.22
Gamma	Shape, α_0	4.00	7.89
X2U	Mixture, π_0	0.50	2.38
	Uniform lower bound, a_0	0.50	
	Uniform upper bound, b_0	50.50	
NM	Mixture, π_0	0.50	17.99
	1st & 2nd Nrv SD, $\sigma_{1,g0}$ & $\sigma_{2,g0}$	15.00	
	Overall SD, σ_0	38.08	

	Parameter values varied by effect size				
		Control/None	Small	Moderate	Large
Truth	Type	$\delta = 0.0$	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
Normal	Location/Mean, μ_{g0}	100.00	98.40	96.00	93.60
	Intended RAR ratio, ρ_d	0.50	0.57	0.66	0.75
Gamma	Scale, β_{g0}	2.50	2.27	1.97	1.72
	Mean, μ_{g0}	10.00	6.89	7.88	9.07
	SD, σ_{g0}	5.00	4.53	3.94	3.44
	Intended RAR ratio, ρ_d	0.50	0.57	0.67	0.75
X2U	X2 DF, k_{g0}	25.00	20.43	12.73	1.12
	Overall mean, μ_{g0}	25.25	22.96	19.11	13.31
	Overall SD, σ_{g0}	11.37	11.45	12.56	15.93
	Intended RAR ratio, ρ_d	0.50	0.53	0.61	0.75
NM	1st Nrv mean, $\mu_{1,g0}$	230.00	222.38	210.96	199.54
	2nd Nrv mean, $\mu_{2,g0}$	300.00	292.38	280.96	269.54
	Overall mean, μ_{g0}	265.00	257.38	245.96	234.54
	Intended RAR ratio, ρ_d	0.50	0.57	0.66	0.75

Nrv: Normal random variable

Distribution-specific parameter values are provided in Table 2.1, and visual displays of each

Figure 2.2: Distribution-specific densities based on intended trial parameters for simulation of treatment and control group response data by effect sizes.



response distribution characterized by these parameter values are provided in Figure 2.2. Without loss of generality, the treatment group was considered superior to the control group when a treatment difference existed ($\delta > 0.0$); as such, the treatment group had a mean response value that was lower than the mean response value of the control group. Also provided in Table 2.1 is the value of τ required to achieve 3:1 randomization ($\rho_d = \rho_0 = 0.75$) when a large treatment difference was observed. This tuning parameter also dictated ρ_d when a large treatment difference was not observed. When no treatment difference existed ($\delta = 0.0$), intended randomization was 0.5 ($\rho_d = 0.5$). Intended randomization ratios varied for small ($\delta = 0.2$) and moderate ($\delta = 0.5$) effect sizes, and are provided in Table 2.1.

Results using preliminary data with 500 trial participants showed no differences in randomization for greater than 200 participants; therefore, the simulation studies were limited to 200 participants in each simulated trial. To produce the previously-described plots of the means and 95% CIs, the RAR ratios across 1,000 iterations of each of the 60 scenarios considered (3 non-normal response distributions $\times$ 4 effect sizes $\times$ 4 methods of randomization + 1 normal response distribution $\times$ 4 effect sizes $\times$ 3 methods of randomization) were collected for each of the 200 participants enrolled. The means and variances of these 60 sets of 1,000 RAR ratios were calculated and 95% CIs were constructed.

The ethical objective of RAR in these trials was to minimize the total expected response value observed in the trial. Therefore, to construct the plot of the $\%_{\text{O-E}}$, group-specific sample means and sample sizes were calculated after the enrollment of each subject into the trial, and their means were calculated across all 1,000 trials. Then, using (1.1) from Section 1.3, the observed ethical objective function was defined as:

$$\overline{O}_i \;=\; \overline{\hat{n}}_{Ti}\,\overline{\hat{\mu}}_{Ti} \;+\; \overline{\hat{n}}_{Ci}\,\overline{\hat{\mu}}_{Ci}, \tag{2.32}$$

where $\overline{\hat{n}}_{gi}$ represents the mean observed number of subjects enrolled in group g from the first to the i^{th} participant, and $\overline{\hat{\mu}}_{gi}$ represents the mean observed sample mean of group g from the first to the i^{th} participant enrolled into the trial over 1,000 simulated trials. Similarly, the expected ethical objective function was defined as:

$$\overline{E}_i \;=\; \overline{E(n_{Ti})}\,\mu_{T0} \;+\; \overline{E(n_{Ci})}\,\mu_{C0}, \tag{2.33}$$

where $\overline{E(n_{gi})}$ represents the mean expected number of subjects enrolled in group g from the first the first to the i^{th} participant enrolled into the trial over 1,000 simulated trials, where:

$$E(n_{gi}) = \begin{cases} 0.5\;i, & \text{for } i = 1, 2, \cdots, L, \\ 0.5\;L \;+\; \rho_d\;(i - L), & \text{for } i > L \text{ and } g = T, \\ 0.5\;L \;+\; (1 - \rho_d)\;(i - L), & \text{for } i > L \text{ and } g = C, \end{cases} \tag{2.34}$$

where L represents the lead-in size of the trial. Lead-ins are described in depth in the following section. When no treatment difference exists, $E(n_{gi}) = 0.5\;i$ for $i = 1, 2, \cdots, 200$. With this, the percent-difference between the observed and expected ethical objective function is defined as:

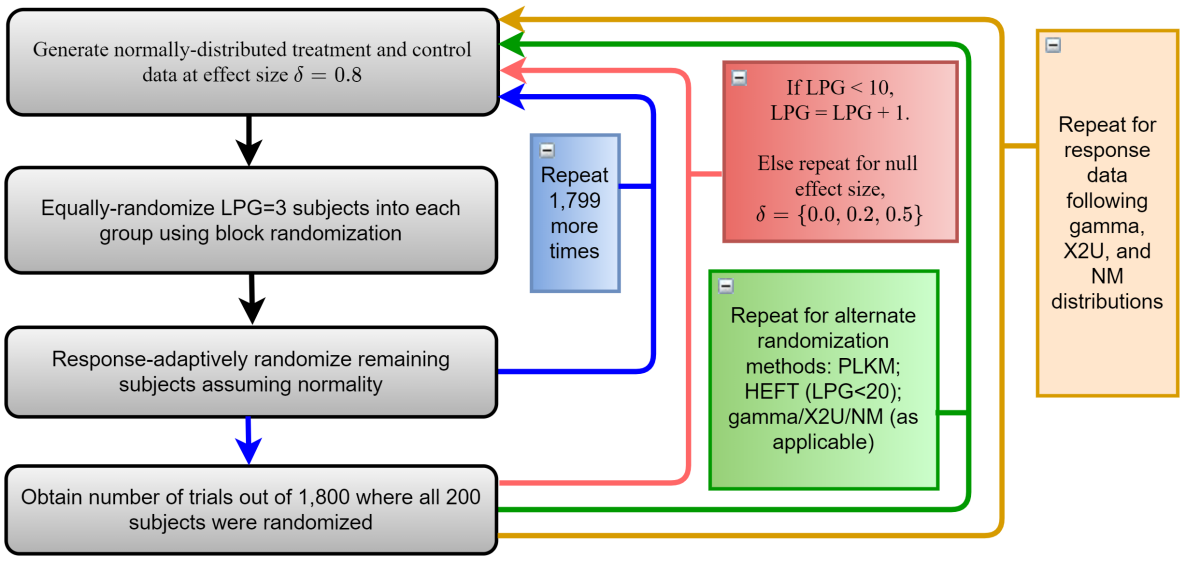
$$\%_{\text{O-E}} \;=\; 100 \;\times\; (\overline{O}_i - \overline{E}_i) \;/\; \overline{E}_i. \tag{2.35}$$

2.2.1 Lead-in Analysis

Figure 2.1 shows that response-adaptive clinical trials require a lead-in period before RAR can begin. During this lead-in period, a small number of trial participants are equally-randomized into either the treatment or control group immediately following trial onset. The purpose of this lead-in period is to accrue enough group-specific response data that the means and SDs that comprise the RAR ratio can be estimated. Since the focus of the present chapter is the estimation of the means and SDs that comprise the RAR ratio, particularly in small samples, initiating RAR as early as possible by minimizing the lead-in period was important to the evaluation of the methods presented

in this chapter. While Haines and Sadiq[43] evaluated start-up duration, a type of lead-in, no other suggestions regarding lead-in size calibration exist in the literature. Therefore, an assessment of possible lead-in lengths was performed in which participants were equally-randomized into the trial using block randomization, selected because it produces an equal number of subjects in each group. Other methods of ER may result in unequal group-specific lead-in sizes, equally-randomizing a greater number of subjects into the trial than desired.

Figure 2.3: Process for selecting lead-in size for simulations used to evaluate the empirical estimation methods using parametric continuous response data.



A visual depiction of the lead-in analysis is provided in Figure 2.3. To conduct the lead-in analysis, 1,800 trials for each of the simulation scenarios described previously (§2.2) were performed for increasing lead-in sizes, beginning with the block randomization of three subjects into each group, i.e., lead-per-group (LPG) of three (LPG=3) for six subjects total (L=6). Due to the complexity of the HEFT method of estimation versus PLKM and MLE estimation, LPG sizes were incremented by one until LPG=20 (L=40) for RAR using the HEFT method of estimation, and until LPG=10 (L=20) for RAR ratios estimated using either the PLKM method of estimation and distribution-specific MLEs. If the means or SDs - and, subsequently, the RAR ratio - could not be

estimated after a participant was enrolled, the simulated trial was stopped. The trial was deemed successful when RAR ratios were estimated for all 200 subjects enrolled into the trial. The percent of successful trials out of the 1,800 performed, denoted %success, was reported by LPG for each simulation scenario.

Reiterating the goal of beginning RAR using the smallest possible lead-in size, a results-driven sensitivity analysis based on the observed %success, denoted $\hat{\%}$success, was performed for scenarios where the $\hat{\%}$success was less than 100 for LPG=3. In these scenarios, the LPGs where $\hat{\%}$success was at least 65, 70, 75, and 80 were of-interest. The minimum %success value of 65 was selected because it represents a majority of trials and each of the HEFT randomization scenarios for LPG=3 could be considered in the sensitivity analysis. From 65, the desired %success value was incremented by five until %success=80, at which point LPGs were too large to be practical.

Table 2.2: Lead-per-group at which at least 65, 70, 75, or 80% of trials were successful when estimating the RAR ratio using HEFT moment estimation and when estimating the RAR ratio using mixture distribution MLEs for mixture-distributed response data by effect size.

		Desired	LPG ($\hat{\%}$ success)			
RAR	Truth	%success	$\delta = 0.0$	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
HEFT	Normal	65	3 (68.9)	3 (71.5)	3 (73.2)	3 (75.8)
		70	14 (72.3)	3 (71.5)	3 (73.2)	3 (75.8)
		75	15 (80.6)	15 (81.5)	14 (77.3)	3 (75.8)
		80	15 (80.6)	15 (81.5)	15 (80.4)	11 (80.3)
	Gamma	65, 70	3 (71.8)	3 (76.1)	3 (77.0)	3 (77.6)
		75	15 (77.1)	3 (76.1)	3 (77.0)	3 (77.6)
		80	16 (89.7)	16 (91.6)	14 (83.0)	12 (80.2)
	X2U	65, 70, 75, 80	3 (83.1)	3 (82.7)	3 (84.3)	3 (89.6)
	NM	65, 70, 75, 80	3 (82.3)	3 (81.6)	3 (82.9)	3 (84.7)
X2U	X2U	65, 70	5 (72.7)	5 (72.2)	5 (74.9)	5 (78.5)
		75	6 (81.7)	6 (82.5)	6 (84.3)	5 (78.5)
		80	6 (81.7)	6 (82.5)	6 (84.3)	6 (85.4)
NM	NM	65, 70	7 (76.7)	7 (77.9)	7 (74.6)	7 (77.1)
		75	7 (76.7)	7 (77.9)	8 (85.9)	7 (77.1)
		80	8 (86.7)	8 (86.3)	8 (85.9)	8 (86.2)

Regardless of the simulated response distribution, all 1,800 trials were successful when subjects were randomized into the trial by estimating the RAR ratio using the PLKM method and normal

MLEs for LPG=3. Likewise, all 1,800 trials were successful for LPG=3 when using gamma MLEs to estimate the RAR ratio for response data simulated to follow the gamma distribution. Results varied for HEFT randomization and mixture randomization of mixture-distributed response data, and are provided in Table A.15 in the Appendix. These scenarios were considered in the sensitivity analysis, and results are provided in Table 2.2.

Specific to HEFT randomization of response data following the normal and gamma distributions, a large increase in LPG was required to observe at least 75 or 80 versus 65 or 70 %success. This represents the difference between equally-randomizing six subjects for the 65-70 %success situation compared to 20-30 for the 75-80 %success situation prior to initiating RAR. Where the present work seeks to evaluate the behavior of various methods of estimating the RAR ratio with an emphasis on small-sample estimation, large lead-in sizes such as these may severely hinder the practicability of the HEFT method and damage the interpretation of its results, particularly when compared to randomization methods requiring much smaller lead-in sizes. Though randomization using larger lead-in sizes may have been more well-behaved in some scenarios (e.g., LPG=12 versus LPG=3 for HEFT randomization of gamma-distributed responses), obtaining empirically-estimated RAR ratios in small samples was more important to the present work than obtaining albeit better-behaved estimated RAR ratios that required three-to-four times the number of subjects in the lead-in period to be estimated. Finally, with respect to mixture randomization of mixture-distributed response data, randomization results based on the LPG required to obtain 65, 70, and 75 %success were nearly identical to the randomization results obtained using the LPG required to observe 80 %success.

Whether HEFT randomization of normal or gamma response data or mixture randomization of mixture-distributed response data, this sensitivity analysis confirmed that overall trial results will not change when randomization is based on the LPG necessary for 65 versus 80 %success: at

Figure 2.4: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where normally-distributed group-specific subject responses reflected a large treatment difference and were randomized using the HEFT empirical estimation method, for LPG=3, whereby at least 65, 70, and 75 %success was observed, and for LPG=11, whereby at least 80 %success was observed.

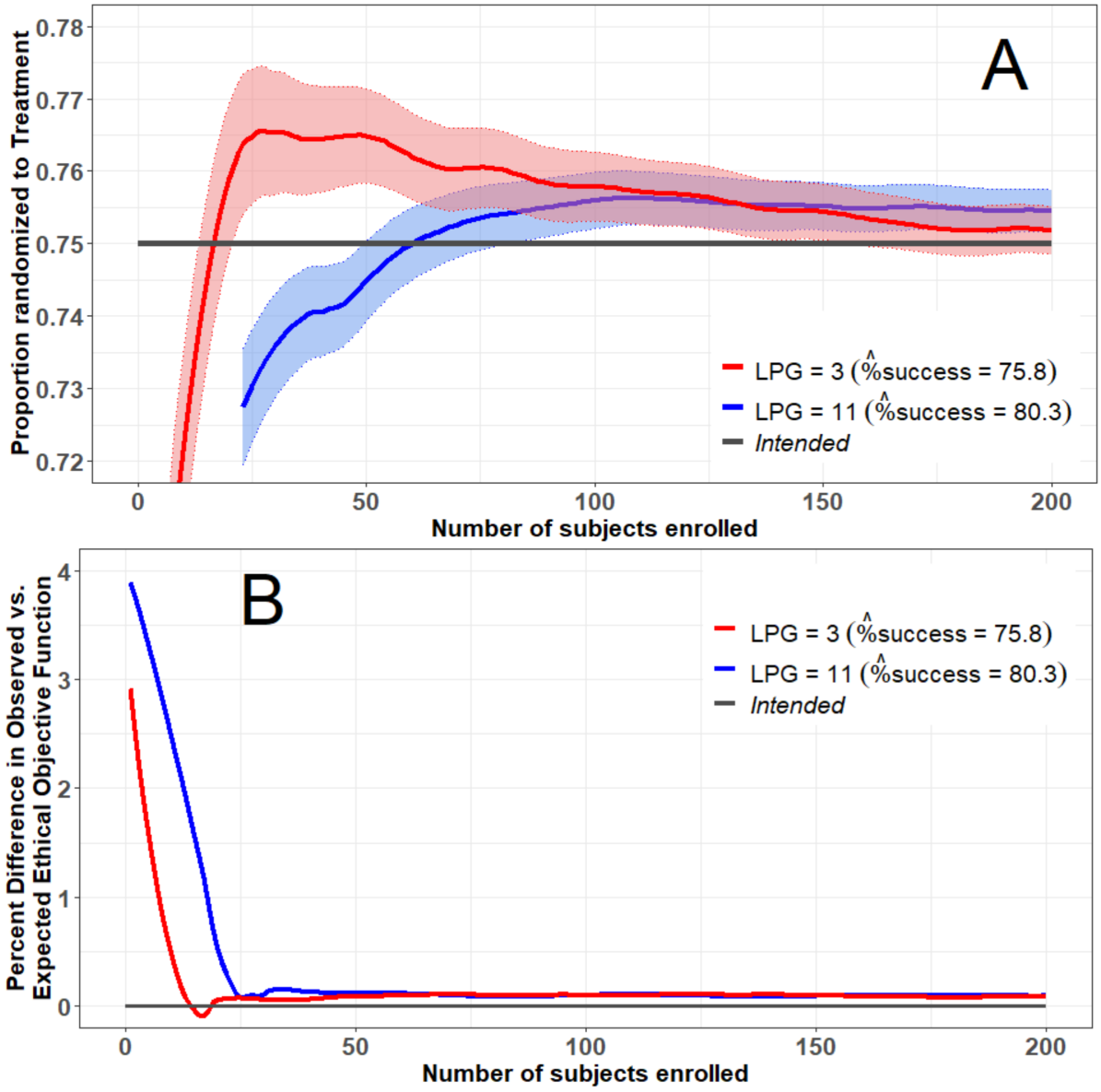


Figure 2.5: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where group-specific subject responses followed gamma distributions that reflected a large treatment difference and were randomized using the HEFT empirical estimation method, for LPG=3, whereby at least 65, 70, and 75 %success was observed, and for LPG=12, whereby at least 80 %success was observed.

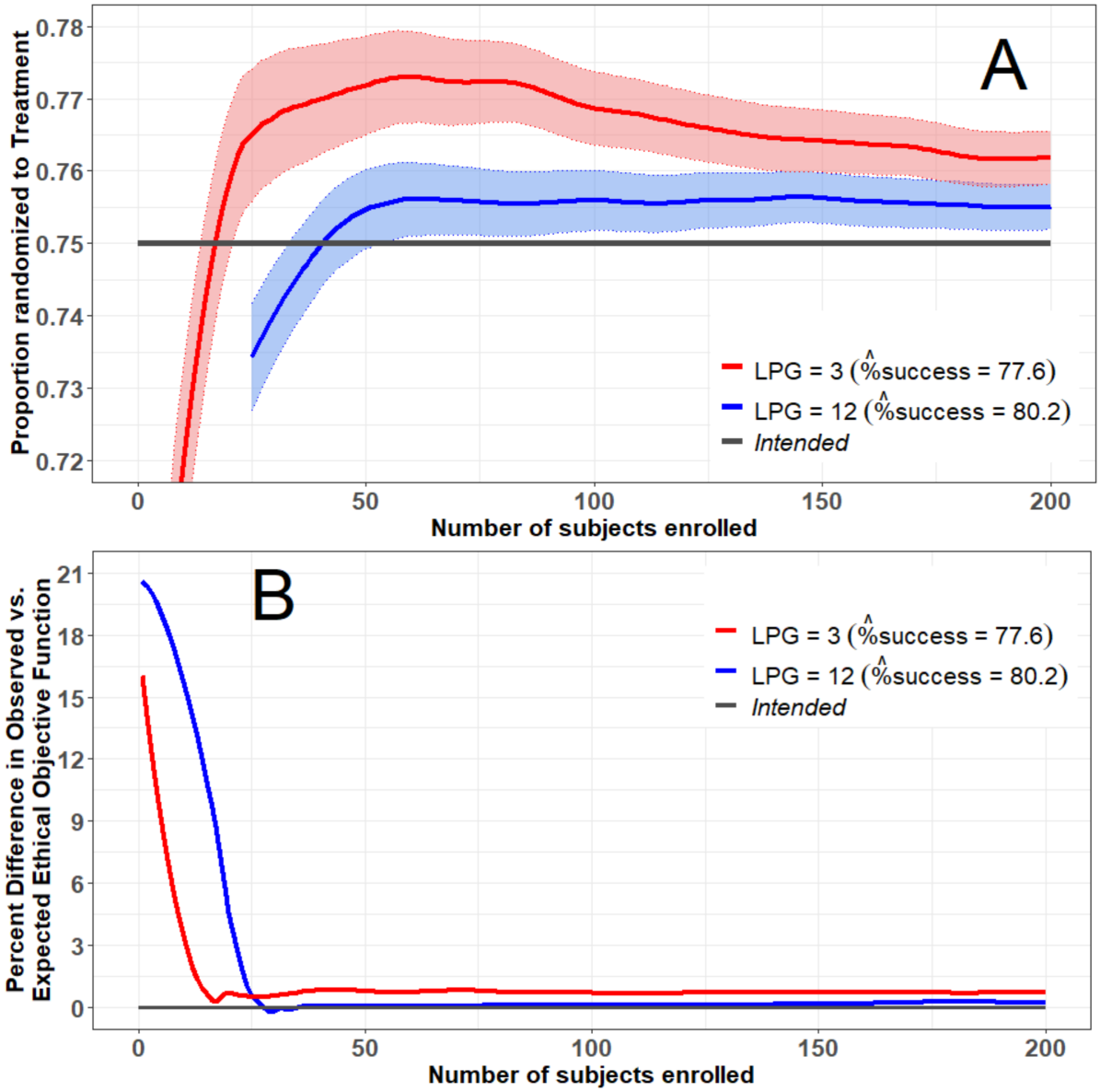


Figure 2.6: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where group-specific subject responses followed X2U distributions that reflected a large treatment difference and were randomized using RAR ratios estimated using X2U MLEs, for LPG=5, whereby at least 65, 70, and 75 %success was observed, and for LPG=6, whereby at least 80 %success was observed.

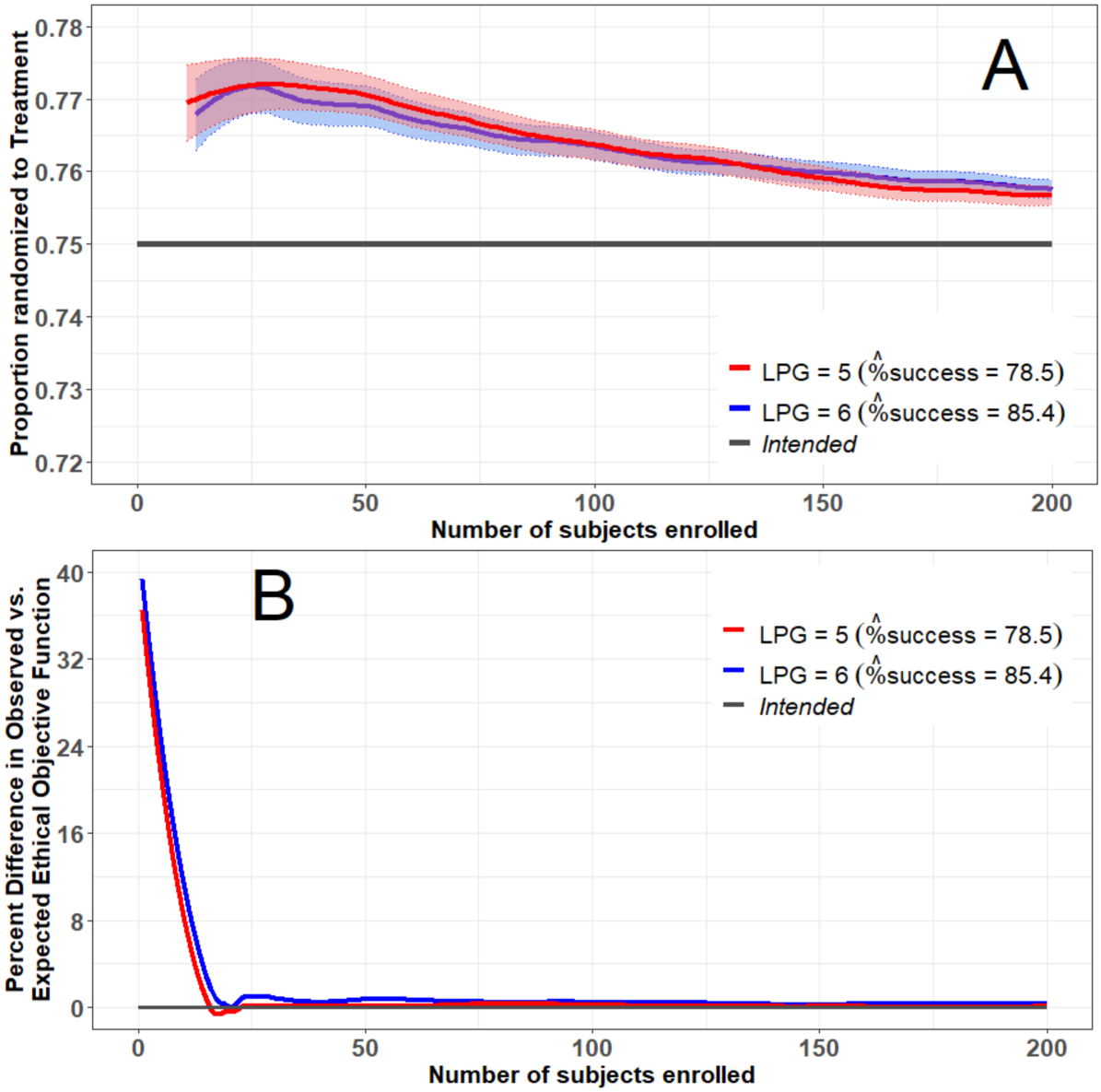
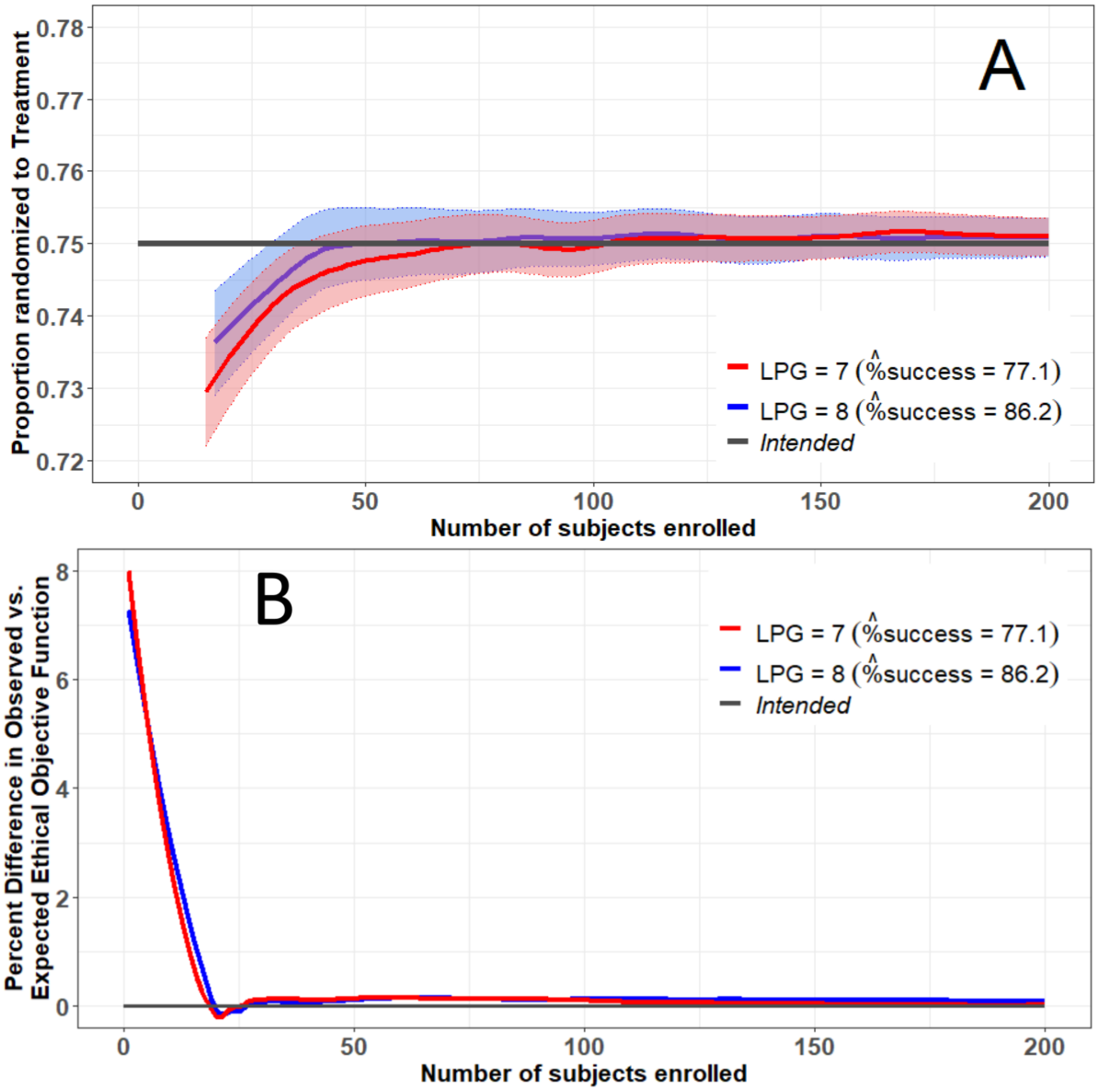


Figure 2.7: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where group-specific subject responses followed NM distributions that reflected a large treatment difference and were randomized using RAR ratios estimated using NM MLEs, for LPG=7, whereby at least 65, 70, and 75 %success was observed, and for LPG=8, whereby at least 80 %success was observed.



most, one or two subjects for every 100 enrolled into the trial may be randomized to the treatment group opposite from the one intended by the trial design. Such a difference is likely insufficient to influence the final results of the clinical trial. And, finally, the purpose of employing RAR is to balance the power of the clinical trial against an ethical objective. When using the LPG required to observe 65 %success, observed randomization minimized the overall mean response value either better than or not dissimilarly to trials using LPG required to observe 80 %success because a larger number of subjects underwent RAR versus ER. For these reasons, lead-in sizes were selected based upon the LPG required to observe at least 65 %success for each scenario. With the exception of mixture randomization of mixture-distributed responses, this was LPG=3. Lead-per-group of five and LPG=7 were used for X2U randomization of X2U response data and NM randomization of NM response data, respectively. With this, corresponding values of L for the calculation of E_i were L=6 save for the former, where E_i was based on L=10 and L=14, respectively.

2.3 Randomization Results

Results for simulated RCTs with continuous outcomes are discussed in what follows. All curves were loess-smoothed using a bandwidth of 0.075. Randomization results for small and moderate effect sizes followed patterns similar to those discussed in the following sections and are, therefore, provided in Appendices A.16-A.23.

2.3.1 Normal Response Data Scenarios

In the baseline scenario ($\delta = 0.0$) where group-specific response data were normally-distributed and reflected no treatment difference (Figure 2.8), randomization behaved as anticipated under RAR: an approximately equal number of participants were randomized to either treatment group (grey line at $\rho_d = 0.5$; plot A) and any deviations from the intended ethical objective function were less than 0.2% (plot B). Furthermore, as displayed in Table 2.3, type I error was well-controlled: no

Figure 2.8: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials when group-specific subject responses were normally-distributed and reflected no treatment difference.

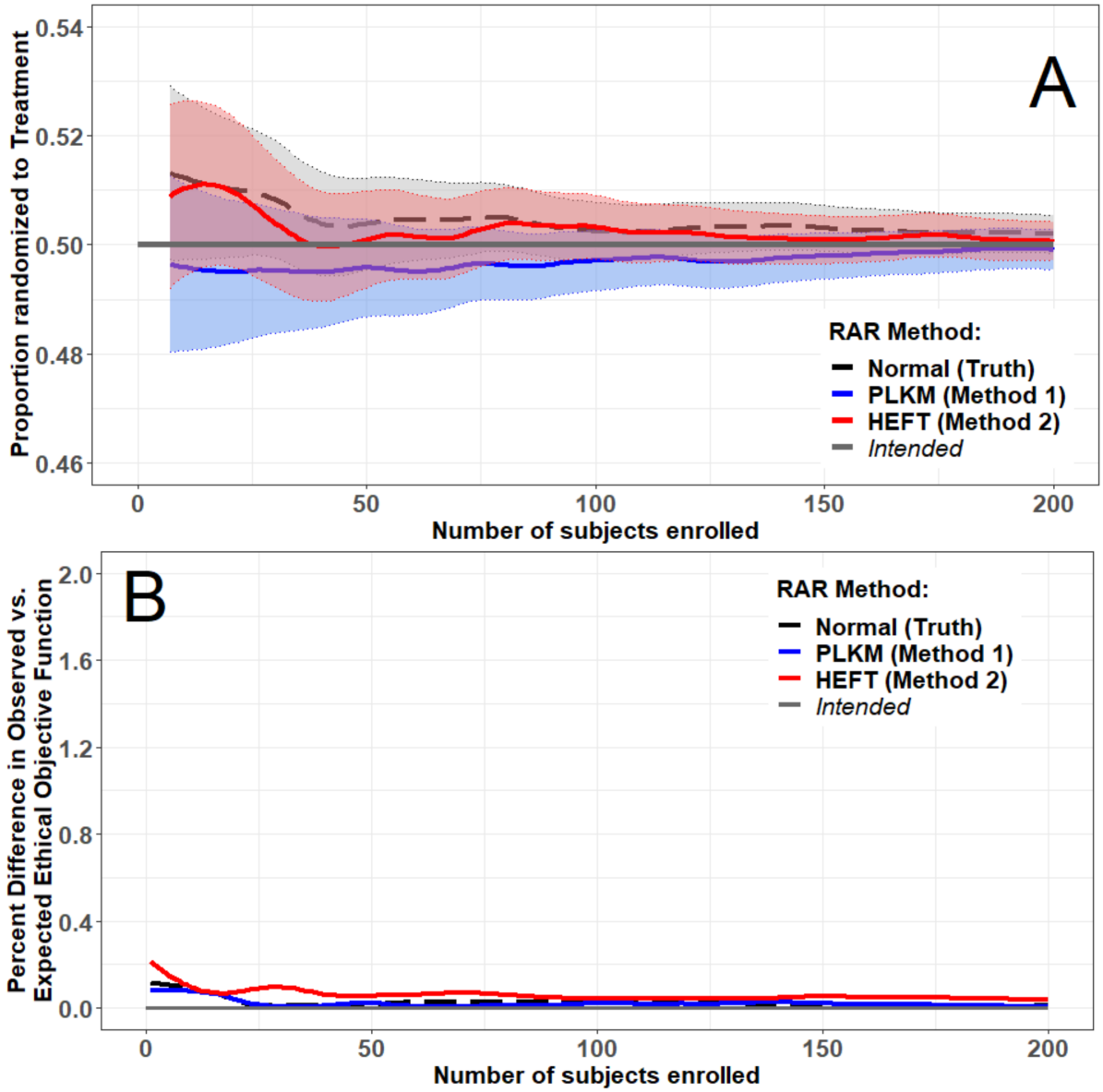


Table 2.3: Assessment of type I error by scenario using t-test based on observed mean response from first 100 trial participants enrolled when no treatment difference existed ($\delta = 0.0$) from 1,000 simulated trials.

Truth	RAR	Treatment Mean	Control Mean	t	p-value
Normal	HEFT	100.09	100.12	-0.498	0.6187
	PLKM	100.14	100.18	-0.746	0.4555
	Normal	100.16	100.17	0.225	0.8219
Gamma	HEFT	10.09	10.12	-0.811	0.4177
	PLKM	9.990	10.01	-0.936	0.3494
	Normal	10.07	10.10	-0.891	0.3733
	Gamma	10.11	10.10	0.447	0.6550
X2U	HEFT	25.33	25.22	1.543	0.1229
	PLKM	25.30	25.18	1.541	0.1235
	Normal	25.30	25.25	0.654	0.5132
	X2U	25.32	25.27	1.111	0.2665
NM	HEFT	266.01	266.06	-0.145	0.8850
	PLKM	265.85	265.96	-0.336	0.7371
	Normal	265.80	265.56	0.824	0.4102
	NM	265.45	265.44	0.075	0.9402

unsubstantiated treatment differences were detected.

When group-specific subject responses were normally-distributed and a large difference existed (Figure 2.9), $\bar{\hat{\rho}}_{\mathrm{H}}$ (red solid curve) was less biased than $\bar{\hat{\rho}}_{\mathrm{P}}$ (blue solid curve) with respect to both $\bar{\hat{\rho}}_{\mathrm{T}}$ (black dashed curve) and ρ_d (grey line at $\rho_d = \rho_0 = 0.75$) (plot A). Specifically, $\bar{\hat{\rho}}_{\mathrm{P}}$ followed the same overall trend as $\bar{\hat{\rho}}_{\mathrm{H}}$, but with an additional bias of approximately 0.02. As well, any deviations from the intended ethical objective of the trial were less than three percent, and existed primarily during the lead-in period (plot B, Figure 2.9).

2.3.2 Non-normal Response Data Scenarios

When response data were not normally-distributed and no treatment difference was observed, randomization using either PLKM or HEFT produced an approximately equal number of subjects in each group (Figure 2.10), and $\%_{\text{O-E}}$ never exceeded two (Figure 2.11). When response data followed the gamma distribution and a large treatment effect was observed, $\bar{\hat{\rho}}_{\mathrm{H}}$, $\bar{\hat{\rho}}_{\mathrm{P}}$, and $\bar{\hat{\rho}}_{\mathrm{T}}$ performed similarly, but did not converge to ρ_d (plot A, Figure 2.12). Furthermore, though $\%_{\text{O-E}}$ was less

Figure 2.9: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials when group-specific subject responses were normally-distributed and a large treatment difference existed.

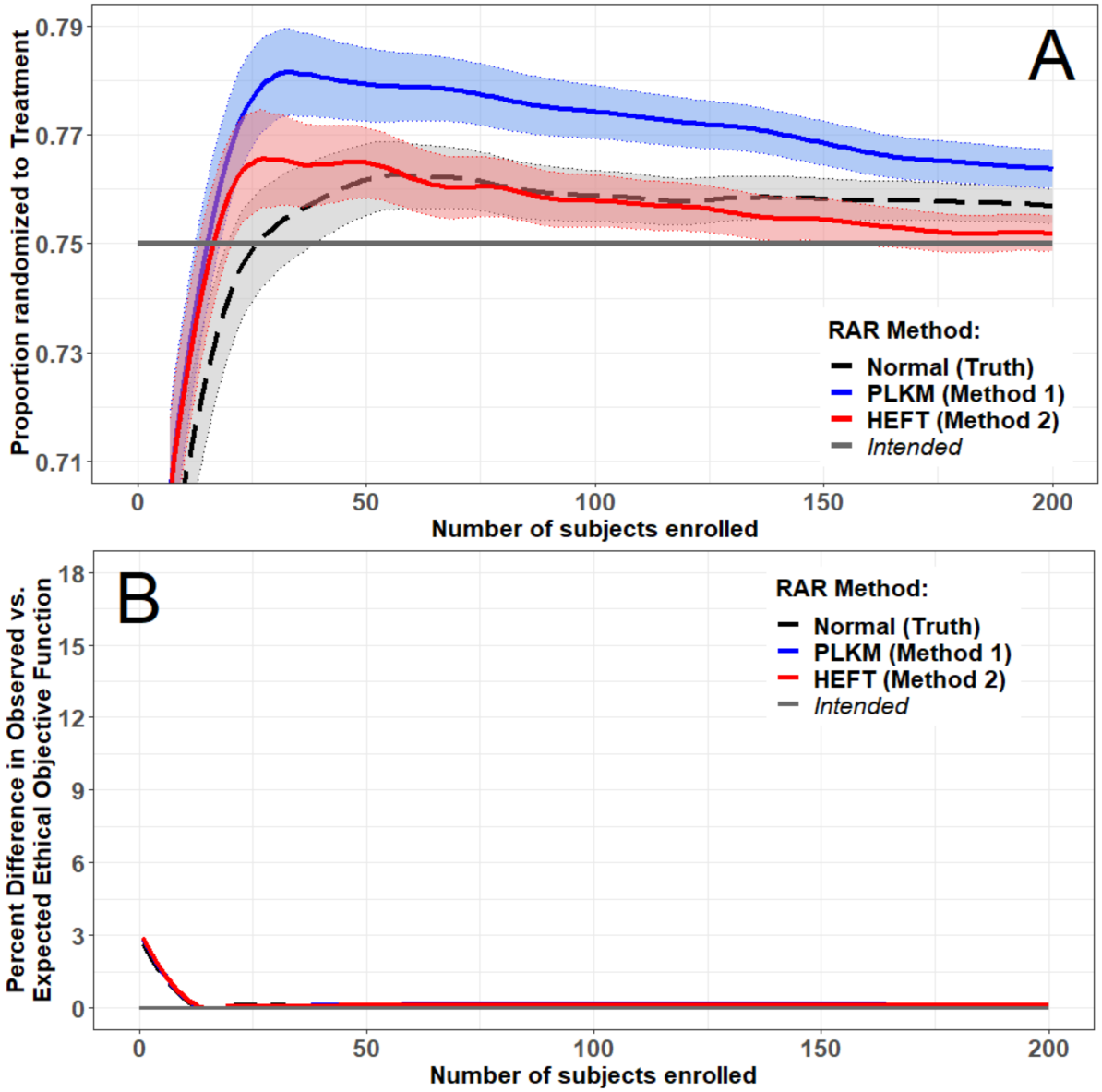


Figure 2.10: Mean of RAR ratios with 95% CI over 1,000 simulated trials when group-specific subject responses were not normally-distributed and reflected no treatment difference.

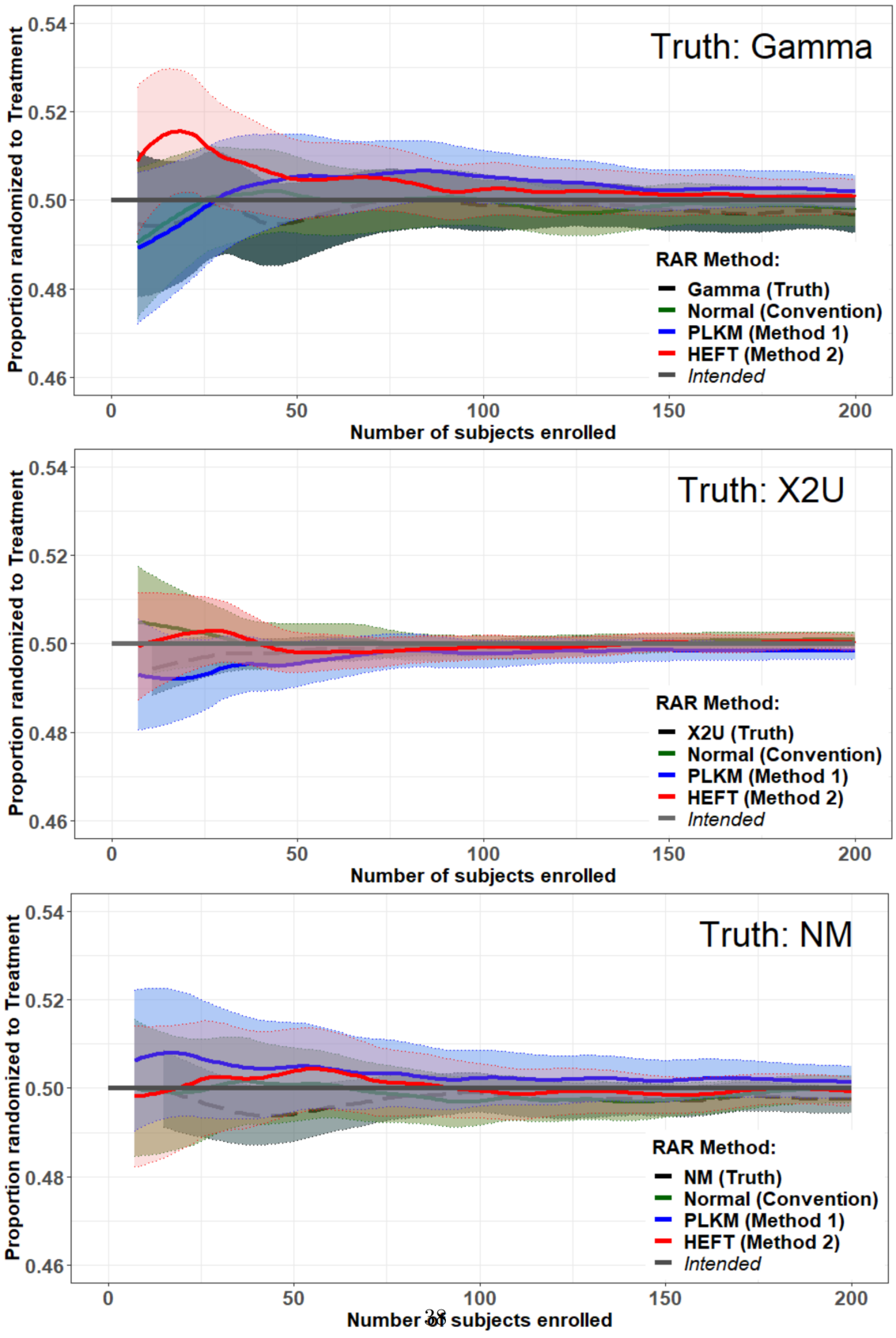


Figure 2.11: Percent-difference between the observed and expected ethical objective function over 1,000 simulated trials when group-specific subject responses were not normally-distributed and reflected no treatment difference.

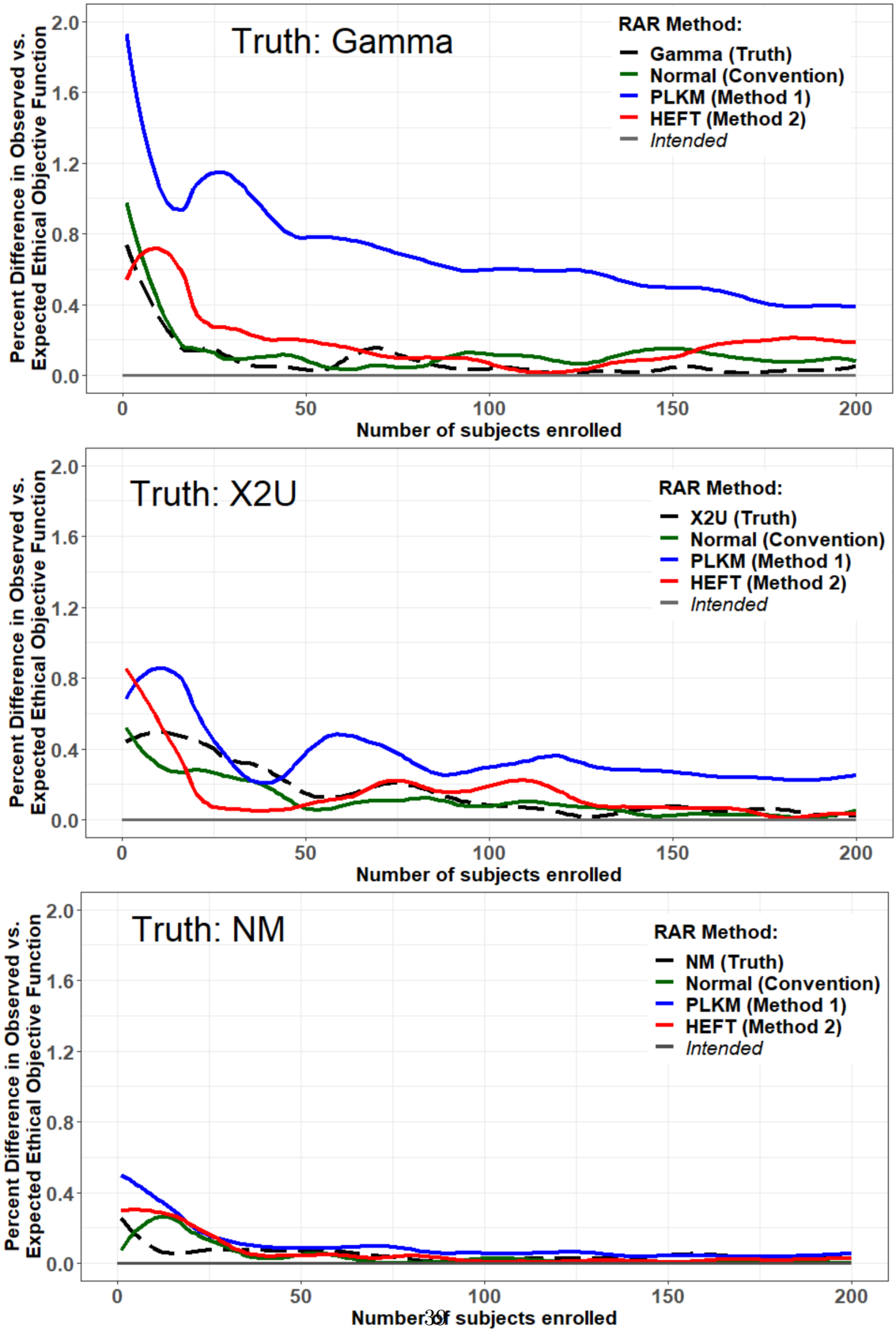
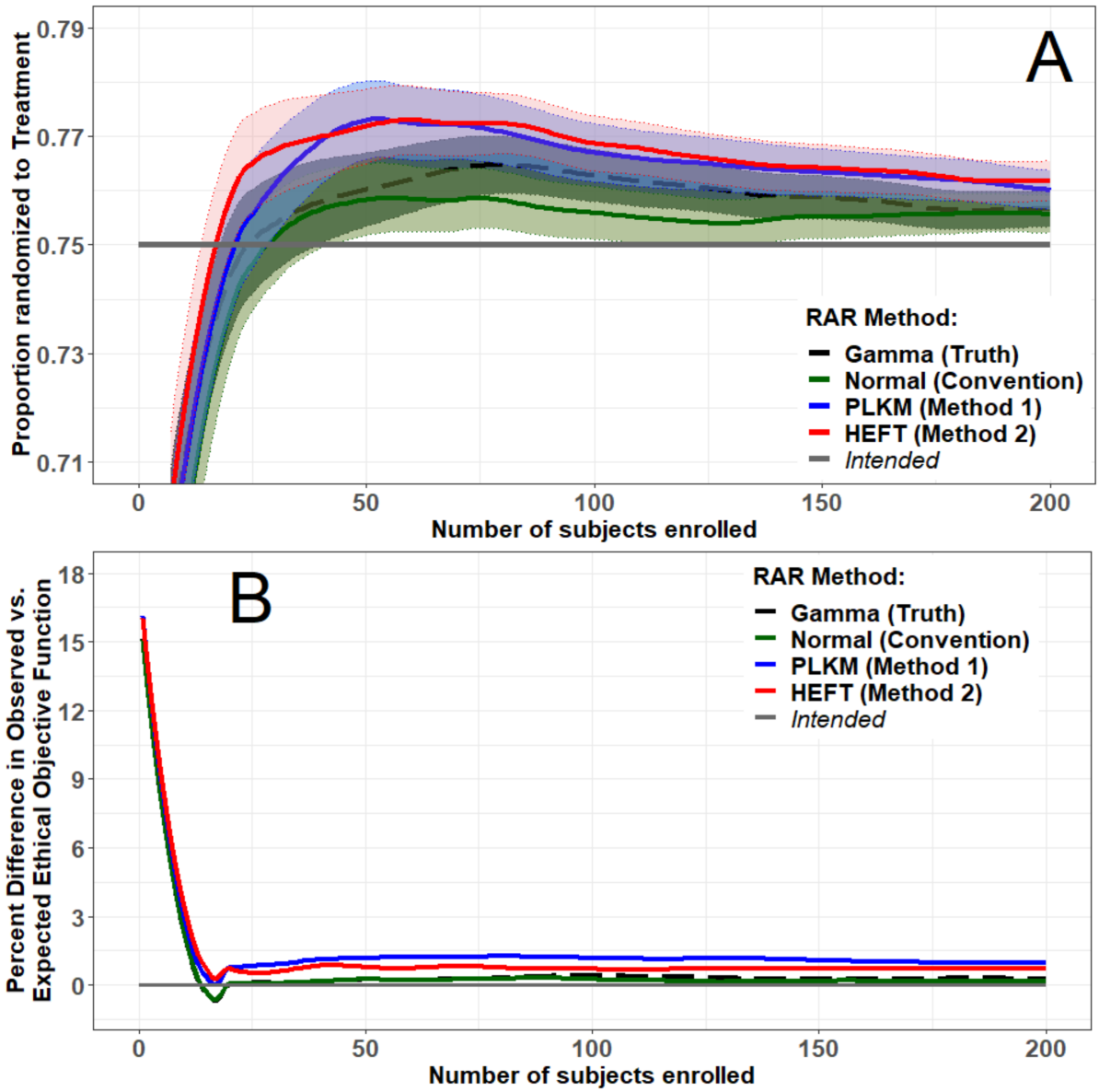


Figure 2.12: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where group-specific subject responses followed the gamma distribution and reflected a large difference.



than two for all methods of randomization following the lead-in period, RAR using PLKM deviated furthest from the intended ethical objective function, followed by randomization under the HEFT method (plot B, Figure 2.12).

A y-axis ranging from $\rho_d \pm 0.05$ was required to display the mean randomization results when group-specific response data followed X2U mixture distributions reflecting a large treatment difference versus the more restricted range of $\rho_d \pm 0.03$ (light-grey dashed lines) sufficient for presenting mean randomization results for all other scenarios (plot A, Figure 2.13). In this scenario, neither empirical estimation method achieved ρ_d and $\bar{\hat{\rho}}_{\mathrm{P}}$ remained dissimilar to $\bar{\hat{\rho}}_{\mathrm{T}}$ for all 200 participants enrolled into the simulated trials while $\bar{\hat{\rho}}_{\mathrm{H}}$ was only marginally similar to $\bar{\hat{\rho}}_{\mathrm{T}}$ for the final 125 participants enrolled.

The y-axis of plot B in Figure 2.13 extends to 40 versus the maximum $\%_{\text{O-E}}$ y-axis value of 18 (light-grey dashed line) sufficient for displaying the $\%_{\text{O-E}}$ results in all other scenarios. This suggests that the means obtained using the X2U MLEs during the lead-in period were considerably different from the intended X2U means in Table 2.1. After the lead-in period, the observed ethical objective function deviated by approximately four percent when the RAR ratio was estimated using the PLKM and HEFT methods and declined to roughly two percent as more participants were enrolled into the trial.

Finally, when response data followed the bimodal mixture of two normal random variables and a large treatment difference was observed, $\bar{\hat{\rho}}_{\mathrm{P}}$ and $\bar{\hat{\rho}}_{\mathrm{H}}$ performed similarly but failed to mirror $\bar{\hat{\rho}}_{\mathrm{T}}$ from the 50^{th} participant onward (plot A, Figure 2.14). As well, neither empirical estimation method converged to ρ_d. However, all methods of randomization adhered to the ethical objective of RAR (plot B, Figure 2.14).

Figure 2.13: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where group-specific subject responses followed the X2U distribution and reflected a large difference.

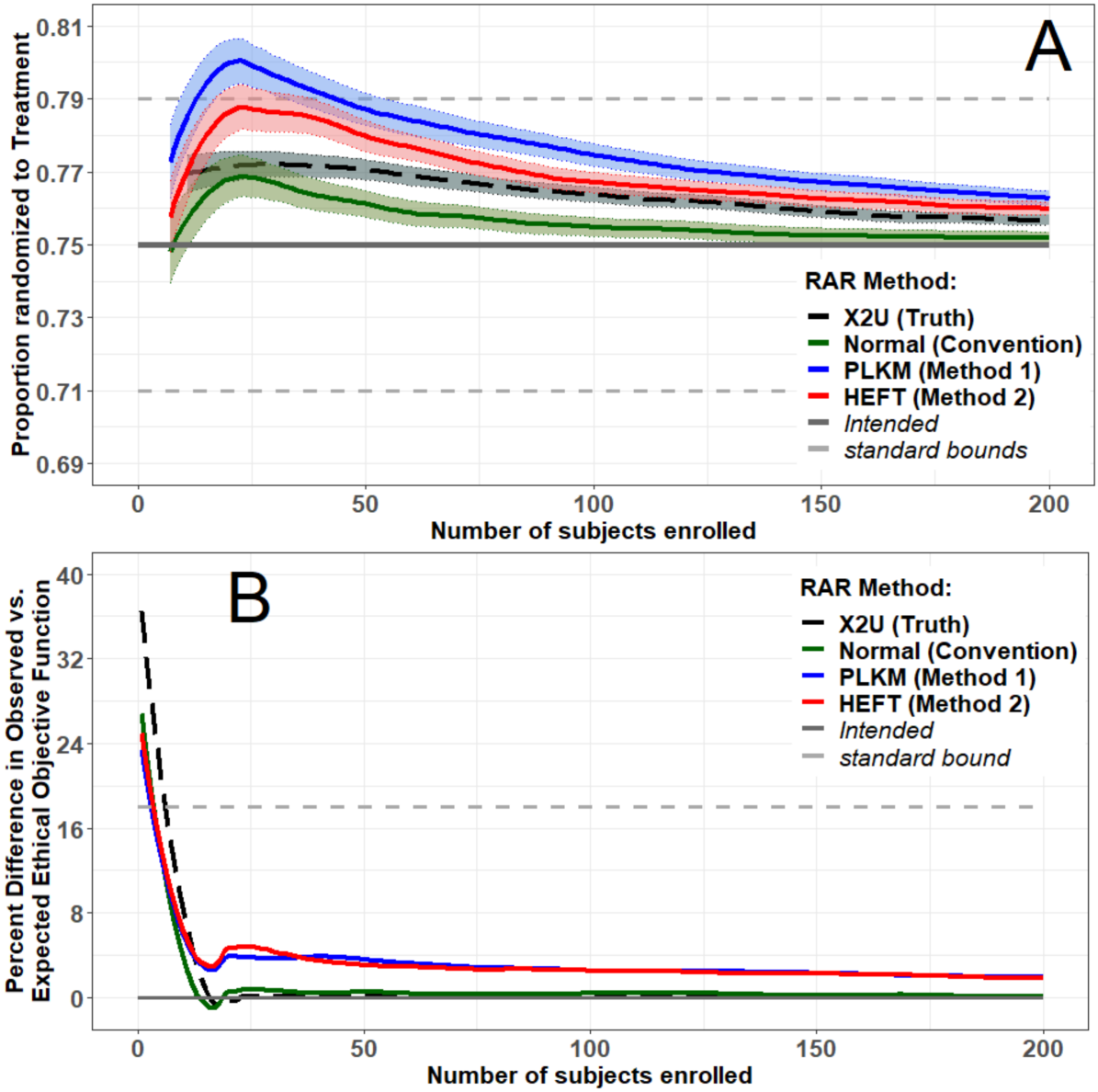
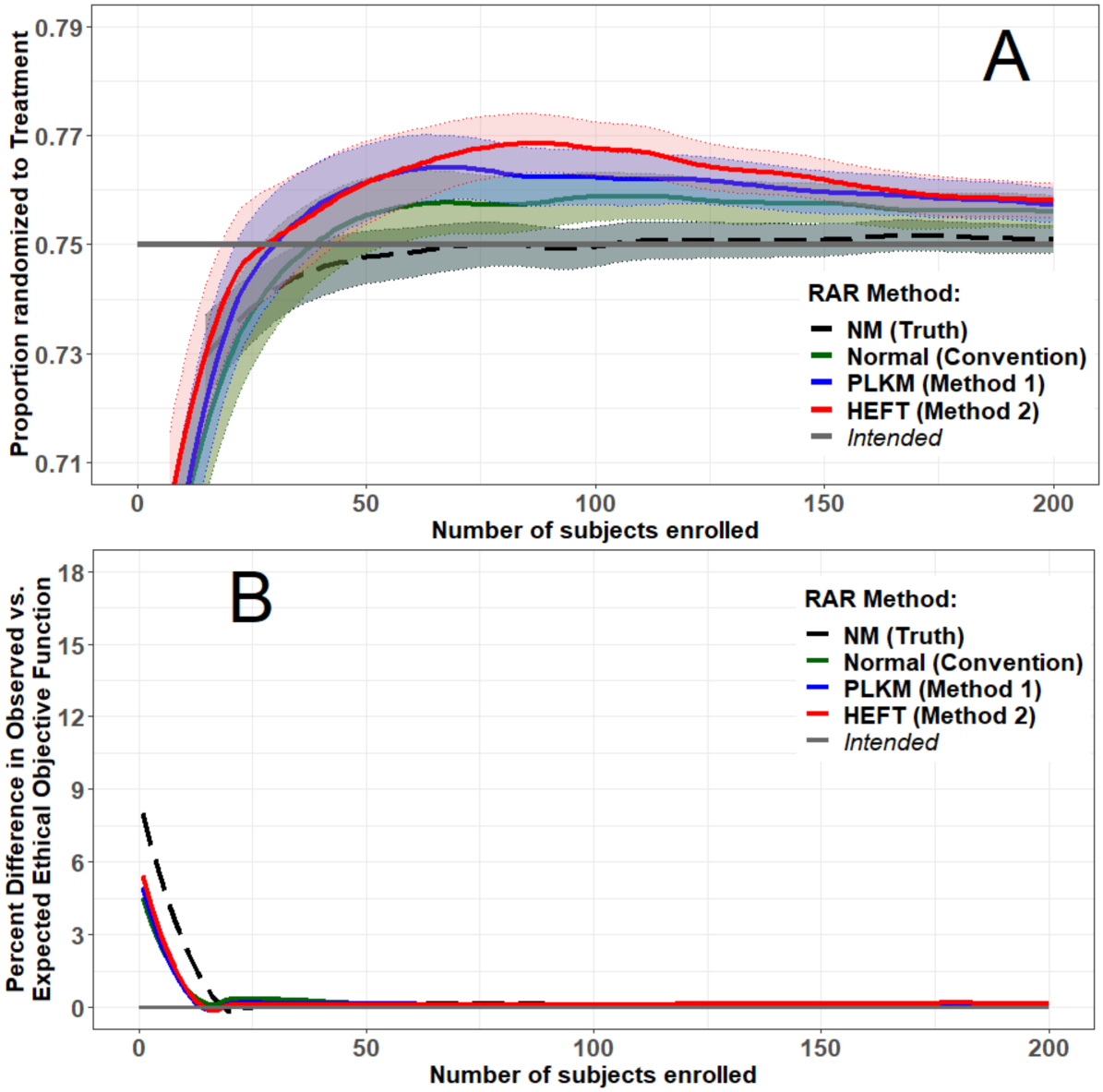


Figure 2.14: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where group-specific subject responses followed the NM distribution and reflected a large difference.



2.4 Conclusion

Evidenced by the lead-in analysis, estimation of the moments for the RAR ratio may be contingent upon the lead-in size, particularly for non-exponential family distributions. Therefore, it is recommended in practice to use the ratio prior to the subject for which the moments cannot be estimated as that subjects RAR ratio, then continue adding blocks of, for example, two or four. Where the focus on the evaluation and implementation of the empirical RAR ratios versus the appropriateness of the lead-in period chosen for the present analysis, the cut-off of a 65% success rate for LPG=3 was selected. For HEFT lead-in, the estimation of moments was more successful for more complex response data (i.e., mixture distributions) versus simple response distributions (i.e., gamma or normal). This may be because HEFT uses a more complex algorithm to select knots, beginning with three knots placements, which may have been harder to place with simpler distributions. For example, if using quadratic splines to model the response data with a single unknown factor, there exists at least seven parameters to estimate, one for each intercept, slope, and quadratic term on either side of the knot, plus a variance term. Adding in the continuity restriction of the mean and its derivative, this becomes a numerical challenge in smaller samples. As anticipated, however, larger effect sizes were more easily detected in smaller samples. [24]

Overall, empirical estimation of the operands of the moments using the HEFT and PLKM methods recreated true randomization and upheld the ethical objective of the trial, yet HEFT estimation produced more well-behaved estimates than the PLKM method. For X2U response data, the normal RAR ratio randomized better than the X2U RAR ratio with respect to achieving and maintaining 3:1 randomization. Because the X2U RAR ratio never converged to 3:1 randomization, it may be true that the BBZ optimal RAR ratio is not appropriate for such heavily skewed response data Even though HEFT most closely matched X2U randomization, $\%_{\text{O-E}}$ showed about 5% deviation for HEFT and PLKM randomization versus approximately zero deviation for

normality, suggesting the normality assumption holds. For NM response data, again, HEFT and PLKM randomization performed well with respect to trial design, but were still outperformed by the normality assumption.

Overall, HEFT and PLKM randomization produced estimates of the RAR ratio that closely adhered to those expected by the trial design. That is, resulting randomization ratios fell to within 0.02 of the intended 3:1 randomization and $\%_{\text{O-E}} \leq 5$ in nearly all scenarios. However, randomization under the normality assumption best-adhered to the trial design regardless of the shape of the response data, even in small samples. Therefore, it is suggested to assume normality in practice.

Of course, though RAR under the normality assumption was comparable in most cases, not every distribution was considered in the present work. Users of these methods may consider situations where there is concern about the use of the sample mean under normality or they may consider alternative methods of approximating the CDF besides HEFT or PLKM, the only two approaches employed in the present work. Alternative distributional assumptions or methods of CDF estimation may out-perform normality; further investigation may be required.

Chapter 3

Robust Randomization of Continuous Response Data Using the Weighted Average

3.1 The Weighed-Average Framework

In Chapter 2, the operands of the robust RAR ratio were estimated empirically. The methodology discussed in the present chapter sought to obtain a robust estimator of the RAR ratio directly by employing a weighted average (WA) framework that automatically adjusted for distributional uncertainty.[42] That is, where many model selection techniques systematically neglect uncertainty, weighted-averaging characteristically incorporates randomness in order to improve model fit by selecting or combining a variety of models in a manner conducive to the observed data.[42] Since averaging over competing models offers results that are more robust than those derived from analyses dependent upon a single selected model, weighted-averaging is an inherently-robust framework.[42] Fragoso[42] and Hoeting[44] contain pertinent examples of the WA framework. Therefore, in the present chapter, participants were randomized into the trial using RAR ratios estimated assuming a set of predetermined continuous distributions, i.e., candidate response distributions (CRD) in lieu of removing the distributional assumption altogether (as in Chapter 2). The resulting set of CRD-specific RAR ratios were then weighted by that CRD's fit to the observed response data, the

combination of which produced a flexible, inherently robust estimator of the RAR ratio.[42,44]

3.1.1 Methodology

For K CRDs, let d_k be the value that represents the $k^{\mathbf{th}}$ $(k = 1, \cdots, K)$ CRD fit to the observed response data based on some measure-of-fit (MOF) where smaller values indicate better fit. Examples of such MOFs include Akaike information criterion (AIC), Bayesian information criterion, Kullback-Leibler distance, the Kolmogorov-Smirnov (KS) test statistic, the Cramer von-Mises criterion, or the Kuiper's test statistic.[45–50] As well, let d_0 be a reference MOF value that serves to scale each of the K MOF values. Letting smaller MOF values indicate a better distributional fit to the data, the minimum d_k value was set as d_0. As such, the scaled MOF values were provided by the difference between each d_k and d_0, denoted d_k^*, i.e.,

$$d_k^* = d_k - d_0. \tag{3.1}$$

Inducing a scale in this manner gives meaning to otherwise-meaningless individual MOF values, allowing direct comparison of the distributional fits produced by each CRD.[45] Since the individual MOF values are positive, real-valued numbers, scaling also alleviates much of the potential computational burden engendered by their direct use.[45] The difference given in (3.1) is the cornerstone of developing the distributional weights for each CRD, denoted w_k:

$$w_k = \frac{exp(-\frac{1}{2}\ d_k^*)}{\sum_{k=1}^{K} exp(-\frac{1}{2}\ d_k^*)}, \tag{3.2}$$

where $w_k \in (0, 1)$ and $\sum_{k=1}^{K} w_k = 1$. These distributional weights may be interpreted as the weight of evidence suggesting that CRD k is the distribution that best-fits the observed response data in the set of K CRDs under consideration, assuming a best-fitting CRD exists in the considered set.[45] Under this weighting scheme, more heavily-weighted CRDs, i.e., CRDs with weights closer to one, were said to better-fit the observed response data than CRDs with w_k closer to zero. Explicitly, the

WA-estimated RAR ratio representing the probability of randomizing subject $i+1$ to the treatment group based on response data observed from participants one to i is given by:

$$\hat{\rho}^{WA}_{T;i+1} = \sum_{k=1}^{K} \hat{w}_{ik}\ \hat{\rho}_k, \tag{3.3}$$

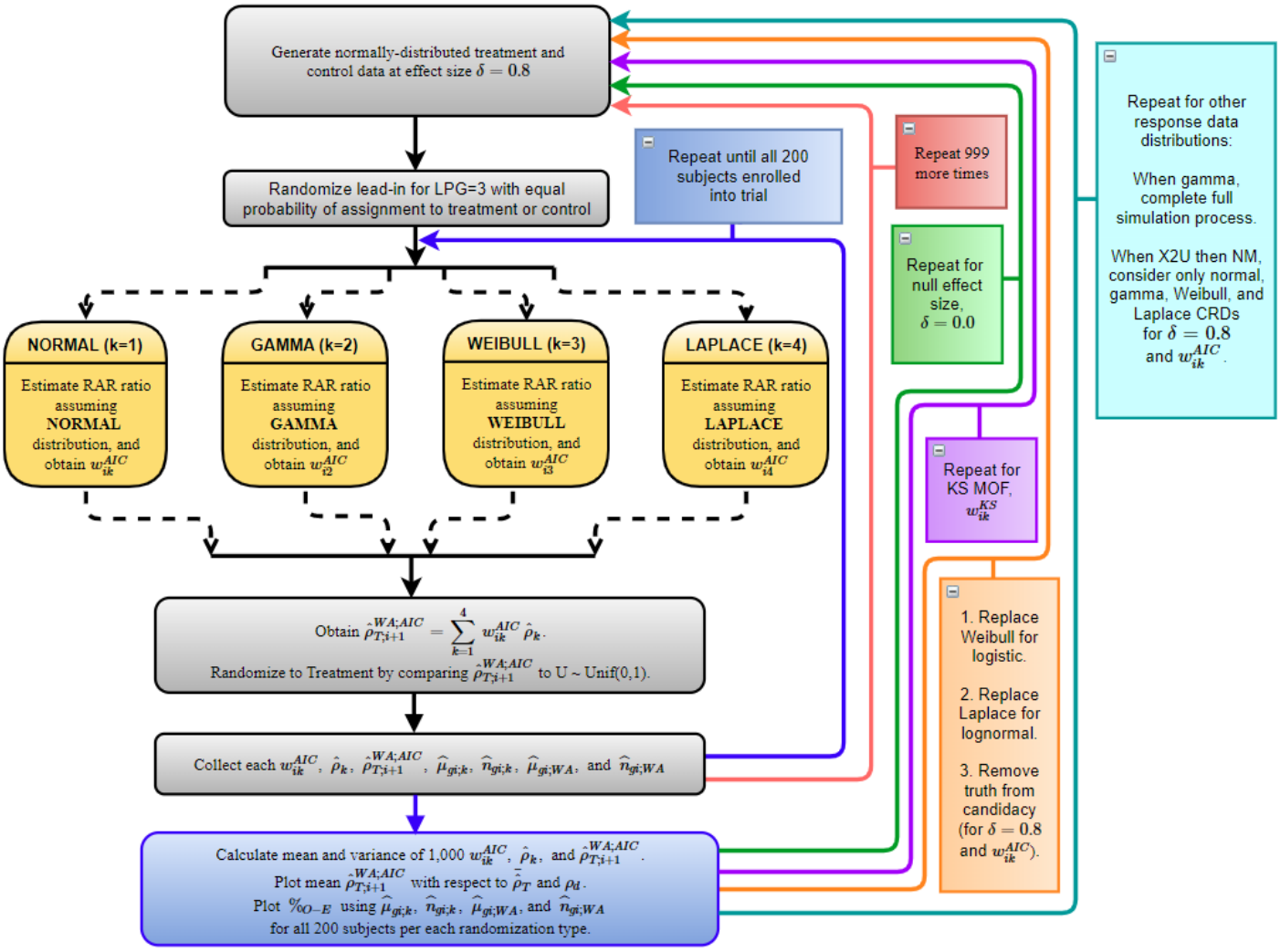
where $\hat{w}_{ik}$ represents the MOF-estimated distributional weight and each $\hat{\rho}_k$ takes the form of 1.18, where $\hat{\sigma}_{gi}$ and $\hat{\mu}_{gi}$ were estimated using the MLEs specific to each CRD k.

3.2 Simulation Process for Evaluation of Weighted-Average Randomization

A simulation study was performed to determine how well the WA method of RAR accounted for distributional misspecification of the RAR ratio. The goal of this simulation study was identical to those discussed in Section 2.2: to evaluate the bias, precision, and ethicality of the method-produced estimator of the RAR ratio using plots of the mean RAR ratios from 1,000 trials and their 95% CI along with the $\%_{O-E}$ plot. Specifically, the mean WA RAR ratio ($\overline{\hat{\rho}}_{WA}$) was compared $\overline{\hat{\rho}}_T$ and ρ_d. Since more heavily-weighted CRDs represent larger proportions of the WA-estimated RAR ratio, the behavior of the CRDs when fit to the observed response data was also of interest, and was assessed using plots of the means of the CRD-specific weights across the 1,000 simulated trials for each of the 200 subjects enrolled into the trial. As well, an assessment of type I error was performed as in Chapter 2, and results were presented by Set (discussed later in this section).

Figure 3.1 provides a visual representation of the WA simulation study. As in Section 2.2, response data were simulated to follow the normal, gamma, X2U, and NM distributions, characterized by the parameters given in Table 2.1. New to the simulation process was the estimation of the RAR ratios corresponding to each CRD within a predetermined set of CRDs. Overall, six CRDs were considered: the normal, Laplace, and logistic distributions, selected for their symmetry, and the gamma, Weibull, and lognormal distributions, selected for their ability to incorporate skewness. Derivation of the MLEs corresponding to the Laplace, logistic, Weibull and lognormal distributions

Figure 3.1: Simulation process for evaluation of WA method of RAR using parametric continuous response data.



are provided in Appendices A.24-A.27, respectively. When the treatment-specific response data were simulated to follow normal and gamma distributions, these CRDs were arranged into distinct sets designed to assess the behavior of WA RAR when multiple hypothetically well-fitting and/or poorly-fitting CRDs were considered within the same set. Specifically, four sets were evaluated:

- Set 1: Truth, two hypothetically well-fitting CRDs, and one hypothetically poorly-fitting CRD,
- Set 2: Truth, one hypothetically well-fitting CRD, and two hypothetically poorly-fitting CRDs,
- Set 3: Truth and three hypothetically poorly-fitting CRDs, and,
- Set 4: One hypothetically well-fitting CRD and two hypothetically poorly-fitting CRDs (i.e., no truth),

where CRDs were hypothesized to fit the simulated response data well or poorly based on the symmetrical or skewed nature of the CRD in relation to the shape of the true distribution of the simulated response data. Because the truth was included in Sets 1-3, it was hypothesized to be the most heavily-weighted, and, thus, best-fitting CRD in these three sets.

Table 3.1: Candidate response distributions within each simulation set when treatment-specific response data were simulated to follow the normal (N), gamma (G), X2U, and NM distributions.

CRDs	Set 1		Set 2		Set 3		Set 4			
Symmetrical:										
Normal	N^*	G	N^*	G	N^*	G		G	$X2U$	NM^*
Laplace	N		N	G		G	N^*	G	$X2U$	NM
Logistic	N					G				
Skewed:										
Gamma	N	G^*	N	G^*	N	G^*	N		$X2U^*$	NM
Weibull		G	N	G	N		N	G^*	$X2U^*$	NM
Lognormal		G			N					

*Hypothesized to be best-fitting CRD(s) within set for response distribution indicated

Table 3.1 details which of the six CRDs falls into each set according to the true response distribution being simulated. Of particular interest was Set 4, designed to illustrate the behavior of WA randomization when the true response distribution was not considered among the candidates. In these scenarios, the best-fitting CRD was hypothesized to be the CRD that most closely matches the known shape of the simulated response data. Best-performing CRDs were indicated in Table 3.1 with an asterisk. Where Sets 1-3 considered scenarios where response data followed only the exponential families of the normal and gamma distributions, Set 4 was also applied to the more complex mixture distributions of X2U and NM. These scenarios were included because the RAR ratio would very likely be misspecified if response data followed distributions such as these in practice. Such distributions would be unlikely to be considered as candidates if employing the WA approach. Thus, Set 4 represents weighted-averaging over a set of simpler CRDs may serve to alleviate the likely misspecification of the RAR ratio that could occur in practice. Table 3.1 shows, two symmetrical and two skewed CRDs were considered when response data were truly X2U- or NM-distributed: the normal and Laplace (symmetrical) and the Weibull and gamma (skewed) distributions. Because the X2U distribution is heavily right-skewed (Figure 2.2), it was anticipated that either the gamma or Weibull distribution, or both, would out-weigh the symmetrical CRDs. And, when response data followed bimodal NM distributions, it was hypothesized that the normal distribution would be the most heavily-weight CRD.

Results from the lead-in analysis discussed in Section 2.2.1 were applied to these simulations such that all CRDs began RAR after LPG=3 and RAR when response data were simulated to follow the X2U and NM distributions, performed for comparison purposes, began after LPG=5 and LPG=7, respectively. Furthermore, two MOFs were used to develop the CRD-specific weights: AIC[45] and the KS test statistic[48]. Finally, simulations were conducted for no treatment difference ($\delta = 0.0$) and a large treatment difference ($\delta = 0.8$). Small and moderate effect sizes were omitted

from the present simulation study based on the results observed in Chapter 2.

3.3 Randomization Results

All curves were loess-smoothed using a bandwidth of 0.2. As illustrated in plot B of both Figures 3.2 and 3.3, the KS test statistic poorly distinguished between CRDs, indicated by mean distributional weights falling within two- and four-thousandths, respectively, of equal weighting (grey horizontal line at 1/K). Therefore, randomization results with respect to AIC-weighting are discussed in this section.

3.3.1 Absence of Treatment Difference

When treatment-specific response data were simulated to follow the normal distribution in the absence of a treatment difference, randomization proceeded as expected for CRD sets containing normality (Sets 1-3). That is, $\overline{\hat{\rho}}_{WA}$, as well as each CRD-specific mean RAR ratio, approximated $\rho_d = 0.5$ (Figure 3.4), and the true response distribution, normality (green curve), was most heavily-weighted across CRD sets (Figure 3.5). The mean distributional weight for the true response distribution, $\overline{\hat{w}}_T$, went to one (i.e., $\overline{\hat{w}}_T \to 1$) almost immediately following the lead-in period for CRD Set 2 (plot B). This suggests that, in addition to the skewed gamma (blue curve) and Weibull (red curve) distributions, the Laplace distribution (purple curve) does not fit normally-distributed response data well either, despite its symmetry. Further evidence to this point is provided in plot A (CRD Set 1), where the mean distributional weight for the Laplace CRD, $\overline{\hat{w}}_{La}$, went to zero, $\overline{\hat{w}}_T \to 0$. When all non-truth CRDs were skewed (plot C, CRD Set 3), the mean distributional weight for the Weibull CRD, $\overline{\hat{w}}_W$, went to zero (i.e., $\overline{\hat{w}}_W \to 0$) while the mean distributional weights for the gamma and lognormal CRDs, $\overline{\hat{w}}_G$ and $\overline{\hat{w}}_{logN}$, respectively, approximated equal weighting at 1/K. Finally, all CRDs and WA (black curve) RAR upheld the ethical objective of RAR to within 0.3% (Figure 3.6), despite the $\%_{\text{O-E}}$ for the Weibull CRD increasing as the number of participants

Figure 3.2: Mean of RAR ratios with 95% CI (A), mean of KS-based distributional weights (B), and percent-difference between the observed and expected ethical objective function (C) over 1,000 simulated trials when treatment-specific participant response data were normally-distributed and reflected a large treatment difference, and CRDs were the normal, gamma, Weibull, and Laplace distributions (i.e., CRD Set 2).

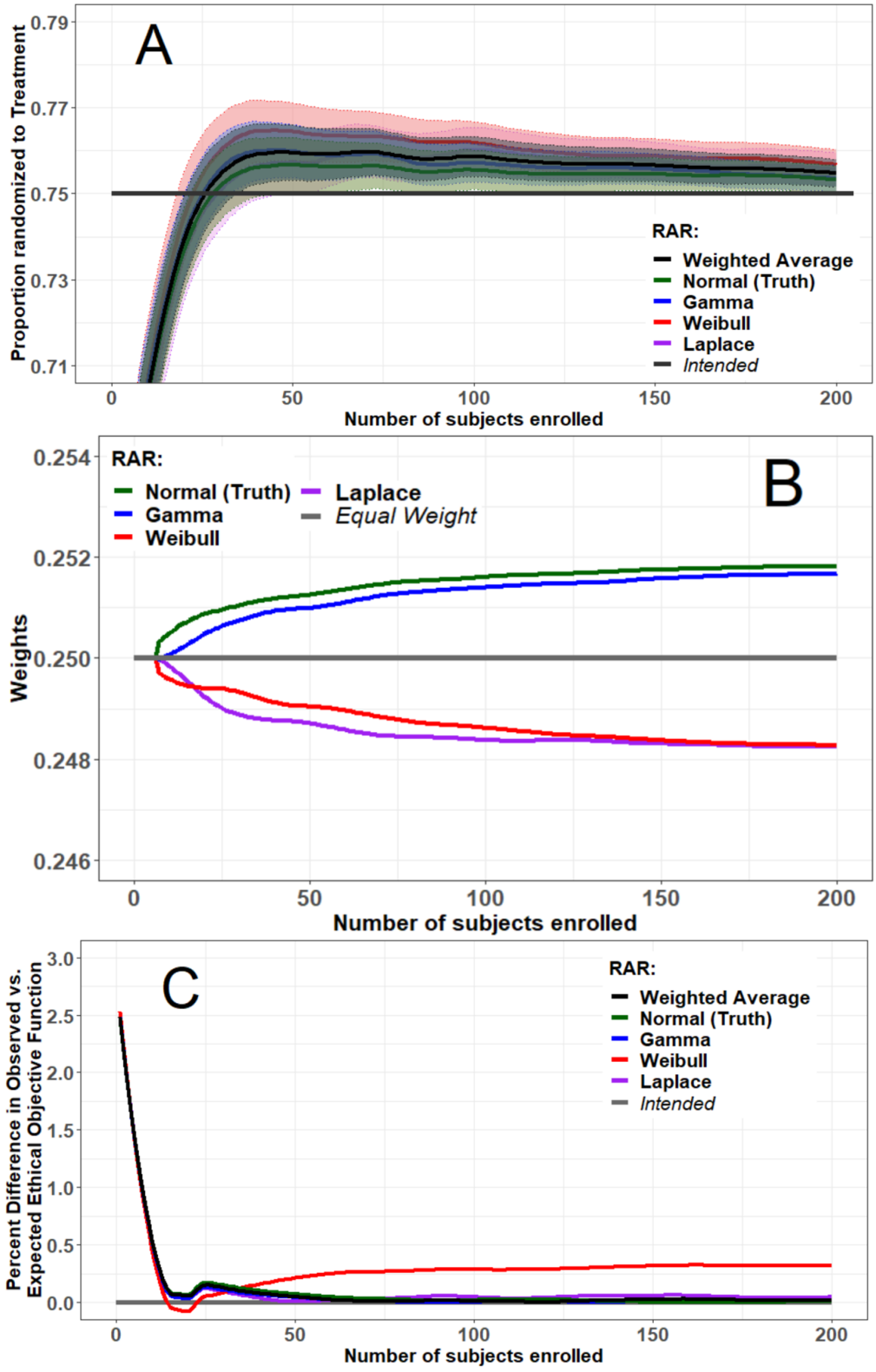


Figure 3.3: Mean of RAR ratios with 95% CI (A), mean of KS-based distributional weights (B), and percent-difference between the observed and expected ethical objective function (C) over 1,000 simulated trials when treatment-specific participant response data were simulated to follow the gamma distribution and reflected a large treatment difference, and CRDs were the normal, gamma, Weibull, and Laplace distributions (i.e., CRD Set 2).

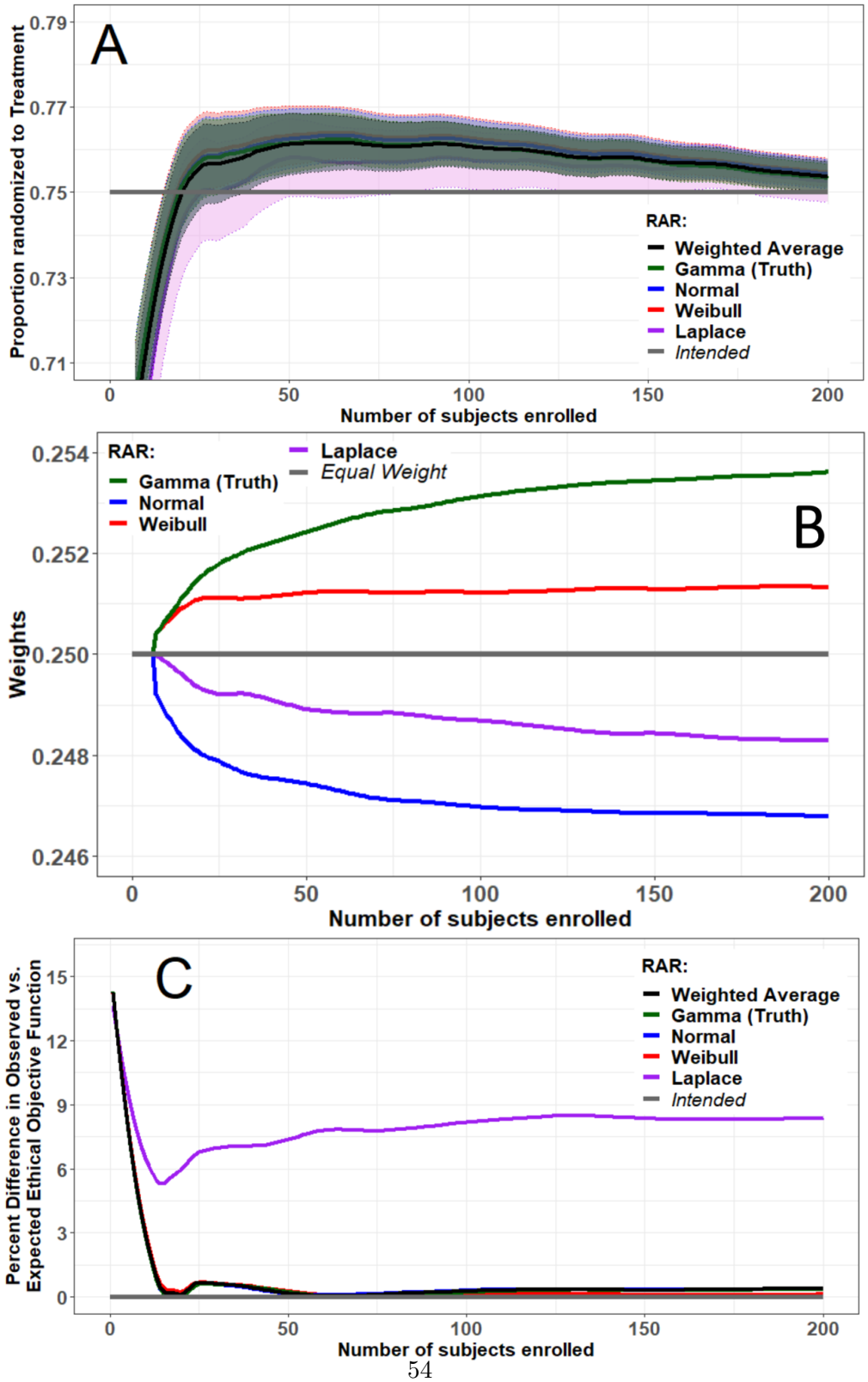


Figure 3.4: Mean of CRD-specific and WA RAR ratios with 95% CI over 1,000 simulated trials when treatment-specific participant responses were normally-distributed and reflected no treatment difference using CRD Sets 1 (A), 2 (B), and 3 (C).

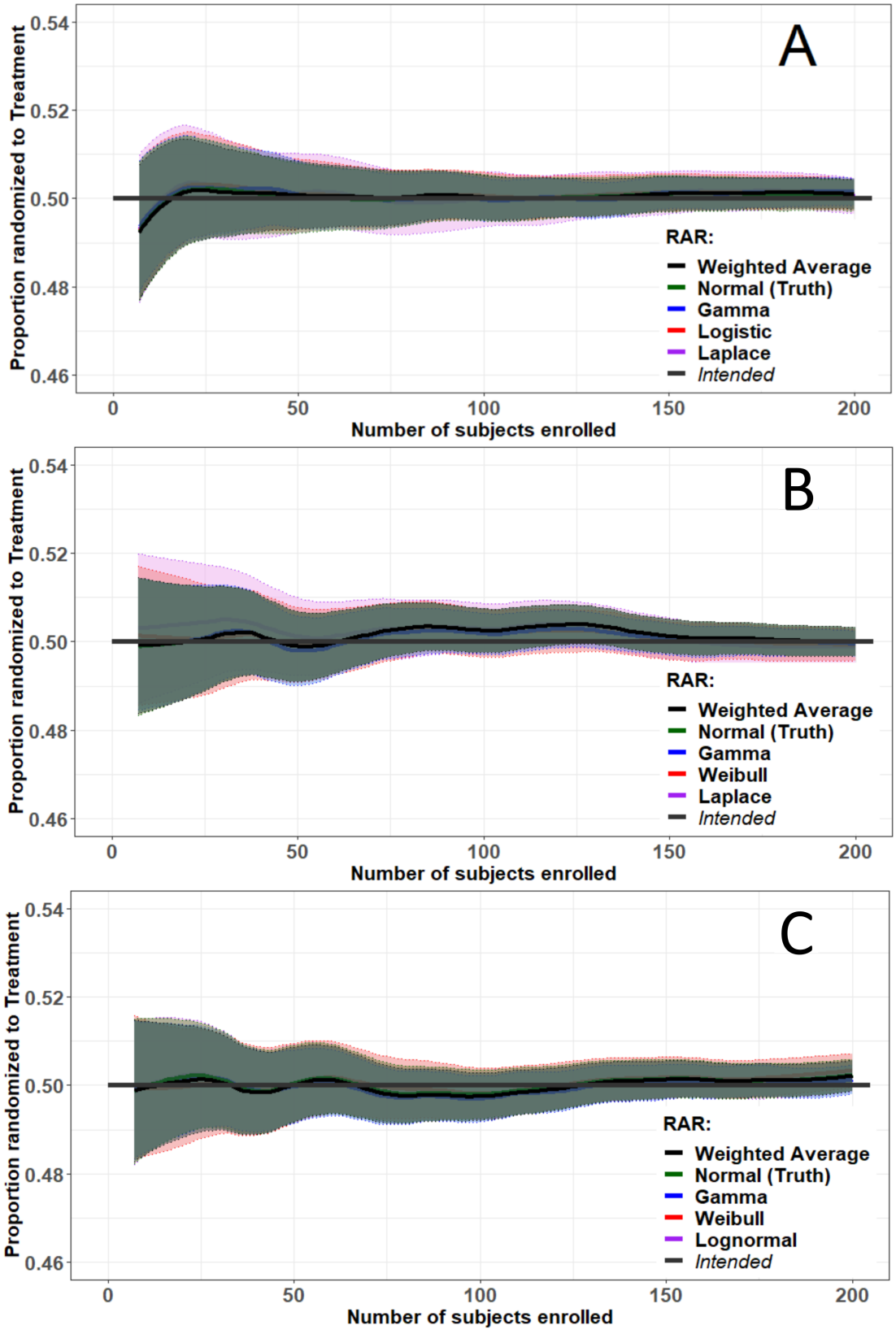


Figure 3.5: Mean of AIC-estimated distributional weights from 1,000 simulated trials when treatment-specific participant responses were normally-distributed and reflected no treatment difference using CRD Sets 1 (A), 2 (B), and 3 (C).

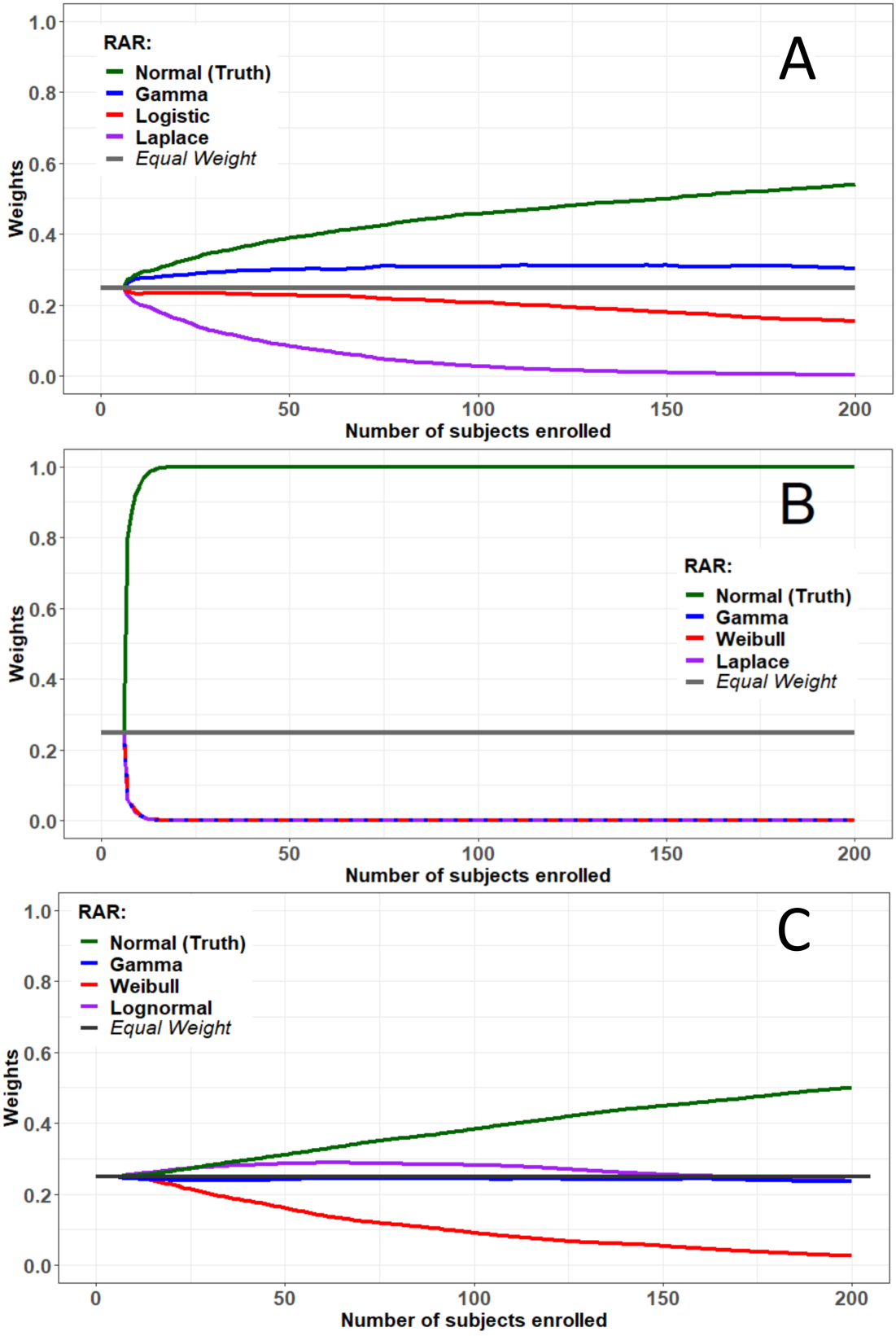
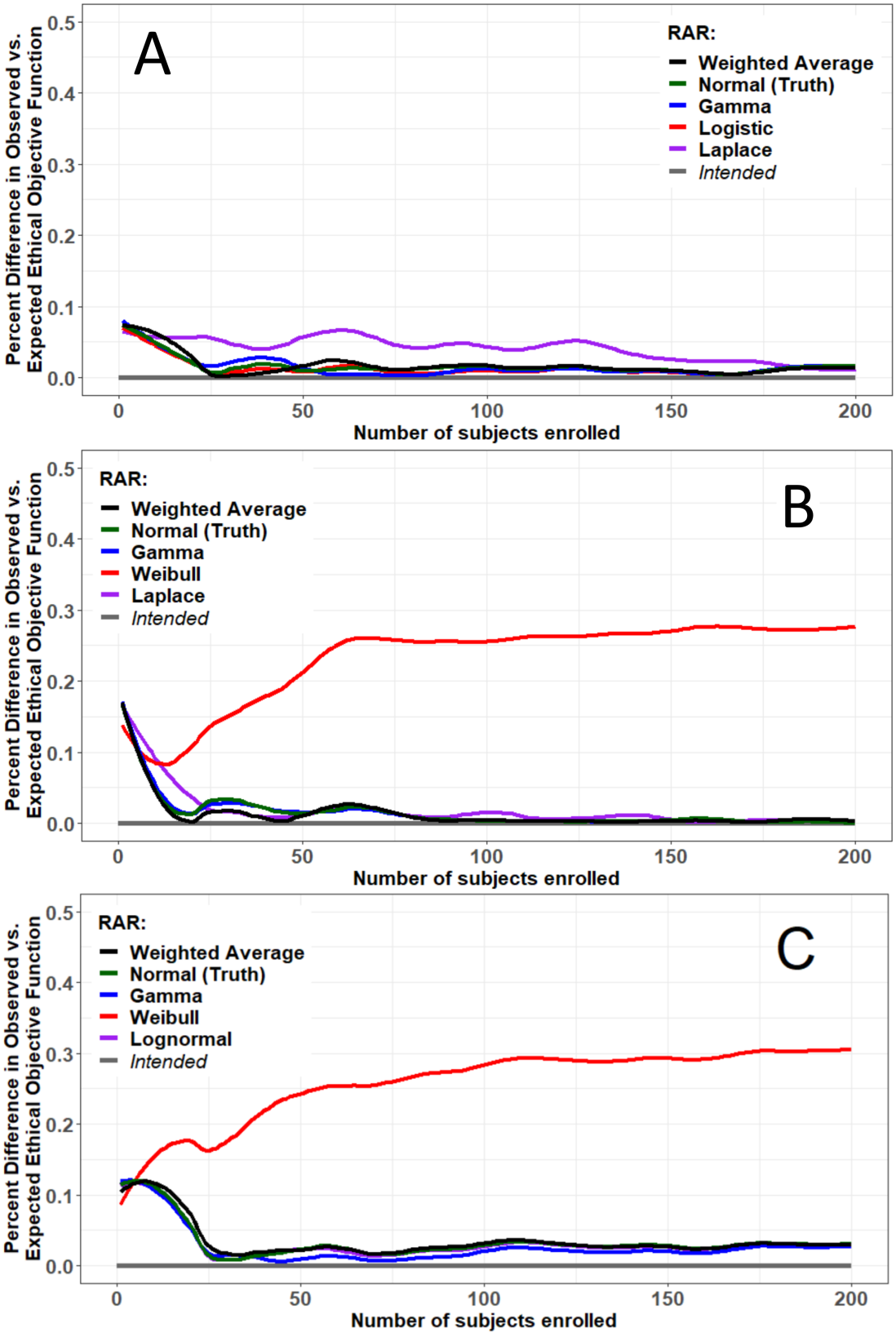


Figure 3.6: Percent-difference between the observed and expected ethical objective function from 1,000 simulated trials when treatment-specific participant responses were normally-distributed and reflected no treatment difference using CRD Sets 1 (A), 2 (B), and 3 (C).



enrolled into the trial increased.

When treatment-specific response data were simulated to follow the gamma distributions in the absence of a treatment difference, neither $\overline{\hat{\rho}}_{WA}$ nor any of the CRD-specific mean RAR ratios deviated from $\rho_d = 0.5$ for any of the 200 participants enrolled into the trial (Figure 3.7). Furthermore, across CRD sets, the gamma distribution (green curve) was consistently weighted heaviest (Figure 3.8), as anticipated. The two skewed CRDs, Weibull (red curve) and lognormal (purple curve), fit the gamma-distributed response data equally well and better than the symmetrical normal distribution (blue curve) for CRD Set 1 (plot A). Maintaining this trend, the skewed Weibull distribution better-fit the skewed gamma-distributed response data than the symmetrical CRDs of normality and the Laplace distribution (purple curve) (plot B, CRD Set 2). Finally, when all non-truth CRDs were symmetrical, weights quickly went to zero following the lead-in period, while $\overline{\hat{w}}_T \to 1$ (plot C, CRD Set 3).

Table 3.2: Assessment of type I error for normal or gamma response data for Sets 1-3 using t-test based on observed mean response from first 100 trial participants enrolled when no treatment difference existed ($\delta = 0.0$) across 1,000 simulated trials.

Truth	Set	Treatment Mean	Control Mean	t	p-value
Normal	1	100.14	100.12	0.396	0.6925
	2	100.10	100.15	-0.977	0.3288
	3	100.09	100.13	-0.774	0.4393
Gamma	1	10.11	10.10	0.052	0.9584
	2	10.10	10.09	0.449	0.6537
	3	10.15	10.09	1.519	0.1290

As indicated by $\overline{\hat{w}}_{La} \to 0$ in plot B and both the mean distributional weight for the logistic CRD, $\overline{\hat{w}}_{lo}$, and $\overline{\hat{w}}_{La}$ going to zero in plot C of Figure 3.8, the WA RAR was not heavily influenced by the Laplace or logistic CRDs, especially after the 50^{th} trial participant was enrolled. Therefore, without loss of generality, $\%_{\text{O-E}}$ for the Laplace CRD alone and both the Laplace and logistic CRDs were omitted from plots B and C, respectively, of Figure 3.9 due to lack of interpretability. Across

Figure 3.7: Mean of CRD-specific and WA RAR ratios with 95% CI over 1,000 simulated trials when treatment-specific participant responses were simulated to follow gamma distributions reflecting no treatment difference using CRD Sets 1 (A), 2 (B), and 3 (C).

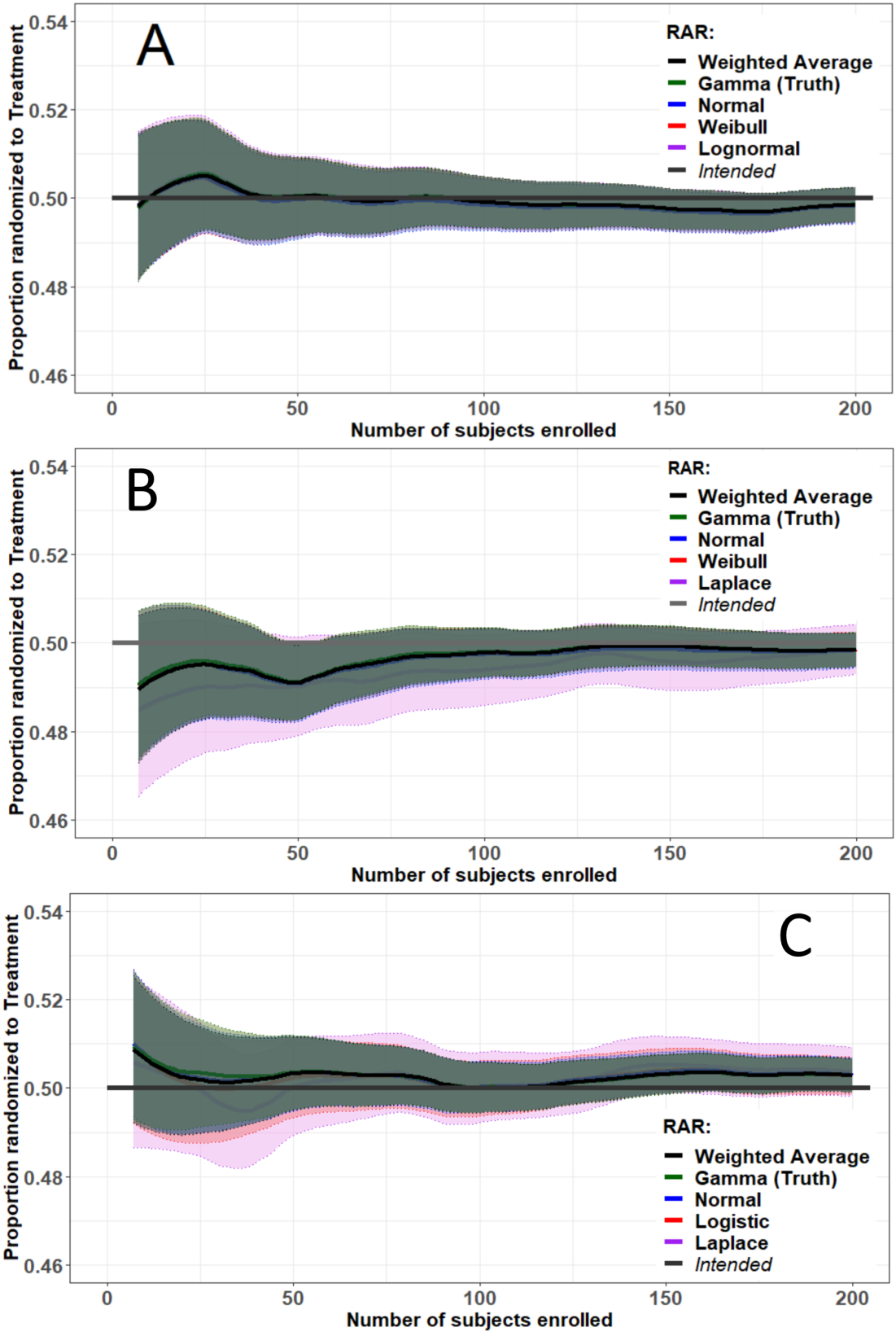


Figure 3.8: Mean of AIC-estimated distributional weights from 1,000 simulated trials when treatment-specific participant responses were simulated to follow gamma distributions reflecting no treatment difference using CRD Sets 1 (A), 2 (B), and 3 (C).

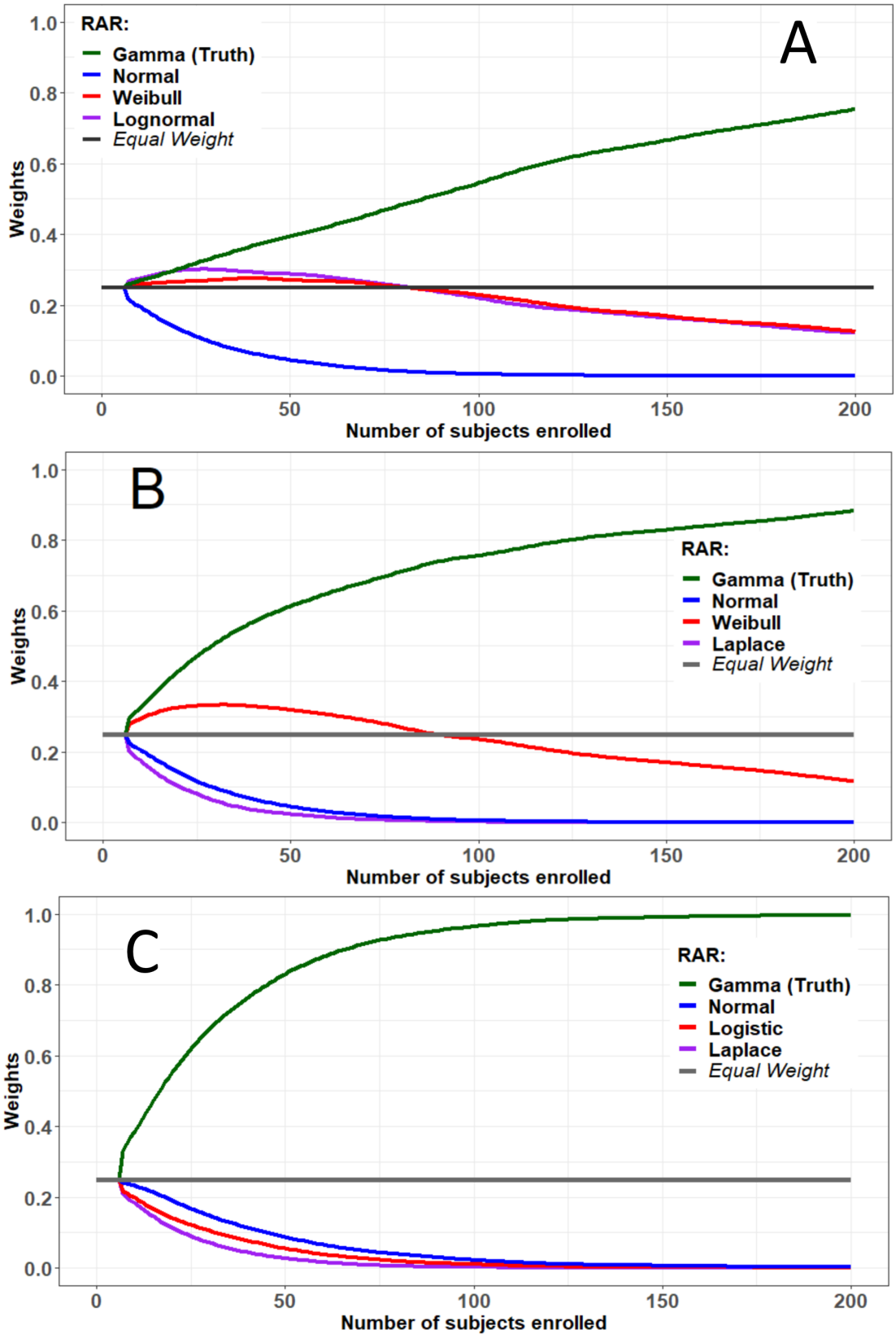
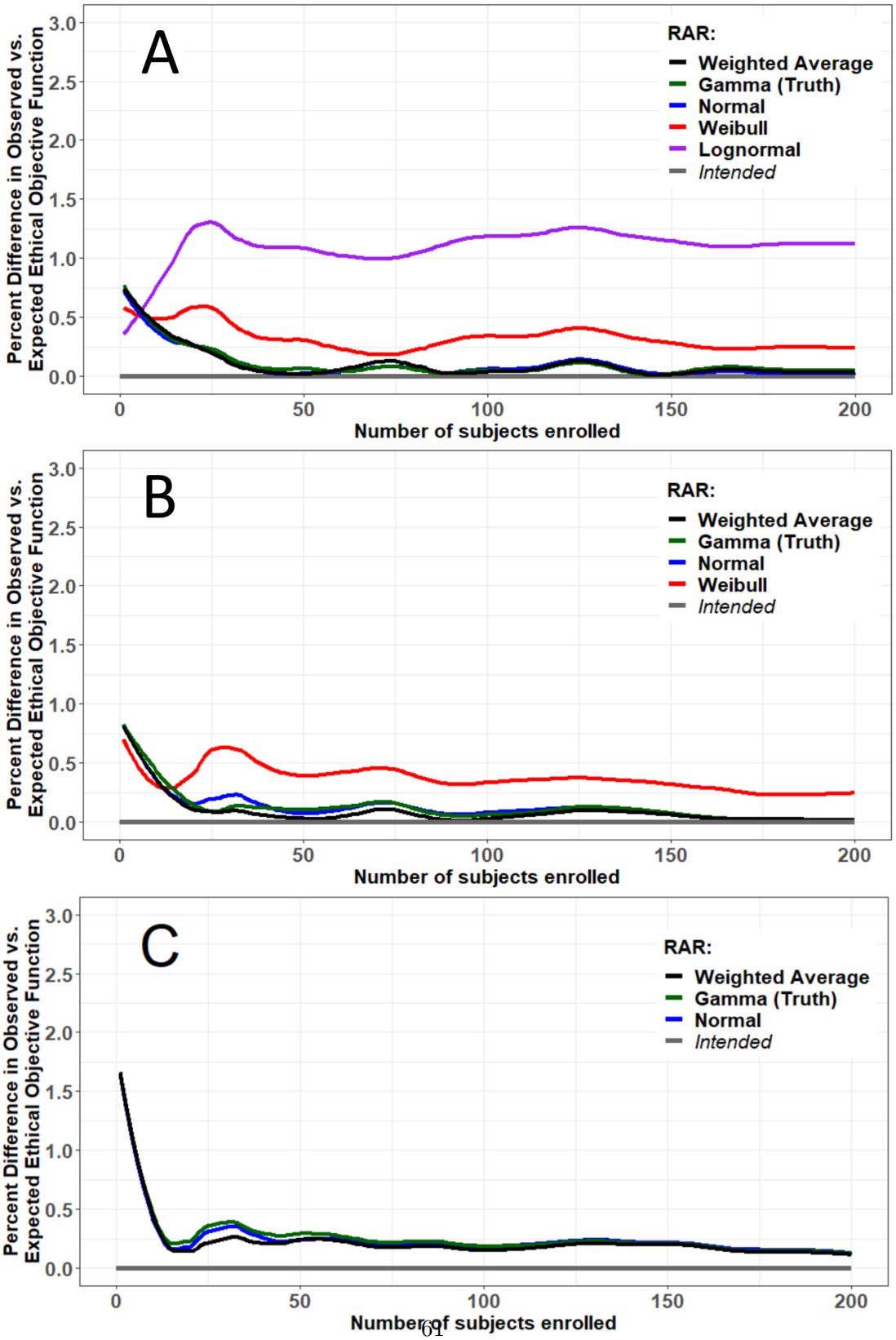


Figure 3.9: Percent-difference between the observed and expected ethical objective function from 1,000 simulated trials when treatment-specific participant responses were simulated to follow gamma distributions reflecting no treatment difference using CRD Sets 1 (A), 2 (B), and 3 (C).



CRD sets, $\%_{\text{O-E}}$ never exceeded 1.75, suggesting that all of the CRD-specific and WA (black curve) RAR ratios successfully adhered to the ethical objective of minimizing the mean response value of the trial. Finally, no type I errors were committed in the scenarios assessed, as evidenced by Table 3.2.

3.3.2 Large Treatment Difference

When treatment-specific response data were simulated to follow normal distributions reflecting a large treatment difference, WA RAR (Figure 3.10, black curve) favored the RAR ratio estimated assuming normality (Figures 3.10 and Figure 3.11, green curve), regardless of the set of CRDs considered. All other distributional weighting trends observed in Figure 3.11 were identical to those observed in Figure 3.5. As well, all CRDs randomized participants similarly, regardless of CRD set, and all mean RAR ratios, both CRD-specific and WA-estimated, nearly converged to ρ_d (Figure 3.10). Finally, all of the CRD-specific RAR ratios as well as the WA method of RAR adhered to the ethical objective of the trial (Figure 3.12). During the lead-in period, all $\%_{\text{O-E}}$ were less than or equal to three, then fell to and remained below 0.5%. Though a small difference, the $\%_{\text{O-E}}$ for the Weibull CRD shows a pattern dissimilar to the $\%_{\text{O-E}}$ for other CRDs and for WA RAR (plots B and C), mirroring the results presented in plots B and C of Figure 3.6.

When treatment-specific response data were simulated to follow gamma distributions reflecting a large treatment difference, each CRD, as well as WA RAR (black curve), randomized trial participants similarly (plots B and C) or nearly identically (plot A), depending on the set of CRDs considered (Figure 3.13). All mean RAR ratios converged to approximately $\rho_d + 0.05$ and over-randomized to the treatment group for all trial participants following the lead-in period, with one exception: when all non-truth CRDs were symmetrical (plot C, CRD Set 3), $\overline{\hat{\rho}}_{La}$ (purple curve) achieved and maintained ρ_d (grey horizontal line) from the 50^{th} participant onward. However, as illustrated by plot C in Figure 3.14, $\overline{\hat{w}}_T \to 1$ (green curve) while all other weights, including the

Figure 3.10: Mean of CRD-specific and WA RAR ratios with 95% CI over 1,000 simulated trials when treatment-specific participant responses were normally-distributed and reflected a large treatment difference using CRD Sets 1 (A), 2 (B), and 3 (C).

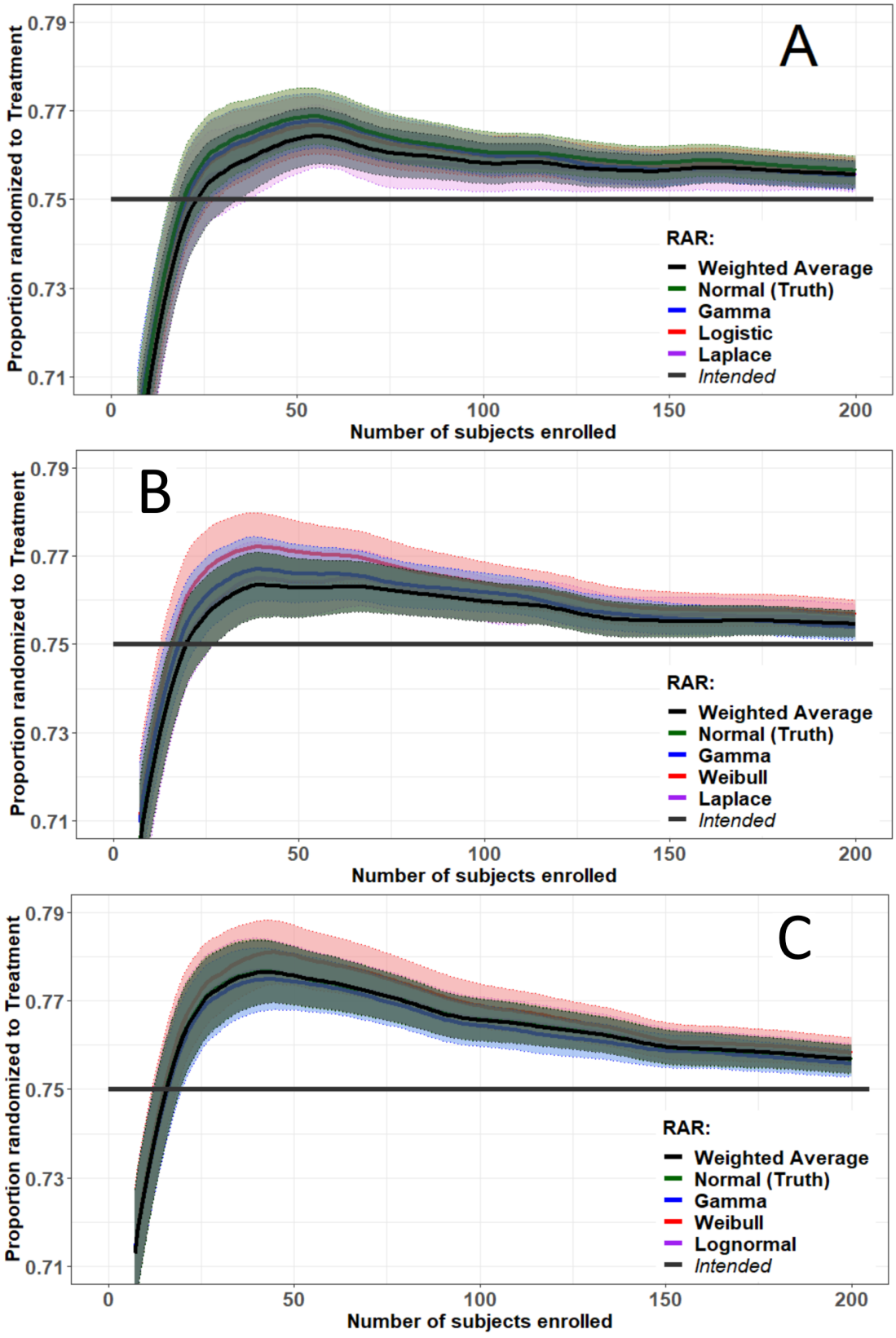


Figure 3.11: Mean of AIC-estimated distributional weights from 1,000 simulated trials when treatment-specific participant responses were normally-distributed and reflected a large treatment difference using CRD Sets 1 (A), 2 (B), and 3 (C).

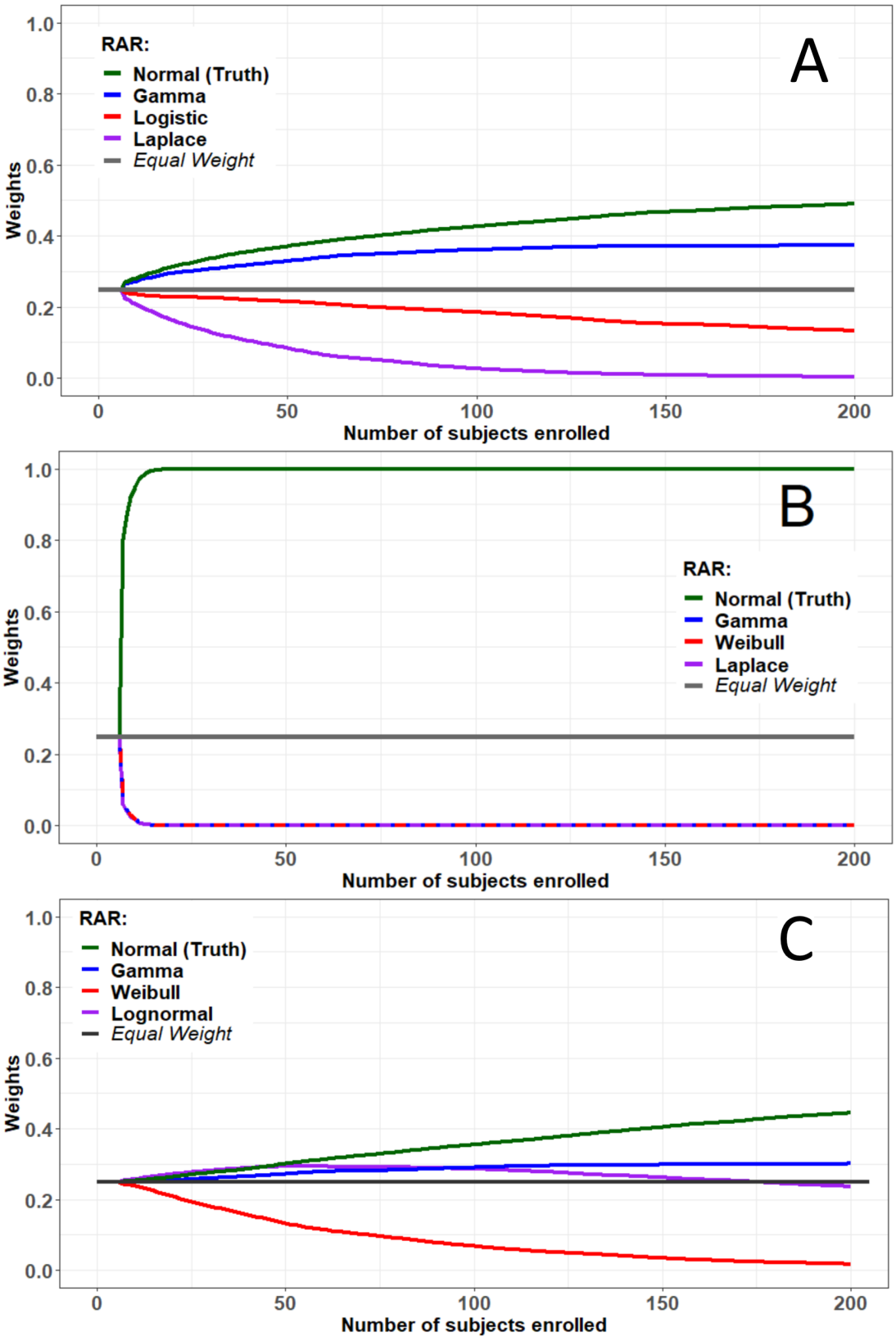


Figure 3.12: Percent-difference between the observed and expected ethical objective function from 1,000 simulated trials when treatment-specific participant responses were normally-distributed and reflected a large treatment difference using CRD Sets 1 (A), 2 (B), and 3 (C).

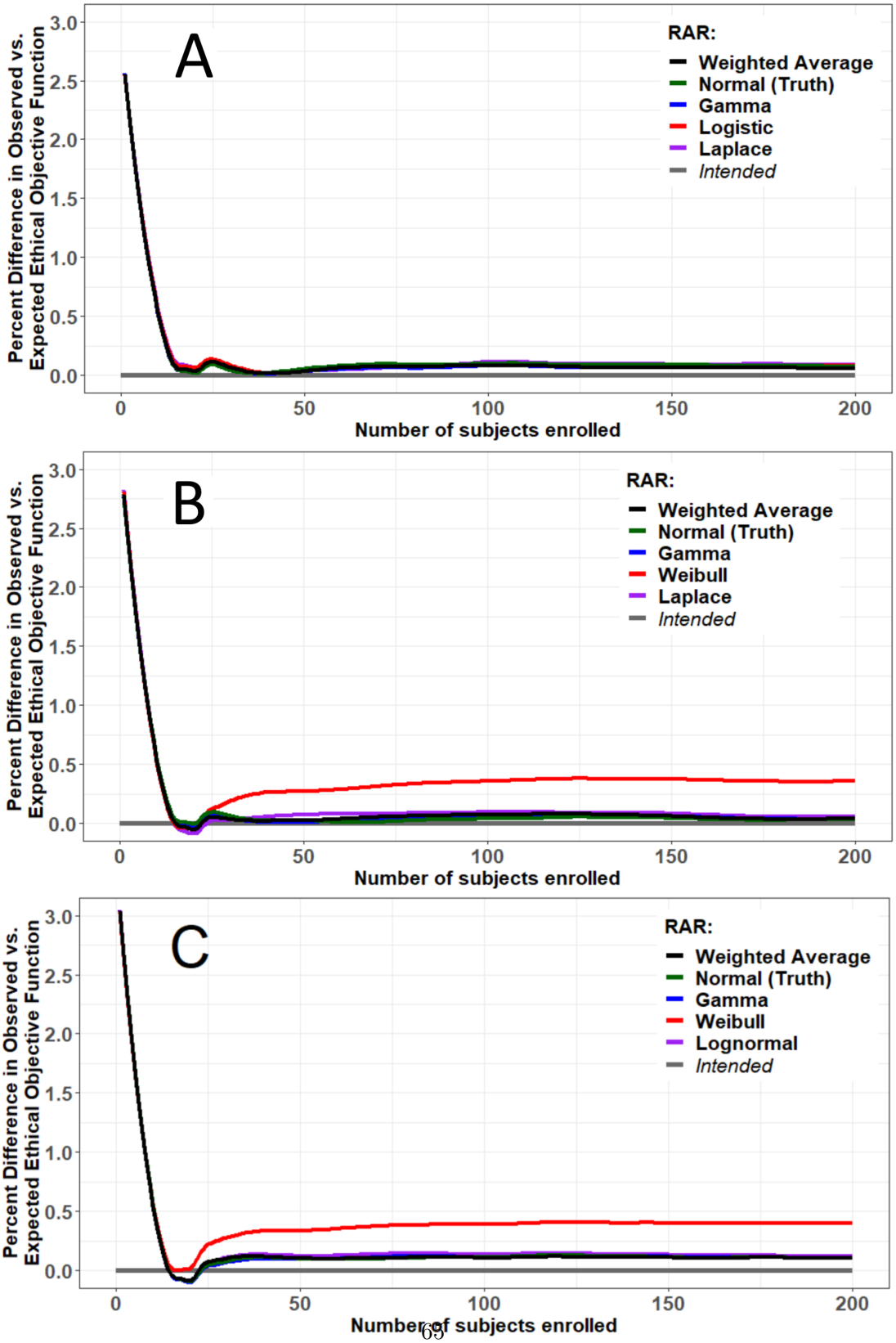


Figure 3.13: Mean of CRD-specific and WA RAR ratios with 95% CI over 1,000 simulated trials when treatment-specific participant responses were simulated to follow gamma distributions reflecting a large treatment difference using CRD Sets 1 (A), 2 (B), and 3 (C).

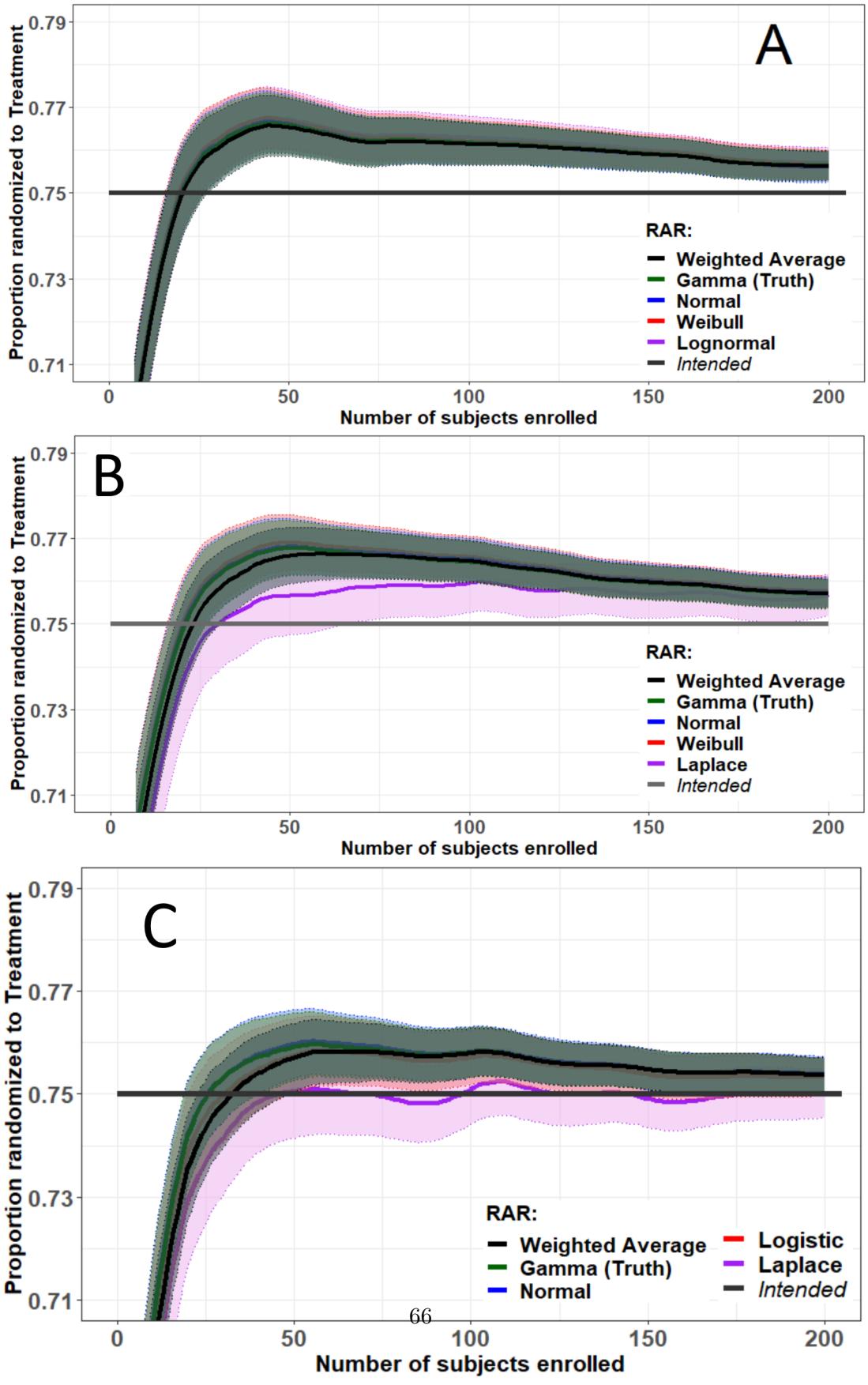
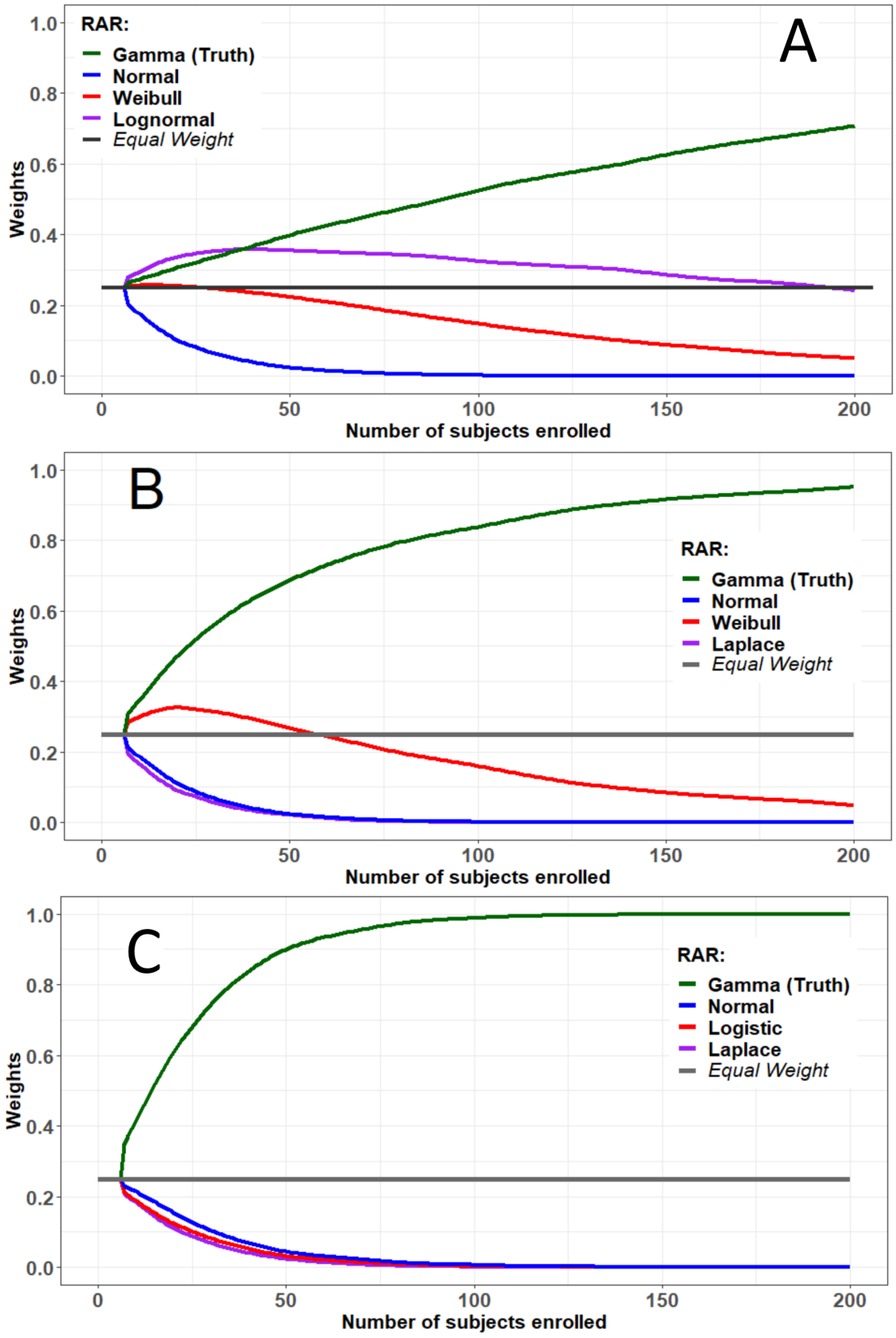


Figure 3.14: Mean of AIC-estimated distributional weights from 1,000 simulated trials when treatment-specific participant responses were simulated to follow gamma distributions reflecting a large treatment difference using CRD Sets 1 (A), 2 (B), and 3 (C).



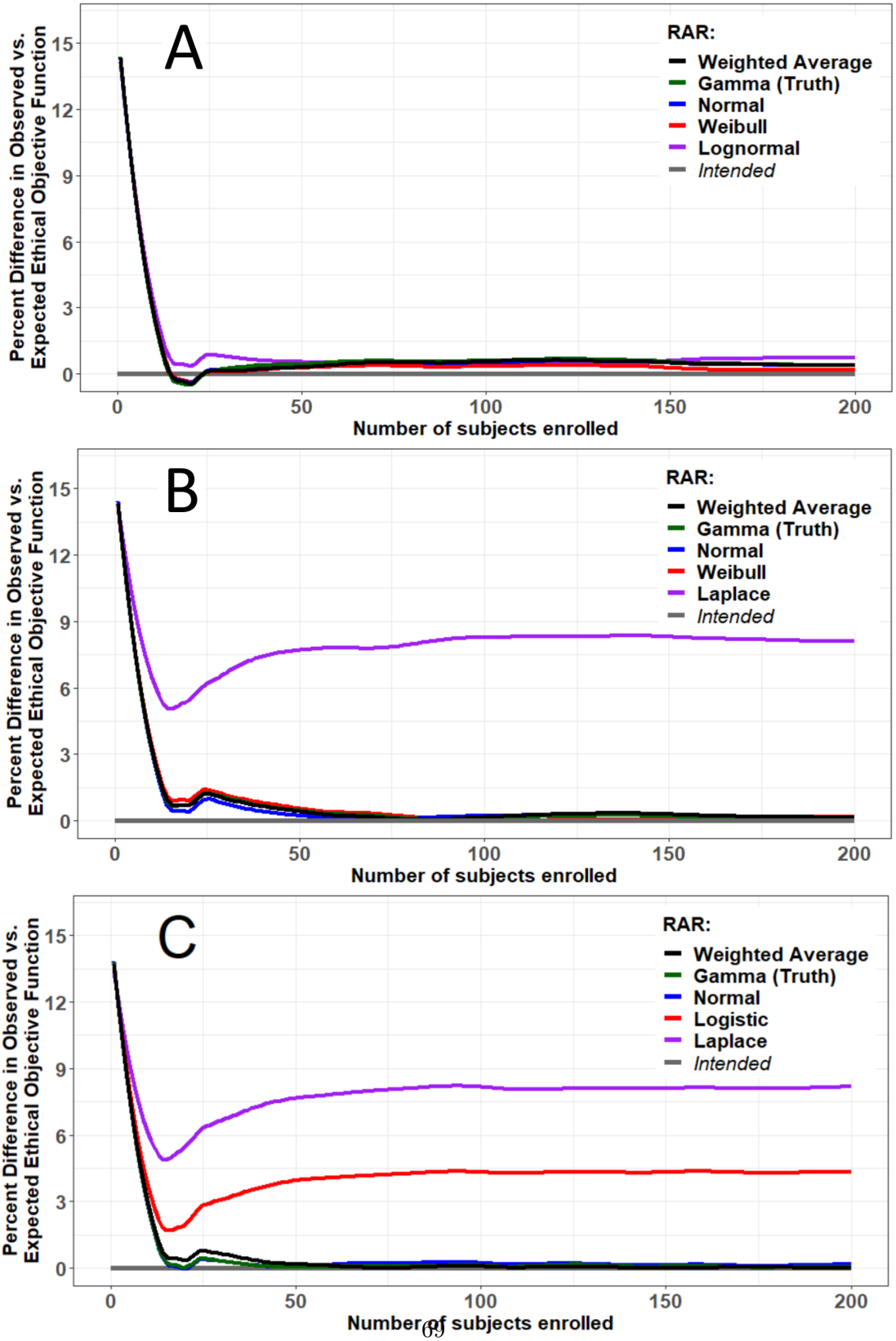
$\overline{\hat{w}}_{La}$, went to zero. Therefore, despite the Laplace-specific RAR ratio's ability to randomize participants as intended by the trial design in this particular setting, it had negligible influence over the WA RAR ratio used to enroll participants into the trial. The remaining distributional weight patterns gleaned in Figure 3.14 were nearly identical to those displayed in the baseline scenario (Figure 3.8).

Laplace RAR produced substantial deviation from the intended ethical objective function when included in the CRD set (Figure 3.15, plots B and C), stabilizing between a 7.5-9% discrepancy. Similarly, RAR assuming the logistic distribution stabilized around a 4.5% deviation between the observed and intended ethical objective function (plot C, red curve). Echoing the baseline scenario (Figure 3.9), since $\overline{\hat{w}}_{La}$ and $\overline{\hat{w}}_{lo}$ went to zero, the $\%_{\text{O-E}}$ results for the Laplace and logistic CRDs were not considered in the interpretation of these results, though they were maintained on plots B and C since their inclusion does not compromise or alter the presentation of the meaningful $\%_{\text{O-E}}$ results. With this, Figure 3.15 illustrates that $\%_{\text{O-E}}$ was less than 15 for all CRDs, as well as for the WA method, during the lead-in period. Subsequently, this deviation diminished to and remained below 1.75%, consistent with the results obtained from the baseline scenario. Therefrom, all estimated RAR ratios fulfilled the ethical objective of the trial.

3.3.3 Truth Removed from Candidacy

Of particular emphasis was the behavior of WA RAR when the true response distribution was not considered within the set of candidates (i.e., Set 4). In the previously-discussed results, the true response distribution was the CRD favored during randomization. This allowed for straightforward results dissemination: overall, CRD-specific RAR and WA RAR behaved similarly and distributional weights identified the true response distribution as the best-fit CRD. When the true response distribution was removed from candidacy, however, randomization results were more complex, requiring more careful visualizations. The plots of the mean RAR ratios were divided into and

Figure 3.15: Percent-difference between the observed and expected ethical objective function from 1,000 simulated trials when treatment-specific participant responses were simulated to follow gamma distributions reflecting a large treatment difference using CRD Sets 1 (A), 2 (B), and 3 (C).



presented using two figures: a plot representing $\overline{\hat{\rho}}_{WA}$ with respect to the hypothetically-unknown $\overline{\hat{\rho}}_T$ and ρ_d (designated plot A-I), and a plot displaying RAR assuming each of the CRDs (designated plot A-II).

When group-specific response data were simulated to follow normal distributions reflecting a large treatment difference and the set of CRDs contained the gamma, Weibull, and Laplace distributions,the resulting WA RAR ratio (dashed green curve) favored normality despite it not being considered among the candidates (Figure 3.16, plot A-I). Moreover, $\overline{\hat{\rho}}_{WA} \approx \rho_d$ (grey horizontal line) for the final 100 participants enrolled into the trial. Though all CRDs randomized similarly (plot A-II), $\overline{\hat{w}}_G \to 1$ while $\overline{\hat{w}}_W$ and $\overline{\hat{w}}_{La}$ went to zero, suggesting that the gamma CRD, versus the hypothesized Laplace CRD, best-fit the normally-distributed response data. Finally, the ethical objective of RAR was upheld, where, following the lead-in period, $\%_{\text{O-E}}$ was less than one-half percent for all RAR ratios (plot C).

When treatment-specific response data were simulated to follow gamma distributions reflecting a large treatment difference, the WA RAR ratio (black solid curve, plot A-I, Figure 3.17), obtained from the skewed Weibull (red curve, plot A-II) and symmetrical normal (blue curve, plot A-II) and Laplace distribution (purple curve, plot A-II) RAR ratios as weighted by AIC, closely mirrored RAR conducted assuming the true gamma distribution (dashed green curve, plot A-I). As anticipated, versus the symmetrical normal or Laplace CRDs, the WA RAR ratio was driven by the fit of the skewed Weibull CRD to the skewed gamma-distributed response data, demonstrated in plot B where $\overline{\hat{w}}_W \to 1$ and both the mean distributional weight for the normal CRD, $\overline{\hat{w}}_N$, and $\overline{\hat{w}}_{La}$ went to zero following the lead-in period. Similar to the scenarios where the gamma distribution was included within the set of CRDs (Figures 3.9 and 3.15), each of the CRD-specific and WA RAR ratios adhere to the ethical objective of the trial with the exception of the uninterpretable Laplace CRD (plot C).

Figure 3.16: Comparison of mean WA RAR ratio against mean true response distribution (non-candidate) RAR ratio (95% CIs) (A-I), mean RAR ratios (95% CIs) for each CRD (A-II), mean of AIC-based distributional weights (B), and percent-difference between the observed and expected ethical objective function (C) over 1,000 simulated trials when treatment-specific participant responses were simulated to follow normal distributions reflecting a large treatment difference, and CRDs were the gamma, Weibull, and Laplace distributions (i.e., CRD Set 4).

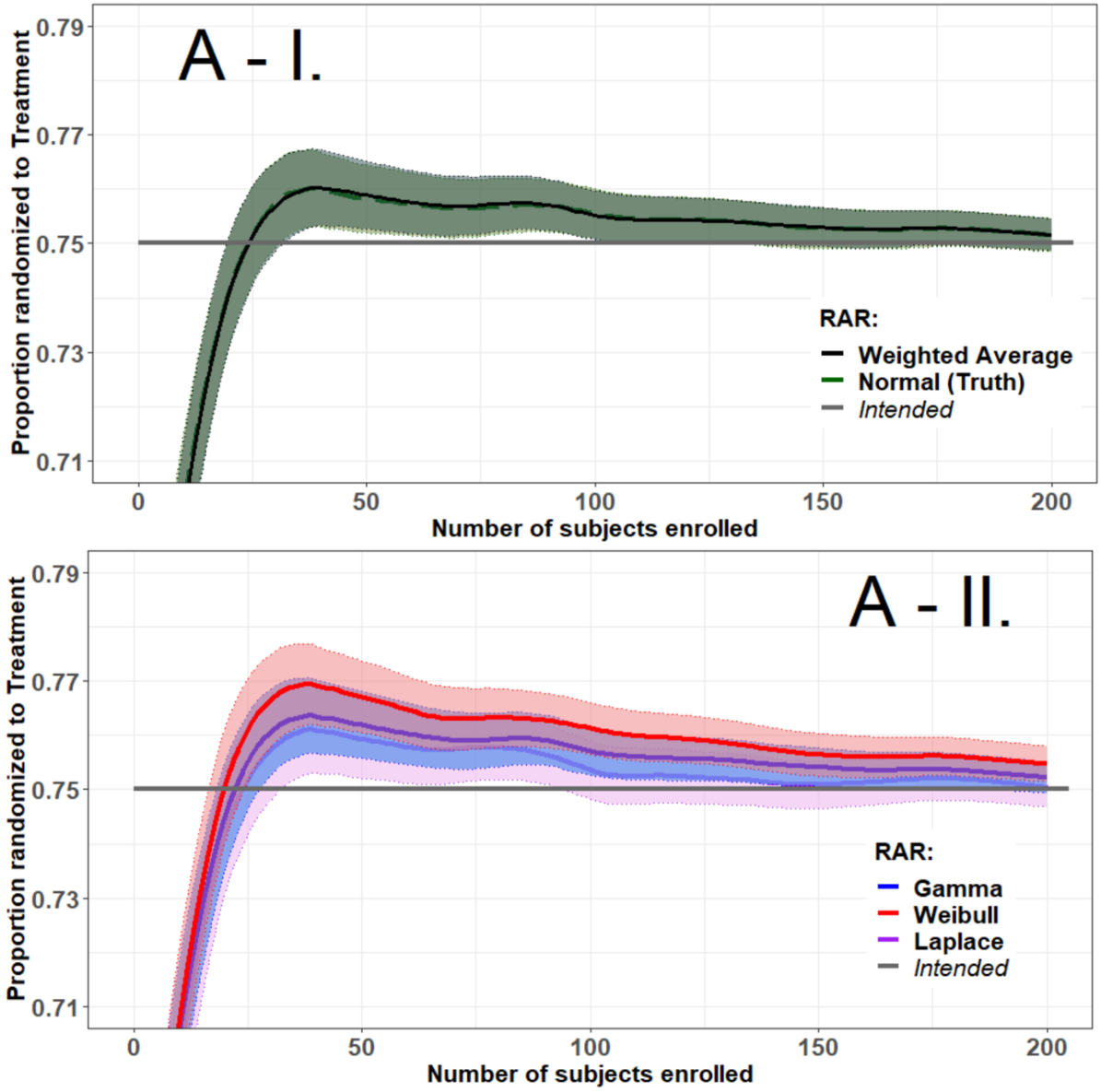
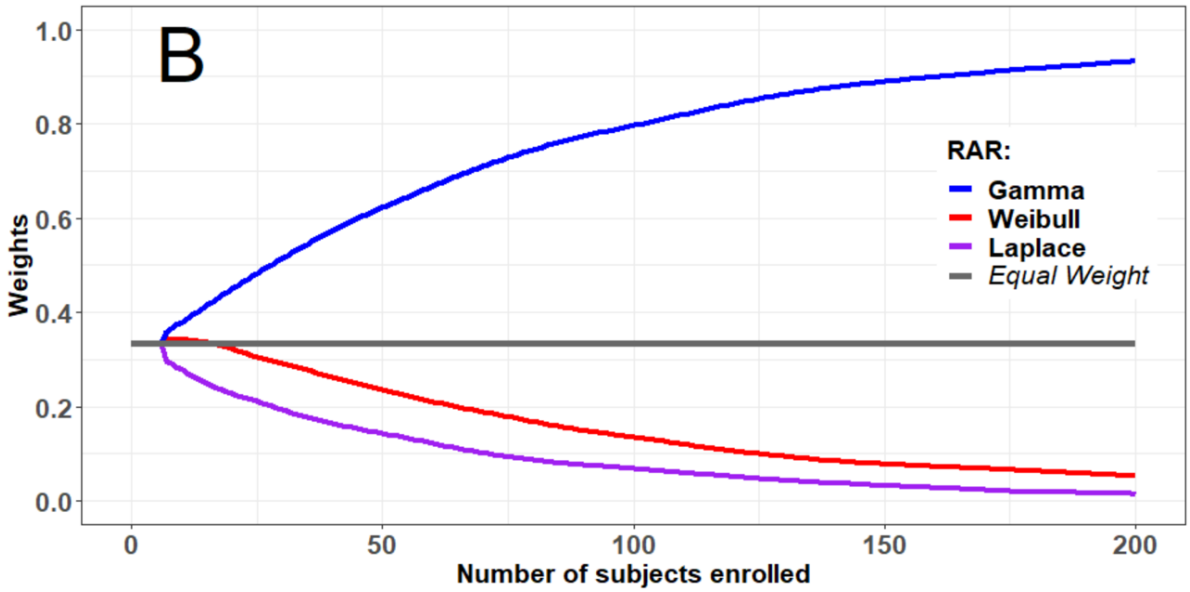


Figure 3.16, cont'd.



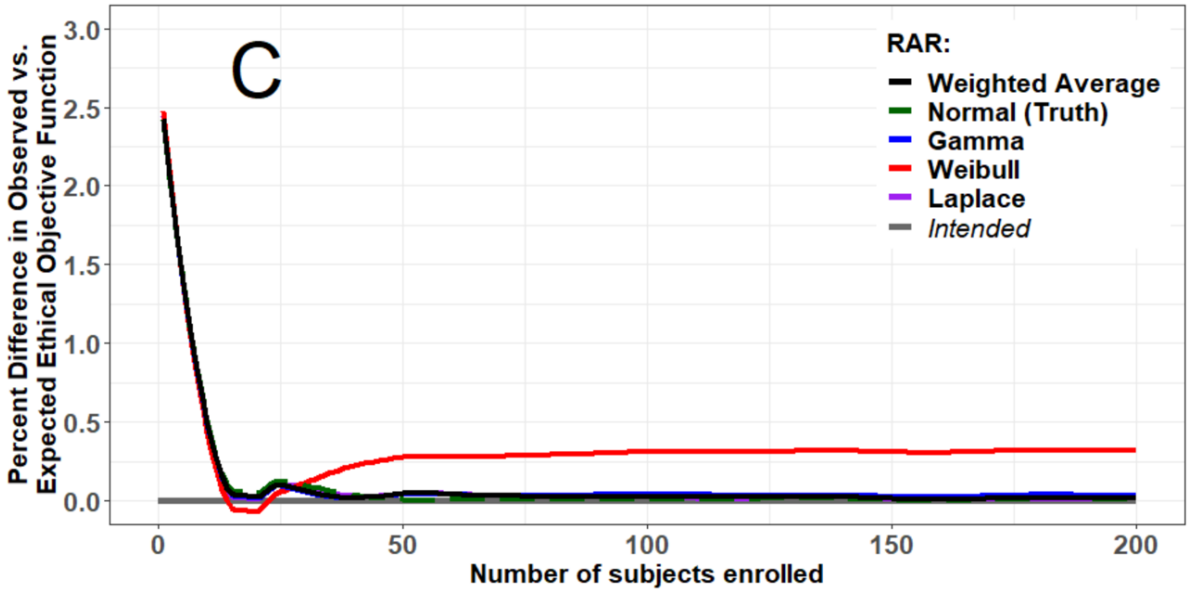


Figure 3.17: Comparison of mean WA RAR ratio against mean true response distribution (non-candidate) RAR ratio (95% CIs) (A-I), mean RAR ratios (95% CIs) for each CRD (A-II), mean of AIC-based distributional weights (B), and percent-difference between the observed and expected ethical objective function (C) over 1,000 simulated trials when treatment-specific participant responses were simulated to follow gamma distributions reflecting a large treatment difference, and CRDs were the normal, Weibull, and Laplace distributions (i.e., CRD Set 4).

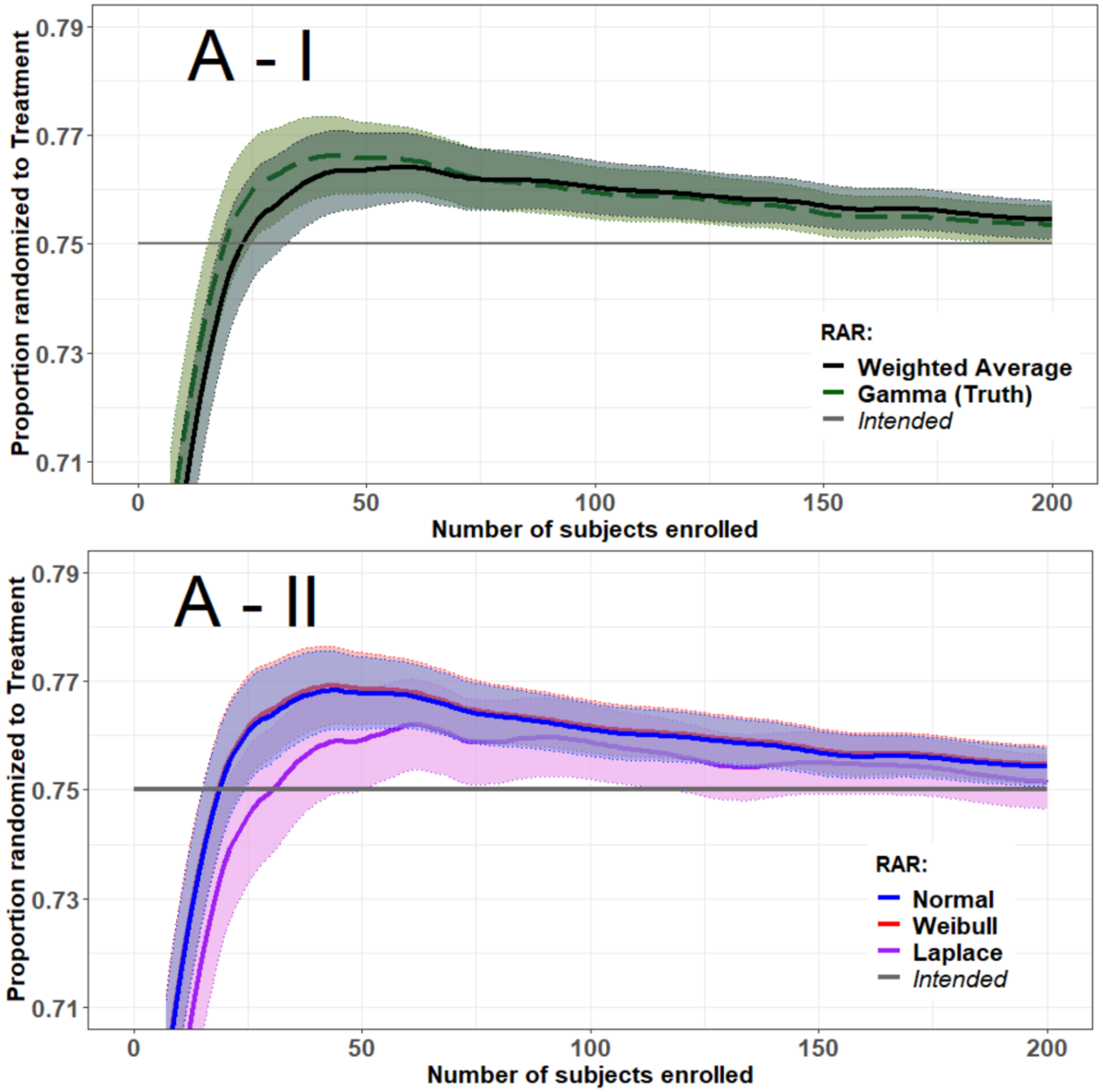
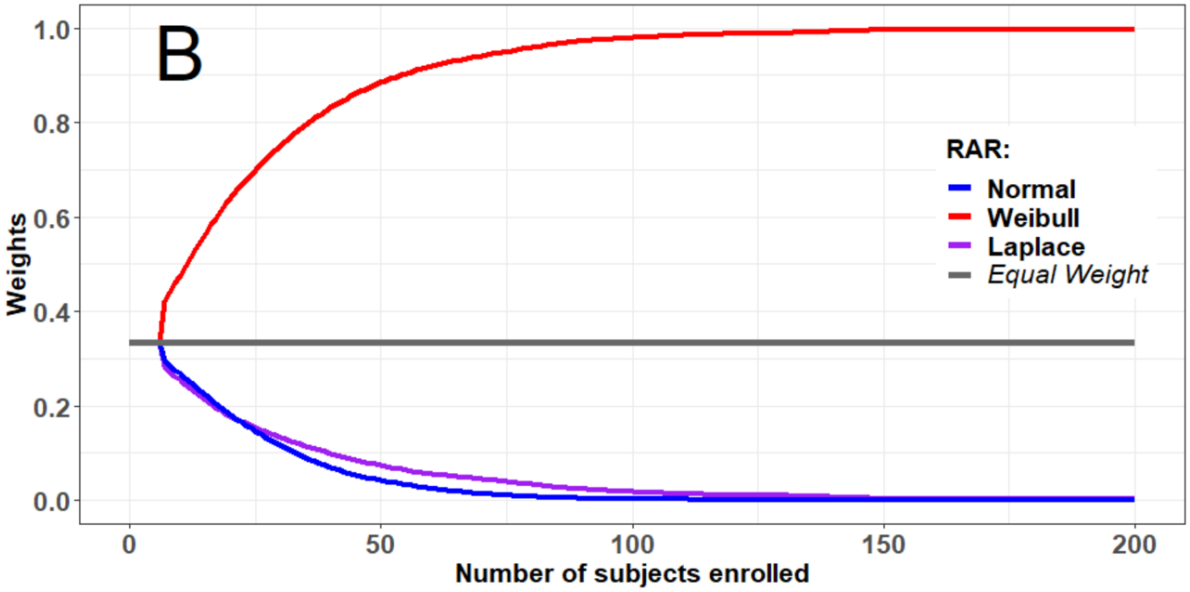


Figure 3.17, cont'd.



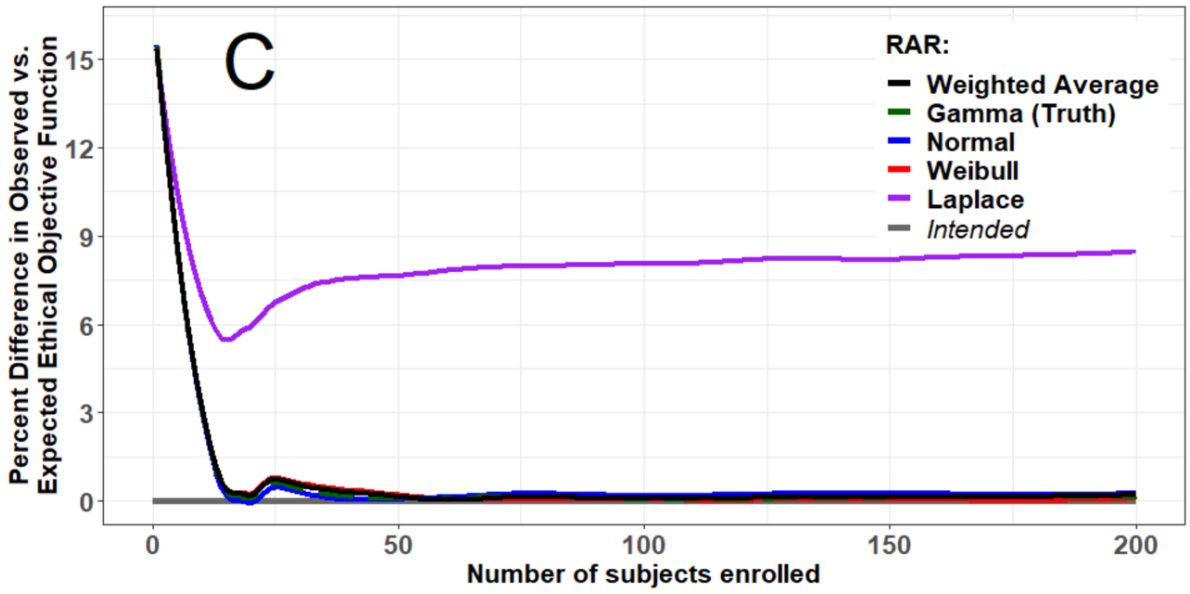


Figure 3.18: Comparison of mean WA RAR ratio against mean true response distribution (non-candidate) RAR ratio (95% CIs) (A-I), mean RAR ratios (95% CIs) for each CRD (A-II), mean of AIC-based distributional weights (B), and percent-difference between the observed and expected ethical objective function (C) over 1,000 simulated trials when treatment-specific participant responses were simulated to follow X2U distributions reflecting a large treatment difference, and CRDs were the normal, gamma, Weibull, and Laplace distributions (i.e., CRD Set 4).

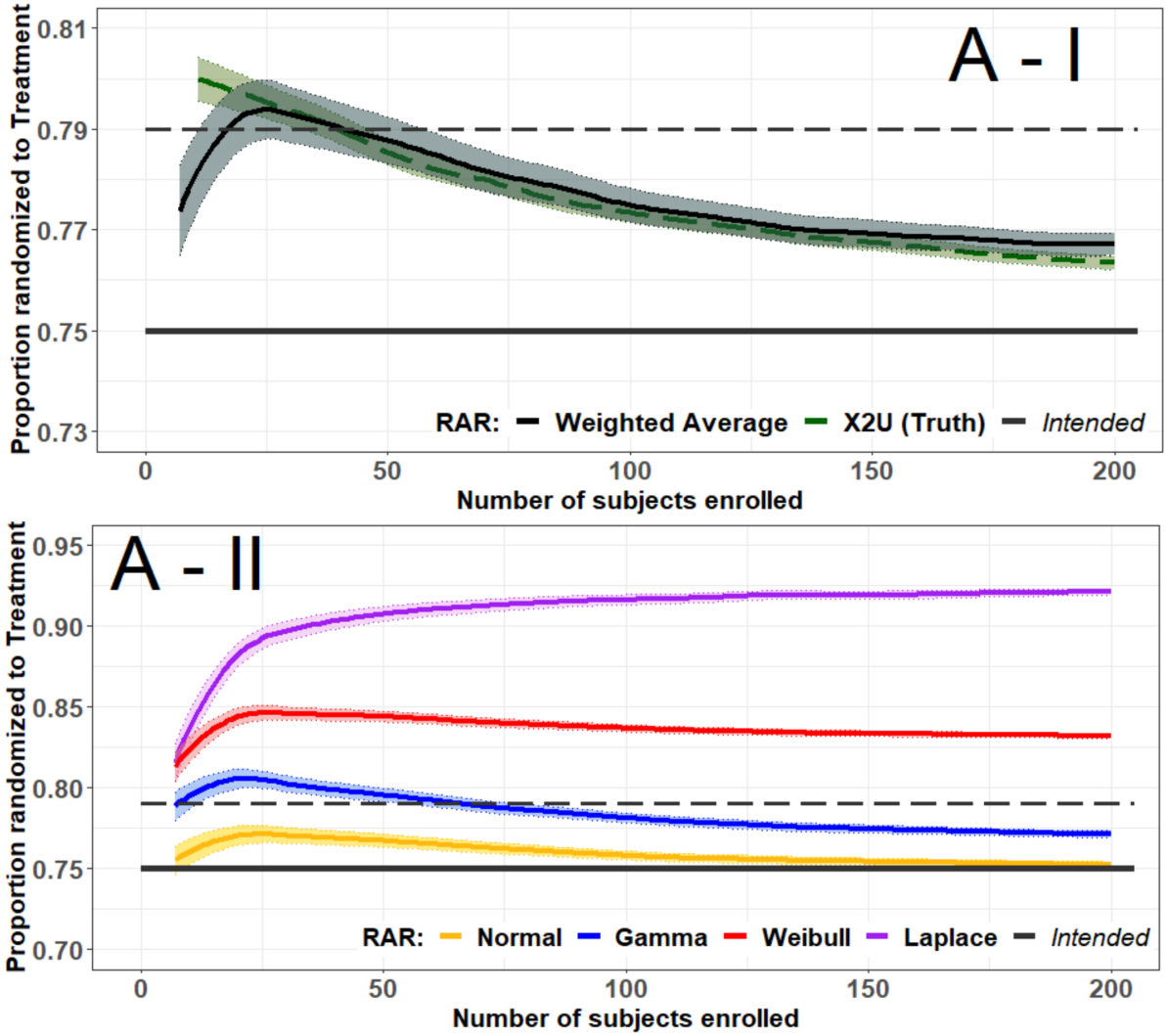
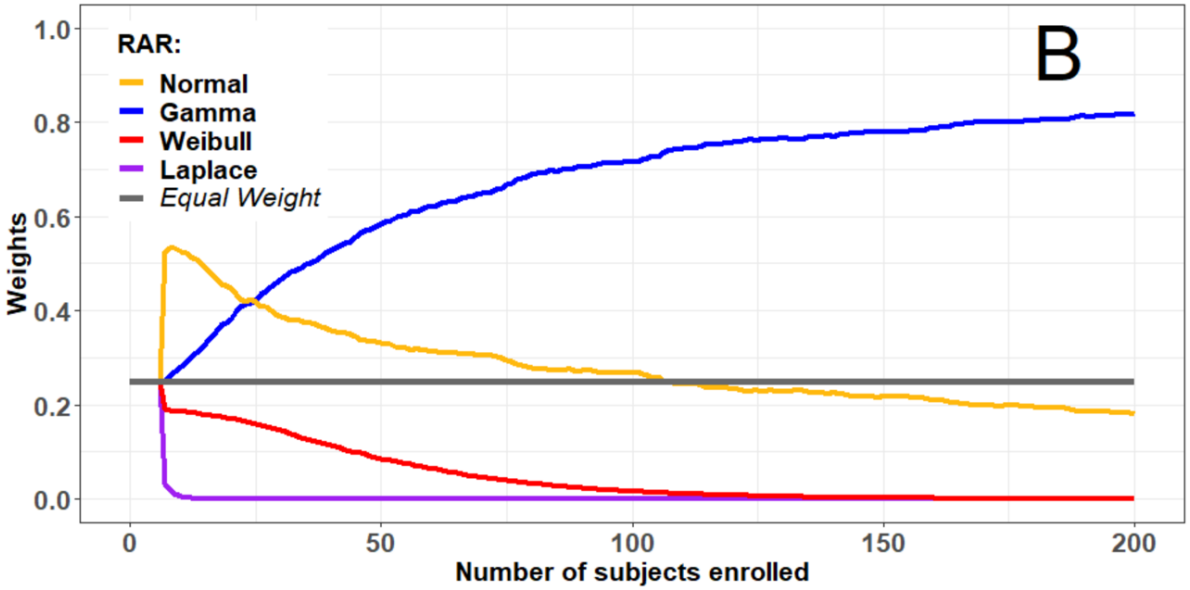
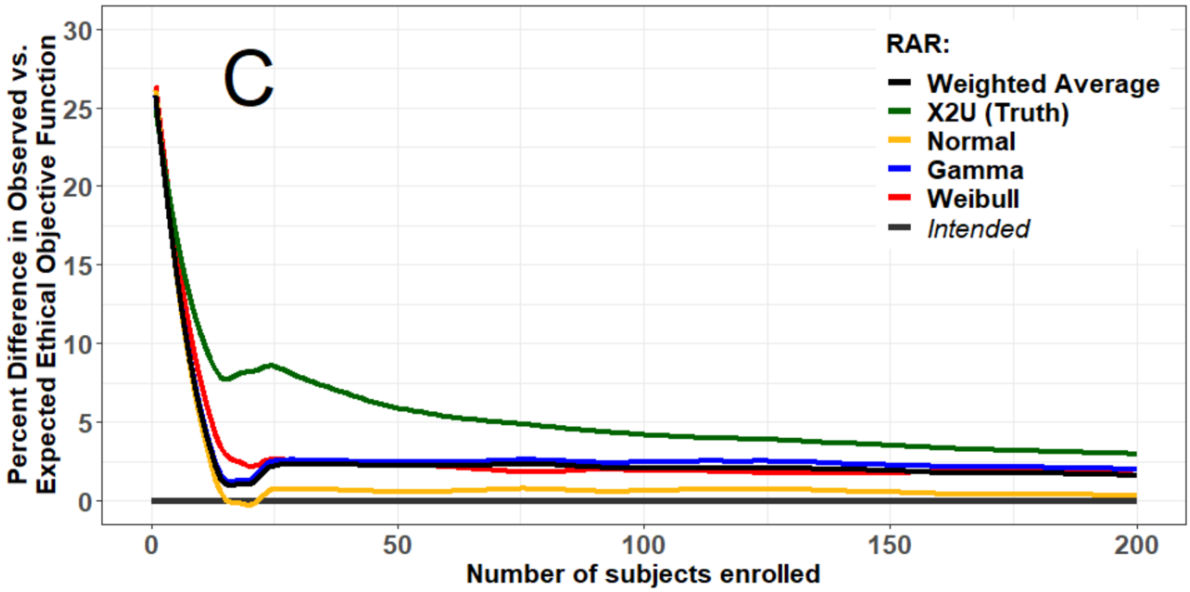


Figure 3.18, cont'd.





When treatment-specific response data were simulated to follow X2U distributions reflecting a large treatment difference, the WA RAR ratio (black solid curve, plot A-I, Figure 3.18) closely recreated X2U RAR (green dashed line, plot A-I) by averaging over AIC-weighted RAR ratios assuming the skewed gamma (blue curve, plot A-II) and Weibull (red curve, plot A-II) and symmetrical normal (gold curve, plot A-II) and Laplace (purple curve, plot A-II) CRDs. The MLEs for the Weibull and Laplace CRDs produce estimated moments dissimilar to those of the X2U-distributed response data, as evidenced by $\overline{\hat{\rho}}_W \in (0.80,\ 0.85)$ and $\overline{\hat{\rho}}_{La} \to 0.9$ in plot A-II. With this, $\overline{\hat{w}}_W$ and $\overline{\hat{w}}_{La} \to 0$, the latter doing so immediately following the lead-in period, suggesting that the Weibull and Laplace CRDs poorly-fit the X2U-distributed response data (plot B). As well, following the lead-in period until the 25^{th} trial participant was enrolled, the normal distribution best-fit the X2U response data; following the 25^{th} trial participant, however, $\overline{\hat{w}}_G \to 1$, as anticipated. Finally, $\%_{\text{O-E}}$ for each of the CRD-specific and WA RAR ratios converged to less than five percent deviation, though the X2U RAR converged slowest (plot C).

When treatment-specific response data were simulated to follow NM distributions reflecting a large treatment difference, $\overline{\hat{\rho}}_{WA}$ (black curve) successfully approximated $\overline{\hat{\rho}}_T$ (green dashed line) (plot A-I, Figure 3.19). And, as anticipated, the symmetrical normal CRD (gold curve) fit the bimodal NM response distribution better than the symmetrical Laplace (purple curve) or skewed gamma (blue curve) or Weibull (red curve) CRDs, indicated by $\overline{\hat{w}}_N \to 1$ and all other distributional weights going to zero immediately following the lead-in period (plot B). Finally, all RAR ratios adhered to the ethical objective of the trial (plot C).

3.4 Discussion

Weighted-average RAR performed well in a number of respects. First, the true response distribution was consistently identified as the best-fit CRD when included in the set of CRDs, and more poorly-fitting CRDs were given less weight, comprising less of the resulting RAR ratio. The latter was

Figure 3.19: Comparison of mean WA RAR ratio against mean true response distribution (non-candidate) RAR ratio (95% CIs) (A-I), mean RAR ratios (95% CIs) for each CRD (A-II), mean of AIC-based distributional weights (B), and percent-difference between the observed and expected ethical objective function (C) over 1,000 simulated trials when treatment-specific participant responses were simulated to follow NM distributions reflecting a large treatment difference, and CRDs were the normal, gamma, Weibull, and Laplace distributions (i.e., CRD Set 4).

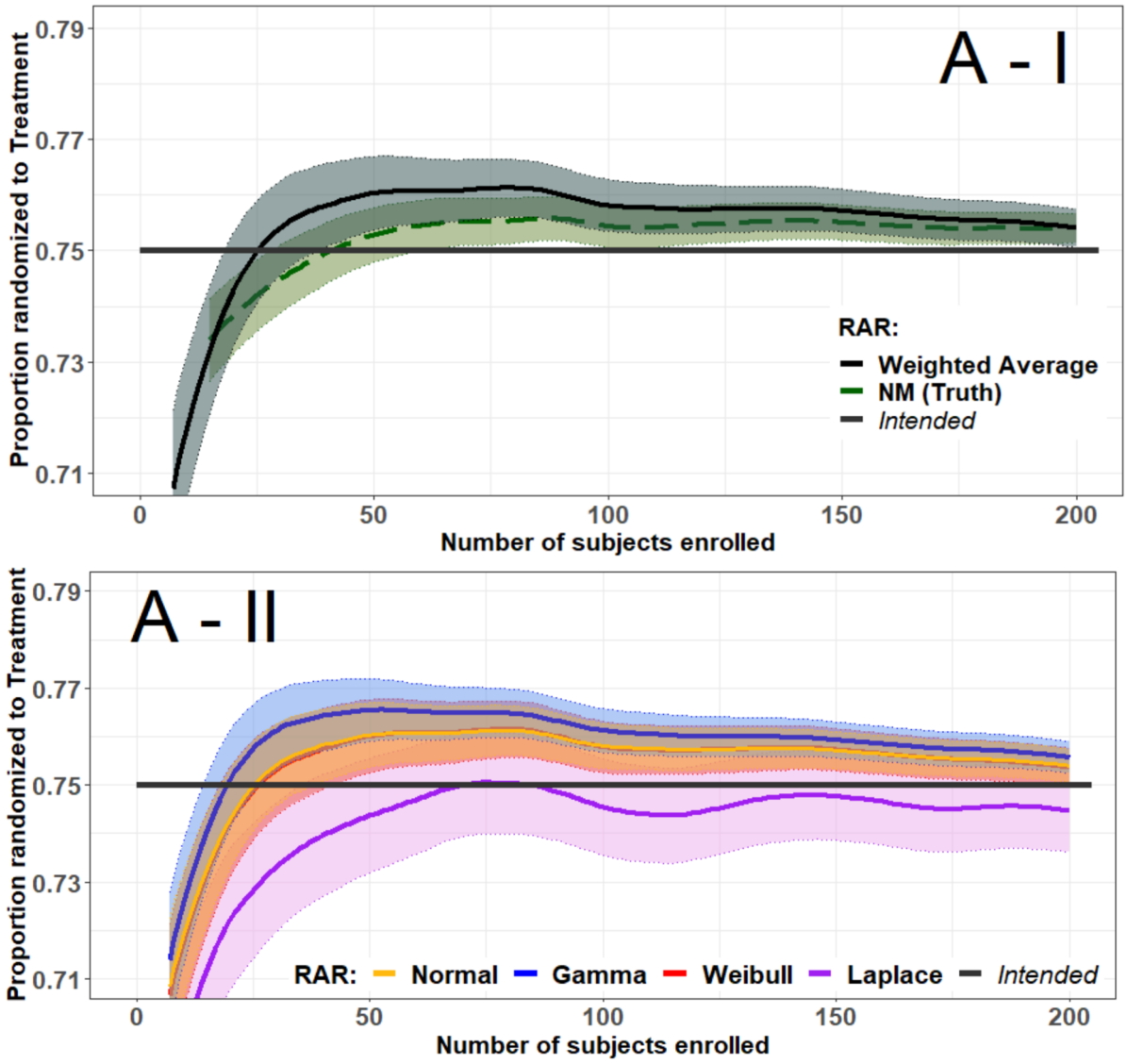
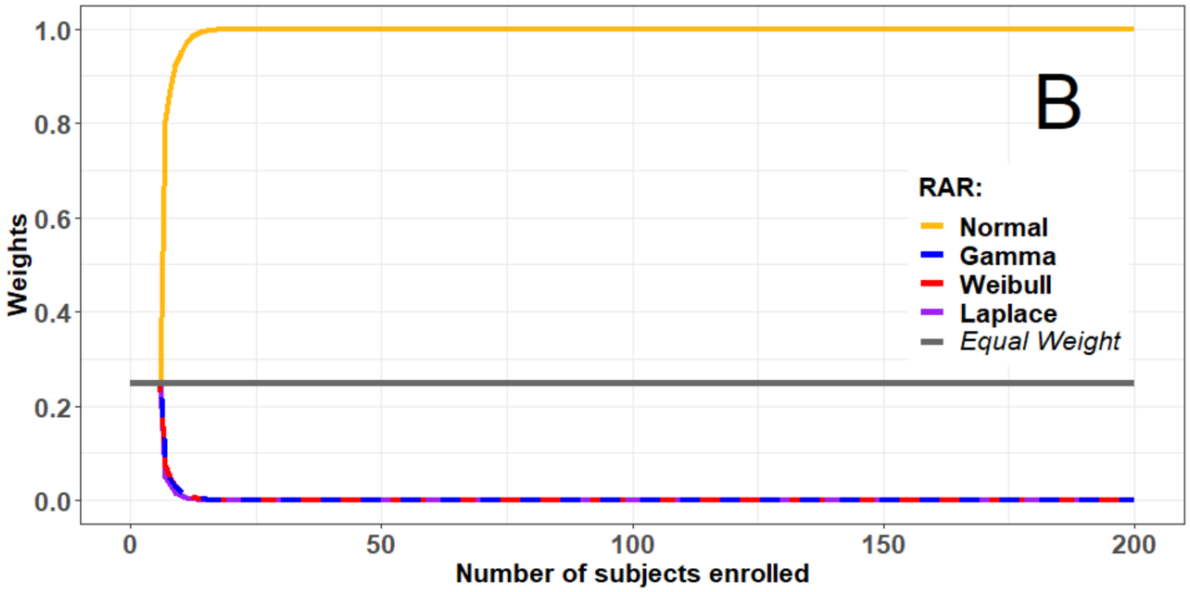
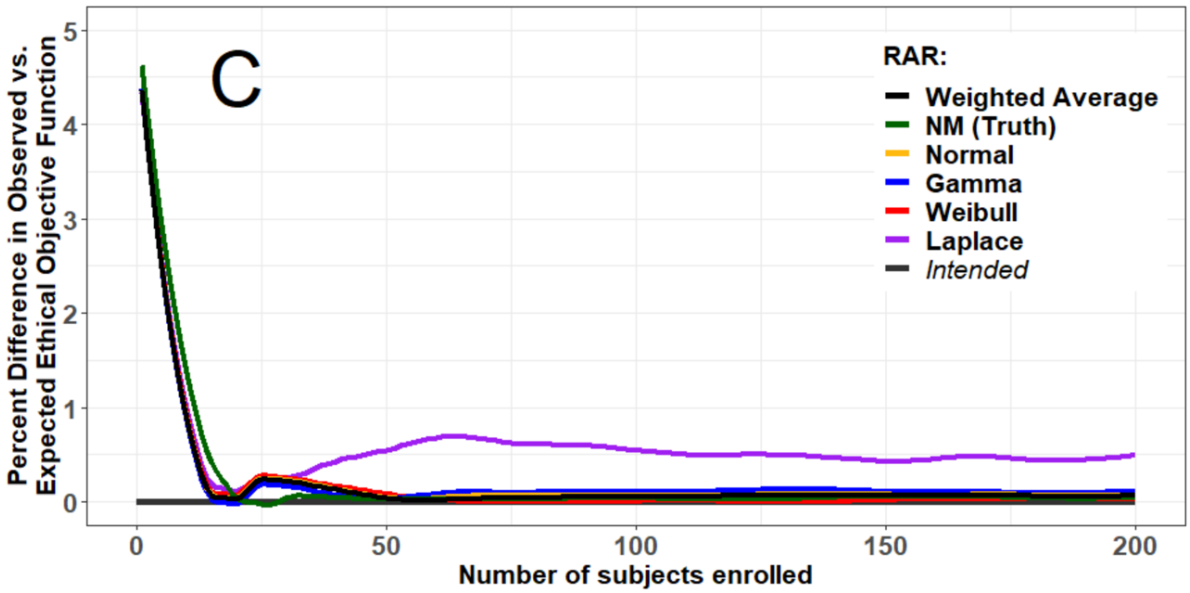


Figure 3.19, cont'd.





demonstrated by the scenario in which response data followed skewed gamma or X2U distributions and the symmetrical Laplace and logistic distributions were considered in the set of CRDs. These CRDs produced misspecified RAR ratios as demonstrated by their respective $\%_{\text{O-E}}$ being markedly larger than that of the other methods of RAR. Consequently, the distributional weights for these CRDs appropriately went to zero. Furthermore, when the truth was removed from the CRD set, WA RAR closely approximated RAR assuming the true response distribution, despite the truth being unknown in the methodology. Thus, the WA approach successfully incorporated distributional uncertainty into the RAR process to produce a RAR ratio that was both ethical and powerful.[7,42]

Ideal randomization notwithstanding, additional considerations are advised if using this framework. For example, the scale of the response data needs to be taken into account when selecting CRDs. Overall, the skewed-versus-symmetrical paradigm of CRD selection produced expected results for unimodal response distributions, meaning that symmetrical CRDs tended to produce heavier weights than skewed CRDs when the observed response data was symmetrical, and vice versa. However, the limiting distribution of a gamma random variable is normality.[24] As such, for symmetrical response data with larger means, as in the normal response data derived in these simulations, the gamma distribution outweighed the symmetrical candidates. Another consideration of this design is the MOF used to produce the distributional weights. Demonstrated by Figures 3.2 and 3.3, weighting based on the KS test statistic failed to markedly discriminate between distributional fits to the observed response data, while AIC did so effectively, showing variability RAR contingent upon the MOF used.

The X2U and NM scenarios were included because they most closely mimicked reality: the true response distribution is complex and, thereby, unlikely to be considered in the set of CRDs. When responses followed X2U distributions, though WA RAR mimicked X2U RAR, the ratios themselves were considerably larger than those produced for the other response distributions, i.e.,

normal, gamma, and NM. This may suggest that the optimal RAR procedure described in Section 1.3 may not be the most appropriate method of randomization for heavily-skewed response data. Moreover, the normal CRD was most heavily-weighted until the 25^{th} trial participant, as well as in scenarios where response data were bimodal, suggesting that RAR under the common normality assumption is innately robust against distributional misspecification, particularly in small samples, and is, therefore, the recommended RAR ratio, overall. If choosing to perform RAR using the WA method, however, selecting one or two symmetrical and skewed CRDs will result in a RAR ratio that can mimic randomization under the true response distribution, even when response data are distributionally complex and the true distribution is unknown, thereby accounting for distributional misspecification.

Chapter 4

Robust Randomization of Survival Response Data

4.1 Optimal RAR for Survival Data

Many RCTs use survival outcomes to measure treatment effectiveness and safety, though little research and application has been conducted in RAR using such outcomes.[25,51,52] Examples of survival response data include time-to-cancer progression in oncology or time-to-graft failure in transplantation.[51] Censoring is a distinguishing characteristic of RCTs with survival outcomes. This occurs when incomplete information regarding survival times exists for some trial participants. Specifically, right-censoring would occur if a participant left the trial before trial completion in such a way that the outcome was not observed and they were lost-to-follow-up. Therefore, RCTs using RAR with right-censored survival outcomes can only rely upon the survival history of patients for whom response data is available.[25] In the subset of RCTs with survival outcomes where recruitment lasts long enough to where some patient outcomes may be observed before all patients have been randomized into the trial, the optimal RAR method can be applied.[25]

Optimal RAR designs for survival outcomes have been developed in the context of interim analyses, multi-arm, and group-sequential RCTs.[51,52,54] These designs are based on the Zhang and Rosenberger (ZR) RAR design as it is convention given its simplicity and its ability to fit neatly into the template proposed by Hu and Rosenberger.[25,53] In this framework, optimality,

variability, and power are considered essential components of the valid RAR process.[25,53]

In the present work, the ZR RAR design with the importance criterion of minimizing the total expected hazard of the trial was employed, versus ZR RAR designs targeting alternative importance criteria. In this framework, survival times were assumed to be parametrically-distributed, following either the exponential and Weibull distributions, where designs developed under the assumption of the Weibull distribution were also applicable to other log-location-scale families.[25] Derivations of the exponential and Weibull ZR RAR ratios are discussed in the following sections.

4.1.1 Optimal ZR RAR for Survival Times Following the Exponential Distribution

Survival times can be approximated by the exponential distribution in many RCTs, leading to straightforward, closed-form theoretical results.[25] Suppose survival time Y follows an exponential distribution with parameter θ, i.e., $Y \sim Exp(\theta)$.[i] It, then, has the density and survival function respectively given by:

$$f_Y(y) = exp(-y/\theta)/\theta, \text{ and,} \tag{4.1}$$

$$S(y) = exp(-y/\theta), \tag{4.2}$$

for $\theta > 0$ and $y > 0$, else $f_Y(y) = 0$. Under this parameterization, Y has the constant hazard rate of $1/\theta$, where θ is the mean survival time.

To derive the optimal RAR ratio, censoring must be incorporated in the modeling of the survival times. Therefore, suppose that patients have survival times following the exponential distribution as previously described, but which are also subject to an independent right-censoring scheme. Let $(y_i, \epsilon_i),\ i = 1, \cdots, n$, be a random sample from such a distribution, where, for the i^{th} patient, $\epsilon_i = 1$

[i] The y versus t notation representing survival times introduced in Chapter 2 is maintained throughout the present chapter, as well.

when y_i is the survival time and $\epsilon_i = 0$ when y_i is the censor time. Then, the likelihood of θ is:

$$\begin{aligned}\ell(\theta|\boldsymbol{y}, \boldsymbol{\epsilon}) &= \prod_{i=1}^{n} \left\{ \frac{1}{\theta} exp\Big(- \frac{y_i}{\theta} \Big) \right\}^{\epsilon_i} exp\Big(- \frac{y_i}{\theta} \Big)^{1-\epsilon_i} \\ &= \prod_{i=1}^{n} \theta^{-\epsilon_i} exp(-y_i/\theta), \end{aligned} \tag{4.3}$$

and the loglikelihood is given by:

$$log\{\ell(\theta|\boldsymbol{y}, \boldsymbol{\epsilon})\} = -log(\theta) r - \frac{1}{\theta} y, \tag{4.4}$$

where $r = \sum_{i=1}^{n} \epsilon_i$ is the cumulative number of events and $y = \sum_{i=1}^{n} y_i$ is the sum of the observed survival times, such that

$$\frac{d}{d\theta} log\{\ell(\theta|\boldsymbol{y}, \boldsymbol{\epsilon})\} = -\frac{r}{\theta} + \frac{y}{\theta^2}. \tag{4.5}$$

Solving $\frac{d}{d\theta} log\{\ell(\theta|\boldsymbol{y}, \boldsymbol{\epsilon})\} = 0$ gives the MLE of θ as $\hat{\theta} = y/r$. As well, the Fisher information for θ, i.e., $I(\theta)$, is given by:

$$\begin{aligned} I(\theta) &= -E\left[\frac{d^2}{d\theta^2} log\{\ell(\theta|\boldsymbol{y}, \boldsymbol{\epsilon})\} \right] \\ &= -\frac{E(r)}{\theta^2} + \frac{2E(y)}{\theta^3}. \end{aligned}$$

Since the sum of n exponential random variables having parameter θ is a gamma random variable with shape parameter n and scale parameter θ, in the absence of censoring, $y \sim \text{Gamma}(n, 1/\theta)$. When there is censoring, however, $y \sim \text{Gamma}(r, 1/\theta)$ and $E(y)$ can be approximated by $\theta E(r)$. Thus,

$$I(\theta) = E(r)/\theta^2, \tag{4.6}$$

and the approximate variance of $\hat{\theta}$ is:

$$Var(\hat{\theta}) = 1/I(\theta) = \theta^2/E(r). \tag{4.7}$$

Response-adaptive randomization is based upon the comparison of the parameters θ_T and θ_C, corresponding to mean survival times for treatment groups T and C, respectively. To derive the optimal RAR ratio, consider two independent samples $\{(y_{gi}, \epsilon_{gi});\ g \in \{T, C\};\ i = 1, \cdots, n\}$, and assume that $\xi_g = E(\epsilon_{gi})$ is the same for all $i = 1, \cdots, n_g$. Then $E(r_g) = n_g E(\epsilon_{g1}) = n_g \xi_g$. With this, the optimization process described in Appendix A.5 can be invoked to solve the following optimization problem to minimize the total expected hazard:

$$\underset{n_T/n_C}{min} \left\{ \frac{n_T}{\theta_T} + \frac{n_C}{\theta_C} \right\} \tag{4.8}$$

subject to:

$$\frac{\theta_T^2}{n_T \xi_T} + \frac{\theta_C^2}{n_C \xi_C} = V, \tag{4.9}$$

to which the solution is:

$$\frac{n_T}{n_C} = \frac{\theta_T^{3/2} \xi_C^{1/2}}{\theta_C^{3/2} \xi_T^{1/2}}, \tag{4.10}$$

resulting in the ZR allocation rule given by:

$$\rho_E = \frac{\sqrt{\theta_T^3 \xi_C}}{\sqrt{\theta_T^3 \xi_C} + \sqrt{\theta_C^3 \xi_T}} \tag{4.11}$$

for survival outcome data with exponentially-distributed survival times. As described in Chapter 1, the parameters used to construct the RAR ratio are supplanted with its corresponding MLE estimated from the response data obtained from patients who have already completed the trial. With this, the probability of enrolling patient $i + 1$ into the treatment group using optimal RAR is given by:

$$\hat{\rho}_{T;i+1} = \frac{\sqrt{\hat{\theta}_{Ti}^3 \hat{\xi}_{Ci}}}{\sqrt{\hat{\theta}_{Ti}^3 \hat{\xi}_{Ci}} + \sqrt{\hat{\theta}_{Ci}^3 \hat{\xi}_{Ti}}}. \tag{4.12}$$

4.1.1.1 Exponential Censoring Scheme

Zhang and Rosenberger employed a censoring scheme first introduced by Rosenberger and Seshaiyer and later expanded by Liu and Coad.[15,25,54] Given a fixed recruitment period R and trial duration D, patient arrival times were assumed to follow an independent uniform distribution on $[0, R]$, i.e., $U[0, R]$, while, independently, patients were subject to a censor time X that followed a $U[0, D]$ distribution. For a patient randomized to group g ($g \in \{T, C\}$) having a survival time Y_g following an exponential distribution with parameter θ_g, let the observed survival outcome be $Q_g = min(Y_g, X, D - R)$. Further, define $W_g = 1$ if $Q_g = Y_g$, otherwise $W_g = 0$. Then, the probability of an event is:

$$\begin{aligned}\xi_g &= E(W_g) \\ &= P(W_g = 1) \\ &= 1 - \frac{\theta_g}{D} + exp\bigg(- \frac{D}{\theta_g}\bigg)\frac{\theta_g}{DR}\bigg\{exp\bigg(\frac{R}{\theta_g}\bigg)(2\theta_g - R) - 2\theta_g\bigg\},\end{aligned} \tag{4.13}$$

the complement of which provides the probability of censoring, denoted η_g, i.e., $\eta_g = 1 - \xi_g$. Details pertaining to the derivation of ξ_g can be found in the Supplementary Materials of Liu and Coad.[54]

With ξ_g, the exponential ZR RAR ratio of (4.11) can be obtained. Then, estimation of ξ_g for the ZR RAR estimator of (4.12) was obtained using the MLE $\hat{\theta}_{gi}$ and the arrival time for patient i enrolled into group g, denoted A_{gi}. With this, the probability of an event for patient i enrolled into group g is given by:

$$\hat{\xi}_{gi} = 1 - \frac{\hat{\theta}_{gi}}{D} + exp\bigg(- \frac{D}{\hat{\theta}_{gi}}\bigg)\frac{\hat{\theta}_{gi}}{DA_{gi}}\bigg\{exp\bigg(\frac{A_{gi}}{\hat{\theta}_{gi}}\bigg)(2\hat{\theta}_{gi} - A_{gi}) - 2\hat{\theta}_{gi}\bigg\}. \tag{4.14}$$

4.1.2 Optimal ZR RAR for Survival Times Following the Weibull Distribution

Exponential survival times assume a constant hazard, which may be unrealistic in practice. [31,32,51,54] Other, more flexible, distributions that can account for changing hazards over time may be used for modeling survival times more accurately. One such distribution is the Weibull distribution, a log-location-scale family distribution.

Suppose survival time Y follows a Weibull distribution with shape parameter α and scale parameter β having the density:

$$f_Y(y) = \frac{\alpha}{\beta} y^{\alpha-1} exp\bigg(- \frac{y^\alpha}{\beta}\bigg), \tag{4.15}$$

for $\alpha, \beta > 0$ and $y > 0$, else $f_Y(y) = 0$. Then $log(Y)$ has an extreme value distribution (EVD) with shape parameter $m = log(\beta)$ and scale parameter $k = 1/\alpha$. In this case, where the observed group-specific survival outcome is $Q_g = min(Y_g, X, D - R)$, let $S_g = log(Q_g)$. Consider the independent random sample $(s_i, \epsilon_i),\ i = 1, \cdots, n$, where s_i represents the survival time for patient i when $\epsilon_i = 1$ and the censor time when $\epsilon_i = 0$. Then, the likelihood function of m and k is:

$$\ell(m, k|\boldsymbol{s}, \boldsymbol{\epsilon}) = \prod_{i=1}^{n} \bigg[\frac{1}{k} exp\bigg(\frac{s_i - m}{k}\bigg) exp\bigg\{ - exp\bigg(\frac{s_i - m}{k}\bigg)\bigg\}\bigg]^{\epsilon_i} exp\bigg\{ - exp\bigg(\frac{s_i - m}{k}\bigg)\bigg\}^{1-\epsilon_i}, \tag{4.16}$$

and the loglikelihood function is:

$$log\{\ell(m, k|\boldsymbol{s}, \boldsymbol{\epsilon})\} = -rlog(k) + \sum_{i=1}^{n} \{\epsilon_i z_i - exp(z_i)\}, \tag{4.17}$$

where $r = \sum_{i=1}^{n} \epsilon_i$, as defined previously, and $z_i = (s_i - m)/k$, following the standard EVD, i.e., $Z \sim EVD(0, 1)$, with density $f_Z(z) = e^z exp(-e^z)$. The MLEs can be obtained numerically by

solving:

$$\frac{\partial}{\partial m} log\{\ell(m,k|\boldsymbol{s},\boldsymbol{\epsilon})\} = -\frac{1}{k}\sum_{i=1}^{n}\{\epsilon_i - exp(z_i)\} = 0, \text{ and,} \tag{4.18}$$

$$\frac{\partial}{\partial k} log\{\ell(m,k|\boldsymbol{s},\boldsymbol{\epsilon})\} = -\frac{r}{k} - \frac{1}{k}\sum_{i=1}^{n}\{\epsilon_i - exp(z_i)\}z_i = 0. \tag{4.19}$$

The approximate variance of the MLE for m, $\hat{m}$, which corresponds to the Weibull scale parameter β, is $Var(\hat{m}) = k^2G/n$, where

$$G = \frac{\xi_i + E(z_i^2 exp(z_i))}{\xi_i^2 + \xi_i E(z_i^2 exp(z_i)) - E(z_i exp(z_i))^2}. \tag{4.20}$$

Estimation of the probability of an event, ξ, is far more intricate in the case of survival times following the Weibull versus the exponential distribution.[25,54] Therefore, Zhang and Rosenberger and others suggest using the simple nonparametric estimator of the group-specific mean number of events for construction of the RAR ratio.

Allow the average hazard to be expressed as the reciprocal of the mean survival time as in the exponential case, i.e., $1/E(S)$, where $E(S) = e^m\Gamma(1+k)$, where $\Gamma(x)$ represents the gamma function such that $\Gamma(x) = (x-1)!$ when x is an integer. Then, the ZR RAR ratio can be constructed based upon the comparison of treatment groups with respect to their average hazards. Consider two independent censored random samples from the previously-described survival distribution, $\{(s_{gi}, \epsilon_{gi});\ g \in \{T,C\};\ i = 1,\cdots,n_g\}$. Then the optimization problem to minimize the total expected hazard is given by:

$$\underset{n_T/n_C}{min}\left\{\frac{n_T}{exp(m_T)\Gamma(1+k_T)} + \frac{n_C}{exp(m_C)\Gamma(1+k_C)}\right\} \tag{4.21}$$

subject to:

$$\frac{k_T^2 G_T}{n_T} + \frac{k_C^2 G_C}{n_C} = V, \tag{4.22}$$

resulting in the ZR RAR ratio given by:

$$\rho_W = \frac{k_T\sqrt{G_T}\left[exp(-m_C)/\Gamma(1+k_C)\right]^{1/2}}{k_T\sqrt{G_T}\left[exp(-m_C)/\Gamma(1+k_C)\right]^{1/2} + k_C\sqrt{G_C}\left[exp(-m_T)/\Gamma(1+k_T)\right]^{1/2}} \tag{4.23}$$

for survival response data having survival times that follow the Weibull distribution. With this, the probability of response-adaptively allocating patient $i+1$ to the treatment group based on survival response obtained from patients one to i is:

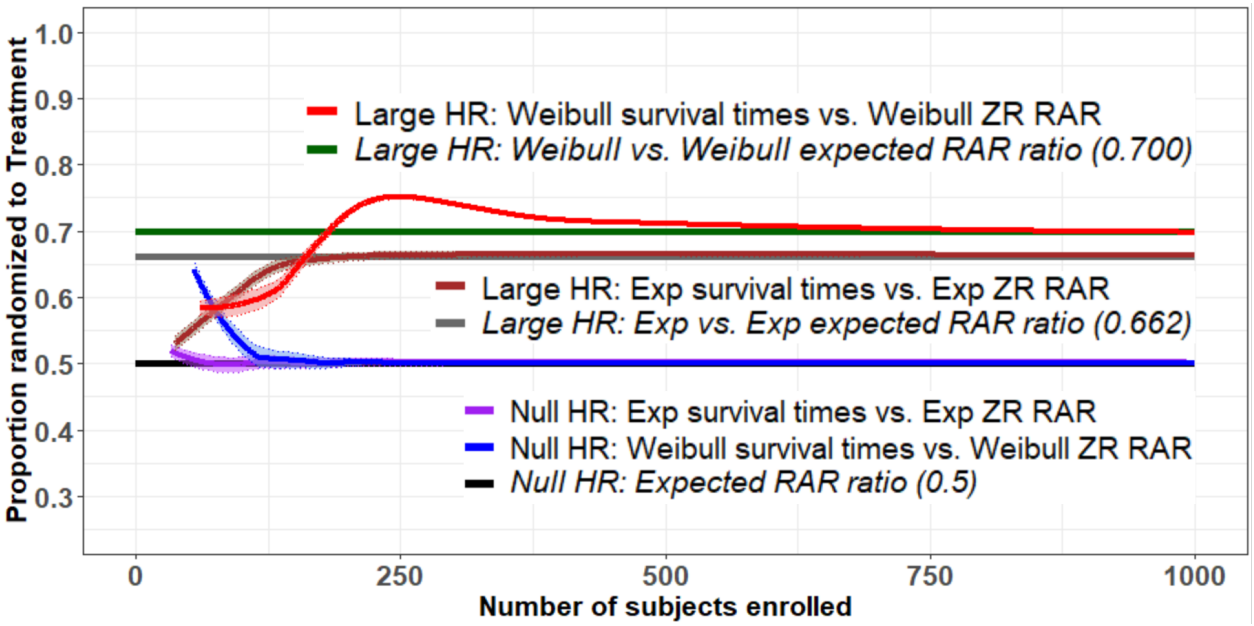
$$\hat{\rho}_{T;i+1} = \frac{\hat{k}_{Ti}\sqrt{\hat{G}_{Ti}}\left[exp(-\hat{m}_{Ci})/\Gamma(1+\hat{k}_{Ci})\right]^{1/2}}{\hat{k}_{Ti}\sqrt{\hat{G}_{Ti}}\left[exp(-\hat{m}_{Ci})/\Gamma(1+\hat{k}_{Ci})\right]^{1/2} + \hat{k}_{Ci}\sqrt{\hat{G}_{Ci}}\left[exp(-\hat{m}_{Ti})/\Gamma(1+\hat{k}_{Ti})\right]^{1/2}}. \tag{4.24}$$

4.2 Motivation: Limitations of the Conventional ZR Approach

When RAR is performed using the ZR RAR ratio provided in (4.12) and observed survival times do, in fact, follow the exponential distribution, then the RAR ratio is said to be correctly-specified. The same is true of observed survival times following the Weibull distribution undergoing RAR using the ZR RAR ratio provided in (4.24).

Correctly-specified ZR RAR is illustrated in Figure 4.1. Here, 1,000 patients are response-adaptively randomized into a simulated RCT where patients survival times followed either the exponential or Weibull distribution, where 20% of patient responses were censored. Response-adaptive randomization was performed using the corresponding ZR RAR provided in either (4.12) or (4.24). Each simulated RCT was replicated 1,000 times, and the resulting 1,000 ZR RAR ratios for each of the 1,000 patients were averaged to create a single mean observed ZR RAR for that patient. Correctly-specified ZR RAR was evaluated for two effects sizes based on the hazard ratio (HR) between treatment groups: the null effect size ($HR = 1.0$) and the large effect size ($HR = 1.5$). For a null effect size, the expected RAR ratio was 0.5; and for a large effect size, the expected RAR ratios were 0.662 and 0.700 when treatment-specific survival times were simulated

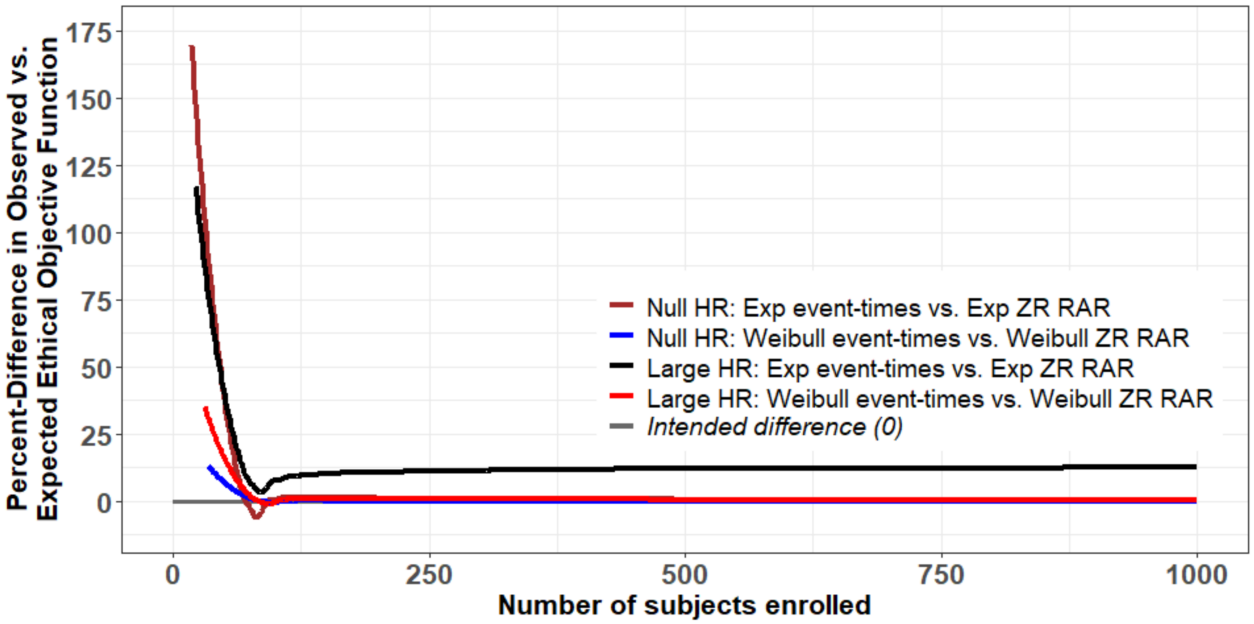
Figure 4.1: Mean of ZR RAR ratios with 95% CI with correctly-specified survival times for null and large treatment differences.



to follow the exponential and Weibull distributions, respectively. Specific details pertaining to simulation construction and data generation, including the parameter values used to target each effect size, are provided later in this chapter. As evidenced by Figure 4.1, when survival times are correctly-specified, regardless of effect size or response distribution, RAR performed using the correct ZR RAR ratio produced randomization results matching those intended by the trial design.

Figure 4.2 indicates how well the ZR RAR ratio maintained the ethical objective of the trial based on the simulated survival outcome response data. Per the ZR construct, the ethical objective of interest is the minimization of the total expected hazard of the trial. Details pertaining to the estimation of the observed and expected objective functions are discussed later in this chapter, but percent-differences closer to zero indicate that randomization is adhering to the ethical design of the trial, as intended by the trial design. When correctly-specified, ZR RAR appeared to adequately minimize the total expected hazard of the trial in all cases with percent-differences near zero, though some deviation (approximately 10%) was observed when survival times followed the exponential

Figure 4.2: Percent-difference between the observed and expected ethical objective function of correctly-specified survival times for null and large treatment differences using ZR RAR.



distribution and a large treatment effect was observed (Figure 4.2).

Although ZR remains the convention for RAR designs for RCTs with survival outcomes, two flaws to the ZR RAR design were identified in the present work. First, the ZR RAR approach relies upon the parametric specification of survival times. Clinical data rarely follow a prescribed distribution.[55] Thus, imposing a parametric definition in practice may induce misspecification error, leading to bias in the RAR ratio, exemplified when survival times are assumed to follow the exponential distribution yet observed patient survival times follow another distribution, e.g., Weibull, or are nonparametric. Second, the ZR RAR allocation rules for parametrically-distributed survival outcomes given in 4.12 and 4.24 can be expressed using the group-specific hazard, defined as the reciprocal of the distribution-specific mean, and the group-specific SDs of the distribution-specific scale parameter. Selecting the SD of the distribution-specific scale parameter as a proxy measure of the variability of the hazard may result in randomization results drastically different

from those intended irrespective of whether or not survival times are correctly specified.

Figure 4.3: Mean of ZR RAR ratios with 95% CI with misspecified survival times for null and large treatment differences.

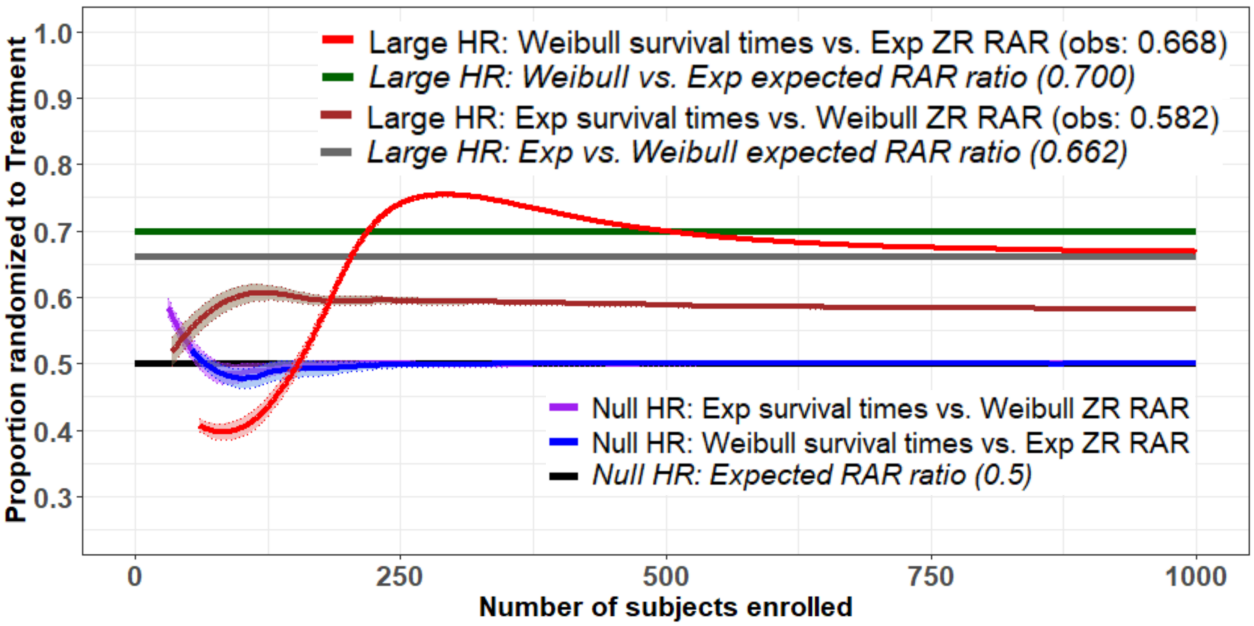


Figure 4.4: Percent-difference between the observed and expected ethical objective function of misspecified survival times for null and large treatment differences using ZR RAR.

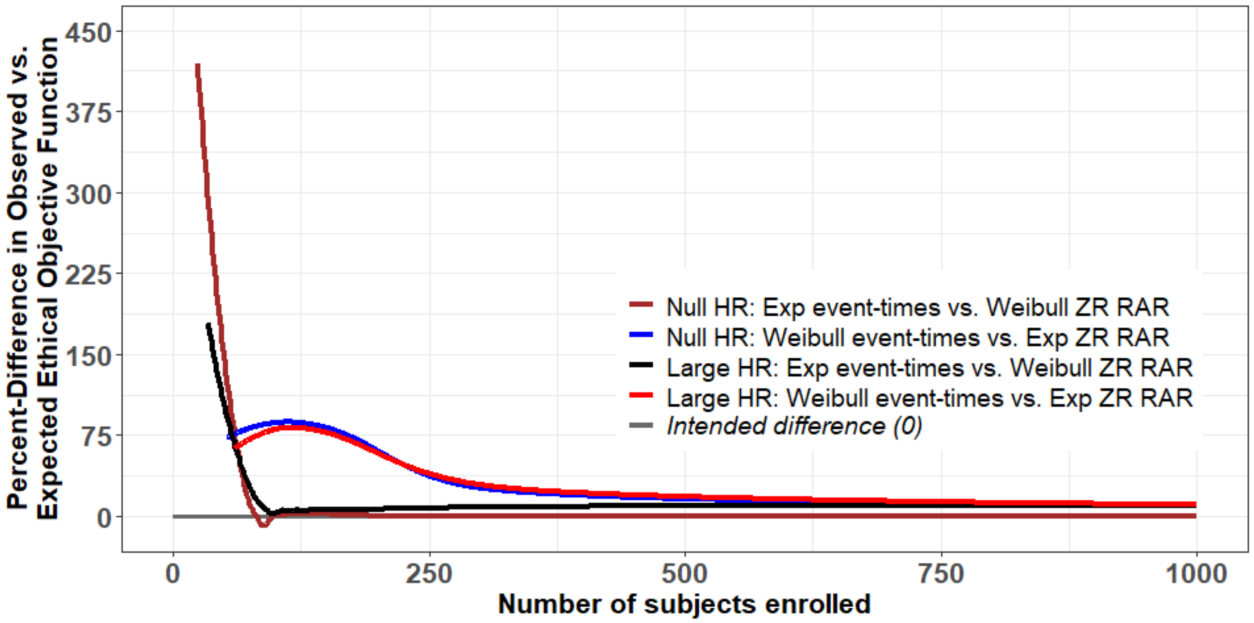


Figure 4.3 demonstrates the effects of misspecification on the ZR RAR ratio. Simulated trials were designed as described previously except misspecification was induced. That is, patient survival times simulated to follow the exponential distribution were randomized using the ZR RAR for survival times assumed to follow the Weibull distribution provided in (4.24). Likewise, survival times simulated to follow the Weibull distribution were randomized assuming an exponential distribution as in (4.12). When no treatment difference existed between groups, randomization remained equal. However, in the presence of a large treatment difference, misspecified ZR RAR ratios for survival times following both the exponential and Weibull distributions under-randomized patients to the treatment group, converging at 0.582 (vs. 0.662) and 0.668 (vs. 0.700), respectively. Furthermore, the misspecified ZR RAR ratio did not adhere to the ethical objective of the trial (Figure 4.4) as well as when it was correctly-specified (Figure 4.2), as evidenced by the approximately 10% difference in the observed versus expected objective function for nearly all scenarios when misspecified.

Misspecification of the ZR RAR ratio may lead to randomization characteristics different from those intended by the trial design. Thus, the two potential drivers of misspecification of the ZR RAR described previously motivated the development of a RAR ratio that (i) does not rely upon the parametric definition of survival times and (ii) obtains measures of the hazards and the variability thereof directly, as opposed to using proxy measures.

4.3 Optimal RAR Based on the Cumulative Hazard

The aim of the present chapter was to construct a RAR ratio for survival data based on the cumulative hazard such that survival times need not adhere to any specific parametric assumption and the measure of the variance of the hazard can be estimated directly. This cumulative hazard-based RAR ratio, denoted H-RAR, is constructed using the group-specific cumulative hazard estimated at time ω, denoted $H_g(\omega)$, and the variance thereof, denoted $Var(H_g(\omega))$. With this, adhering to the optimal RAR framework constructed by ZR as explained in Appendix A.5, the H-RAR method

of optimal randomization sought to minimize the total cumulative hazard of the trial at a given time y via the objective function:

$$\underset{n_T/n_C}{min}\Big\{n_T\ H_T(\omega)\ +\ n_C\ H_C(\omega)\Big\}, \tag{4.25}$$

subject to the variance constraint imposed by:

$$\frac{Var(H_T(\omega))}{n_T} + \frac{Var(H_C(\omega))}{n_C} = V. \tag{4.26}$$

Taking $SD(H_g(\omega)) = Var(H_g(\omega))^{1/2}$, the following allocation rule based on the cumulative hazard is produced:

$$\rho_H = \frac{SD(H_C(\omega))\ H_T(\omega)^{1/2}}{SD(H_C(\omega))\ H_T(\omega)^{1/2}\ +\ SD(H_T(\omega))\ H_C(\omega)^{1/2}}. \tag{4.27}$$

With this, the probability of randomizing patient $i+1$ to T based on information obtained from the first i patients enrolled into the trial is:

$$\rho_{H;i+1} = \frac{SD(H_{Ci}(\omega))\ H_{Ti}(\omega)^{1/2}}{SD(H_{Ci}(\omega))\ H_{Ti}(\omega)^{1/2}\ +\ SD(H_{Ti}(\omega))\ H_{Ci}(\omega)^{1/2}}. \tag{4.28}$$

Basing this RAR ratio on the cumulative hazard was done so that a parametric definition for survival times was not required. However, this RAR ratio could be used in the case of survival times being assumed to follow a specified parametric distribution. In this case, MLEs of the distribution-specific parameters would be obtained and used to estimate the distribution-specific $H_{gi}(\omega)$. The variance thereof may be obtained using the delta method.[24] Though the H-RAR is applicable to either parametric or non-parametric survival times, it was of-interest not to assume a parametric distribution in the present work. Therefore, $H_{gi}(\omega)$ and the variance thereof were estimated empirically. It should be noted, however, that if a parametric assumption was made for the survival times, and this assumption was correct, the resulting variance of the estimated cumulative hazard would be equal to or smaller than the empirically-estimated value.

Empirical estimation of the cumulative hazard and its variability can be obtained in many ways, for example, by HEFT estimation or via transformation of the Kaplan-Meier survivor function. [34,39] In the present chapter the Nelson-Aalen estimator was employed to demonstrate randomization using H-RAR, selected for its simplicity versus other methods and its ability to produce an estimate from a small number of events. [32] The Nelson-Aalen empirical estimator of the cumulative hazard function, denoted $H_{gi}^{NA}(\omega)$, and its variance, denoted $Var(H_{gi}^{NA}(\omega))$ are given by:

$$H_{gi}^{NA}(\omega) = \sum_{y_{gi} \leq \omega} \frac{\epsilon_{gi}}{\gamma_{gi}}, \text{ and} \tag{4.29}$$

$$Var(H_{gi}^{NA}(\omega)) = \sum_{y_{gi} \leq \omega} \frac{\epsilon_{gi}}{\gamma_{gi}^2}, \tag{4.30}$$

where ϵ_{gi} and γ_{gi} respectively represent the group-specific number of events having occurred and the number of patients still at-risk among the first i patients enrolled into the trial until a specified time ω. [32] Using these definitions of $H_{gi}^{NA}(\omega)$ and $Var(H_{gi}^{NA}(\omega))$, the variance constraint given in (4.26) was modified slightly in order to maintain the ZR optimal RAR construct. That is, defining $v(H_{gi}^{NA}(\omega)) = n_g \times Var(H_{gi}^{NA}(\omega))$, then taking $V = 1$:

$$Var(H_{Ti}^{NA}(\omega)) + Var(H_{Ci}^{NA}(\omega)) = 1 \tag{4.31}$$

$$\Leftrightarrow \frac{v(H_{Ti}^{NA}(\omega))}{n_T} + \frac{v(H_{Ci}^{NA}(\omega))}{n_C} = 1. \tag{4.32}$$

Then, letting $SD(H_{gi}^{NA}(\omega)) = v(H_{gi}^{NA}(\omega))^{1/2}$, the H-RAR ratio given in (4.28) is given as:

$$\rho_{H^{NA};i+1} = \frac{SD(H_{Ci}^{NA}(\omega))\ H_{Ti}^{NA}(\omega)^{1/2}}{SD(H_{Ci}^{NA}(\omega))\ H_{Ti}^{NA}(\omega)^{1/2}\ +\ SD(H_{Ti}^{NA}(\omega))\ H_{Ci}^{NA}(\omega)^{1/2}}. \tag{4.33}$$

4.3.1 H-RAR Performance Assessment

Performance of the H-RAR was assessed by comparison to the ZR convention. This was done in three ways. The first assessment compared the behavior of the H-RAR ratio against that of the ZR RAR ratio when survival times were correctly-specified. In this case, ideally, trends in H-RAR

and ZR RAR matched. Second, H-RAR was compared to ZR RAR when the ZR RAR ratio was misspecified using the well-behaved and -understood exponential and Weibull distributions. Finally, of-interest was the behavior of the H-RAR ratio compared to the conventional ZR RAR ratio when the ZR RAR ratio was severely misspecified using the exponential change-point hazard (ECPH) distribution. Ideally the H-RAR behaved as intended by the trial design.

4.3.1.1 Exponential Change-Point Hazard Distribution

Change-point hazard functions are commonly used in survival analyses due to their simplicity, flexibility, and ability to model complex time-to-event patterns.[55] Consider the following piecewise constant hazard function of $\lambda(y)$ of time $y \in (0, \zeta]$ with P discrete values such that:

$$\lambda(y) = \lambda_p,\ \zeta_{p-1} < y \leq \zeta_p\ (p = 1, \cdots, P), \tag{4.34}$$

where $0 = \zeta_0 < \zeta_1 < \cdots < \zeta_p = \zeta$ specify P time intervals in $(0, \zeta]$. Consider n observations from this model. For the $i^{\mathbf{th}}$ $(i = 1, \cdots, n)$ observation, let y_i $(y_i \leq \zeta)$ denote the survival time when the event indicator $\epsilon_i = 1$ or the censor time when $\epsilon_i = 0$. In order to discern the likelihood of an observed outcome, the data pair $(y_i,\ \epsilon_i)$ generated by the observed survival time and event indicator must be segmented into P intervals:

$$y_{ip} = \begin{cases} 0, & y_i \leq \zeta_{p-1} \\ y_i - \zeta_{p-1}, & \zeta_{p-1} < y_i \leq \zeta_p \\ \zeta_p - \zeta_{p-1}, & y_i > \zeta_p \end{cases}, \text{ and } \epsilon_{ip} = \epsilon_i\ \mathbf{1}\{\zeta_{p-1} < y_i \leq \zeta_p\} \tag{4.35}$$

for $p = 1, \cdots, P$, where $\mathbf{1}\{\cdot\}$ is the indicator function with the value 1 if the inequality is met and 0 otherwise. The vectors $(y_{i1}, \cdots, y_{ip})'$ and $(\epsilon_{i1}, \cdots, \epsilon_{ip})'$ are the decompositions of y_i and ϵ_i into the P time intervals;in particular, $\sum_{p=1}^{P} y_{ip} = y_i$ and $\sum_{p=1}^{P} \epsilon_{ip} = \epsilon_i$. Then the likelihood function

can be expressed as:

$$\ell(\lambda_1, \cdots, \lambda_P | \mathbf{y},\ \epsilon) = \prod_{i=1}^{n} \prod_{p=1}^{P} \lambda_p^{\epsilon_{ip}} exp(-\lambda_p y_{ip}), \tag{4.36}$$

making the loglikelihood:

$$log\{\ell(\lambda_1, \cdots, \lambda_P | \mathbf{y},\ \epsilon)\} = \sum_{i=1}^{n} \sum_{p=1}^{P} \Big\{ \epsilon_{ip} log(\lambda_p) - \lambda_p y_{ip} \Big\}, \tag{4.37}$$

such that the MLE of the interval-specific hazards for $p = 1, \cdots, P$ are:

$$\hat{\lambda}_p = \frac{\sum_{i=1}^{n} \epsilon_{ip}}{\sum_{i=1}^{n} y_{ip}}. \tag{4.38}$$

with SD $\lambda_p / \sqrt{\epsilon_p}$.

For the present scenarios, survival outcomes having survival times following the ECPH distribution were segmented into three time intervals, $P = 3$, such that $\{\zeta_0,\ \zeta_1,\ \zeta_2,\ \zeta_3\} = \{0.0,\ 0.2,\ 0.3,\ D\}$ for both groups. With this, the cumulative hazard may be expressed as:

$$H(y_i) = \begin{cases} \lambda_1 y_i, & 0 \leq y_i < \zeta_1 \\ \lambda_1 \zeta_1 + \lambda_2 (y_i - \zeta_1), & \zeta_1 \leq y_i < \zeta_2 \\ \lambda_1 \zeta_1 + \lambda_2 (\zeta_1 - \zeta_2) + \lambda_3 (y - i - \zeta_2), & \zeta_2 \leq y_i < \zeta_3 \end{cases} \tag{4.39}$$

In the present work, $\lambda(y_{p,g,i})$ and $H(y_{p,g,i})$ represent the group-specific ($g \in \{T, C\}$) hazard and cumulative hazard, respectively, estimated over time interval p ($p = 1, \cdots, P$) based on survival outcomes observed from the first i patients enrolled into the trial. Thus, these estimators ($\lambda(y_{p,g,i})$, $H(y_{p,g,i})$, and $SD(\lambda(y_{p,g,i}))$) were used to construct the RAR ratios for the ECPH distribution based on the piecewise-hazard function for comparison to the ZR RAR and based on the cumulative hazard for comparison to H-RAR. Results were used to discern an appropriate tuning value, discussed in later sections.

4.3.1.2 Tuning of the Optimal RAR Ratio Based on the Treatment Difference

The optimal ZR design for survival outcomes detailed here does not offer a mechanism for directly tuning the resulting distribution-specific optimal RAR ratios toward a desired ratio once a desired treatment effect is observed. Biswas and Bhattacharya modified the ZR RAR ratio for clinical trials producing continuous response data by including a tuning parameter, τ, as defined in previous chapters, on the group-specific means in order to target a prespecified randomization ratio, ρ_0, when a desired treatment effect was observed.[6,7] This procedure is detailed in Chapter 1. Prior to the present work, tuning has not been directly implemented in optimal RAR ratios intended for use in clinical trials with survival outcomes.

For such RAR ratios, let the treatment difference of interest be the hazard ratio (HR) defined as:

$$HR = \lambda_{C0} \Big/ \lambda_{T0}, \tag{4.40}$$

where λ_{g0} for $g \in \{T, C\}$ represents the group-specific measure of the hazard, be it the reciprocal of the mean as in the ZR parametric cases or a function of the cumulative hazard as in the H-RAR case. Then, using an approach similar to the framework particularized in Section 1.3.1, tuning of the survival RAR ratio was achieved by raising λ_{g0} to the τ^{th} power. To demonstrate this, consider a generalization of the survival RAR ratio based on λ_{g0} where $SD(\lambda_{g0})$ represents the measure of variability used to construct the ratio, e.g., the SD of the scale parameter in ZR RAR or a function of the variance of the cumulative hazard when using H-RAR. Then, when tuned, the exponential and Weibull ZR RAR ratios provided in (4.11) and (4.23), respectively, and the H-RAR ratio in (4.27) can all be expressed as:

$$\rho = \frac{SD(\lambda_{T0})\ \lambda_{C0}^{\tau/2}}{SD(\lambda_{T0})\ \lambda_{C0}^{\tau/2}\ +\ SD(\lambda_{C0})\ \lambda_{T0}^{\tau/2}}. \tag{4.41}$$

As such, τ is defined as:

$$\tau = 2 \times \frac{log\left(\frac{SD(\lambda_{T0})}{SD(\lambda_{C0})}\right) + log\left(\frac{1}{\rho_0} - 1\right)}{log\left(\lambda_{T0} \Big/ \lambda_{C0}\right)}. \tag{4.42}$$

In practice, when used to randomize patient $i+1$ into the trial, the estimators of the survival RAR ratios given in (4.12), (4.24), and (4.33) are expressed explicitly by:

$$\hat{\rho}_{T;i+1} = \frac{\frac{\hat{\theta}_{Ti}}{\sqrt{\hat{\xi}_{Ti}}}\left(\frac{1}{\hat{\theta}_{Ci}}\right)^{\tau/2}}{\frac{\hat{\theta}_{Ti}}{\sqrt{\hat{\xi}_{Ti}}}\left(\frac{1}{\hat{\theta}_{Ci}}\right)^{\tau/2} + \frac{\hat{\theta}_{Ci}}{\sqrt{\hat{\xi}_{Ci}}}\left(\frac{1}{\hat{\theta}_{Ti}}\right)^{\tau/2}}, \tag{4.43}$$

$$\hat{\rho}_{T;i+1} = \frac{\hat{k}_{Ti}\sqrt{\hat{G}_{Ti}}\left[exp(-\hat{m}_{Ci})/\Gamma(1+\hat{k}_{Ci})\right]^{\tau/2}}{\hat{k}_{Ti}\sqrt{\hat{G}_{Ti}}\left[exp(-\hat{m}_{Ci})/\Gamma(1+\hat{k}_{Ci})\right]^{\tau/2} + \hat{k}_{Ci}\sqrt{\hat{G}_{Ci}}\left[exp(-\hat{m}_{Ti})/\Gamma(1+\hat{k}_{Ti})\right]^{\tau/2}}, \tag{4.44}$$

$$\hat{\rho}_{T;i+1} = \frac{\hat{SD}(\hat{H}_{Ti}^{NA}(\omega))\ \hat{H}_{Ci}^{NA}(\omega)^{\tau/2}}{\hat{SD}(\hat{H}_{Ci}^{NA}(\omega))\ \hat{H}_{Ti}^{NA}(\omega)^{\tau/2} + \hat{SD}(\hat{H}_{Ti}^{NA}(\omega))\ \hat{H}_{Ci}^{NA}(\omega)^{\tau/2}}. \tag{4.45}$$

for exponential ZR RAR, Weibull ZR RAR, and H-RAR, respectively. It was of-interest in the present work to evaluate and compare the behavior of each of these optimal RAR ratio estimators when survival times followed various response distributions.

4.4 Evaluation of H-RAR Performance

A simulation study was performed as a means of measuring the bias, precision, and ethicality of the ZR RAR and empirical H-RAR ratios. For all scenarios, the means and 95% CI of the ZR or H-RAR ratios averaged over 1,000 simulated trials were plotted with respect to the RAR ratio intended by the trial design, denoted ρ_d. When tuned, the intended RAR ratio was 1:1 ($\rho_d = 0.5$) when group-specific survival outcomes were generated to reflect no treatment difference ($HR = 1.0$) and 3:1 ($\rho_d = \rho_0 = 0.75$) when a large treatment difference was generated ($HR = 1.5$).[32] Means of the averaged RAR ratios over 1,000 simulated trials were plotted to measure bias such that smaller distances from ρ_d represented less bias, and, thereby, suggested that that RAR ratio more

accurately enrolled patients into the trial according to the trial design. Precision was measured using the 95% CI about the mean of the averaged RAR ratios such that CI not containing ρ_d may suggest that these RAR ratios produce power-versus-ethics dynamics that differ from those intended by the trial design.

Adherence to the ethical objective of the RAR design was measured using the plot of the percent-difference between the observed and expected ethical objective function ($\%_{\text{O-E}}$). If the RAR ratio under investigation adhered to the ethical objective of the trial, the observed ethical objective function should match the intended ethical objective function. With this, RAR ratio estimators producing smaller absolute $\%_{\text{O-E}}$ (i.e., closer to zero) performed best according to this metric. Details pertaining to the calculation of the $\%_{\text{O-E}}$ per each of the three H-RAR performance assessments are discussed in the following section. Finally, type I error was evaluated using the Z-score introduced by Zhang and Rosenberger for each of the exponential or Weibull survival times based on observed survival responses from the first 500 participants enrolled into the simulated trials where no treatment effect was generated (HR=1.0).

4.5 Simulation Details

All simulated trials enrolled 1,000 patients. When the RAR ratio under investigation was the ZR RAR ratio constructed using Weibull parameters, simulated trials were performed over 1,500 replicates. Otherwise, 1,000 replicates were performed. This difference was due to greater complexity of moment estimation for the Weibull ZR RAR versus the exponential ZR RAR or H-RAR, and its impact on the size of the lead-in group prior to the initiation of RAR. Additional details are discussed later in this chapter.

The treatment group was considered superior to the control group when a large treatment difference existed. As such, survival outcomes observed from the treatment group were generated to reflect longer survival times, thus corresponding to a lower hazard of experiencing an event.

Table 4.1: Distribution-specific parameter values for simulation of treatment and control group survival outcomes by effect size where trial recruitment time $R = 85/85 = 1$ and trial duration $D = 100/85 \approx 1.176$.

Survival Times	Parameter	20% Censoring	
		Control/Null HR	Large HR
Exponential	Mean/Scale, θ_{g0}	0.189	0.283
	Hazard, λ_{g0}	5.292	3.528
	SD of scale, $SD(\theta_{g0})$	0.211	0.339
	Intended censoring, η_{g0}	0.200	0.299
	Intended ρ_d without τ	0.500	0.662
	τ for 3:1 using ZR RAR	3.093	3.093
	τ for 3:1 using H RAR	3.625	3.625
Weibull	Shape, α_{g0}	3.500	2.000
	Scale, β_{g0}	0.128	0.195
	EVD scale, k_{g0}	0.286	0.500
	EVD location, m_{g0}	-2.058	-0.711
	Mean, θ_{g0}	0.115	0.172
	Hazard, λ_{g0}	8.700	5.800
	Intended censoring, η_{g0}	0.800	0.700
	${G_{g0}}^*$	1.449	1.716
	SD Weibull scale, $SD(m_{g0})$	0.344	0.655
	Intended ρ_d without τ	0.500	0.700
	τ for 3:1 using ZR RAR	2.242	2.242
	τ for 3:1 using H RAR	3.700	3.700
	$^*\mathrm{E}(\mathrm{ze}^{\mathrm{z}})=0.423$, $\mathrm{E}(\mathrm{z}^2\mathrm{e}^{\mathrm{z}})=0.827$ $\forall$ g		
Exponential Change-Point Hazard	Hazard over $(\zeta_0,\ \zeta_1]^*$, $\lambda_{1,g0}$	7.000	4.667
	Hazard over $(\zeta_1,\ \zeta_2]^*$, $\lambda_{2,g0}$	30.000	20.000
	Hazard over $(\zeta_2,\ \zeta_3]^*$, $\lambda_{3,g0}$	5.000	3.333
	Intended censoring, η_{g0}	0.800	0.750
	τ for 3:1 using ZR RAR	11.000	11.000
	τ for 3:1 using H RAR	3.963	3.963
	$^*\zeta_0$=0.0, ζ_1=0.2, ζ_2=0.3, ζ_3=D $\forall$ g		

Control group survival outcomes where survival times followed an exponential distribution with mean parameter θ_{C0} were developed as follows. First, the recruitment time and trial duration parameters were fixed as R and D, respectively, and the desired probability of event, ξ_{C0}, taking the value 0.8 to reflect 20% censoring was used in (4.14) to obtain θ_{C0}. The hazard of the control group, λ_{C0}, was obtained by taking the reciprocal of the mean parameter. The resulting SD of the control group mean, $SD(\theta_{C0})$, was obtained using $\theta_{C0}/\sqrt{\xi_{C0}}$. Next, the hazard of the treatment group, λ_{T0}, was obtained by dividing the control group hazard by the desired HR (either 1.0 or 1.5). Then, the treatment group mean, θ_{T0}, was found by taking the reciprocal thereof, i.e., $1/\lambda_{T0}$. Fixed R and D and θ_{T0} were used to obtain ξ_{T0} using (4.14) and $SD(\theta_{T0})$. With this, all parameters necessary to derive the exponential survival times reflecting varying treatment difference and censoring probabilities were calculated. Table 4.1 contains the specific values for each of these parameters, as well as those for the subsequently-described Weibull and ECPH survival times.

Simulations wherein survival times following the Weibull distribution were derived were more involved. First, the Weibull shape parameter for the control group, α_{C0}, was selected. Then, a simulation of 1,000 replicates with 1,000 patients was performed in order to select the Weibull scale parameter, β_{C0}, that produced the desired ξ_{C0}. Because Weibull survival times are log-transformed for ZR RAR, the EVD parameters corresponding to each Weibull parameter was obtained: the EVD scale parameter, k_{C0}, was obtained by taking the reciprocal of α_{C0} and the EVD location parameter, m_{C0}, was obtained by taking the logarithm of β_{C0}. The mean control group survival time was defined as $\theta_{C0} = exp(m_{C0})\Gamma(1 + k_{C0})$, the reciprocal of which provided the hazard of the control group, λ_{C0}, as defined by Zhang and Rosenberger. To find the variance of these survival times, the quantities $E(z_i exp(z_i))$ and $E(z_i^2 exp(z_i))$ were obtained in R by numerically integrating $zf(z)$ and $z^2 f(z)$, respectively, from 0 to 75. These values were held constant for both the control and treatment groups, and were used along with ξ_{C0} to obtain G_{C0}. With this, $SD(m_{C0})$, the SD

of the control group EVD location parameter (corresponding to β_{C0} was obtain by $k_{C0} \times \sqrt{G_{C0}}$.

The hazard of the treatment group, λ_{T0}, was obtained by taking the ratio of the λ_{C0} and HR, and the mean survival time, θ_{T0}, was the reciprocal thereof. Upon selecting the Weibull shape parameter for the treatment group, α_{T0}, the Weibull scale parameter for the treatment group was obtained using the Weibull expression of the mean such that, $\beta_{T0} = \theta_{T0}/\Gamma(1 + 1/\alpha_{T0})$. The probability of an event occurring in the treatment group, ξ_{T0}, was discerned by generating 1,000 iterations of survival outcome data having Weibull survival times with parameters α_{T0} and β_{T0} for 1,000 recruited patients. The mean of the sum of events scaled by 1,000 provided the expected ξ_{T0}. The treatment group location and scale parameters for the EVD, m_{T0} and k_{T0}, respectively, were obtained, and G_{T0} was estimated using the previously fixed $E(z_i exp(z_i))$, $E(z_i^2 exp(z_i))$, and ξ_{T0} such that $SD(m_{T0})$ was obtained.

Finally, change-points and hazards parametrizing survival times following the ECPH distribution for the control group were selected using the `cpsurvsim::exp_cdfsim` in R.[55] These parameters were selected such that the desired censoring probability was observed over 1,000 replicated datasets enrolling 1,000 patients. Change-points were held constant for both groups so that the assumption of proportional hazards could be maintained within time intervals. This allowed for the obtainment of the interval-specific treatment group hazards by dividing the interval-specific control group hazards by the intended HR. A simulation was performed where 1,000 replicates of 1,000 patients whose survival outcomes consisted of survival times that followed the ECPH distribution parametrized according to the interval-specific treatment group hazards and their corresponding change-points. The censoring probability was discerned by taking the complement of the number of observed events scaled by 1,000, the number of replicated trials.

4.5.1 Survival Data Generation

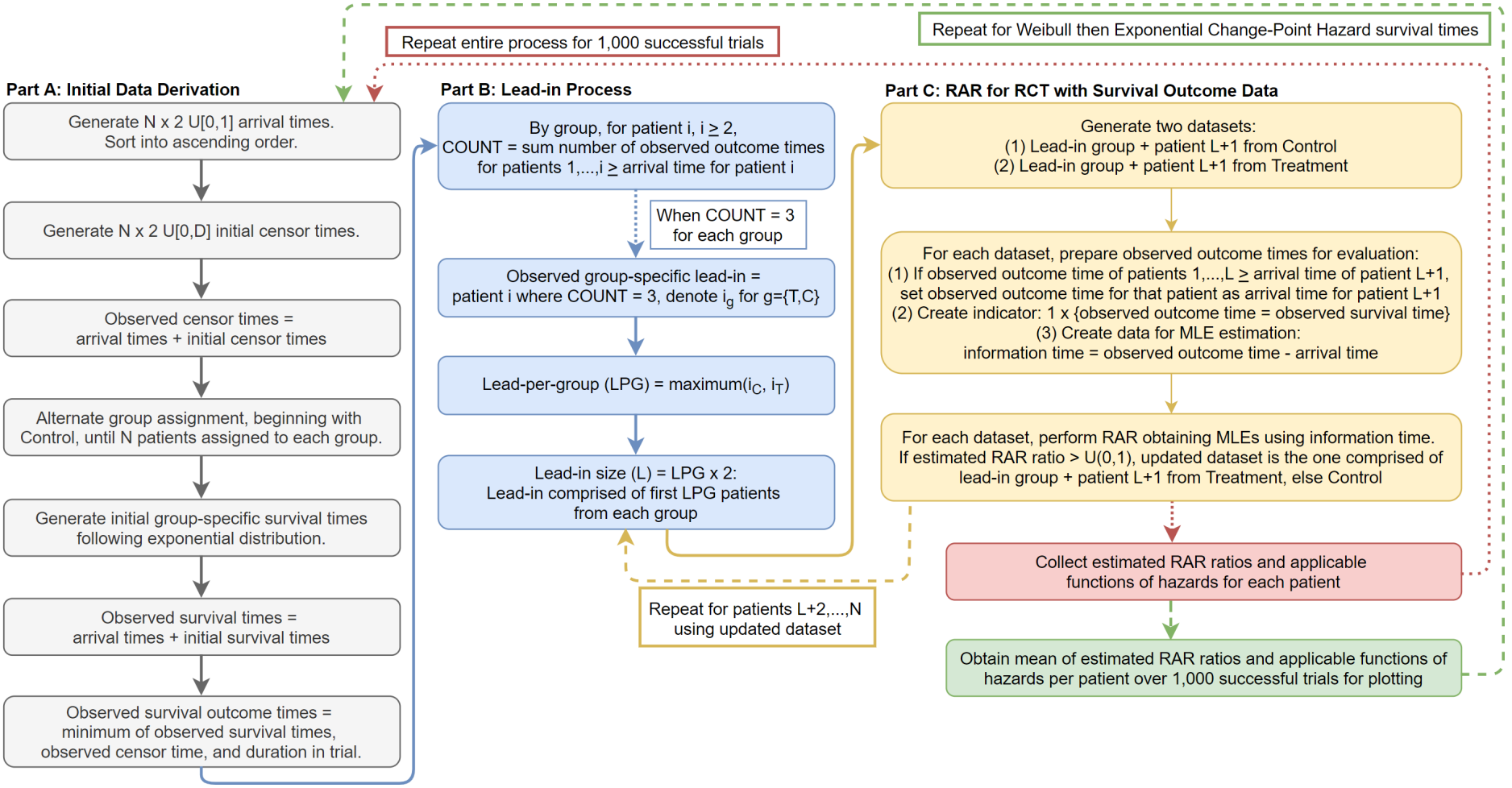
It was shown by Zhang and Rosenberger and Hu and Zhang that a moderate delay in censored survival responses had only a marginal effect on the asymptotic properties of these RAR procedures.[6,8] Therefore, survival responses were artificially treated as immediate during the simulation process, mimicking the simulation design of others.[8,25,54] It is worth noting that relying on these asymptotic properties is a reasonable assumption for short-term survival outcomes, but not when measuring survival in terms of months or years. The immediate measure assumption may no longer be valid in such scenarios because there would be a high level of censoring at the beginning of the trial. Without having observed an event, there would not be enough survival information to initiate RAR estimation, as such, the asymptotic assumptions would fail.

General survival data derivation and the censoring scheme employed in the present simulation were based on the Rosenberg and Seshaiyer dynamic discussed in Section 4.1.1.1. Trial details were based loosely on the motivating example of Jones et al., explored in both Zhang and Rosenberger and Liu and Coad.[15,25,54,56] This was a phase III clinical trial comparing docetaxel and paclitaxel, two drugs approved for use in patients with metastatic breast cancer, where docetaxel showed more favorable chemical and biological outcomes than paclitaxel. With this, it would be ideal to allocate as many patients to the docetaxel group as possible. For this reason, the docetaxel group was considered the treatment group and the paclitaxel group was the control. As in both Zhang and Rosenberger and Liu and Coad, though Jones et al. conducted the trial at multiple sites, RAR was assumed to be centralized in the present simulation.

Where an event was defined as either tumor progression, unacceptable toxicity, or consent withdrawal, the ideal survival outcome was censoring due to having completed the study having neither experienced any of the previously-described events nor being censored for any other reason resulting in a trial participant being lost-to-follow-up. Therefore, it was hypothesized that the docetaxel

group would have a larger censoring probability than the paclitaxel group. The recruitment time of the Jones et al. docetaxel-versus-paclitaxel trial was 84 months and the duration of the trial was 102 months. For the present simulation, these values were rounded to reflect a recruitment time of 85 and trial duration of 100. As in Zhang and Rosenberger, since the ZR RAR ratios are not dependent upon the unit of the parameters of interest, the recruitment time and trial duration parameters were scaled by the recruitment time, i.e., $R = 85/85 = 1$ and $D = 100/85 \approx 1.18$.[25,54] The parameter values described in the previous section were scaled similarly.

Figure 4.5: Survival Simulation Flowchart



The arrival time for a given patient and the trial duration start from the beginning of the study, whereas the survival time and censor time commence from the arrival of that patient (Part A of Figure 4.5). At the patient level, a random sample of arrival times following $U[0, 1]$ were generated for $N \times 2$ (N patients in two treatment groups) recruited patients, then sorted into ascending order. Independently, $N \times 2$ initial censor times were generated using $U[0, D]$. Observed patient censor times were then the sum of that patient's arrival time and initial censor time.[25,54]

To aid in the construction of the lead-in (details in next section), the first pair of arrival and observed censor times were assigned to the control group, then the second pair to the treatment group, and so forth following this pattern until N arrival and observed censor times were assigned to each group. By group, initial survival times were generated and summed with their corresponding arrival times to produce the observed survival time.[25,54] Finally, the observed survival outcome times for a given patient was the minimum of that patient's observed survival time, observed censor time, and the length of time that that patient was in the study, represented by the difference between D and that patient's arrival time.[25,54]

4.5.2 Lead-in Simulation Process

As discussed in previous chapters, lead-ins are necessary in response-adaptive CTs because RAR assigns patients according to already-observed response data.[3,6,7,25,54] Therefore, the present simulation followed the work of Liu and Coad by using the permuted-block design to construct the lead-in from which RAR would begin.[54] A block-size of one (i.e., alternating allocation) was used to ensure the fewest number of patients were equally-randomized into the trial during the lead-in period. Part B of Figure 4.5 demonstrates the lead-in process used in the present simulation study.

The ZR RAR ratios are based on the parameters of the assumed survival time distributions. Therefore, for a more stable start to MLE estimation, the present simulation designated that the lead-in period consist of at least three observed survival outcomes, meaning either an event or

Table 4.2: Percent-convergence at three observed survival outcomes per group and median lead-in sizes per simulated RCT scenario.

20% Censor		Exp ZR RAR:		Weibull ZR RAR:		H-based RAR:
	Survival	%conv (median)		%conv (median)		%conv (median)
HR	Times	Not tuned	Tuned	Not tuned	Tuned	Tuned
1.0	Exp	100.0 (72)	100.0 (72)	99.6 (72)	99.7 (72)	100.0 (72)
	Weibull	100.0 (134)	100.0 (134)	99.8 (132)	100.0 (136)	100.0 (136)
	ECPH		100.0 (64)		97.2 (64)	100.0 (64)
1.5	Exp	100.0 (78)	100.0 (76)	99.0 (78)	99.7 (78)	100.0 (78)
	Weibull	100.0 (128)	100.0 (128)	99.8 (128)	99.7 (128)	100.0 (130)
	ECPH		100.0 (70)		98.3 (72)	100.0 (70)

%conv: percent of trials with all patients randomized; Exp: Exponential; ECPH: Exponential change-point hazard

censor time beyond censoring for the reason that the subject was still participating in the study was observed. At least one of these observed survival outcomes must have been an observed survival time. Three observed outcomes was selected for the initiation of RAR because this was the smallest number of observations to begin variance estimation. Table 4.2 demonstrates that, across scenarios, greater than 98% of simulated trials were successful, where simulated trials were deemed successful when all 1,000 patients were randomized into the trial following at least three observed survival outcomes during the lead-in period. The resulting median lead-in sizes per scenario are provided in Table 4.2, as well. To account for sequentially-increasing arrival times (i.e., staggered entry), the number of occurrences where the observed survival outcome for patients one to $i-1$ were greater than the arrival time for patient i were counted, beginning with the second patient in each group. The patient number i at which it first occurs that at least three survival outcomes have been observed, one of which was an event, was recorded as the lead-in size for each group. Then, the trial lead-in size, L, was taken as the maximum of the two group-specific lead-in sizes.

4.5.3 RAR for Simulated RCT with Survival Outcomes

Following the generation of data (§4.5.1) and the determination of the lead-in group (§4.5.2), RAR was initiated via the randomization of patient $L+1$ (Part C of Figure 4.5). Before MLE estimation

could begin, two options for patient $L+1$ were considered. These two options were characterized as the first patient in each group whose arrival time was greater than the arrival time belonging to the final patient enrolled during the lead-in, or patient L. Thus, two sets of survival data were created where both were comprised of the arrival and observed survival outcome times collected during the lead-in period, but where one dataset appended the first applicable record for patient $L+1$ from the control group, and the other from the treatment group. For each group-specific set of survival data, any lead-in patients' observed outcome survival times that were greater than the arrival time for patient $L+1$ (i.e., the maximum arrival time) were reset to be the maximum arrival time per patient $L+1$. With this, MLE estimation at the time of enrollment for subject $L+1$ could be performed.

Two additional variables were created in order to execute MLE estimation. First was an indicator variable taking the value of one when the observed survival outcome time for a given patient is that patient's survival time. Second was information time, the difference between each patient's observed survival outcome time and the arrival time for that patient. Maximum likelihood estimation was performed using information time and the event indicator at the patient level. These estimates were used to construct the RAR ratio under investigation for that simulation. The resulting RAR ratio was compared to a uniform random variate using $U(0,1)$ such that patient $L+1$ was randomized to the treatment group if the estimated RAR ratio was greater than the uniform random variate. If group assignment was treatment, the survival data belonging to patient $L+1$ from the treatment group was appended to the lead-in survival data as a part of the overall trial response data. Put another way, the dataset reflecting the lead-in survival data and the survival information for patient $L+1$ from the treatment group becomes the set of survival data resulting from the simulated trial thus far, containing the survival data upon which the randomization of patient $L+2$ will be based. The same is true of the control group survival data if the RAR ratio

was less than the uniform random variate. This process was repeated until all N patients were enrolled into the trial. This particular approach was taken in an effort to mimic staggered entry into the trial where the arrival time of a given patient i was always greater than the arrival time of patient $i-1$, regardless of group affiliation.

To produce the previously-described plots of the means and 95% CIs, the RAR ratios across 1,000 iterations of each of the ZR RAR vs. H-RAR scenarios considered were collected for each of the 1,000 patients enrolled. The means and variances of these sets of 1,000 RAR ratios were calculated and 95% CIs were constructed. The ethical objective of RAR in these trials was to minimize either the total expected hazard or mean cumulative hazard of patients enrolled into the trial. Therefore, to construct the plot of the $\%_{\text{O-E}}$, group-specific sample sizes and hazards or cumulative hazards were collected after the enrollment of each patient. Their means were then calculated across 1,000 trials.

Table 4.3: Observed versus expected hazard or cumulative hazard for estimation of percent-difference ($\%_{\text{O-E}}$) in ethicality of RAR by scenario.

Scenario	Survival Times	RAR	Observed	Expected
1: Correct Specification	Exp	Exp ZR	$\hat{\lambda}_{gi} = 1/\hat{\theta}_{gi}$	$\lambda_{g0} = 1/\theta_{g0}$
	Weibull	Weibull ZR	$\hat{\lambda}_{gi} = \frac{exp(-\hat{m}_{gi})}{\Gamma(1+\hat{k}_{gi})}$	$\lambda_{g0} = \frac{exp(-m_{g0})}{\Gamma(1+k_{g0})}$
2: Misspec-ification	Exp	Weibull ZR	$\hat{\lambda}_{gi} = \frac{exp(-\hat{m}_{gi})}{\Gamma(1+\hat{k}_{gi})}$	$\lambda_{g0} = 1/\theta_{g0}$
	Weibull	Exp ZR	$\hat{\lambda}_{gi} = 1/\hat{\theta}_{gi}$	$\lambda_{g0} = \frac{exp(-m_{g0})}{\Gamma(1+k_{g0})}$
3: H-RAR Using Nelson-Aalen	Exp	H-RAR	$\hat{H}^{NA}_{gi}(y)$ of (4.29)	Exp $\hat{H}^{E}_{gi}(y) = y_g/\hat{\theta}_{gi}$
	Weibull	H-RAR	$\hat{H}^{NA}_{gi}(y)$ of (4.29)	Weibull $\hat{H}^{W}_{gi}(y) = (y_g/\hat{\beta}_{gi})^{\hat{\alpha}_{gi}}$

The simulation design varied for each of the simulation scenarios, as did the estimation of the $\%_{\text{O-E}}$. Borrowing the general notation introduced in Section 4.3.1.2, the ethical objective functions can be expressed in a more general form:

$$\min_{n_T/n_C} \Big\{ n_T \lambda_{T0} + n_C \lambda_{C0} \Big\}, \tag{4.46}$$

With this, the observed, $\overline{O}_i$, and expected, $\overline{E}_i$, ethical objective functions were defined as:

$$\overline{O}_i = \overline{\hat{n}}_{Ti}\,\overline{\hat{\lambda}}_{Ti} \;+\; \overline{\hat{n}}_{Ci}\,\overline{\hat{\lambda}}_{Ci}, \text{ and,} \tag{4.47}$$

$$\overline{E}_i = \overline{E(n_{Ti})}\,\lambda_{T0} \;+\; \overline{E(n_{Ci})}\,\lambda_{C0}, \tag{4.48}$$

where $\bar{\hat{n}}_{gi}$ and $\overline{E(n_{gi})}$ were defined as in Section 2.2, and $\bar{\hat{\lambda}}_{gi}$ represents the mean observed hazard or cumulative hazard of group g from the first to the i^{th} patient enrolled into the trial, averaged over 1,000 simulated trials. In most cases, λ_{g0}, the expected group-specific hazard or cumulative hazard, was constant, determined by parameter values set at the start of the simulation. However, when RAR was performed using H-RAR, λ_{g0} was the cumulative hazard estimated for each patient enrolled into the trial and averaged over 1,000 trials. This is described more fully by scenario in Table 4.3. With this, the percent-difference between the observed and expected ethical objective of the trial is defined as in (2.35), i.e., $\%_{\text{O-E}} = 100 \times (\overline{O}_i - \overline{E}_i)/\overline{E}_i$.

4.6 Randomization Results

Results for simulated RCTs with survival outcomes are discussed in what follows. All curves were loess-smoothed using a bandwidth of 0.15. As discussed in Section 4.3.1, the performance of the H-RAR was assessed in comparison to ZR RAR when survival times were (a) correctly-specified, (b) incorrectly-specified, and (c) followed neither the exponential nor Weibull distribution. When tuned and correctly-specified, randomization using the ZR RAR ratio behaved precisely as intended by the trial design. That is, mean randomization converges to 3:1 when a large HR is observed and stays at 0.5 when an HR of 1.0 is observed (Figure 4.6), and the ethical objective of the trial is well-maintained (Figure 4.7). When misspecified, save for the null case where randomization remains equal, the previously-described behavior changes. As opposed to converging to the intended 3:1 randomization ratio (Figure 4.8, green curve), patients are over-randomized to treatment (0.808) when survival times followed the exponential distribution but patients were randomized into the

Figure 4.6: Mean of ZR RAR ratios with 95% CI with correctly-specified survival times for null and large treatment differences when tuned to target desired 3:1 ratio.

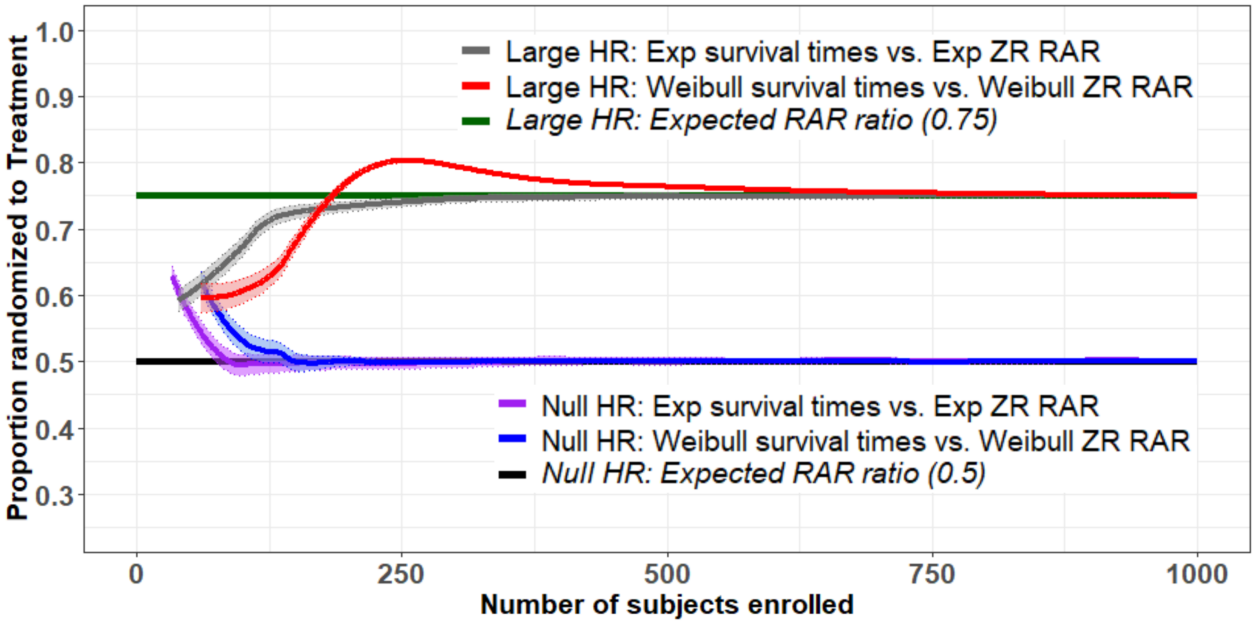


Figure 4.7: Percent-difference between the observed and expected ethical objective function of correctly-specified survival times for null and large treatment differences using ZR RAR when tuned to target desired 3:1 ratio.

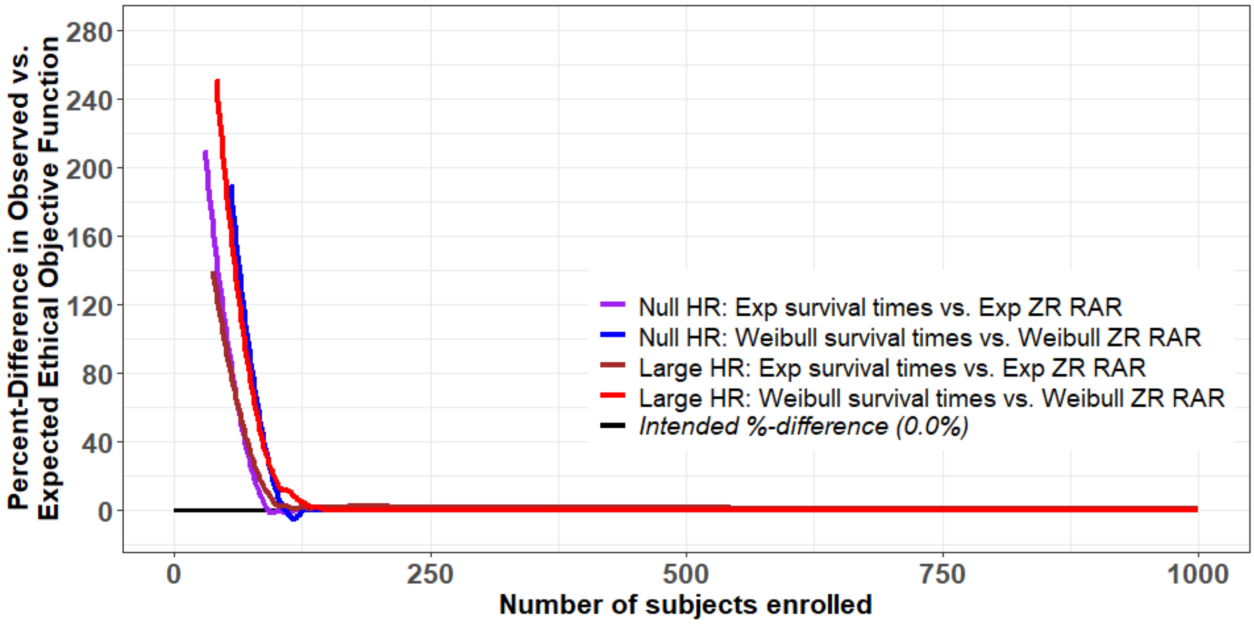


Figure 4.8: Mean of ZR RAR ratios with 95% CI with incorrectly-specified survival times for null and large treatment differences when tuned to target desired 3:1 ratio.

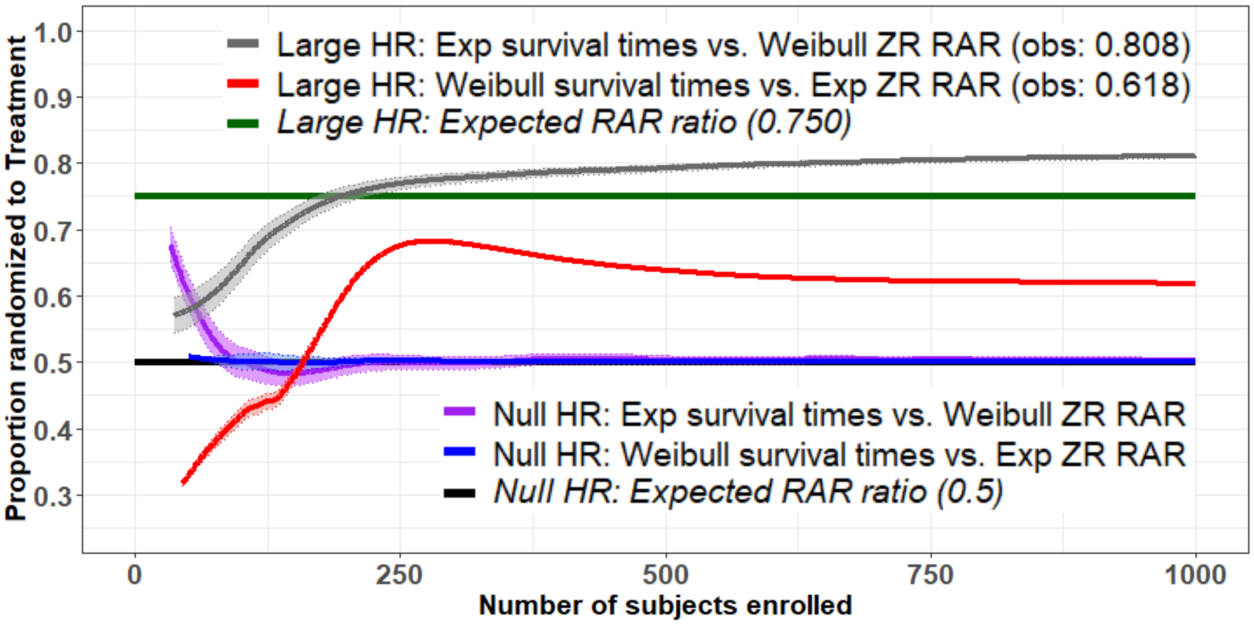
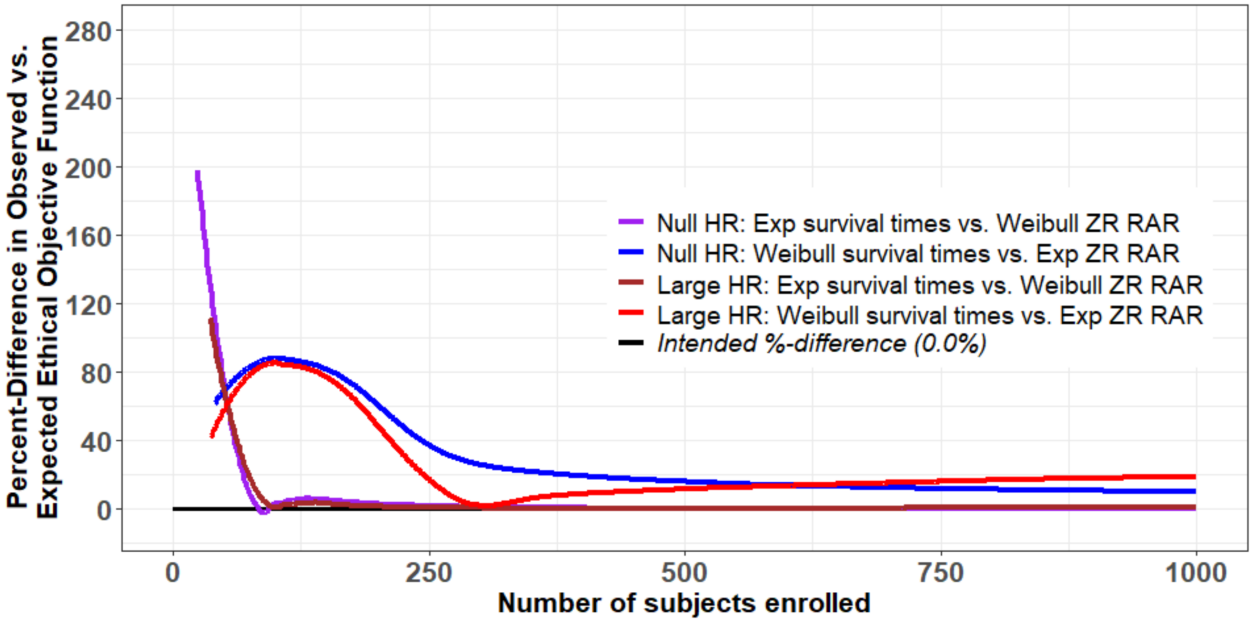


Figure 4.9: Percent-difference between the observed and expected ethical objective function of incorrectly-specified survival times for null and large treatment differences using ZR RAR when tuned to target desired 3:1 ratio.



trial using the Weibull ZR RAR ratio (grey curve). By contrast, when the opposite was true (red curve), patients were under-randomized to treatment (0.618). When survival times followed the Weibull distribution but were randomized using the exponential ZR RAR ratio, the ethical objective of the trial was not always maintained (Figure 4.9, red and blue curves).

Figure 4.10: Mean of H-RAR ratios with 95% CI for null and large treatment differences when tuned to target desired 3:1 ratio.

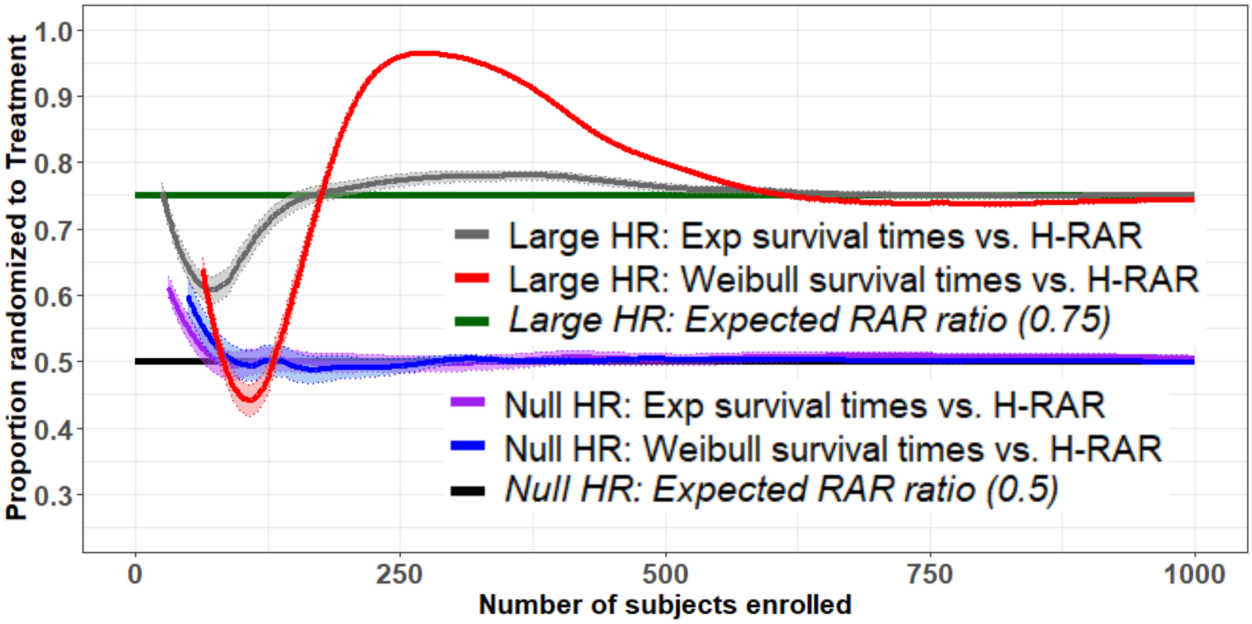


Table 4.4: Assessment of type I error using the ZR-defined Z-score for exponential and Weibull survival times based on observed survival information from first 500 participants enrolled into the simulated trial when no treatment difference existed (HR=1.0) across 1,000 simulated trials.

Truth	RAR	Z-score (ZR)	p-value
Exponential	Exponential	-0.001	0.9499
	Weibull	0.002	0.9952
	H-RAR	-0.002	0.9105
Weibull	Exponential	0.000	0.9771
	Weibull	0.002	0.9614
	H-RAR	0.002	0.9981

Randomization using the H-RAR method (Figure 4.10) converged to 3:1 randomization, both behaving in a manner similar to the ZR RAR conventional approach when correctly-specified (Fig-

Figure 4.11: Percent-difference between the observed and expected ethical objective function of H-RAR for null and large treatment differences when tuned to target desired 3:1 ratio.

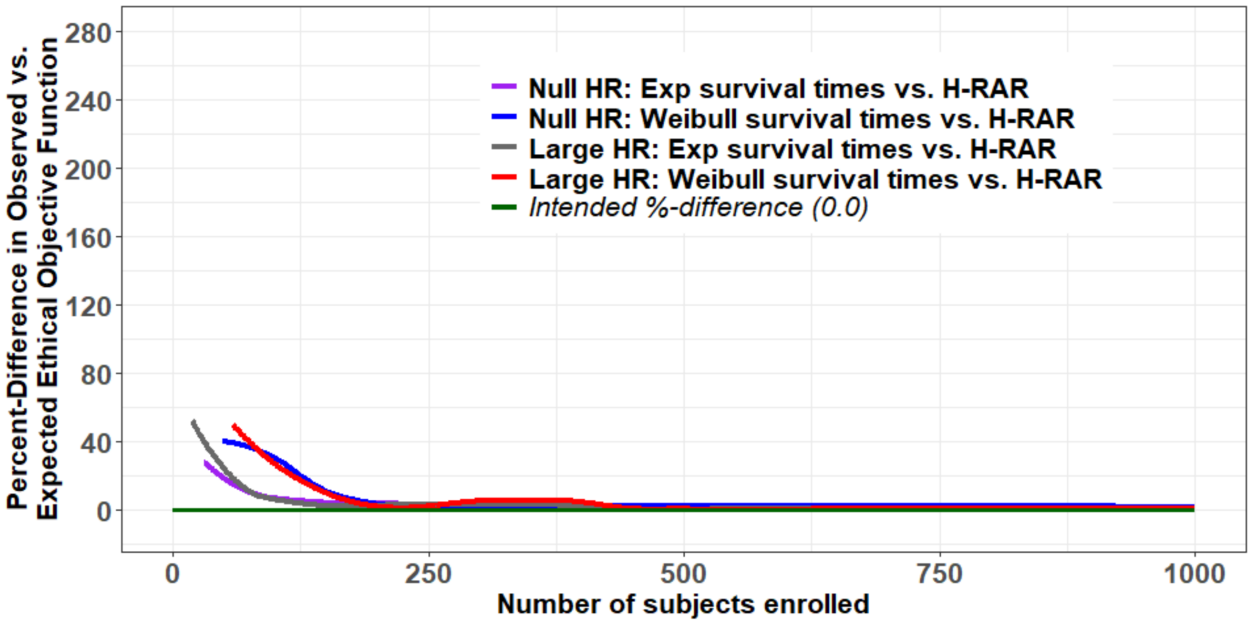
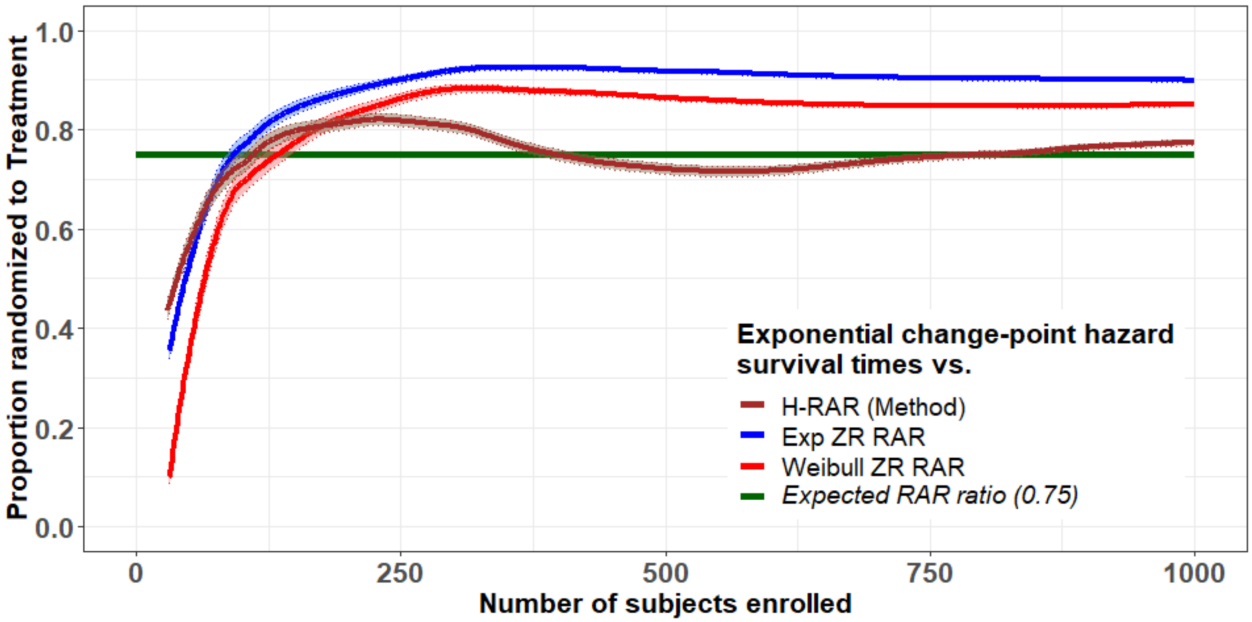


Figure 4.12: Mean of H-RAR ratios with 95% CI for a large treatment difference when tuned to target desired 3:1 ratio.



ure 4.6) and outperforming the same when misspecified (Figure 4.8). When a large treatment difference was observed and survival times followed the Weibull distribution (red curve), ZR RAR more closely adhered to the trial design in samples less than approximately 500 patients when correctly-specified and between approximately 250 and 400 patients when misspecified. Nonetheless, H-RAR adhered to the ethical objective of randomization than ZR RAR regardless of distributional specification (Figure 4.11). Finally, when survival times followed neither the exponential nor the Weibull distributions and were tuned to target 3:1 randomization (Figure 4.12, brown curve), H-RAR most closely adhered to randomization desired by the trial design than did either of the knowingly-misspecified ZR RAR ratios (blue or red curves). Finally, as evidenced by Table 4.4, no erroneous treatment differences were observed.

4.7 Discussion

Response-adaptive randomization based on the cumulative hazard presents as a viable alternative to the conventional optimal RAR approach set forth by Zhang and Rosenberger for survival outcomes.[25] However, when fewer than 500 subjects have been enrolled into the trial, ZR RAR is suggested for the sake of simplicity if the distribution of the survival times is either exponential or Weibull and known to be correct. If the ethical goal of randomization is the primary focus of the trial or the distribution of the survival times cannot be assumed as either exponential or Weibull, H-RAR may be the most appropriate option overall.

In order to compare the H-RAR and the ZR RAR methods, tuning was required. Tuning relied upon the proportional hazard assumption at 30 days, the time at which the large effect size was assumed to be observed. Furthermore, the simulation itself was based on the Jones et al. trial reconstructed in other works discussing RAR with survival outcomes. There were some flaws to this design, however. As opposed to considering the trial duration (D) a single unit, the trial recruitment period (R) was the single unit ($R = 85/85 = 1$ and $D = 100/85 > 1$). If R were made

much closer to zero, a shorter lead-in would have been observed, and RAR would have performed better because there would have been less of a lag between randomization and the outcome.

Chapter 5

Conclusion

5.1 Misspecification for Continuous Response Data

When designing a clinical trial with continuous response outcomes, most investigators assume that subject responses are going to follow the normal distribution. Using this normality assumption, ER is typically employed. Equal randomization is the most powerful type of randomization, requiring the fewest number of subjects to be enrolled into the trial to discern a clinically-meaningful difference between group than any other type of randomization. By the time a drug or intervention has progressed to a phase III clinical trial, there should be very strong evidence supporting its effectiveness and efficacy over the active control (e.g., standard of care). Therefore, despite its power, in the context of a phase III clinical trial, ER is unethical because, by definition, approximately half of all trial participants will be randomized to an inferior treatment group. Thus, RAR should be used to maintain the power of the trial and ethical treatment of the subjects in the trial by skewing randomization toward the better-performing treatment group based on responses obtained from subjects who have already completed their intervention. Because RAR is based upon the mean and SD of the response distribution of the observed subject responses, it was speculated that the normality assumption cannot always be justified in small sample.

Therefore, two methods were developed to obtain a RAR ratio robust against distributional

misspecification, particularly deviations from normality, when response values were continuous in nature. The first method estimated the RAR ratio using empirical estimation of the CDF obtained using either the PLKM or HEFT methodologies (Chapter 2). The second method of obtaining a RAR ratio that is robust to distributional misspecification employed a weighted-average approach (Chapter 3). With this, a variety of RAR ratios corresponding to a variety of potential response distributions (e.g., normal, gamma, logistic) were weighted based on how well that response distribution fit the observed response data. Weights were obtained two ways, using the AIC or KS metrics, and summed to one. Once the weights and randomization ratios per each candidate response distribution were obtained, their products are summed.

It was found that, in general, neither the empirical nor the weight-average randomization methods outperformed randomization under the normality assumption. Nonetheless, the goal of obtaining estimators of the RAR ratio either independent of any distributional assumption or considering a variety of response distributions was achieved. These estimators were well-behaved with respect to adhering to and maintaining randomization as intended by the trial design. With respect to the empirical estimators, in general, HEFT outperformed PLKM randomization, but at the price of a substantially larger lead-in period, something that is not always practical in real-world trials. Nonetheless, the results observed in Chapters 2 and 3 pertain only the scenarios investigated herein. Results may vary for alternative distributions or methods of estimation. Appendices A.28 and A.29 hold examples of the R codes used to simulate the trials in Chapters 2 and 3, respectively.

5.2 Misspecification for Survival Response Data

Little research with respect to RAR has been done in RCTs with survival outcomes. Zhang and Rosenberger developed an optimal RAR framework for survival outcomes that is convention, but is insufficient for robust estimation. The ZR RAR approach relies upon the correct distributional specification of survival times and obtains measures of the hazards and the variance thereof via proxies.

Accuracy of the distributional assumption is not guaranteed with real-world data. Therefore, the ZR RAR framework was used to develop an estimator of the RAR ratio for survival outcomes using the cumulative hazard, the H-RAR (Chapter 4). The H-RAR was shown to be quite flexible and robust to distributional misspecification of survival times. That is, H-RAR nearly recreated the correctly-specified ZR RAR and outperformed the misspecified ZR RAR when survival times followed either the well-behaved exponential or Weibull response distributions. When survival times followed a poorly-behaved response distribution, i.e., the ECPH distribution, H-RAR achieved and maintained randomization as intended by the trial design more quickly than ZR RAR intended for survival times following either exponential or Weibull survival times (R code for simulated trial provided in Appendix A.30). Results thus far utilize the most simplistic estimator of the cumulative hazard function, the Nelson-Aalen estimator. Future work may focus on moving into more complex estimation methods and considering alternative motivating data beyond the oft-used Jones et al. trial.

5.2.1 Advantages and Disadvantages of Optimal RAR

There are a number of advantages and disadvantages to employing RAR.[57,58] Advantages to using RAR center about its ability to balance ethics and power. Optimal RAR is as powerful as the commonly-implemented ER, but targets an ethical objective. With this, optimal RAR is able to discern a better treatmet group while participants are being randomized into the trial, such that, if neither treatment is proving to be superior, randomization simply remains equal. With this, ideally, participant enrollment will be higher and attrition will be lower because a larger number of participants are undergoing the better treatment if one is being evidenced.

Disadvantages of the optimal RAR design include its requirement of a lead-in group to be equally-randomized into the study prior to the initiation of RAR. In some cases, the lead-in period may be small (e.g., three participants per group). Other times, the lead-in may be quite large,

depending on the method of optimal RAR being employed. For this reason, RCTs using RAR may only be advantageous when many participants can be recruited quickly. In this same vein, RAR should only be adopted when the period of time between participant enrollment and the obtainment of response information is small. Otherwise, the lead-in period will inherently be extended: a large number of participants will be enrolled into the trial under ER while the clinical trialists are waiting to accrue enough results information from those enrolled to begin RAR. Finally, estimation of the RAR ratio is performed after each subject enrolled into the trial. Such constant estimation procedures may be computationally and logistically taxing, requiring complex computing methods and more personnel training − all of which adds financial expense to the trial design.

5.2.2 Limitations and Future Work

In the present work, unimodal response distributions were evaluated for each method. This assumption may not be appropriate for any of these methods. For example, consider the scenario in which the habits of the medical professionals providing the trial intervention (e.g., training exercise, medications, etc.) to patients change over time. This is a likely scenario given human behaviors change over time, without excepting behaviors pertaining to the progress of the RCT. With this, greater than one mode may be observed from the first to the i^{th} participant enrolled into the trial. Where this is a moderately common occurrence, the assumption of unimodality collapses.[58] Furthermore, the present work assumed both the treatment and control groups followed the same response distribution versus situations where the treatment followed one response distribution and the control response data followed another. This is another distributional assumption that may not be appropriate.

In future work, the assumptions of distributional unimodality and between-group similarity may need to be re-addressed. Other limitations to the present work to be addressed in future iterations include the incorporation of multi-site context, the consideration of participant drop-out or missing

information, the incorporation of a lag between trial enrollment and response data obtainment. And, as alluded to previously, alternative estimation methods and/or distributional assumptions may need to be considered in an attempt to replicate or clarify the results presented herein.

Appendix

A.1 Link Function-Based Randomization Designs

Response-adaptive randomized designs estimate a treatment difference based on available data.[3] This treatment difference plays an integral role in the link function, denoted Λ, a function that maps an estimated treatment difference to the inclusive unit such that a treatment difference of zero elicits a randomization probability of one-half and mapping is symmetric.[3] If an estimated treatment difference, d, yields an allocation probability of ρ_d, then $-d$ should yield an allocation probability of $1 - \rho_d$.

Any CDF that is symmetric about zero would offer a reasonable link function. The normal and the logistic CDFs are two convenient examples since both have preexisting, well-defined link functions (the probit link and the logit link, respectively), both are symmetric, and both may be symmetric about zero by simple manipulation of their respective location parameters.

Under the link function-based design, subject $\{i+1\}$ is allocated to T with probability $\Lambda(\hat{\mu}_{Ti} - \hat{\mu}_{Ci})$ and to C with probability $1 - \Lambda(\hat{\mu}_{Ti} - \hat{\mu}_{Ci}) = \Lambda(\hat{\mu}_{Ci} - \hat{\mu}_{Ti})$, where $\hat{\mu}_{gi}$ is the mean of the first $\{1, 2, \cdots, i\}$ subjects that were allocated to treatment g.[3] Then, the treatment with the more desirable responses is favored during randomization.

A.2 Continuous Adaptation of Binary-Outcome Randomization Designs

The Doubly-adaptive Biased Coin Design (DBCD) is a family of RAR schemes designed to achieve a desired randomization ratio using two arguments, the current allocation ratio as determined by the treatment responses from subjects who have already completed the trial, and the target allocation ratio.[1,3,8,9] Though originally conceived for trials evaluating binary outcomes, the DBCD can be applied for continuous or time-to-event treatment responses.[10]

In the same vein, the Continuous Drop-the-Loser (CDL) design developed by Ivanova et al.[11] is an extension of the urn-based, binary Drop-the-Loser design by Ivanova.[12] Under the CDL design, entering subjects are randomized according to a process based on draws from an urn containing $(B+1)$ balls of either type 0 or type b, where $b = 1, \cdots, B$, where B represents the number of treatments in the trial. If the response to treatment b for subject i exceeds a predefined threshold, then the type b ball used to allocate subject i is returned to the urn; otherwise, it is removed. Type 0 balls are returned immediately to the urn if drawn, and B balls are added to the urn. Over time, randomization is skewed toward the superior treatment.

Yao and Wei[13] developed a similar design for time-to-event outcomes called Randomized Play-the-Winner (RPW). The RPW scheme was shown to randomize more subjects to the better treatment with small loss of power in simulated RCTs where survival times were dichotomized.[13].

A.3 Designs Using Treatment Effect Mapping

Assuming two treatments, Treatment Effect Mapping (TEM) is defined by letting $\eta(\cdot)$ be a continuous function that maps from the real line to the inclusive unit, i.e., $\mathcal{R} \mapsto [0,1]$, such that $\eta(0) = 0.5,\ \eta(x) > 0.5$ for $x > 0$, and $\eta(x) < 0.5$ for $x < 0$.[14] If Δ denotes a measure of the true treatment effect, where $\hat{\Delta}_j$ is the observed value of Δ as determined by the treatment responses obtained from subjects $\{1, \cdots, j\}$, then the superiority of treatment over control is evidenced by $\hat{\Delta}_j > 0$, i.e., $\eta(\hat{\Delta}_j) > 0.5$. If $\hat{\Delta}_j = 0$, then the two treatments are deemed equivalent.

With this, Rosenberger[14] proposed a RAR design in which subject j is allocated to the superior treatment with probability $\eta(\hat{\Delta}_{j-1})$, where Δ is the normalized linear rank test statistic and the TEM function is given by $\eta(x) = (1+x)/2$. Under this design, if Y_i represents the response from subject i, $i = 1, \cdots, n$, among j responses, $j = i, \cdots, n$, for n subjects total, then, the rank of Y_i is given by R_{ij}, and the score function for R_{ij} is given by a_{ij}. After allocating the first two subjects according to a fair coin toss, all subsequent subjects are allocated to T with probability

$\left\{\frac{1}{2}+\frac{\sum_{j=1}^{i}a_{ji}(\delta_j-\frac{1}{2})}{2\sum_{j=1}^{i}a_{ji}^{+}}\right\}$, where $\delta_j=1$ if subject j was allocated to T, and $a_{ij}^{+}=a_{ij}I(a_{ij}>0)$, where $I(\cdot)$ is the indicator function. The values of a_{ji}^{+} for $\delta_j=1$ will be larger when subjects in T produce larger responses than subjects in C, resulting in an elevated probability for subject $\{i+1\}$ to be allocated to T, since larger ranks indicate better responses.

Rosenberger and Seshaiyer[15] also developed a TEM design using the log-rank statistic while focusing attention on TEM for survival times that were assumed to follow an exponential, Weibull, and lognormal distribution. Hallstrom et al.[16] utilized TEM using a statistic which is relative to the proportion of uncensored subjects over all subjects regardless of treatment group affiliation. This method was found to increase ethical gain, but had little effect on power.[16]

A.4 Wilcoxon-Mann-Whitney Adaptive Design

Let the responses to treatments T and C be real-valued random variables such that $Y_T\sim F(y_T)$ and $Y_C\sim G(y_C)$, $\epsilon\in[0,1]$ is the unknown location shift pairing the CDFs $F(y_T)$ and $G(y_C)$ with the relationship $G(y_C)=F(y_T-\epsilon)$. The Wilcoxon-Mann-Whitney (WMW) test[18,19] evaluates the hypothesis $H_0:\epsilon=0$ against $H_1:\epsilon>0$, where $\epsilon\in[0,1]$ is defined as follows. Define $Z_i=\epsilon_i Y_{Ti}+(1-\epsilon_i)Y_{Ci}$, where Y_{gi} represents the response to treatment g for subject i. Then, define the score function for Z_i as $\phi(Z_i,Z_j)=1$ if $(Z_i>Z_j)$ and $(\epsilon_i>\epsilon_j)$, and zero otherwise. The WMW-type test statistic[18,19] is then given by $W_i=\sum_{s=1}^{i}\sum_{j=1}^{i}\phi(Z_s,Z_j)$.

The FW urn design[20–22] process begins with an urn containing 2α balls, α balls for each treatment T and C. If a T ball is drawn from the urn, it is immediately replaced and β C balls are added. Thus, after the first two subjects are equally-allocated, the randomization probability to T for the third subject onward is given by $\left\{\frac{\alpha+\beta W_i}{2\alpha+\beta N_{Ti}N_{Ci}}\right\}$, where N_{gi} represents the number of times subjects currently allocated to treatment g. Allocating according to this process[21,22] skews the urn in favor of the superior treatment. Bandyopadhyay and Biswas[17] bridged these two concepts

by incorporating the WMW-type[18,19] test statistic W into the FW urn design[20–22], resulting in the WAD test statistic, $U_n = \frac{T_n}{N_{Tn}\, N_{Cn}}$.

A.5 Zhang and Rosenberger Derivation of Optimal RAR Ratio

Let n_g be the target sample size for treatment group g, where $g \in \{T, C\}$ such that $n_T + n_C = N$, where N represents the total number of subjects enrolled into the trial, indexed by i ($i = 1, 2, \cdots, N$). As well, let Y_g be the responses of patients assigned to treatment group g, where $Y_g \sim N(\mu_g, \sigma_g^2)$. Assuming smaller response values are more desirable, researchers may wish to minimize the total expected response value observed from patient data, thus-producing the following optimization problem of subjecting the objective function of the form:

$$\begin{aligned} R^* &= \underset{n_T/n_C}{min} \left\{ E\left[\sum_{g=T}^{C} \sum_{i=1}^{n_g} Y_{gi} \right] \right\} \\ &= \underset{n_T/n_C}{min} \left\{ E\left[\sum_{i=1}^{n_T} Y_{Ti} + \sum_{i=1}^{n_C} Y_{Ci} \right] \right\} \\ &= \underset{n_T/n_C}{min} \Big\{ n_T\, \mu_T + n_C\, \mu_C \Big\}, \end{aligned} \tag{1}$$

to a constraint on the variance:

$$V = \frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}, \tag{2}$$

for some constant V. The solution of which provides the optimal RAR ratio. To solve:

1. Set $R = n_T/n_C$ and $R^* = \overset{min}{\text{R}} \{n_T\mu_T + n_C\mu_C\}$.

2. Re-write (1) in terms of R to minimize R^* with respect to R.

   (a) Solve R for n_T and n_C:

i. For n_T:

$$R = \frac{n_T}{n_C}$$

$$Rn_C = n_T$$

$$R(N - n_T) = n_T$$

$$RN = n_T(1 + R)$$

$$\frac{R}{1+R}N = n_T \tag{3}$$

ii. For n_C:

$$R = \frac{n_T}{n_C}$$

$$\frac{n_T}{R} = n_C$$

$$\left(\frac{R}{1+R}N\right)\frac{1}{R} = n_C$$

$$\frac{1}{1+R}N = n_C \tag{4}$$

(b) Update R^*:

$$R^* = \underset{\mathrm{R}}{^{min}} \left\{ \frac{R}{1+R}N\mu_T + \frac{1}{1+R}N\mu_C \right\} \tag{5}$$

3. Solve for N by inserting (3) and (4) into (2), then update (5). Biswas and Bhattacharya suggest taking $V = 1$ for simplicity and without loss of generality.

(a) Solve for N:

$$\begin{aligned} V &= \frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C} \\ 1 &= \sigma_T^2 \frac{1+R}{RN} + \sigma_C^2 \frac{1+R}{N} \\ &= \frac{1+R}{N}\left(\frac{\sigma_T^2}{R} + \sigma_C^2\right) \\ N &= (1+R)\left(\frac{\sigma_T^2}{R} + \sigma_C^2\right) \end{aligned} \tag{6}$$

(b) Update R^*:

$$\begin{aligned} R^* &=^{min}_{\text{R}} \left\{ \frac{R}{1+R}(1+R)\left(\frac{\sigma_T^2}{R} + \sigma_C^2\right)\mu_T + \frac{1}{1+R}(1+R)\left(\frac{\sigma_T^2}{R} + \sigma_C^2\right)\mu_C \right\} \\ &=^{min}_{\text{R}} \left\{ \left(\frac{\sigma_T^2}{R} + \sigma_C^2\right)(R\mu_T + \mu_C) \right\} \end{aligned} \tag{7}$$

4. Minimize objective criterion within (7) by taking derivative with respect to R and setting equal to zero.

(a) Derivative:

$$\frac{\partial}{\partial R}\left\{ \left(\frac{\sigma_T^2}{R} + \sigma_C^2\right)(R\mu_T + \mu_C) \right\} = -\frac{\mu_C \sigma_T^2}{R^2} + \mu_T \sigma_C^2$$

(b) Set equal to zero:

$$\begin{aligned} \frac{\mu_C \sigma_T^2}{R^2} &= \mu_T \sigma_C^2 \\ R &= \frac{\sigma_T \sqrt{\mu_C}}{\sigma_C \sqrt{\mu_T}} \end{aligned} \tag{8}$$

Since $R = n_T/n_C$ and the value of R that minimizes the total expected response from all patients is $R = \frac{\sigma_T\sqrt{\mu_C}}{\sigma_C\sqrt{\mu_T}}$, it is plain to see that $n_T = \sigma_T\sqrt{\mu_C}$ and $n_C = \sigma_C\sqrt{\mu_T}$, i.e.,

$$R = \frac{n_T}{n_C} = \frac{\sigma_T\sqrt{\mu_C}}{\sigma_C\sqrt{\mu_T}} \Rightarrow \begin{cases} n_T = \sigma_T\sqrt{\mu_C} \\ \\ n_C = \sigma_C\sqrt{\mu_T} \end{cases} \tag{9}$$

Where ρ is the proportion of patients randomized to the treatment group T, using (9),

$$\begin{aligned}\rho &= \frac{n_T}{N} \\ &= \frac{n_T}{n_T + n_C} \\ &= \frac{\sigma_T\sqrt{\mu_C}}{\sigma_T\sqrt{\mu_C} + \sigma_C\sqrt{\mu_T}}.\end{aligned}$$

A.6 Jennison and Turnbill Design

Assume two treatment groups T and C, where groups are denoted by g, i.e., $g = \{T, C\}$, and group-specific responses have a mean μ_g and variance σ_g^2. Jennison and Turnbill[23] developed a constrained optimal RAR framework assuming treatment variances, σ_g^2, were known. The goal of this design was to minimize the expected value of the following objective function:

$$\underset{n_T/n_C}{min} \left\{ n_T a^{max(\mu_T - \mu_C, 0)/\delta} + n_C a^{max(\mu_C - \mu_T, 0)/\delta} \right\}, \tag{10}$$

where a is a predetermined constant, δ is the treatment difference for which the study was powered, and n_g is the number of subjects allocated to treatment g. Solving this according to the variance of the estimated treatment difference in (1.2) gives the following optimal randomization ratio to T:

$$\rho_{JT} = R/(1 + R), \tag{11}$$

where,

$$R = \frac{\sigma_T}{\sigma_C}\sqrt{\frac{a^{max(\mu_C - \mu_T, 0)/\delta}}{a^{max(\mu_T - \mu_C, 0)/\delta}}}. \tag{12}$$

Atkinson and Biswas[3] identified that, though this designis straightforward, it is based upon a poorly-defined importance criterion. As a result, it is unclear what the objective function (10) aims to minimize, damaging design interpretation.

A.7 Biswas and Mandal Design

For treatments T and C, Biswas and Mandal[5] assumed treatment-specific responses have a mean μ_g and variance σ_g^2, where σ_T^2 and σ_C^2 are assumed to be unknown and smaller responses more desirable. Furthermore, Biswas and Mandal[5] sought to minimize the total number of responses greater than an investigator-defined threshold constant, y_0. A large-enough response could be interpreted as a failure. Therefore, the goal of this design is to minimize the total number of expected failures. Therefore, subjecting the objective function:

$$\underset{n_T/n_C}{min} \left[n_T \, \Phi\left(\frac{\mu_T - y_0}{\sigma_T}\right) + n_C \, \Phi\left(\frac{\mu_C - y_0}{\sigma_C}\right)\right] \tag{13}$$

to the variance constraint of (1.2) produces the optimal randomization rule to T given by:

$$\rho_{BM} = \frac{\sigma_T \sqrt{\Phi\left(\frac{\mu_C - y_0}{\sigma_C}\right)}}{\sigma_T \sqrt{\Phi\left(\frac{\mu_C - y_0}{\sigma_C}\right)} + \sigma_C \sqrt{\Phi\left(\frac{\mu_T - y_0}{\sigma_T}\right)}}. \tag{14}$$

A.8 Algorithm 1: Trial Generation.

Trial generation:

1. Generate N random variates per group (Y_{Ti} and Y_{Ci}, $i = \{1, \cdots, N\}$).

2. Generate N uniform random variates ($u_i \sim U(0,1)$, $i = \{1, \cdots, N\}$).

3. For lead-in period $i \in \{1, \cdots, L\}$, if $u_i < 0.5$, subject response data was Y_{Ti}, else Y_{Ci}.

4. For subjects $\{L+1, \cdots, N\}$, obtain estimates for randomization: because smaller responses are more desirable, if $u_i < \hat{\rho}$, subject responses was Y_{Ti}, else Y_{Ci}.

A.9 Algorithm 2: X2U Response Data.

Simulation of X2U response data:

1. Generate N random variates per distribution ($X_i \sim \chi^2(k)$ and $U_i \sim U(a, b)$, for $i = \{1, \cdots, N\}$).

2. Generate N uniform random variates ($u_i \sim U(0, 1)$, for $i = \{1, \cdots, N\}$).

3. For mixing parameter π, if $u_i < \pi$, subject response was X_i, else U_i.

A.10 Algorithm 3: NM Response Data.

Simulation of response data following a mixture of two normal random variables, $N_1 \sim N(\mu_1,\ \sigma_1^2)$ and $N_2 \sim N(\mu_2,\ \sigma_2^2)$, for a treatment difference, δ:

1. Denote control group response data N_C such that it is comprised of $N_{1C} \sim N(\mu_{1C},\ \sigma_{1C}^2)$ and $N_{2C} \sim N(\mu_{2C},\ \sigma_{2C}^2)$, and treatment group response data as N_T such that it is comprised of $N_{1T} \sim N(\mu_{1T},\ \sigma_{1T}^2)$ and $N_{2T} \sim N(\mu_{2T},\ \sigma_{2T}^2)$.

2. Set control group mean parameters, μ_{1C} and μ_{2C}. Set all SD for equal, $\sigma_{1T} = \sigma_{2T} = \sigma_{1C} = \sigma_{2C}$. And, set the mixing parameter $\pi = 0.5$.

3. Find control group mean and variance:

   $E(N_C) = \pi\ \mu_{1C} +\ (1 - \pi)\ \mu_{2C}$,

   $E(N_C^2) = \pi\ (\sigma_{1C}^2\ +\ \mu_{1C}^2)\ +\ (1 - \pi)\ (\sigma_{2C}^2\ +\ \mu_{2C}^2)$, and,

   $Var(N_C) = E(N_C^2) - E(N_C)^2$.

4. Set $Var(N_T) = Var(N_C)$.

5. Solve for $E(N_T)$ using (2.31).

6. Set $\epsilon = 0.5 \times (\mu_{2C} - \mu_{1C})$.

7. Then, $\mu_{1T}\ =\ E(N_T)\ -\ \epsilon$ and $\mu_{2T}\ =\ E(N_T)\ +\ \epsilon$.

A.11 Derivation of Normal MLEs

Let Y be normally-distributed with unknown mean μ and variance σ^2. The Y has the following PDF:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} exp\Big[-\frac{1}{2}\Big(\frac{y-\mu}{\sigma}\Big)^2\Big],\ y \in \mathbb{R},$$

Consider y_i, $i = 1, \cdots, n$, an independent random sample from a normal distribution. The loglikelihood of μ and σ given the observed data is maximized to obtain the MLEs for μ and σ, $\bar{y}$ and s, respectively.[24] The likelihood is given by:

$$\ell(\mu, \sigma|\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\Big[-\frac{1}{2}\Big(\frac{y_i-\mu}{\sigma}\Big)^2\Big].$$

Taking the logarithm gives:

$$log\ell(\mu, \sigma|\mathbf{y}) = -\frac{n}{2}log(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Setting $\frac{\partial}{\partial\mu} log\ell(\mu, \sigma|\mathbf{y}) = 0$ gives:

$$\begin{aligned}\hat{\mu} &= \sum_{i=1}^{n} y_i/n \\ &= \bar{y},\end{aligned}$$

and setting $\frac{\partial}{\partial\sigma} log\ell(\mu, \sigma|\mathbf{y}) = 0$ gives:

$$\begin{aligned}\hat{\sigma} &= \sqrt{\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{n}}\Bigg|_{\mu=\hat{\mu}} \\ &= \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}} \\ &= s.\end{aligned}$$

With this, for observed continuous response data Y_g ($g = \{T, C\}$), the treatment-specific estimates of the mean and standard deviation are:

$$\{\hat{\mu}_g, \hat{\sigma_g}\} = \{\bar{y}_g, s_g\},$$

when response data follow the normal distribution.

A.12 Derivation of Gamma MLEs

Let Y follow a gamma distribution with shape parameter α and scale parameter β. The PDF of Y is:

$$f_Y(y) = \frac{y^{\alpha-1}exp(-y/\beta)}{\Gamma(\alpha)\beta^\alpha},\ y \in \mathbb{R}^+,\ 0 \text{ otherwise.}$$

For y_i, $i = 1, \cdots, n$, an independent gamma-distributed random sample, the MLEs for α and β, $\hat{\alpha}$ and $\hat{\beta}$, respectively, are obtained by maximizing the loglikelihood of the parameters given the observed data.[24] The likelihood of α and β is:

$$\ell(\alpha, \beta|\mathbf{y}) = \prod_{i=1}^{n} \frac{y_i^{\alpha-1}exp(-y/\beta)}{\Gamma(\alpha)\beta^\alpha},$$

and the loglikelihood is:

$$log\ell(\alpha, \beta|\mathbf{y}) = n(\alpha - 1)\overline{log(y)} - \frac{n}{\beta}\bar{y} - nlog\Gamma(\alpha) - n\alpha log(\beta).$$

Setting $\frac{\partial}{\partial\beta} log\ell(\alpha, \beta|\mathbf{y}) = 0$ gives:

$$\hat{\beta} = \bar{y}/\alpha\Big|_{\alpha=\hat{\alpha}},$$

while $\frac{\partial}{\partial\alpha} log\ell(\alpha, \beta|\mathbf{y}) = 0$ gives the equation:

$$0 = \overline{log(y)} - log(\beta) - \frac{\partial}{\partial\alpha} log\Gamma(\alpha)\Big|_{\beta=\hat{\beta}},$$

where $\hat{\alpha}$ and $\hat{\beta}$ are solved numerically. With this, for observed continuous response data Y_g ($g = \{T, C\}$), the treatment-specific estimates of the mean and standard deviation are:

$$\{\hat{\mu}_g, \hat{\sigma_g}\} = \{\hat{\alpha}_g\hat{\beta}_g,\ \hat{\alpha}_g\hat{\beta}_g^2\},$$

when response data follow a gamma distribution.

A.13 Derivation of χ^2-Uniform Mixture MLEs

Let $X \sim \chi^2_k$ and $U \sim U(a, b)$ ($k, a, b \in \mathbb{R}^+$) where $E(X) = k$ and $Var(X) = 2k$ and $E(U) = \frac{a+b}{2}$ and $Var(U) = \frac{(b-a)^2}{12}$ and $X \perp U$. Now take $Y = \pi X + (1-\pi)U$ for mixing parameter π ($\pi \in (0, 1)$). Then,

$$\begin{aligned} E(Y) &= \pi E(X) + (1-\pi)E(U) \\ &= \pi k + (1-\pi)\frac{a+b}{2}, \end{aligned}$$

and,

$$\begin{aligned} Var(Y) &= \pi^2 Var(X) + (1-\pi)^2 Var(U) + 2\pi(1-\pi)Cov(X, U) \\ &= \pi^2 2k + (1-\pi)^2\frac{(b-a)^2}{12},\ \text{because } X \perp U \\ &= \pi^2\left(2k + \frac{(b-a)^2}{12}\right) + (1-2\pi)\frac{(b-a)^2}{12}. \end{aligned}$$

For a random sample y_i, $(i = 1, \cdots, n)$, estimates of the parameters characterizing the χ^2-Uniform mixture distribution, i.e., $\boldsymbol{\theta} = (\pi, k, a, b)$, given by $\hat{\boldsymbol{\theta}}$, are obtained through the maximization of the likelihood. Where Y has the PDF:

$$f_Y(y) = \pi f_X(x|k) + (1-\pi)f_U(u|a, b), \tag{15}$$

for $x \in \mathbb{R}^+$ and $u \in \mathbb{R}$ (0, otherwise), the corresponding likelihood is given by:

$$\ell(k, a, b|\boldsymbol{y}) = \prod_{i=1}^{n}\left[\pi f_X(x_i|k) + (1-\pi)f_U(u_i|a, b)\right] \tag{16}$$

such that

$$\ell(k|\boldsymbol{y}) = \prod_{i=1}^{n}\Big[\pi f_X(x_i|k)\Big]$$
$$\ell(a,b|\boldsymbol{y}) = \prod_{i=1}^{n}\Big[(1-\pi) f_U(u_i|a,b)\Big].$$

With this, the loglikelihood of k is given by:

$$\begin{aligned} log\ell(k|\boldsymbol{y}) &= \sum_{i=1}^{n} log\left(\pi\frac{x_i^{k/2-1}exp(-x_i/2)}{2^{k/2}\Gamma(k/2)}\right) \\ &= \sum_{i=1}^{n}\left[log(\pi) + \left(\frac{k}{2}-1\right)log(x_i) - \frac{x_i}{2} - \frac{k}{2}log(2) - log(\Gamma(k/2))\right] \\ &= \sum_{i=1}^{n}\left[\frac{k}{2}log(x_i) - \frac{k}{2}log(2) - log(\Gamma(k/2))\right]. \end{aligned}$$

Setting

$$\frac{d}{dk}log\ell(k|\boldsymbol{y}) = \frac{1}{2}\left[\sum_{i=1}^{n} log(x_i) - nlog(2)\right] - n\frac{d}{dk}log(\Gamma(k/2))$$

equal to zero gives

$$\frac{1}{2}\left[\overline{log(x)} - log(2)\right] = \frac{d}{dk}log(\Gamma(k/2))$$

to be solved numerically to determine $\hat{k}$. Likewise, the loglikelihood of a and b is given by:

$$log\ell(a,b|\boldsymbol{y}) = \sum_{i=1}^{n} log\left(\frac{1-\pi}{b-a}\right)\mathbf{1}_{(a\leq u_i\leq b)}.$$

Setting

$$\frac{\partial}{\partial a}log\ell(a,b|\boldsymbol{y}) = \frac{n}{b-a}$$
$$\frac{\partial}{\partial b}log\ell(a,b|\boldsymbol{y}) = -\frac{n}{b-a},$$

equal to zero gives $\hat{a} = min(u_i)$ and $\hat{b} = max(u_i)$. Finally, for $\hat{\pi}$, take p to be a Bernoulli random variable with probability parameter π such that:

$$p = \begin{cases} 1 \text{ with probability } \pi, & \text{if } y_i \sim \chi^2_k, \\ 0 \text{ with probability } 1-\pi, & \text{if } y_i \sim U(a,b). \end{cases}$$

With this, the PDF of Y given by (15) can be re-written as:

$$f_Y(y) = \Big[\pi f_X(x|k)\Big]^p + \Big[(1-\pi) f_U(u|a,b)\Big]^{1-p},$$

such that the likelihood given in (16) may take the form:

$$\ell(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{p}) = \prod_{i=1}^{n}\Big[\pi f_X(x_i|k)\Big]^{p_i} + \prod_{i=1}^{n}\Big[(1-\pi) f_U(u_i|a,b)\Big]^{1-p_i},$$

and the loglikelihood is given by:

$$log\ell(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{p}) = \sum_{i=1}^{n}\Big\{p_i log(\pi f_X(x_i|k)) + (1-p_i) log((1-\pi) f_U(u_i|a,b))\Big\},$$

such that the loglikelihood of π is given by:

$$log\ell(\pi|\boldsymbol{y},\boldsymbol{p}) = \sum_{i=1}^{n}\Big\{p_i log(\pi) + (1-p_i) log(1-\pi)\Big\}.$$

Where $q = \sum_{i=1}^{n} p_i$, setting

$$\frac{d}{d\pi} log\ell(\pi|\boldsymbol{y},\boldsymbol{p}) = \frac{q}{\pi} - \frac{n-q}{1-\pi}$$

equal to zero gives $\hat{\pi} = q/n$. From this, group-specific estimates of the mean and standard deviations of the χ^2-Uniform mixture distribution based upon the observed continuous response data Y_g ($g = \{T, C\}$) are:

$$\hat{\mu}_g = \hat{\pi}_g \hat{k}_g + (1-\hat{\pi}_g)\frac{\hat{a}_g + \hat{b}_g}{2}$$
$$\hat{\sigma_g} = \left\{\hat{\pi}_g^2\left(2\hat{k}_g + \frac{(\hat{b}_g - \hat{a}_g)^2}{12}\right) + (1-2\hat{\pi}_g)\frac{(\hat{b}_g - \hat{a}_g)^2}{12}\right\}^{1/2}.$$

A.14 Derivation of Normal-Mixture MLEs

Let $X_j \sim N(\mu_j, \sigma_j^2)$ for $j = 1, 2$ where $E(X_j) = \mu_j$ ($\mu_j \in \mathbb{R}$) and $Var(X_j) = \sigma_j^2$ ($\sigma_j \in \mathbb{R}^+$) and $X_1 \perp X_2$. Now let $Y = \alpha X_1 + (1-\alpha)X_2$ for mixing parameter α ($\alpha \in (0,1)$). Then,

$$\begin{aligned} E(Y) &= \alpha E(X_1) + (1-\alpha)E(X_2) \\ &= \alpha\mu_1 + (1-\alpha)\mu_2, \end{aligned}$$

and,

$$\begin{aligned} Var(Y) &= \alpha^2 Var(X_1) + (1-\alpha)^2 Var(X_2) + 2\alpha(1-\alpha)Cov(X_1, X_2) \\ &= \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2, \text{ because } X_1 \perp X_2 \\ &= \alpha^2(\sigma_1^2 + \sigma_2^2) + \sigma_2^2(1-2\alpha). \end{aligned}$$

For a random sample y_i, $(i = 1, \cdots, n)$, estimates of the parameters characterizing the mixture of normals distribution, i.e., $\boldsymbol{\theta} = (\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, given by $\hat{\boldsymbol{\theta}}$, are obtained through the maximization of the likelihood. Where Y has the PDF:

$$f_Y(y) = \frac{\alpha}{\sqrt{2\pi}\sigma_1} exp\bigg[-\frac{1}{2}\bigg(\frac{x_1-\mu_1}{\sigma_1}\bigg)^2\bigg] + \frac{1-\alpha}{\sqrt{2\pi}\sigma_2} exp\bigg[-\frac{1}{2}\bigg(\frac{x_2-\mu_2}{\sigma_2}\bigg)^2\bigg],\ x_j \in \mathbb{R};\ 0, \text{ otherwise}, \tag{17}$$

loglikelihoods are given by:

$$\begin{aligned} log\ell(\mu_1, \sigma_1^2|\boldsymbol{y}) &= \sum_{i=1}^{n}\bigg[log(\alpha) - \frac{1}{2}log(2\pi) - log(\sigma_1) - \frac{1}{2}\bigg(\frac{x_{1i}-\mu_1}{\sigma_1}\bigg)^2\bigg] \\ &= -\sum_{i=1}^{n}\bigg[log(\sigma_1) + \frac{1}{2}\bigg(\frac{x_{1i}-\mu_1}{\sigma_1}\bigg)^2\bigg], \text{ when } y_i \sim f_{X_1}, \text{ and,} \\ log\ell(\mu_2, \sigma_2^2|\boldsymbol{y}) &= -\sum_{i=1}^{n}\bigg[log(\sigma_2) + \frac{1}{2}\bigg(\frac{x_{2i}-\mu_2}{\sigma_2}\bigg)^2\bigg], \text{ when } y_i \sim f_{X_2}. \end{aligned}$$

Setting $log\ell(\mu_1, \sigma_1^2|\boldsymbol{y})$ and $log\ell(\mu_2, \sigma_2^2|\boldsymbol{y})$ equal to zero gives the MLEs $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ derived in Appendix A.11. For the derivation of $\hat{\alpha}$, take p to be a Bernoulli random variable with probability

parameter α, i.e., $p \sim Bern(\alpha)$, such that:

$$p = \begin{cases} 1 \text{ with probability } \alpha, & \text{if } y_i \sim f_{X_1}, \\ 0 \text{ with probability } 1-\alpha, & \text{if } y_i \sim f_{X_2}. \end{cases}$$

Then re-write the PDF of (17) as:

$$f_Y(y) = \Big[\alpha f_{X_1}(x_1)\Big]^p + \Big[(1-\alpha) f_{X_2}(x_2)\Big]^{1-p},$$

such that the likelihood may take the form:

$$\ell(\alpha|\boldsymbol{y},\boldsymbol{p}) = \prod_{i=1}^{n}\Big[\alpha f_{X_1}(x_{1i})\Big]^{p_i} + \prod_{i=1}^{n}\Big[(1-\alpha) f_{X_2}(x_{2i})\Big]^{1-p_i},$$

and the loglikelihood is given by:

$$\begin{aligned} log\ell(\alpha|\boldsymbol{y},\boldsymbol{p}) &= \sum_{i=1}^{n}\Big\{p_i log(\alpha f_{X_1}(x_{1i})) + (1-p_i) log((1-\alpha) f_{X_2}(x_{2i}))\Big\} \\ &= \sum_{i=1}^{n}\Big\{p_i log(\alpha) + (1-p_i) log(1-\alpha)\Big\}. \end{aligned}$$

Then, where

$$\frac{d}{d\alpha} log\ell(\alpha|\boldsymbol{y},\boldsymbol{p}) = \frac{\sum_{i=1}^{n} p_i}{\alpha} - \frac{\sum_{i=1}^{n}(1-p_i)}{1-\alpha},$$

setting equal to zero gives $\hat{\alpha} = \sum_{i=1}^{n} p_i/n$. From this, treatment-specific estimates of the mean and standard deviation of the NM distribution based upon observed continuous response data Y_g ($g = \{T, C\}$) are:

$$\hat{\mu}_g = \hat{\alpha}_g \hat{\mu}_{1g} + (1-\hat{\alpha}_g)\hat{\mu}_{2g}$$

$$\hat{\sigma}_g = \{\hat{\alpha}_g(\hat{\sigma}^2_{1g} + \hat{\sigma}^2_{2g}) + \hat{\sigma}^2_{2g}(1-2\hat{\alpha}_g)\}^{1/2}.$$

A.15 Lead-in Analysis Results for Continuous Response Data

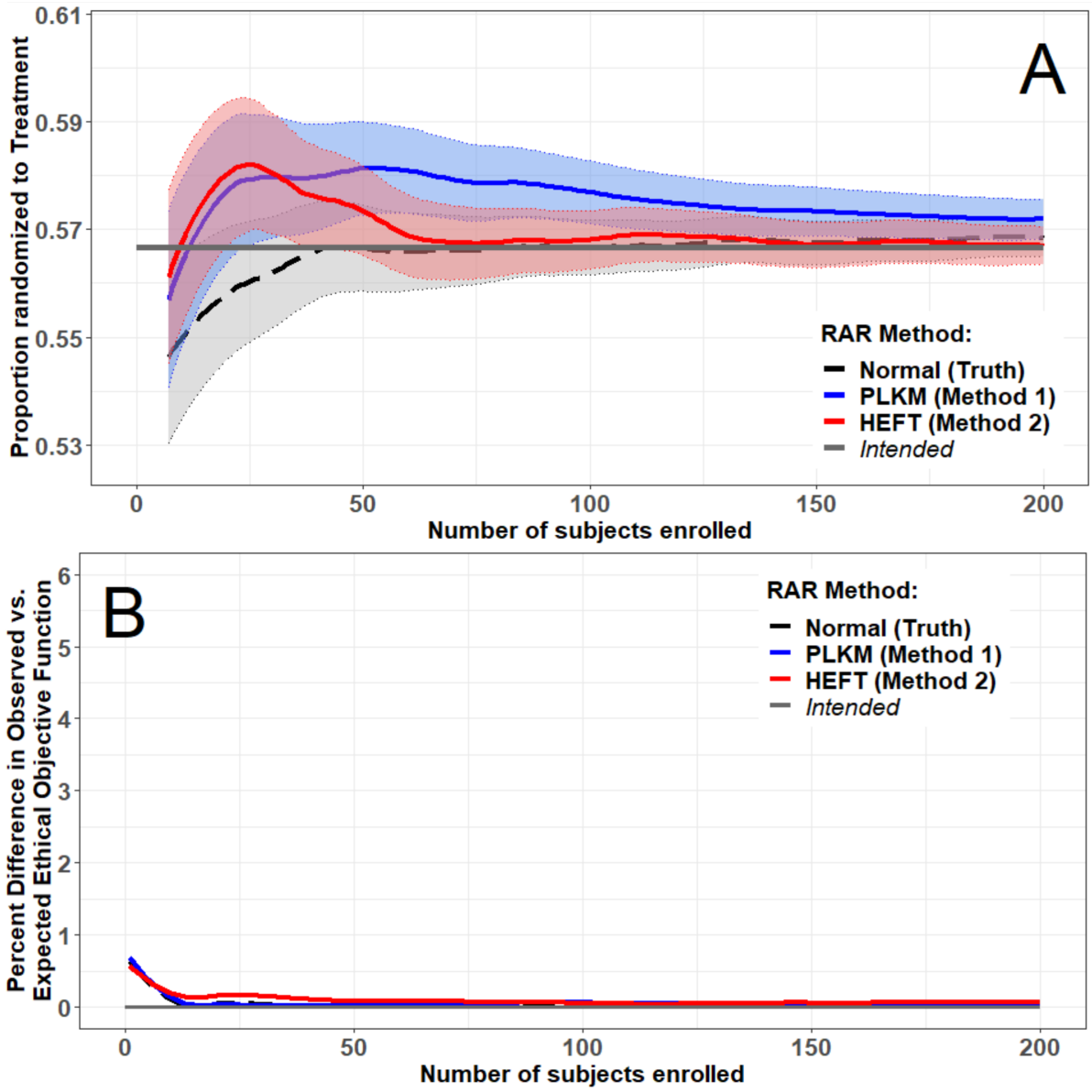
Table A.15: Observed percent of successful trials over increasing LPGs for HEFT randomization of Normal and Gamma response data and mixture randomization of mixture-distributed response data.

	$\hat{\%}$success for HEFT RAR when Truth = Normal				$\hat{\%}$success for HEFT RAR when Truth = Gamma			
LPG	$\delta = 0.0$	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$	$\delta = 0.0$	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
3	68.9	71.5	73.2	75.8	71.8	76.1	77.0	77.6
4	65.7	65.6	70.0	74.7	69.7	70.8	72.7	77.1
5	63.3	62.7	68.3	72.1	64.6	67.3	70.2	76.0
6	61.0	61.3	67.2	70.9	62.6	67.6	70.3	74.7
7	62.3	66.5	66.8	69.9	69.8	69.4	73.3	74.9
8	64.7	68.5	70.4	75.4	71.9	73.6	74.4	74.7
9	64.0	64.3	68.7	75.7	70.1	72.2	75.9	79.9
10	64.1	65.0	69.9	77.2	69.7	69.6	75.7	79.8
11	61.8	64.7	69.9	80.3	65.7	68.4	72.5	78.3
12	64.5	64.5	71.7	80.4	68.1	65.5	73.8	80.2
13	65.7	67.8	70.9	83.3	64.0	67.7	74.3	82.8
14	72.3	70.3	77.3	83.5	66.3	66.9	83.0	86.4
15	80.6	81.5	80.4	86.1	77.1	76.3	84.9	89.5
16	89.3	89.7	91.3	91.9	89.7	91.6	92.4	91.7
17	88.1	88.7	90.8	89.3	88.7	88.1	89.0	93.8
18	91.5	91.5	93.1	92.5	93.7	93.8	94.5	93.5
19	90.1	90.1	91.1	91.8	93.3	100.0	100.0	100.0
20	90.7	92.4	93.4	92.4	92.5	93.1	93.1	92.0

	$\hat{\%}$success for X2U RAR when Truth = X2U				$\hat{\%}$success for NM RAR when Truth = NM			
LPG	$\delta = 0.0$	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$	$\delta = 0.0$	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
3	35.4	39.1	41.9	43.9	0.3	1.2	0.9	1.3
4	57.8	60.5	62.5	64.9	16.8	17.3	15.7	15.1
5	72.7	72.2	74.9	78.5	41.4	41.7	40.7	38.0
6	81.7	82.5	84.3	85.4	60.0	62.3	62.4	62.0
7	88.8	88.1	89.1	91.6	76.7	77.9	74.6	77.1
8	94.1	92.5	94.3	95.1	86.7	86.3	85.9	86.2
9	94.5	93.5	96.1	96.1	92.2	93.2	92.1	91.3
10	95.9	95.7	97.2	98.1	96.1	95.4	96.7	96.9

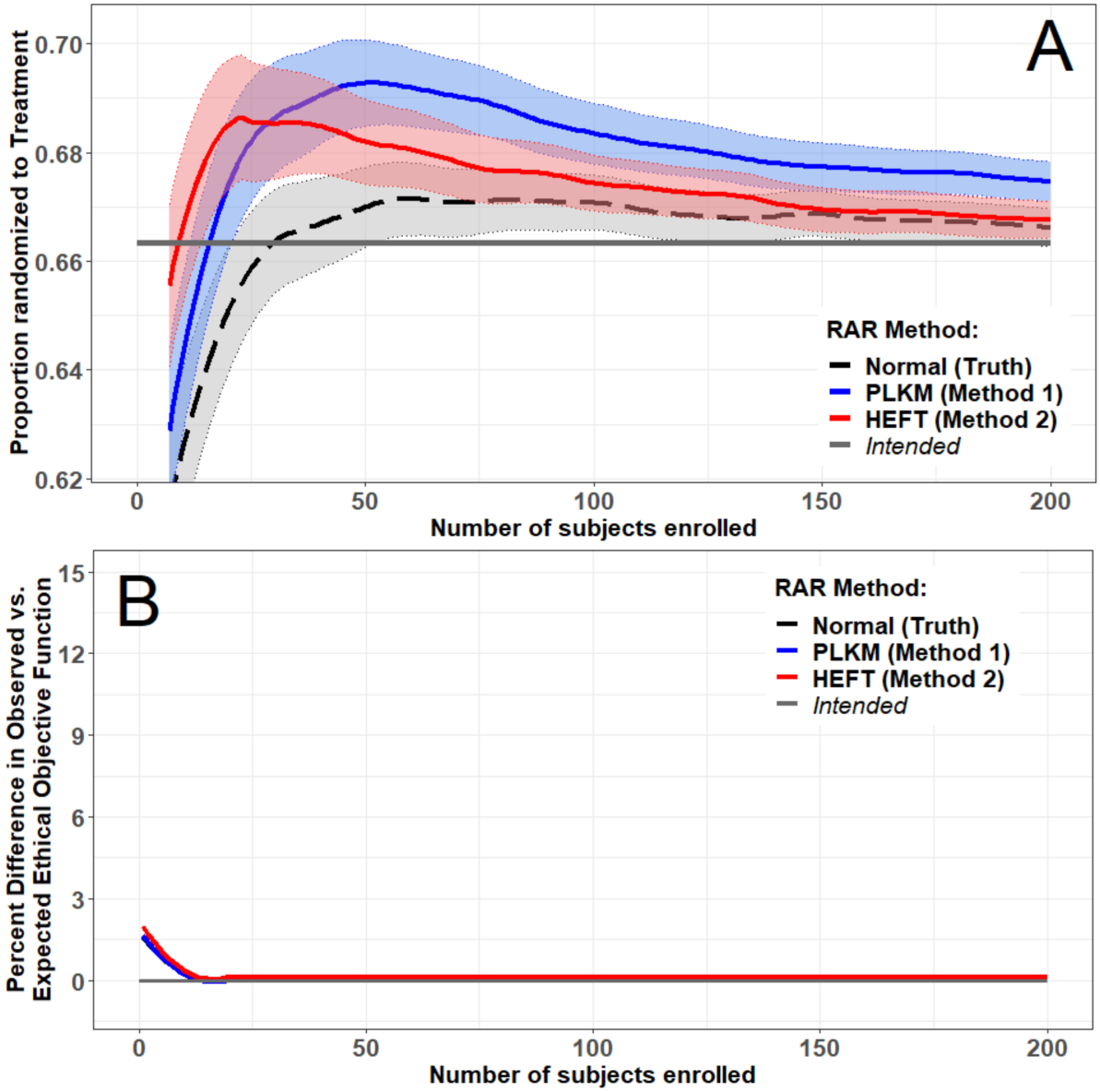
A.16 Randomization results when Truth=Normal and $\delta = 0.2$

Figure A.16: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials when the normality assumption was true and a small treatment difference existed.



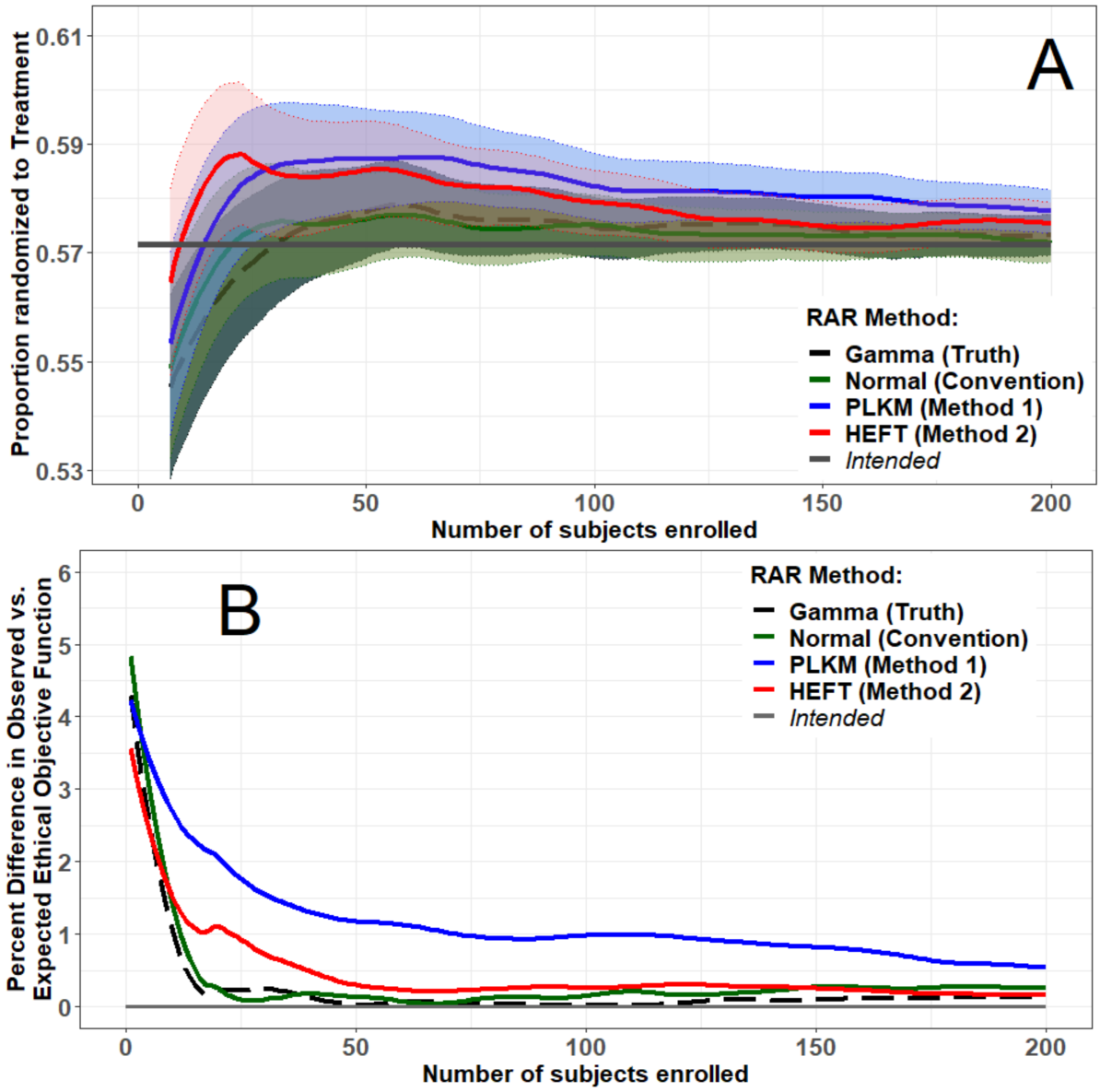
A.17 Randomization results when Truth=Normal and $\delta = 0.5$

Figure A.17: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials when the normality assumption was true and a moderate treatment difference existed.



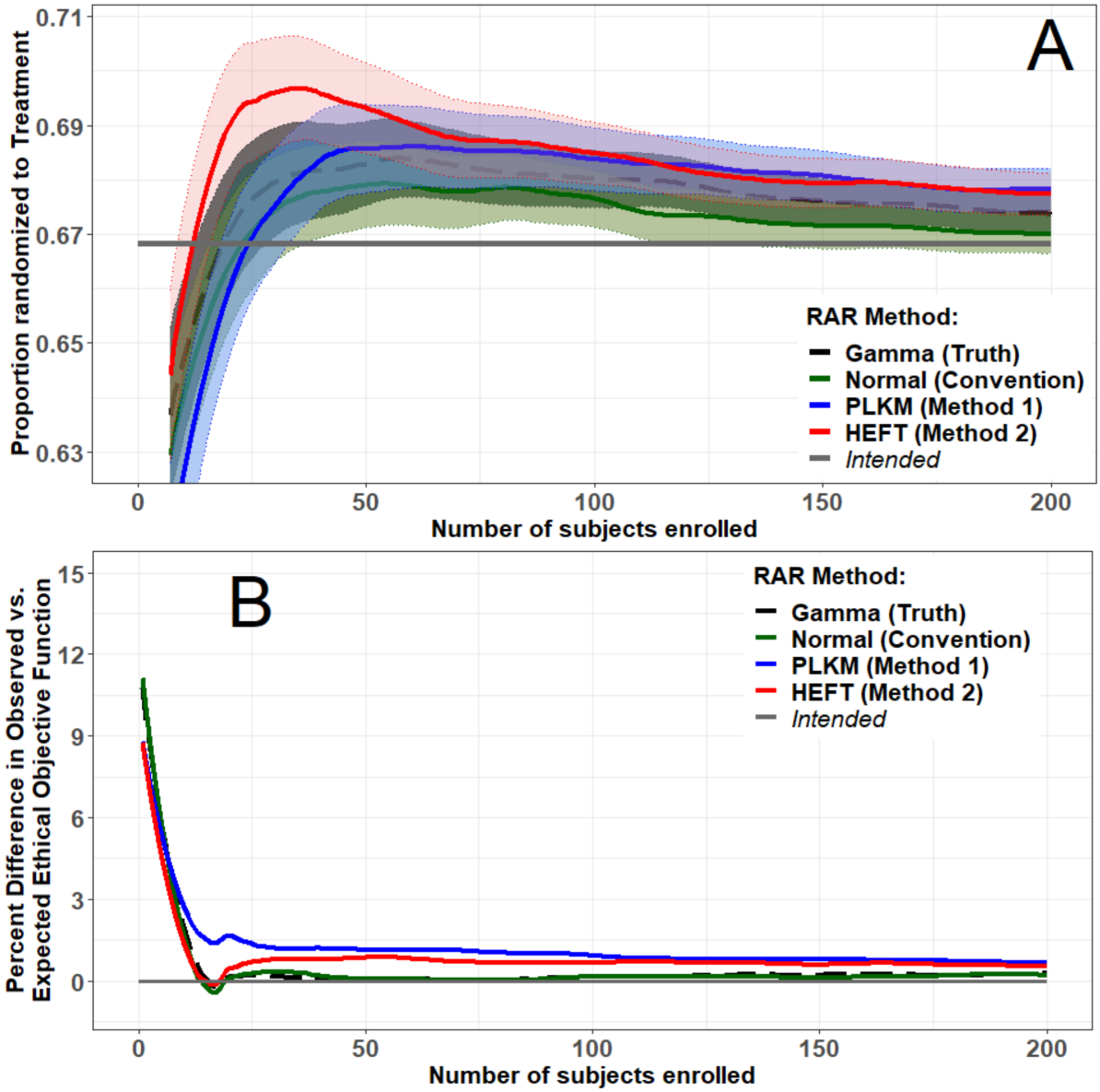
A.18 Randomization results when Truth=Gamma and $\delta = 0.2$

Figure A.18: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where group-specific subject responses followed the gamma distribution and reflected a small difference.



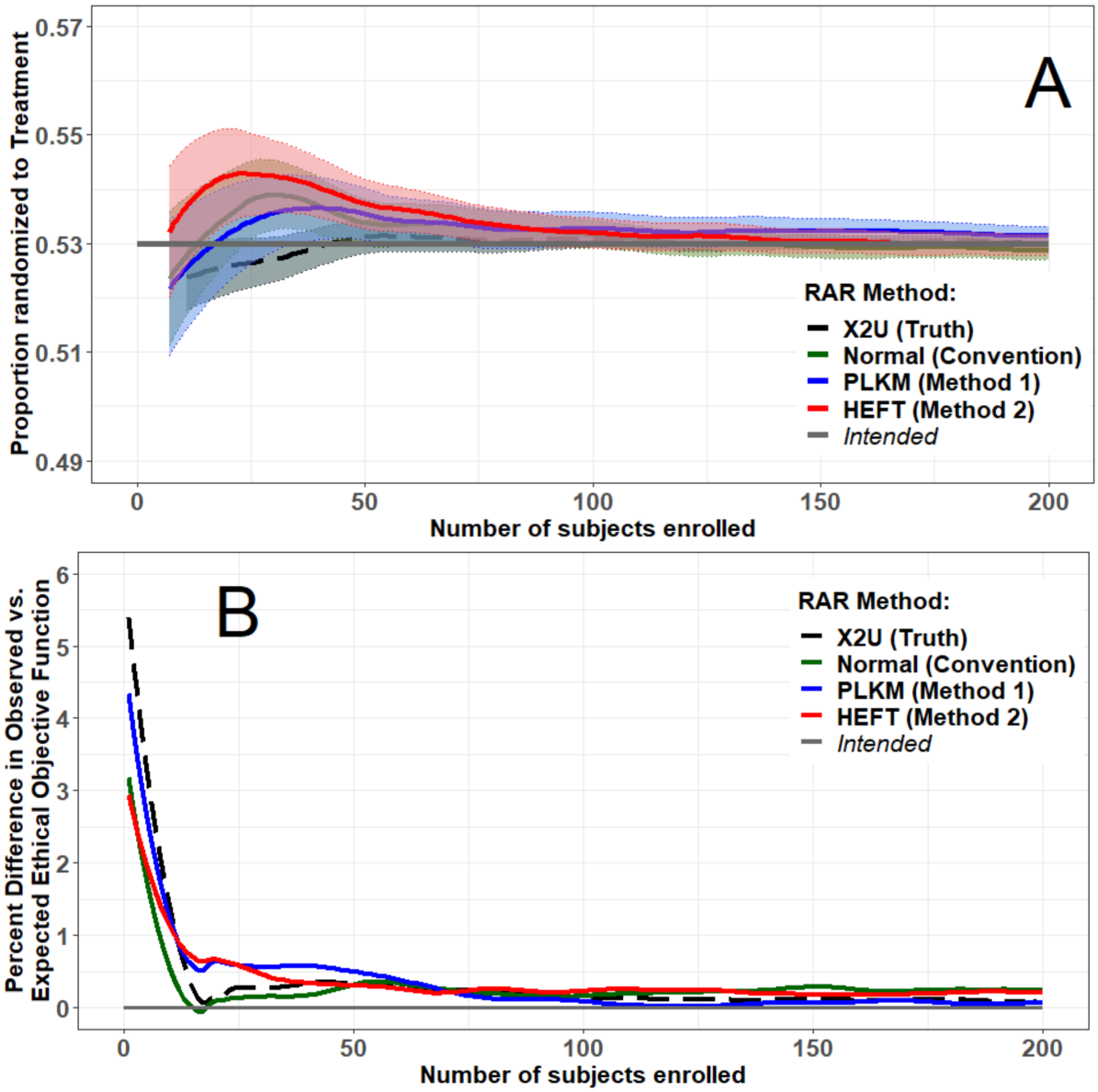
A.19 Randomization results when Truth=Gamma and $\delta = 0.5$

Figure A.19: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where group-specific subject responses followed the gamma distribution and reflected a moderate difference.



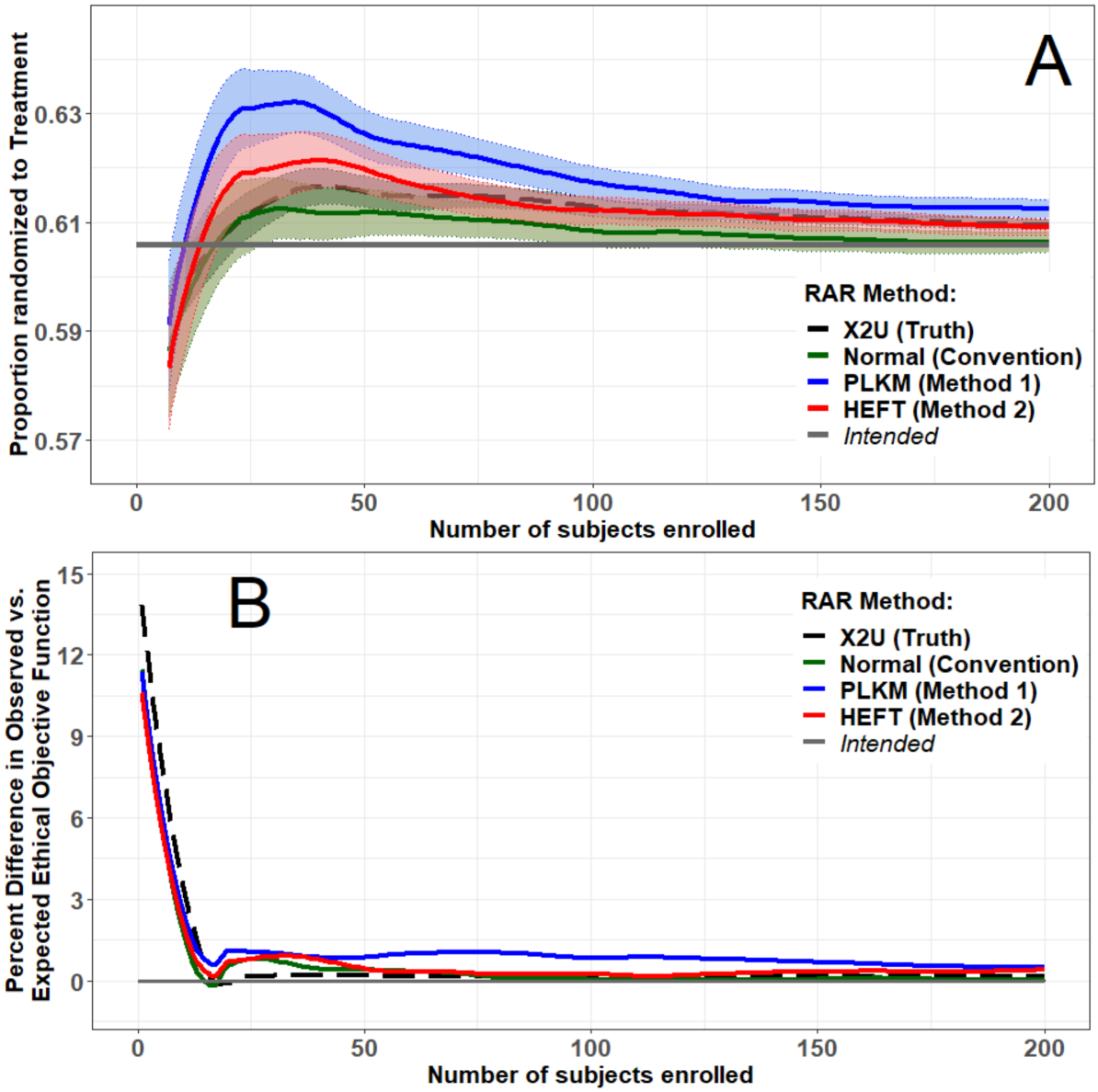
A.20 Randomization results when Truth=X2U and $\delta = 0.2$

Figure A.20: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where group-specific subject responses followed the X2U distribution and reflected a small difference.



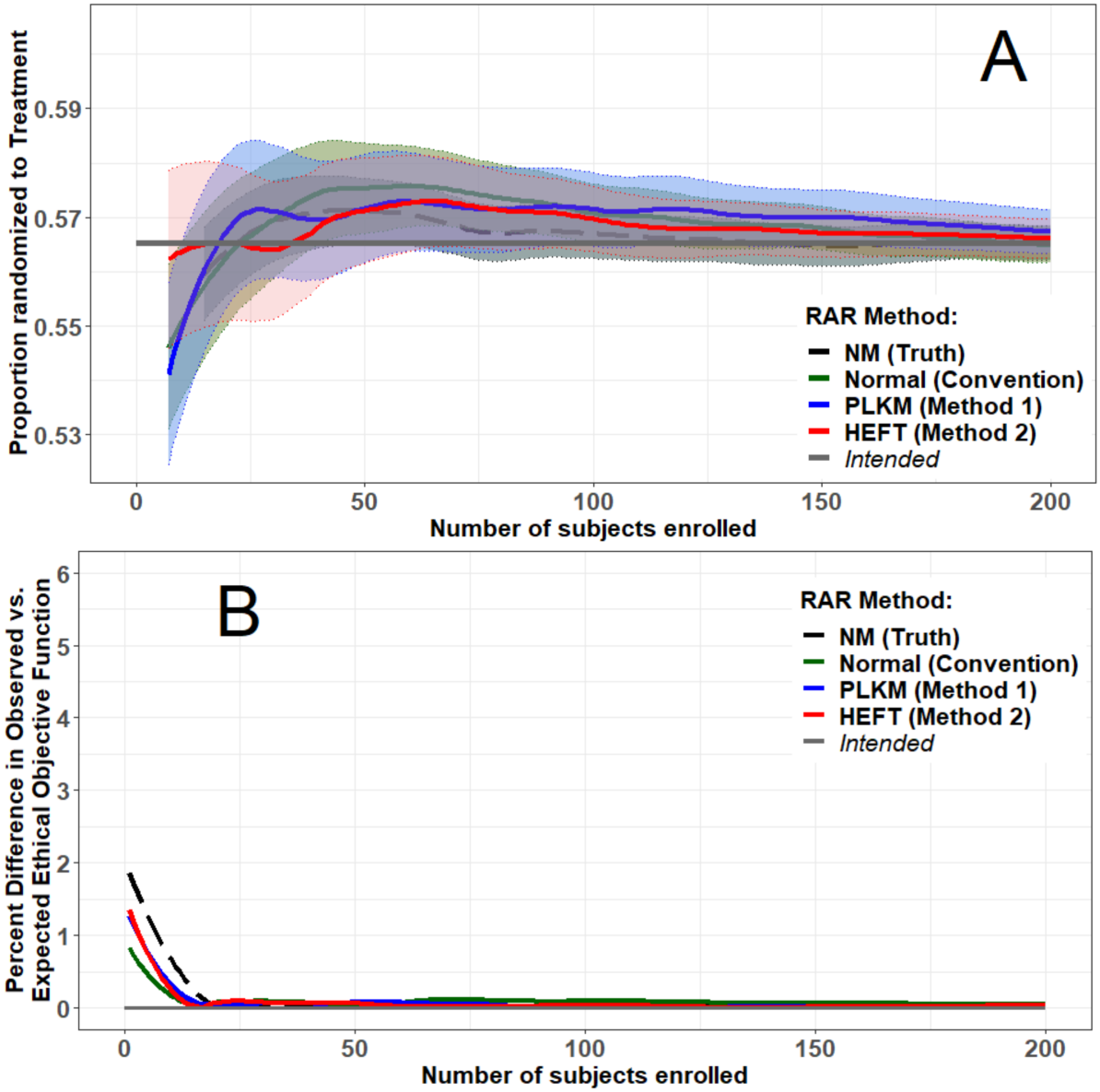
A.21 Randomization results when Truth=X2U and $\delta = 0.5$

Figure A.21: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where group-specific subject responses followed the X2U distribution and reflected a moderate difference.



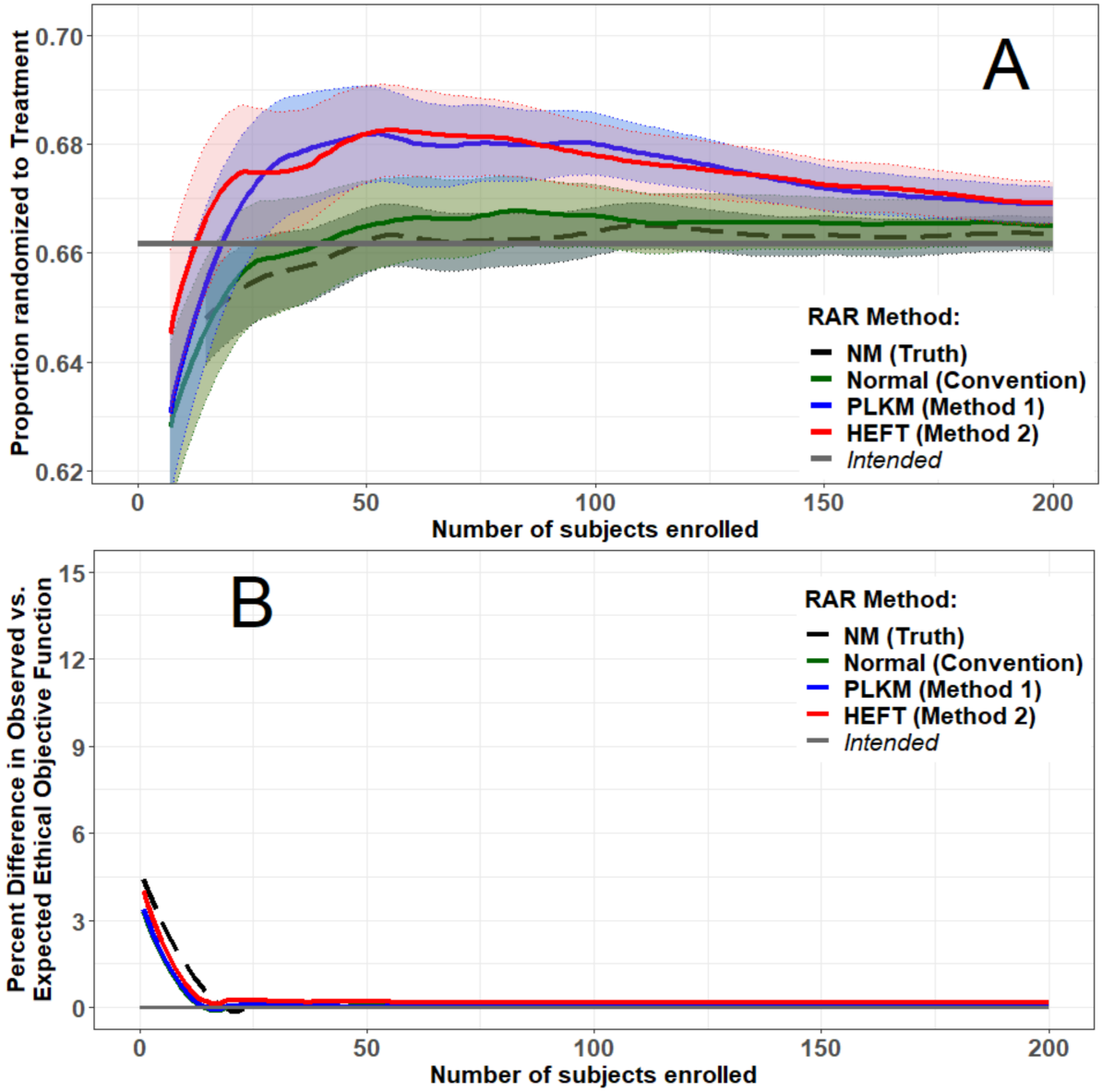
A.22 Randomization results when Truth=NM and $\delta = 0.2$

Figure A.22: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where group-specific subject responses followed the NM distribution and reflected a small difference.



A.23 Randomization results when Truth=NM and $\delta = 0.5$

Figure A.23: Mean of RAR ratios with 95% CI (A) and percent-difference between the observed and expected ethical objective function (B) over 1,000 simulated trials where group-specific subject responses followed the NM distribution and reflected a moderate difference.



A.24 Derivation of Laplace MLEs

Let Y follow a Laplace distribution with location parameter a and scale parameter b. The PDF of Y is:

$$\begin{aligned} f_Y(y) &= \frac{1}{2b} exp(-|y_i - a|/b) \\ &= \frac{1}{2b} exp\left(-1/b\left[(a-y)I_{\{y<a\}} + (y-a)I_{\{y\geq a\}}\right]\right), y \in \mathbb{R}, \ \ 0 \text{ otherwise.} \end{aligned}$$

For y_i, $i = 1, \cdots, n$, an independent Laplace-distributed random sample, the MLEs for a and b, $\hat{a}$ and $\hat{b}$, respectively, are obtained by maximizing the loglikelihood of the parameters given the observed data.[24] The likelihood of a and b is:

$$\ell(a, b|\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{2b} exp\left(-1/b\left[(a-y_i)I_{\{y_i<a\}} + (y_i-a)I_{\{y_i\geq a\}}\right]\right),$$

and the loglikelihood is:

$$\begin{aligned} log\ell(a, b|\mathbf{y}) &= -nlog(2b) - 1/b \sum_{i=1}^{n} \left[(a-y_i)I_{\{y_i<a\}} + (y_i-a)I_{\{y_i\geq a\}}\right] \\ &= \begin{cases} -nlog(2b) - 1/b \sum_{i=1}^{n}(a-y_i), \text{ if } y_i < a, \\ -nlog(2b) - 1/b \sum_{i=1}^{n}(y_i-a), \text{ if } y_i \geq a. \end{cases} \end{aligned}$$

Take the partial derivative with respect to a of $log\ell(a, b|\mathbf{y})$ yields:

$$\frac{\partial}{\partial a} log\ell(a, b|\mathbf{y}) = \begin{cases} -1/b \sum_{i=1}^{n}(a-y_i), \text{ if } y_i < a, \\ -1/b \sum_{i=1}^{n}(y_i-a), \text{ if } y_i \geq a. \end{cases}$$

Setting $\frac{\partial}{\partial a} log\ell(a, b|\mathbf{y}) = 0$ gives:

$$\begin{cases} 0 = \frac{\partial}{\partial a} \sum_{i=1}^{n}(a-y_i), \text{ if } y_i < a, \\ 0 = \frac{\partial}{\partial a} \sum_{i=1}^{n}(y_i-a), \text{ if } y_i \geq a. \end{cases}$$

Therefore, the MLE of a is the median.[24] Now, taking the derivative of $log\ell(a, b|\mathbf{y})$ with respect to b gives:

$$\frac{\partial}{\partial b} - nlog(2b) - 1/b\sum_{i=1}^{n}|y_i - a| = \frac{-n}{b} + \frac{\sum_{i=1}^{n}|y_i - a|}{b^2}.$$

Setting this equal to zero yields:

$$\hat{b} = \sum_{i=1}^{n}|y_i - a|/n\Big|_{a=\hat{a}},$$

solved numerically. With this, for observed continuous response data Y_g ($g = \{T, C\}$) following a Laplace distribution, the treatment-specific estimates of the mean and standard deviation are:

$$\{\hat{\mu}_g, \hat{\sigma_g}\} = \{\hat{a}_g, \hat{b}_g\sqrt{2}\}.$$

A.25 Derivation of Logistic MLEs

Let Y follow the logistic distribution with location parameter a and scale parameter b, having the following PDF:

$$f_Y(y) = \frac{exp\Big(-\frac{y-a}{b}\Big)}{\sigma\Big[1 + exp\Big(-\frac{y-a}{b}\Big)\Big]^2}, y \in \mathbb{R}^+, \ 0 \text{ otherwise.}$$

For y_i ($i = 1, \cdots, n$), an independent logistic-distributed random sample, the MLEs of a and b, given as $\hat{a}$ and $\hat{b}$, respectively, are obtained my maximizing the loglikelihood of the parameters given the observed response data. Where the likelihood of a and b is:

$$\ell(a, b|\boldsymbol{y}) = \prod_{i=1}^{n}\frac{exp\Big(-\frac{y_i-a}{b}\Big)}{b\Big[1 + exp\Big(-\frac{y_i-a}{b}\Big)\Big]^2},$$

the loglikelihood is:

$$log\ell(a, b|\boldsymbol{y}) = \sum_{i=1}^{n}\Big\{-\frac{y_i - a}{b} - log(b) + 2log\Big[1 + exp\Big(-\frac{y_i - a}{b}\Big)\Big]\Big\}.$$

To find $\hat{a}$ and $\hat{b}$, respectively, $\frac{\partial}{\partial a} log\ell(a, b|\boldsymbol{y})$ and $\frac{\partial}{\partial b} log\ell(a, b|\boldsymbol{y})$ are:

$$\frac{\partial}{\partial a} log\ell(a, b|\boldsymbol{y}) = \frac{1}{b}\sum_{i=1}^{n}\left[n + 2\frac{exp(-\frac{y_i-a}{b})}{1+exp(-\frac{y_i-a}{b})}\right], \text{ and,}$$
$$\frac{\partial}{\partial b} log\ell(a, b|\boldsymbol{y}) = \frac{1}{b^2}\sum_{i=1}^{n}\left[(y_i - a)\left(1 + 2\frac{exp(-\frac{y_i-a}{b})}{1+exp(-\frac{y_i-a}{b})}\right)\right].$$

Then, setting $\frac{\partial}{\partial a} log\ell(a, b|\boldsymbol{y})$ and $\frac{\partial}{\partial b} log\ell(a, b|\boldsymbol{y})$ equal to zero gives the following equations to be solved numerically to derive $\hat{a}$ and $\hat{b}$, respectively:

$$-\frac{n}{2} = \sum_{i=1}^{n}\frac{exp(-\frac{y_i-a}{b})}{1+exp(-\frac{y_i-a}{b})}\Bigg|_{a=\hat{a},b=\hat{b}}, \text{ and,}$$
$$\bar{y} - a = b - \frac{2}{n}\left[\sum_{i=1}^{n} y_i\frac{exp(-\frac{y_i-a}{b})}{1+exp(-\frac{y_i-a}{b})} - a\sum_{i=1}^{n}\frac{exp(-\frac{y_i-a}{b})}{1+exp(-\frac{y_i-a}{b})}\right]\Bigg|_{a=\hat{a},b=\hat{b}}.$$

With this, for observed continuous response data Y_g ($g = \{T, C\}$), the group-specific estimates of the mean and SD are:

$$\{\hat{\mu}_g, \hat{\sigma}_g\} = \{\hat{a}_g, \hat{b}_g\pi/\sqrt{3}\},$$

when responses follow the logistic distribution.

A.26 Derivation of Weibull MLEs (Continuous)

Let Y follow a Weibull distribution with shape parameter α and scale parameter β. The PDF of Y is:

$$f_Y(y) = \frac{\alpha}{\beta^\alpha}y^{\alpha-1}exp\left[-\left(\frac{y}{\beta}\right)^\alpha\right],\ y \in \mathbb{R}^+,\ 0 \text{ otherwise.}$$

For y_i, $i = 1, \cdots, n$, an independent Weibull-distributed random sample, the MLEs for α and β, $\hat{\alpha}$ and $\hat{\beta}$, respectively, are obtained by maximizing the loglikelihood of the parameters given the observed data.[24] The likelihood of α and β is:

$$\ell(\alpha, \beta|\mathbf{y}) = \prod_{i=1}^{n}\frac{\alpha}{\beta^\alpha}y_i^{\alpha-1}exp\left[-\left(\frac{y_i}{\beta}\right)^\alpha\right],$$

and the loglikelihood is:

$$log\ell(\alpha,\beta|\mathbf{y}) = nlog(\alpha) - n\alpha log(\beta) + \alpha\sum_{i=1}^{n} log(y_i) - log(y_i) - \sum_{i=1}^{n}\frac{y_i^\alpha}{\beta^\alpha}$$

Setting $\frac{\partial}{\partial\beta}log\ell(\alpha,\beta|\mathbf{y}) = 0$ gives:

$$\hat{\beta} = \left[\frac{\sum_{i=1}^{n}(y_i^\alpha)}{n}\right]^{1/\alpha}\Bigg|_{\alpha=\hat{\alpha}},$$

while $\frac{\partial}{\partial\alpha}log\ell(\alpha,\beta|\mathbf{y}) = 0$ gives the equation:

$$0 = \frac{n}{\alpha} - nlog(\beta) + log(y_i) - \left(\frac{y_i}{\beta}\right)^\alpha log\left(\frac{y_i}{\beta}\right)|_{\alpha=\hat{\alpha},\beta=\hat{\beta}},$$

where $\hat{\alpha}$ and $\hat{\beta}$ are solved numerically. With this, for observed continuous response data Y_g ($g = \{T, C\}$) following a Weibull distribution, the treatment-specific estimates of the mean and standard deviation are:

$$\hat{\mu}_g = \hat{\beta}_g\Gamma(1 + 1/\hat{\alpha}_g),$$
$$\hat{\sigma}_g = \hat{\beta}_g\left[\Gamma\left(1 + \frac{2}{\hat{\alpha}_g}\right) - \Gamma\left(1 + \frac{1}{\hat{\alpha}_g}\right)^2\right]^{1/2}$$

A.27 Derivation of Lognormal MLEs (Continuous)

Let Y follow a lognormal distribution with location parameter a and scale parameter b, having PDF:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}\frac{1}{yb}exp\left[-\frac{1}{2}\left(\frac{log(y) - a}{b}\right)^2\right],\ y \in \mathbb{R}^+,\ 0 \text{ otherwise.}$$

For y_i, $(i = 1, \cdots, n)$, an independent lognormally-distributed random sample, the MLEs of a and b, given as $\hat{a}$ and $\hat{b}$, respectively, are obtained my maximizing the loglikelihood of the parameters given the observed response data. Where the likelihood of a and b is:

$$\ell(a,b|\boldsymbol{y}) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\frac{1}{y_ib}exp\left[-\frac{1}{2}\left(\frac{log(y_i) - a}{b}\right)^2\right]$$

the loglikelihood is:

$$
\begin{aligned}
log\ell(a,b|\boldsymbol{y}) &= \sum_{i=1}^{n}\left[-\frac{1}{2}log(2\pi)-log(y_i)-log(b)-\frac{1}{2}\left(\frac{log(y_i)-a}{b}\right)^2\right] \\
&= \sum_{i=1}^{n}\left[log(b)+\frac{1}{2}\left(\frac{log(y_i)-a}{b}\right)^2\right].
\end{aligned}
$$

To find $\hat{a}$ and $\hat{b}$, respectively, $\frac{\partial}{\partial a}log\ell(a,b|\boldsymbol{y})$ and $\frac{\partial}{\partial b}log\ell(a,b|\boldsymbol{y})$ are:

$$
\begin{aligned}
\frac{\partial}{\partial a}log\ell(a,b|\boldsymbol{y}) &= \frac{1}{b^2}\left[\sum_{i=1}^{n}log(y_i)-na\right], \text{ and,} \\
\frac{\partial}{\partial b}log\ell(a,b|\boldsymbol{y}) &= \frac{1}{b}\left[\frac{1}{b^2}\sum_{i=1}^{n}\{log(y_i)-a\}^2-n\right].
\end{aligned}
$$

Then, setting $\frac{\partial}{\partial a}log\ell(a,b|\boldsymbol{y})$ and $\frac{\partial}{\partial b}log\ell(a,b|\boldsymbol{y})$ equal to zero gives the following equations:

$$
\begin{aligned}
\hat{a} &= \overline{log(y)} \\
\hat{b} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{log(y_i)-a\}\Big|_{a=\hat{a}} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{log(y_i)-\overline{log(y)}\}.
\end{aligned}
$$

With this, for observed continuous response data Y_g, $(g=\{T,C\})$, the group-specific estimates of the mean and SD are:

$$
\begin{aligned}
\hat{a}_g &= exp\left(\hat{a}_g+\frac{\hat{b}_g^2}{2}\right), \\
\hat{b}_g &= exp(\hat{a}_g+\hat{b}_g^2/2)\sqrt{exp(\hat{b}_g^2)-1},
\end{aligned}
$$

when responses follow the lognormal distribution.

A.28 Chapter 2 Simulated Trial R Code

```
library(polspline)

library(pracma)


seed <- as.integer(runif(1, 1, 10^7))

set.seed(seed)


type1     <- 'N'

type2     <- 'H'


N         <- 500

n.trials  <- 1000


del       <- 8

delta     <- del/10

rho.0     <- 0.75

lpg       <- 3


#############################################################################


mu.B0    <- 100

sigma.0  <- 8

mu.A0    <- mu.B0 - 0.8*sigma.0

mu.A1    <- mu.B0 - delta*sigma.0
```

```
tau    <- 2 * log(1/rho.0 - 1) / log(mu.A0 / mu.B0)

tt1    <- mu.B0^(tau/2)

tt2    <- mu.A1^(tau/2)

rho.d  <- tt1/(tt1+tt2)


##############################################################################


## 1. lead-in function -- block randomization

ld.in <- function(){

A.dat    <- rnorm(lpg*2, mean=mu.A1, sd=sigma.0)

B.dat    <- rnorm(lpg*2, mean=mu.B0, sd=sigma.0)

resp     <- c(A.dat[1:lpg], B.dat[(lpg+1):(lpg*2)])

grp      <- c(rep('A', lpg), rep('B', lpg))

rho.hat  <- 0.5

U        <- runif(lpg*2)   ## just to fill these in bc don't feel like having 'NA'


df00  <- cbind.data.frame(A.dat, B.dat, U, rho.hat, grp, resp)

lead  <- nrow(df00)


xbar.A <- xbar.B <- NULL

for (x in 1:lead){

xbar.A[x] <- mean(df00$A.dat[1:x])

xbar.B[x] <- mean(df00$B.dat[1:x])
```

```
}

s.A <- s.B <- rep(NA, lead)
for (s in 2:lead){
s.A[s] <- sd(df00$A.dat[1:s])
s.B[s] <- sd(df00$B.dat[1:s])
}

n.A <- n.B <- NULL
for (a in 1:lead){
n.A[a] <- sum(df00$grp[1:a]=='A')
n.B[a] <- sum(df00$grp[1:a]=='B')
}

df0 <- cbind.data.frame(df00, xbar.A, xbar.B, s.A, s.B, n.A, n.B)

df.NA <- as.data.frame(matrix(NA, nrow=N-lead, ncol=ncol(df0)))
colnames(df.NA) <- colnames(df0)

df <- as.data.frame(rbind(df0, df.NA))

df$A.dat[(lead+1):N] <- rnorm(N-lead, mean=mu.A1, sd=sigma.0)
df$B.dat[(lead+1):N] <- rnorm(N-lead, mean=mu.B0, sd=sigma.0)
df$U[(lead+1):N] <- runif(N-lead)
```

```
return(list(df=df, lead=lead))
}



## 2. randomization process
rand <- function(mean.A, sd.A, mean.B, sd.B, cc=0.01){
psi.A   <- mean.A^(tau/2)
psi.B   <- mean.B^(tau/2)

rho.obs   <- (sd.A * psi.B) / ((sd.A * psi.B) + (sd.B * psi.A))

rho.hat   <- ifelse(psi.A > 0 & psi.B > 0 & rho.obs < cc,
cc,
ifelse(psi.A > 0 & psi.B > 0 & cc <= rho.obs & rho.obs <= 1-cc,
rho.obs,
ifelse(psi.A > 0 & psi.B > 0 & rho.obs > 1-cc,
1-cc,
ifelse(psi.A > 0 & psi.B < 0,
cc,
ifelse(psi.A < 0 & psi.B > 0,
1-cc,
ifelse(psi.A < 0 & psi.B < 0
& sd.A/sd.B < sqrt(psi.A/psi.B),
```

```
1-cc,

cc))))))

return(rho.hat)

}




ld <- ld.in()

#dat <- ld$df$resp[which(ld$df$grp=='A')]

#### 3. HEFT modeling and estimation

HEFT.fn    <- function(dat){

hft.fit  <- heft(dat)


inf.sub  <- 1000

intv     <- seq(0, max(dat), length.out = inf.sub)


heft.Ft <- try(pheft(q=intv, fit=hft.fit), silent=T)

if(is(heft.Ft, "try-error")){

ex.heft <- sd.heft <- NA

}

else {

y1  <- 1-heft.Ft

y2  <- intv*y1


area <- area2 <- NULL
```

```
for (i in 1:(inf.sub-1)) {

xt1       <- c(intv[i], intv[i+1])

area[i]   <- trapz(x = xt1, y = c(y1[i], y1[i+1]))

area2[i]  <- trapz(x = xt1, y = c(y2[i], y2[i+1]))

}

ex.heft <- sum(area)

sd.heft <- sqrt(2*sum(area2) - ex.heft^2)

}

return(list(ex.heft=ex.heft, sd.heft=sd.heft))

}
```

```
##############################################################################

##############################################################################

##############################################################################
```

```
rh <- Gp <- Obs <- N.A <- N.B <- matrix(NA, ncol=n.trials, nrow=N)

SD.A <- SD.B <- Mean.A <- Mean.B <- matrix(NA, ncol=n.trials, nrow=N)


for (j in 1:n.trials) {


ld <- ld.in()
```

```
## completing the trial data-frame for remaining N-lead subjects
for (i in (ld$lead+1):N) {

A.df <- subset(ld$df, grp=="A")$A.dat
B.df <- subset(ld$df, grp=="B")$B.dat

logA <- capture.output({
A.heft <- HEFT.fn(dat=A.df)
})
logB <- capture.output({
B.heft <- HEFT.fn(dat=B.df)
})

if(is.na(A.heft$ex.heft) | is.na(A.heft$ex.heft)){
break
} else {
ld$df$rho.hat[i] <- rand(mean.A=A.heft$ex.heft, sd.A=A.heft$sd.heft,
mean.B=B.heft$ex.heft, sd.B=B.heft$sd.heft)

ld$df$grp[i]   <- with(ld$df, ifelse(U[i] < rho.hat[i], "A", "B"))
ld$df$resp[i]  <- with(ld$df, ifelse(grp[i]=="A", A.dat[i], B.dat[i]))

ld$df$n.A[i]  <- sum(ld$df$grp[1:i]=='A', na.rm=T)
ld$df$n.B[i]  <- sum(ld$df$grp[1:i]=='B', na.rm=T)
```

```
ld$df$s.A[i]  <- A.heft$sd.heft

ld$df$s.B[i]  <- B.heft$sd.heft


ld$df$xbar.A[i]  <- A.heft$ex.heft

ld$df$xbar.B[i]  <- B.heft$ex.heft

}

}


rh[,j]   <- ld$df$rho.hat

Gp[,j]   <- ld$df$grp

Obs[,j]  <- ld$df$resp


N.A[,j]  <- ld$df$n.A

N.B[,j]  <- ld$df$n.B


Mean.A[,j]  <- ld$df$xbar.A

Mean.B[,j]  <- ld$df$xbar.B


SD.A[,j]  <- ld$df$s.A

SD.B[,j]  <- ld$df$s.B


print(j)

}
```

A.29 Chapter 3 Simulated Trial R Code

```
library(stats4)


seed <- as.integer(runif(1, 1, 10^7))

set.seed(seed)


true      <- 'N'

crd1      <- 'La'

crd2      <- 'Lo'

crd3      <- 'G'

crd4      <- 'N'

MOF       <- 'AIC'


K         <- 4   # number of candidate response distribns for weighted average

k.mat     <- 1:K


del       <- 8

delta     <- del/10

lpg       <- 3


N         <- 200

n.trials  <- 1100
```

```
rho.0     <- 0.75

##############################################################################################

### 1. normal parameters
mu.B0     <- 100
sigma.0   <- 8
mu.A0     <- mu.B0-0.8*sigma.0#/sqrt(rho.0*(1-rho.0))
mu.A1     <- mu.B0-delta*sigma.0#/sqrt(rho.0*(1-rho.0))

rho.term  <- (1-rho.0)/rho.0
mu.term   <- mu.A0/mu.B0
sig.term  <- sigma.0/sigma.0
tau       <- 2*(log(rho.term)+log(sig.term))/log(mu.term); #tau

tt1    <- mu.B0^(tau/2)
tt2    <- mu.A1^(tau/2)
rho.d  <- tt1/(tt1+tt2)

##############################################################################################

### 2. gamma parameters
alpha.A1  <- (mu.A1/sigma.0)^2
beta.A1   <- sigma.0^2/mu.A1
```

```
alpha.B0  <- (mu.B0/sigma.0)^2
beta.B0   <- sigma.0^2/mu.B0

#############################################################################################
#############################################################################################
#############################################################################################

## 1. lead-in function true data
ld.in <- function(){
A.dat    <- rnorm(lpg*2, mean=mu.A1, sd=sigma.0)
B.dat    <- rnorm(lpg*2, mean=mu.B0, sd=sigma.0)

df00 <- cbind.data.frame(A.dat, B.dat)

eval(parse(text=paste0(paste0('df00$grp.k', k.mat, sep=' <- ', collapse=''),
"df00$grp.wa <- c(rep('A', lpg), rep('B', lpg))",
sep='', collapse='')))
eval(parse(text=paste0(paste0('df00$resp.k', k.mat, sep=' <- ', collapse=''),
"df00$resp.wa <- c(A.dat[1:lpg], B.dat[(lpg+1):(lpg*2)])",
sep='', collapse='')))
eval(parse(text=paste0('df00$U.k', k.mat, ' <- runif(lpg*2)', sep='', collapse='; ')))
df00$U.wa <- runif(lpg*2)
lead  <- nrow(df00)
```

```
xbar.A <- xbar.B <- NULL
for (x in 1:lead){
xbar.A[x] <- mean(df00$A.dat[1:x])
xbar.B[x] <- mean(df00$B.dat[1:x])
}

s.A <- s.B <- rep(NA, lead)
for (s in 2:lead){
s.A[s] <- sd(df00$A.dat[1:s])
s.B[s] <- sd(df00$B.dat[1:s])
}

n.A <- n.B <- NULL
for (a in 1:lead){
n.A[a] <- sum(df00$grp.wa[1:a]=='A')
n.B[a] <- sum(df00$grp.wa[1:a]=='B')
}

eval(parse(text=paste0('df00$xbar.', c('A', 'B'), '.wa <- xbar.', c('A', 'B'), sep='', collapse
eval(parse(text=paste0('df00$s.', c('A', 'B'), '.wa <- s.', c('A', 'B'), sep='', collapse='; '
eval(parse(text=paste0('df00$n.', c('A', 'B'), '.wa <- n.', c('A', 'B'), sep='', collapse='; '

grd2 <- cbind.data.frame(gp=c(rep('A', K), rep('B', K)), nm=rep(k.mat, 2))
```

```
eval(parse(text=paste0('df00$xbar.', grd2$gp, '.k', grd2$nm, ' <- xbar.', grd2$gp, sep='', coll
eval(parse(text=paste0('df00$s.', grd2$gp, '.k', grd2$nm, ' <- s.', grd2$gp, sep='', collapse=
eval(parse(text=paste0('df00$n.', grd2$gp, '.k', grd2$nm, ' <- n.', grd2$gp, sep='', collapse=

df.NA <- as.data.frame(matrix(NA, nrow=N-lead, ncol=ncol(df00)))
colnames(df.NA) <- colnames(df00)
df <- as.data.frame(rbind(df00, df.NA))

df$A.dat[(lead+1):N] <- rnorm(N-lead, mean=mu.A1, sd=sigma.0)
df$B.dat[(lead+1):N] <- rnorm(N-lead, mean=mu.B0, sd=sigma.0)
df$U.wa[(lead+1):N]  <- runif(length((lead+1):N))
eval(parse(text=paste0('df$U.k', k.mat, '[(lead+1):N] <- runif(N-lead)')))
df$d.0 <- c(rep(0, lead), rep(NA, N-lead))

eval(parse(text=paste0(paste0('df$rho.hat.k', k.mat, sep='', collapse=' <- '),
' <- c(rep(0.5, lead), rep(NA, N-lead))')))
eval(parse(text=paste0(paste0('df$d.k', k.mat, sep='', collapse=' <- '), ' <- ',
paste0('df$d.k', k.mat, '.star', sep='', collapse=' <- '),
' <- d.0 <- c(rep(0, lead), rep(NA, N-lead))')))
eval(parse(text=paste0('df$w.numr.k', k.mat, ' <- exp(-0.5*df$d.k', k.mat, '.star)', sep='', co
eval(parse(text=paste0('df$w.den <- with(df, ',
paste0('w.numr.k', k.mat, sep='', collapse=' + '), ')')))
eval(parse(text=paste0('df$w.k', k.mat, ' <- with(df, w.numr.k', k.mat, '/w.den)', sep='', coll
eval(parse(text=paste0('df$rho.hat.wa <- with(df, ',
```

```
paste0('rho.hat.k', k.mat, '*w.k', k.mat, sep='', collapse=' + '), ')')))


return(list(df=df, lead=lead))

}


##############################################################################################


## 6. randomization process

rand <- function(mean.A, sd.A, mean.B, sd.B, cc=0.01){

psi.A   <- mean.A^(tau/2)

psi.B   <- mean.B^(tau/2)


rho.obs   <- (sd.A * psi.B) / ((sd.A * psi.B) + (sd.B * psi.A))


rho.hat   <- ifelse(psi.A > 0 & psi.B > 0 & rho.obs < cc,

cc,

ifelse(psi.A > 0 & psi.B > 0 & cc <= rho.obs & rho.obs <= 1-cc,

rho.obs,

ifelse(psi.A > 0 & psi.B > 0 & rho.obs > 1-cc,

1-cc,

ifelse(psi.A > 0 & psi.B < 0,

cc,

ifelse(psi.A < 0 & psi.B > 0,

1-cc,
```

```
ifelse(psi.A < 0 & psi.B < 0 &

sd.A/sd.B < sqrt(psi.A/psi.B),

1-cc,

cc))))))

return(rho.hat)

}


##############################################################################################


#### 7. gamma MLE function (k=1)

gam.mle.fn <- function(grp.dat, start.alpha, start.beta){


logL.gam <- function(alpha, beta) {

-sum(suppressWarnings(dgamma(grp.dat, shape=alpha, scale=beta, log=T)))

}


mle.gam1 <- try(mle(minuslogl = logL.gam,

start     = list(alpha=start.alpha, beta=start.beta),

method = "L-BFGS-B",

lower = rep(1e-05, 2), upper=rep(10^5, 2)), silent=T); #mle.gam1


if(is(mle.gam1, "try-error")){

mle.gam.mean <- mle.gam.sd <- aic.gam <- NA

}
```

```
else {

aic.gam <- AIC(mle.gam1)

mle.gam <- mle.gam1@coef


mle.gam.mean <- as.numeric(prod(mle.gam))

mle.gam.sd <- as.numeric(sqrt(mle.gam[1]*mle.gam[2]^2))

}

return(list(mle.gam.mean=mle.gam.mean, mle.gam.sd=mle.gam.sd, aic.gam=aic.gam))

}


##############################################################################################


#### 8. normal MLE function (k2)

norm.mle.fn <- function(grp.dat, start.mu, start.sigma) {


logL.norm <- function(mu, sigma){

-sum(dnorm(grp.dat, mean=mu, sd=sigma, log=T))

}


mle.norm <- try(mle(minuslogl = logL.norm,

start = list(mu = start.mu, sigma = start.sigma),

method = "L-BFGS-B",

lower = c(-10^5, 1e-05), upper=c(10^5, 10^5)), silent=T)
```

```
if(is(mle.norm, "try-error")){
mle.norm.mean <- mle.norm.sd <- aic.norm <- NA
}
else {
aic.norm <- AIC(mle.norm)
coef.norm <- as.vector(mle.norm@coef)

mle.norm.mean  <- coef.norm[1]
mle.norm.sd    <- coef.norm[2]

if(mle.norm.mean<=0 | mle.norm.sd<=0){
mle.norm.mean <- mle.norm.sd <- NA
}
}
return(list(mle.norm.mean=mle.norm.mean, mle.norm.sd=mle.norm.sd, aic.norm=aic.norm))
}

##############################################################################

#### 9. logistic MLE function (k2)
logis.mle.fn <- function(grp.dat, start.loc, start.sc){

logL.logis <- function(loc, sc){
-sum(suppressWarnings(dlogis(grp.dat, location=loc, scale=sc, log=T)))
```

```
}


mle.logis <- try(mle(minuslogl=logL.logis,

start=list(loc=start.loc, sc=start.sc),

method='BFGS'), silent=T); #mle.logis


if(is(mle.logis, 'try-error')){

mle.logis.mean <- mle.logis.sd <- aic.logis <- NA

}

else{

logis.coef <- as.vector(mle.logis@coef)

aic.logis <- AIC(mle.logis)


mle.logis.mean <- logis.coef[1]

mle.logis.sd <- logis.coef[2]*pi/sqrt(3)

}

return(list(mle.logis.mean=mle.logis.mean, mle.logis.sd=mle.logis.sd, aic.logis=aic.logis))

}

strt.scl.logis <- sigma.0*pi/sqrt(3)


##############################################################################


#### 10. laplace MLE function (k1)

lap.mle.fn <- function(grp.dat){
```

```
mle.lap.mean <- median(grp.dat)

b.hat <- sum(abs(grp.dat-mle.lap.mean))/length(grp.dat)

mle.lap.sd <- sqrt(2) * b.hat

lnlik <- -length(grp.dat)*log(2*b.hat) - sum(abs(grp.dat-mle.lap.mean))/b.hat

aic.lap <- (4-2*lnlik)


return(list(mle.lap.mean=mle.lap.mean, mle.lap.sd=mle.lap.sd, aic.lap=aic.lap))

}


##############################################################################################

##############################################################################################

##############################################################################################


## Running trial:


Rh.wa <- Obs.wa <- Grp.wa <- Unf.wa <- matrix(NA, ncol=n.trials, nrow=N)

Min.MOF <- Wt.denr <- matrix(NA, ncol=n.trials, nrow=N)

N.A.wa <- N.B.wa <- Mean.A.wa <- Mean.B.wa <- SD.A.wa <- SD.B.wa <- matrix(NA, ncol=n.trials, n


eval(parse(text=paste0(paste0('Rh.k', k.mat,

' <- MOF.k', k.mat, ' <- MOF.star.k', k.mat,

' <- Wt.numr.k', k.mat, ' <- Weight.k', k.mat,

' <- Obs.k', k.mat, ' <- Unf.k', k.mat,
```

```
' <- N.A.k', k.mat, ' <- N.B.k', k.mat,

' <- SD.A.k', k.mat, ' <- SD.B.k', k.mat,

' <- Mean.A.k', k.mat, ' <- Mean.B.k', k.mat,

sep='', collapse=' <- '),

' <- matrix(NA, ncol=n.trials, nrow=N)')))



for (j in 1:n.trials) {


#### normal randomization (k=1)

ld <- ld.in()


for (i in (ld$lead+1):N) {


A.df <- subset(ld$df, grp.wa=="A")$A.dat

B.df <- subset(ld$df, grp.wa=="B")$B.dat



##### for normal rho-hat (k=4) #####

norm.mle.A <- norm.mle.fn(grp.dat=A.df, start.mu=mu.A1, start.sigma=sigma.0)

norm.mle.B <- norm.mle.fn(grp.dat=B.df, start.mu=mu.B0, start.sigma=sigma.0)


if(is.na(norm.mle.A$mle.norm.mean) | is.na(norm.mle.A$mle.norm.sd) |

is.na(norm.mle.B$mle.norm.mean) | is.na(norm.mle.B$mle.norm.sd)) {
```

```
break
} else {
ld$df$rho.hat.k4[i] <- rand(mean.A = norm.mle.A[[1]], sd.A = norm.mle.A[[2]],
mean.B = norm.mle.B[[1]], sd.B = norm.mle.B[[2]])

ld$df$xbar.A.k4[i] <- norm.mle.A[[1]]
ld$df$xbar.B.k4[i] <- norm.mle.B[[1]]

ld$df$s.A.k4[i] <- norm.mle.A[[2]]
ld$df$s.B.k4[i] <- norm.mle.B[[2]]

ld$df$grp.k4[i]   <- with(ld$df, ifelse(U.k4[i] < rho.hat.k4[i], "A", "B"))
ld$df$resp.k4[i]  <- with(ld$df, ifelse(grp.k4[i]=="A", A.dat[i], B.dat[i]))

ld$df$n.A.k4[i]     <- sum(ld$df$grp.k4[1:i]=='A', na.rm=T)
ld$df$n.B.k4[i]     <- sum(ld$df$grp.k4[1:i]=='B', na.rm=T)
}
ld$df$d.k4[i] <- norm.mle.fn(grp.dat=ld$df$resp.wa[1:(i-1)], start.mu=(mu.A1+mu.B0)/2,
start.sigma=sigma.0)[[3]]



##### for gamma rho-hat (k=3) #####
gam.mle.A <- gam.mle.fn(grp.dat=A.df, start.alpha=alpha.A1, start.beta=beta.A1)
```

```
gam.mle.B <- gam.mle.fn(grp.dat=B.df, start.alpha=alpha.B0, start.beta=beta.B0)


if(is.na(gam.mle.A$mle.gam.mean) | is.na(gam.mle.A$mle.gam.sd) |
is.na(gam.mle.B$mle.gam.mean) | is.na(gam.mle.B$mle.gam.sd)) {
break
} else {
ld$df$rho.hat.k3[i] <- rand(mean.A=gam.mle.A$mle.gam.mean, sd.A=gam.mle.A$mle.gam.sd,
mean.B=gam.mle.B$mle.gam.mean, sd.B=gam.mle.B$mle.gam.sd)


ld$df$xbar.A.k3[i] <- gam.mle.A$mle.gam.mean
ld$df$xbar.B.k3[i] <- gam.mle.B$mle.gam.mean


ld$df$s.A.k3[i] <- gam.mle.A$mle.gam.sd
ld$df$s.B.k3[i] <- gam.mle.B$mle.gam.sd


ld$df$grp.k3[i]   <- with(ld$df, ifelse(U.k3[i] < rho.hat.k3[i], "A", "B"))
ld$df$resp.k3[i]  <- with(ld$df, ifelse(grp.k3[i]=="A", A.dat[i], B.dat[i]))


ld$df$n.A.k3[i]     <- sum(ld$df$grp.k3[1:i]=='A', na.rm=T)
ld$df$n.B.k3[i]     <- sum(ld$df$grp.k3[1:i]=='B', na.rm=T)
}
ld$df$d.k3[i] <- gam.mle.fn(grp.dat=ld$df$resp.wa[1:(i-1)],
start.alpha=(alpha.A1+alpha.B0)/2,
start.beta=(beta.A1+beta.B0)/2)[[3]]
```

```
##### for logistic rho-hat (k=2) #####
logis.mle.A <- logis.mle.fn(grp.dat=A.df, start.loc=mu.A1, start.sc=strt.scl.logis)
logis.mle.B <- logis.mle.fn(grp.dat=B.df, start.loc=mu.B0, start.sc=strt.scl.logis)

if(is.na(logis.mle.A$mle.logis.mean) | is.na(logis.mle.A$mle.logis.sd) |
is.na(logis.mle.B$mle.logis.mean) | is.na(logis.mle.B$mle.logis.sd)) {
break
} else {
ld$df$rho.hat.k2[i] <- rand(mean.A = logis.mle.A[[1]], sd.A = logis.mle.A[[2]],
mean.B = logis.mle.B[[1]], sd.B = logis.mle.B[[2]])

ld$df$xbar.A.k2[i] <- logis.mle.A[[1]]
ld$df$xbar.B.k2[i] <- logis.mle.B[[1]]

ld$df$s.A.k2[i]    <- logis.mle.A[[2]]
ld$df$s.B.k2[i]    <- logis.mle.B[[2]]

ld$df$grp.k2[i]    <- with(ld$df, ifelse(U.k2[i] < rho.hat.k2[i], "A", "B"))
ld$df$resp.k2[i]   <- with(ld$df, ifelse(grp.k2[i]=="A", A.dat[i], B.dat[i]))

ld$df$n.A.k2[i]    <- sum(ld$df$grp.k2[1:i]=='A', na.rm=T)
```

```
ld$df$n.B.k2[i]    <- sum(ld$df$grp.k2[1:i]=='B', na.rm=T)
}
ld$df$d.k2[i] <- logis.mle.fn(grp.dat=ld$df$resp.wa[1:(i-1)],
start.loc=(mu.A1+mu.B0)/2,
start.sc=strt.scl.logis)[[3]]




##### for laplace rho-hat (k=2) #####
lap.mle.A <- lap.mle.fn(grp.dat=A.df)
lap.mle.B <- lap.mle.fn(grp.dat=B.df)

if(is.na(lap.mle.A$mle.lap.mean) | is.na(lap.mle.A$mle.lap.sd) |
is.na(lap.mle.B$mle.lap.mean) | is.na(lap.mle.B$mle.lap.sd)){
break
} else {
ld$df$rho.hat.k1[i] <- rand(mean.A = lap.mle.A[[1]], sd.A = lap.mle.A[[2]],
mean.B = lap.mle.B[[1]], sd.B = lap.mle.B[[2]])

ld$df$xbar.A.k1[i] <- lap.mle.A[[1]]
ld$df$xbar.B.k1[i] <- lap.mle.B[[1]]

ld$df$s.A.k1[i]    <- lap.mle.A[[2]]
ld$df$s.B.k1[i]    <- lap.mle.B[[2]]
```

```
ld$df$grp.k1[i]    <- with(ld$df, ifelse(U.k1[i] < rho.hat.k1[i], "A", "B"))
ld$df$resp.k1[i]   <- with(ld$df, ifelse(grp.k1[i]=="A", A.dat[i], B.dat[i]))

ld$df$n.A.k1[i]    <- sum(ld$df$grp.k1[1:i]=='A', na.rm=T)
ld$df$n.B.k1[i]    <- sum(ld$df$grp.k1[1:i]=='B', na.rm=T)
}
ld$df$d.k1[i] <- lap.mle.fn(grp.dat=ld$df$resp.wa[1:(i-1)])[[3]]



##### developing weights #####
eval(parse(text=paste0('ld$df$d.0[i] <- min(', paste0('ld$df$d.k', k.mat, '[i]',
sep='', collapse=', '), ', na.rm=T)')))
eval(parse(text=paste0('ld$df$d.k', k.mat, '.star[i] <- with(ld$df, d.k', k.mat,
'[i]-d.0[i])', sep='', collapse='; ')))
eval(parse(text=paste0('ld$df$w.numr.k', k.mat, '[i] <- exp(-0.5*ld$df$d.k', k.mat,
'.star[i])', sep='', collapse='; ')))
eval(parse(text=paste0('ld$df$w.den[i] <- with(ld$df, ',
paste0('w.numr.k', k.mat, '[i]', sep='', collapse=' + '), ')')))
eval(parse(text=paste0('ld$df$w.k', k.mat, '[i] <- with(ld$df, w.numr.k', k.mat,
'[i]/w.den[i])', sep='', collapse='; ')))
```

```
##### weighted-average randomization #####
eval(parse(text=paste0('ld$df$rho.hat.wa[i] <- with(ld$df, ',
paste0('rho.hat.k', k.mat, '[i]*w.k', k.mat, '[i]',
sep='', collapse=' + '), ')')))

ld$df$xbar.A.wa[i]   <- mean(A.df)
ld$df$xbar.B.wa[i]   <- mean(B.df)

ld$df$s.A.wa[i]      <- sd(A.df)
ld$df$s.B.wa[i]      <- sd(B.df)

ld$df$grp.wa[i]   <- with(ld$df, ifelse(U.wa[i]<rho.hat.wa[i], 'A', 'B'))
ld$df$resp.wa[i]  <- with(ld$df, ifelse(grp.wa[i]=='A', A.dat[i], B.dat[i]))

ld$df$n.A.wa[i]      <- sum(ld$df$grp.wa[1:i]=='A', na.rm=T)
ld$df$n.B.wa[i]      <- sum(ld$df$grp.wa[1:i]=='B', na.rm=T)
}

Obs.wa[,j]     <- ld$df$resp.wa
Rh.wa[,j]      <- ld$df$rho.hat.wa
Grp.wa[,j]     <- ld$df$grp.wa
Unf.wa[,j]     <- ld$df$U.wa
```

```
Min.MOF[,j]    <- ld$df$d.0

Wt.denr[,j]    <- ld$df$w.den


N.A.wa[,j]     <- ld$df$n.A.wa

N.B.wa[,j]     <- ld$df$n.B.wa


Mean.A.wa[,j]  <- ld$df$xbar.A.wa

Mean.B.wa[,j]  <- ld$df$xbar.B.wa


SD.A.wa[,j]    <- ld$df$s.A.wa

SD.B.wa[,j]    <- ld$df$s.B.wa



eval(parse(text=paste0('Rh.k', k.mat, '[,j] <- ld$df$rho.hat.k', k.mat, sep='', collapse='; ')

eval(parse(text=paste0('Unf.k', k.mat, '[,j] <- ld$df$U.k', k.mat, sep='', collapse='; ')))

eval(parse(text=paste0('Obs.k', k.mat, '[,j] <- ld$df$resp.k', k.mat, sep='', collapse='; ')))


eval(parse(text=paste0('MOF.k', k.mat, '[,j] <- ld$df$d.k', k.mat, sep='', collapse='; ')))

eval(parse(text=paste0('MOF.star.k', k.mat, '[,j] <- ld$df$d.k', k.mat, '.star', sep='', colla

eval(parse(text=paste0('Wt.numr.k', k.mat, '[,j] <- ld$df$w.numr.k', k.mat, sep='', collapse='

eval(parse(text=paste0('Weight.k', k.mat, '[,j] <- ld$df$w.k', k.mat, sep='', collapse='; ')))


eval(parse(text=paste0('N.A.k', k.mat, '[,j] <- ld$df$n.A.k', k.mat, sep='', collapse='; ')))

eval(parse(text=paste0('N.B.k', k.mat, '[,j] <- ld$df$n.B.k', k.mat, sep='', collapse='; ')))
```

```
eval(parse(text=paste0('SD.A.k', k.mat, '[,j] <- ld$df$s.A.k', k.mat, sep='', collapse='; ')))

eval(parse(text=paste0('SD.B.k', k.mat, '[,j] <- ld$df$s.B.k', k.mat, sep='', collapse='; ')))


eval(parse(text=paste0('Mean.A.k', k.mat, '[,j] <- ld$df$xbar.A.k', k.mat, sep='', collapse=';

eval(parse(text=paste0('Mean.B.k', k.mat, '[,j] <- ld$df$xbar.B.k', k.mat, sep='', collapse=';


print(j)

}
```

A.30 Chapter 4 Simulated Trial R Code

```
library(survival)


set.seed(1250)


rho.1     <- 0.75

rho.0     <- 0.5

HR.1      <- 1.5


type1     <- 'E'    ## data: exp

type2     <- 'NA'   ## RAR: nelson-aalen cumulative hazard

hr.sz     <- 'L'    ## large HR

pct       <- 20     ## heavy censoring


lpg       <- 3

N         <- 1000

n.trials  <- 1000

Ntr <- '1k'

tauu <- 8.1


#=============================================================================================


rcu.pd    <- 85

R         <- 85  /rcu.pd        ### recruitment period (85 days)  / 85
```

```
D         <- 100 /rcu.pd        ### trial duration (100 days)     / 85

days      <- 30  /rcu.pd        ### interested in 30-day survival / 85

### ^^^ idk if 30/85 is really looking at 30 days


#=============================================================================================


grp.labb  <- c('C','T')

vrrs      <- c('tm.g', 'ldin', 'grp', 'A0', 'C0', 'C1', 'S0', 'S1', 'Z0')

pwr.vrs.df <- c("z.stat", "pval", "rej", "HR.obs.NA", 'HR.obs.exp')

pwr.vrs.est <- c('zz.stat', 'pvalu', 'rejj', 'hr.obs.na', 'hr.obs.exp')


#=============================================================================================


## 1. censoring function

xi.fn <- function(thta, DD, RR){

1 - thta/DD + exp(-DD/thta) * (thta/(DD*RR)) * (exp(RR/thta) * (2*thta-RR) - 2*thta)

}


#=============================================================================================


### to select control group parameters

tht.vec   <- seq(from=15.9, to=16.1, by=0.00001) /rcu.pd

xi.thtv   <- xi.fn(tht.vec, D, R)

xi.C.df0  <- data.frame(tht.vec, xi.thtv, bias=xi.thtv-(1-pct/100))
```

```
### setting control group parms
tht.C1.init  <- xi.C.df0$tht.vec[which.min(abs(xi.C.df0$bias))]
xi.C1.init   <- xi.fn(tht.C1.init, D, R)
haz.C1.init  <- 1/tht.C1.init
sd.tht.C1    <- tht.C1.init/sqrt(xi.C1.init)

### treatment group parameters
haz.T1.init  <- haz.C1.init/HR.1
tht.T1.init  <- 1/haz.T1.init
xi.T1.init   <- xi.fn(tht.T1.init, D, R)
sd.tht.T1    <- tht.T1.init/sqrt(xi.T1.init)

##=============================================================================================
##=============================================================================================

## 2. survival randomization process
rand <- function(haz.T, haz.C, sd.tht.T, sd.tht.C, cc=0.01){

psi.T <- haz.T^(tauu/2)
psi.C <- haz.C^(tauu/2)

tt1 <- sd.tht.T*psi.C
tt2 <- sd.tht.C*psi.T
```

```
rho.obs <- tt1 / (tt1 + tt2)


rho.hat <- ifelse(psi.T > 0 & psi.C > 0 & rho.obs < cc,

cc,

ifelse(psi.T > 0 & psi.C > 0 & cc <= rho.obs & rho.obs <= 1-cc,

rho.obs,

ifelse(psi.T > 0 & psi.C > 0 & rho.obs > 1-cc,

1-cc,

ifelse(psi.T > 0 & psi.C < 0,

cc,

ifelse(psi.T < 0 & psi.C > 0,

1-cc,

ifelse(psi.T < 0 & psi.C < 0

& sd.tht.T/sd.tht.C < sqrt(psi.T/psi.C),

1-cc,

cc))))))


return(rho.hat)

}


#=============================================================================================

#=============================================================================================
```

```
## 3. deriving arrival & censoring data for both groups
# step 1: generating arrival and censoring data
# for both groups st arrival times < censor time
# and adding 50/50 C/T group variable for PBD(1) lead-in
ac.dat  <- function(){
A0    <- sort(runif(N*2, 0, R))  ## U(0,1) arrival times
C0    <- runif(N*2, 0, D)
C1    <- C0+A0
grp   <- rep(grp.labb, N)
df01  <- data.frame(A0, C0, C1, grp)
return(df01)
}

#=============================================================================================
#=============================================================================================

## 4. prepping the Z1/W0/T0 matrices
byt.df <- function(df01.spl, matr, clnm.txt){
dff <- as.data.frame(matr)
clnms <- paste0('c(', paste0("'", clnm.txt, ".t", 1:(nrow(df01.spl)-1), "'", sep='', collapse=
", '", clnm.txt, ".D')", sep='', collapse='')
colnames(dff) <- eval(parse(text=clnms))
col.sums <- colSums(dff, na.rm=T)
return(list(dff=dff, col.sums=col.sums))
```

```
}

#=============================================================================================

#=============================================================================================

## 5. deriving survival times

# step 2: splitting data by group to pull grp-specific survival time

# st survival times > its corresponding arrival time

df.split <- function(df01, grp.lab, grp.tht){


df01.gp        <- subset(df01, grp==grp.lab)

df01.gp$S0     <- rexp(nrow(df01.gp), rate=1/grp.tht)

df01.gp$S1     <- with(df01.gp, S0+A0)

df01.gp$Z0     <- with(df01.gp, pmin(S1, C1, D))

df01.spl       <- df01.gp #rbind.data.frame(df01.gp, c(D, rep(NA, ncol(df01.gp)-1)))

df01.spl$tm.g  <- 1:nrow(df01.spl)


Z1.mat  <- W0.mat <- matrix(NA, nrow=nrow(df01.spl)-1, ncol=nrow(df01.spl))

Cnt.mat <- T0.mat <- matrix(NA, nrow=nrow(df01.spl)-1, ncol=nrow(df01.spl))

for(n. in 2:nrow(df01.spl)){

cnt0  <- with(df01.spl, 1*(A0[n.]>=Z0[1:(n.-1)]))

z1    <- with(df01.spl, ifelse(A0[n.]<Z0[1:(n.-1)], A0[n.], Z0[1:(n.-1)]))

w0    <- with(df01.spl, 1*(z1==S1[1:(n.-1)]))

t0    <- with(df01.spl, z1-A0[1:(n.-1)])
```

```
Cnt.mat[,n.]  <- c(cnt0, rep(NA, nrow(df01.spl)-n.))

Z1.mat[,n.]   <- c(z1,   rep(NA, nrow(df01.spl)-n.))

W0.mat[,n.]   <- c(w0,   rep(NA, nrow(df01.spl)-n.))

T0.mat[,n.]   <- c(t0,   rep(NA, nrow(df01.spl)-n.))

}

Cnt.mat[1,1] <- Z1.mat[1,1] <- W0.mat[1,1] <- T0.mat[1,1] <- 0


Cntt              <- byt.df(df01.spl, Cnt.mat, 'cnt')

Cnt.df            <- Cntt$dff

df01.spl$cnt      <- Cntt$col.sums


ZZ1               <- byt.df(df01.spl, Z1.mat, 'Z1')

Z1.df             <- ZZ1$dff


WW0               <- byt.df(df01.spl, W0.mat, 'W0')

W0.df             <- WW0$dff

df01.spl$W0.sums  <- WW0$col.sums


TT0               <- byt.df(df01.spl, T0.mat, 'T0')

T0.df             <- TT0$dff


# step 5: creating indicator for number of observed outcomes

# up to lpg-many for lead-in group
```

```
df01.spl$ldin00 <- 0

df01.spl$ldin00[which(df01.spl$cnt<lpg)] <- 1


return(list(df01.spl=df01.spl, Cnt.df=Cnt.df, Z1.df=Z1.df, W0.df=W0.df, T0.df=T0.df))

}


#=============================================================================================

#=============================================================================================


## 6. data derivation, full trial

full.dat <- function(){


df01 <- ac.dat()

repeat{

# step 6: df.split function above to generate full group-specific data

dfspl.C <- df.split(df01, 'C', tht.C1.init)

dfspl.T <- df.split(df01, 'T', tht.T1.init)


df01.C <- dfspl.C$df01.spl

df01.T <- dfspl.T$df01.spl


# step 7: ensuring equal-sized lead-in by taking the max lpg from each group (df01.g$ldin)

ld.sz.C   <- sum(df01.C$ldin00)

ld.sz.T   <- sum(df01.T$ldin00)
```

```
ld.sz.gp  <- max(ld.sz.C, ld.sz.T)


# step 8: pulling the group-specific lead-in data (df02.g)
# step 9: ensuring at least one event in each lead-in group (resulting data: df02.chk.g)
# else, re-run until true (hence repeat loop) [shouldn't be a problem most of the time]
tmp2 <- paste0('df01.', grp.labb, '$ldin <- 0; ',
'df01.', grp.labb, '$ldin[df01.', grp.labb, '$tm.g %in% 1:ld.sz.gp] <- 1; ',
'df02.', grp.labb, ' <- df01.', grp.labb, '[which(df01.', grp.labb, '$ldin==1),]; ',
'df02.chk.', grp.labb, ' <- df02.', grp.labb, '[with(df02.', grp.labb,
', which(A0[nrow(df02.', grp.labb, ')]>=Z0)),]')
eval(parse(text=tmp2))
if(sum(df02.chk.T$W0.sums)>=1 & sum(df02.chk.C$W0.sums)>=1){
break
}
}


eval(parse(text=paste0('Cnt.df.', grp.labb, ' <- dfspl.', grp.labb, '$Cnt.df; ',
'Z1.df.',  grp.labb, ' <- dfspl.', grp.labb, '$Z1.df; ',
'W0.df.',  grp.labb, ' <- dfspl.', grp.labb, '$W0.df; ',
'T0.df.',  grp.labb, ' <- dfspl.', grp.labb, '$T0.df')))


df01.C$ldin00  <- df01.T$ldin00  <- NULL
df01.C$cnt     <- df01.T$cnt     <- NULL
df01.C$W0.sums <- df01.T$W0.sums <- NULL
```

```
df02 <- rbind.data.frame(df01.C, df01.T)

df03 <- df02[order(df02$A0),]


df04 <- df03[which(df03$ldin==1),]

df04.NA <- as.data.frame(matrix(NA, nrow=N-nrow(df04), ncol=ncol(df04)))

colnames(df04.NA) <- colnames(df04)


df05 <- rbind.data.frame(df04, df04.NA)

df05b <- df05[vrrs]

df06a <- cbind.data.frame(tm=1:nrow(df05), df05b)

df06a$rho.hat <- ifelse(df06a$ldin==1, 0.5, NA)

df06a$U <- rep(NA, nrow(df06a))

df06a$U[which(is.na(df06a$ldin))] <- runif(nrow(df04.NA))

df06a$ldin[which(is.na(df06a$ldin))] <- 0


tmg.value <- df06a$tm.g[sum(df06a$ldin)]+1

df06a$tm.g[which(df06a$ldin==0)] <- tmg.value:(tmg.value+sum(df06a$ldin==0)-1)


haz.T  <- haz.C <- n.T <- n.C <- xii.T <- xii.C <- rep(NA, nrow(df06a))

z.stat <- pval  <- rej <- HR.obs.NA <- HR.obs.exp <- rep(NA, nrow(df06a))

tht.T  <- tht.C <- sd.tht.T <- sd.tht.C <- rep(NA, nrow(df06a))

z1.sum.C <- w0.sum.C <- t0.sum.C <- z1.sum.T <- w0.sum.T <- t0.sum.T <- rep(NA, nrow(df06a))

Ht.exp.C <- Ht.exp.T <- Ht.NA.C <- Ht.NA.T <- se.Ht.NA.C <- se.Ht.NA.T <- rep(NA, nrow(df06a))
```

```
tmp3 <- data.frame(z1.sum.C, w0.sum.C, t0.sum.C, z1.sum.T, w0.sum.T, t0.sum.T,
tht.T, tht.C, xii.T, xii.C, haz.T, haz.C, sd.tht.T, sd.tht.C,
n.T, n.C, z.stat, pval, rej, HR.obs.NA, HR.obs.exp,
Ht.exp.C, Ht.exp.T, Ht.NA.C, Ht.NA.T, se.Ht.NA.C, se.Ht.NA.T)
df06 <- cbind.data.frame(df06a, tmp3)

return(list(ld.sz.C=ld.sz.C, ld.sz.T=ld.sz.T, ld.sz.gp=ld.sz.gp,
df01.T=df01.T, df01.C=df01.C, df03=df03, df06=df06))
}

#############################################################################################

set.seed(220)
trial.dat  <- full.dat()
df03       <- trial.dat$df03
Z1.C       <- with(df03, ifelse(A0[nrow(df03)]<Z0[1:(nrow(df03)-1)], A0[nrow(df03)], Z0[1:(nrow
W0.C       <- with(df03, 1*(Z1.C==S1[1:(nrow(df03)-1)]))
T0.C       <- with(df03, Z1.C-A0[1:(nrow(df03)-1)])
cdf0       <- cbind.data.frame(Z1.C, W0.C, T0.C)
cdf0$da.C  <- da.C <- 1*(Z1.C<days)

#############################################################################################
```

```
set.seed(220)

fitKM.C1   <- survfit(Surv(T0.C, W0.C) ~ 1)

tm.untl1    <- sum(fitKM.C1$time<days)

sum(da.C); sum(tm.untl1)


ht.NA.fit.C1      <- with(fitKM.C1, n.event/n.risk)

Ht.NA.fit.C1      <- cumsum(ht.NA.fit.C1)


var.Ht.NA.fit.C1  <- cumsum(with(fitKM.C1, n.event/n.risk^2))

v.Ht.NA.fit.C1    <- var.Ht.NA.fit.C1*length(T0.C)

se.Ht.NA.fit.C1   <- sqrt(v.Ht.NA.fit.C1)


haz.C.df1         <- data.frame(time=fitKM.C1$time, ht.NA.fit.C1, Ht.NA.fit.C1,

var.Ht.NA.fit.C1, v.Ht.NA.fit.C1, se.Ht.NA.fit.C1)

# tm.untl=1*(fitKM.C1$time<days))

tm.dat.C1         <- haz.C.df1[tm.untl1,]

Ht.NA.C1          <- tm.dat.C1$Ht.NA.fit.C1

se.Ht.NA.C1       <- tm.dat.C1$se.Ht.NA.fit.C1


haz.exp.C.df1     <- data.frame(time=fitKM.C1$time, Ht.exp.C=fitKM.C1$time/tht.C1.init,

tm.untl1=1*(fitKM.C1$time<days))

tm.dat.C1         <- haz.exp.C.df1[tm.untl1,]

Ht.exp.C1         <- tm.dat.C1$Ht.exp.C
```

```
Ht.NA.C1; se.Ht.NA.C1; Ht.exp.C1



#############################################################################################

set.seed(220)
fitKM.C2   <- survfit(Surv(Z1.C, W0.C) ~ 1)
tm.untl2    <- sum(fitKM.C2$time<days)
sum(da.C);
sum(tm.untl2)

ht.NA.fit.C2      <- with(fitKM.C2, n.event/n.risk)
Ht.NA.fit.C2      <- cumsum(ht.NA.fit.C2)

var.Ht.NA.fit.C2  <- cumsum(with(fitKM.C2, n.event/n.risk^2))
v.Ht.NA.fit.C2    <- var.Ht.NA.fit.C2*length(Z1.C)
se.Ht.NA.fit.C2   <- sqrt(v.Ht.NA.fit.C2)

haz.C.df2         <- data.frame(time=fitKM.C2$time, ht.NA.fit.C2, Ht.NA.fit.C2,
var.Ht.NA.fit.C2, v.Ht.NA.fit.C2, se.Ht.NA.fit.C2)
#tm.untl=1*(fitKM.C2$time<days))
tm.dat.C2         <- haz.C.df2[tm.untl2,]
Ht.NA.C2          <- tm.dat.C2$Ht.NA.fit.C2
se.Ht.NA.C2       <- tm.dat.C2$se.Ht.NA.fit.C2
```

```
haz.exp.C.df2     <- data.frame(time=fitKM.C2$time, Ht.exp.C=fitKM.C2$time/tht.C1.init,
tm.untl2=1*(fitKM.C2$time<days))
tm.dat.C2         <- haz.exp.C.df2[tm.untl2,]
Ht.exp.C2         <- tm.dat.C2$Ht.exp.C

Ht.NA.C1; se.Ht.NA.C1; Ht.exp.C1
Ht.NA.C2; se.Ht.NA.C2; Ht.exp.C2

##-----------------------------------------------------------------------------------------
##-----------------------------------------------------------------------------------------
##-----------------------------------------------------------------------------------------
##-----------------------------------------------------------------------------------------

Rh  <- Gp  <- Uf   <- Tm.Grp <- matrix(NA, ncol=n.trials, nrow=N)
N.T <- N.C <- Xi.T <- Xi.C   <- matrix(NA, ncol=n.trials, nrow=N)

HR.Obs.NA <- HR.Obs.Exp <- matrix(NA, ncol=n.trials, nrow=N)
Z.Stat <- Pval   <- Rej    <- matrix(NA, ncol=n.trials, nrow=N)
```

```
Haz.T  <- Haz.C  <- Tht.T  <- Tht.C  <- SD.Tht.T <- SD.Tht.C <- matrix(NA, ncol=n.trials, nrow=

A0.obs <- C0.Obs <- C1.Obs <- S0.Obs <- S1.Obs   <- Z0.Obs   <- matrix(NA, ncol=n.trials, nrow=


Z1.Sums.C <- W0.Sums.C <- T0.Sums.C <- matrix(NA, ncol=n.trials, nrow=N)

Z1.Sums.T <- W0.Sums.T <- T0.Sums.T <- matrix(NA, ncol=n.trials, nrow=N)

lead.vec  <- ld.C.vec  <- ld.T.vec  <- rep(NA, n.trials)


CumHaz.NA.C <- CumHaz.NA.T <- CumHaz.Exp.C <- CumHaz.Exp.T <- matrix(NA, ncol=n.trials, nrow=N)

SE.CumHaz.NA.C <- SE.CumHaz.NA.T <- matrix(NA, ncol=n.trials, nrow=N)



for(j in 1:n.trials){


trial.dat    <- full.dat()


ld.sz.T      <- trial.dat$ld.sz.T

ld.sz.C      <- trial.dat$ld.sz.C

ld.sz.gp     <- trial.dat$ld.sz.gp


df03         <- trial.dat$df03

df06         <- trial.dat$df06


lead         <- ld.sz.gp*2

remainder    <- nrow(df06)-lead
```

```
## initial saving

lead.vec[j]  <- lead

ld.C.vec[j]  <- ld.sz.C

ld.T.vec[j]  <- ld.sz.T


Uf[,j]       <- df06$U

Tm.Grp[,j]   <- df06$tm.g



for(rem. in 1:remainder){


ii       <- lead+rem.

tm.i     <- ii

tm.grp   <- df06$tm.g[ii]



###### control data and MLE estimation

C.df.06 <- df06[which(df06$grp=='C'), vrrs]

C.df.03 <- df03[with(df03, which(grp=='C' & tm.g==tm.grp)), vrrs]

C.df.0  <- rbind.data.frame(C.df.06, C.df.03)


Z1.C <- with(C.df.0, ifelse(A0[nrow(C.df.0)]<Z0[1:(nrow(C.df.0)-1)],

A0[nrow(C.df.0)], Z0[1:(nrow(C.df.0)-1)]))
```

```
W0.C <- with(C.df.0, 1*(Z1.C==S1[1:(nrow(C.df.0)-1)]))

T0.C <- with(C.df.0, Z1.C-A0[1:(nrow(C.df.0)-1)])

da.C <- 1*(Z1.C<days)









tht.C        <- sum(T0.C)/sum(W0.C)

xii.C        <- xi.fn(tht.C, D, R)

haz.C        <- 1/tht.C

sd.tht.C     <- tht.C/sqrt(xii.C)


fitKM.C          <- survfit(Surv(T0.C, W0.C) ~ 1)

tm.untl          <- sum(fitKM.C$time<days)

ht.NA.fit.C      <- with(fitKM.C, n.event/n.risk)

Ht.NA.fit.C      <- cumsum(ht.NA.fit.C)#/length(T0.C)


# var.Ht.NA.fit.C  <- cumsum(with(fitKM.C, ((n.risk-n.event)/(n.risk-1))*(n.event/n.risk^2)))

var.Ht.NA.fit.C  <- cumsum(with(fitKM.C, n.event/n.risk^2))#/length(T0.C)

v.Ht.NA.fit.C    <- var.Ht.NA.fit.C*length(T0.C)

se.Ht.NA.fit.C   <- sqrt(v.Ht.NA.fit.C)
```

```
haz.C.df         <- data.frame(time=fitKM.C$time, ht.NA.fit.C, Ht.NA.fit.C,
var.Ht.NA.fit.C, v.Ht.NA.fit.C, se.Ht.NA.fit.C,
tm.untl=1*(fitKM.C$time<days))
tm.dat.C         <- haz.C.df[tm.untl,]
Ht.NA.C          <- tm.dat.C$Ht.NA.fit.C
se.Ht.NA.C       <- tm.dat.C$se.Ht.NA.fit.C

haz.exp.C.df     <- data.frame(time=fitKM.C$time, Ht.exp.C=fitKM.C$time/tht.C,
tm.untl=1*(fitKM.C$time<days))
tm.dat.C         <- haz.exp.C.df[tm.untl,]
Ht.exp.C         <- tm.dat.C$Ht.exp.C

# windows(height=5, width=8)
# plot(x=haz.C.df$time, y=haz.C.df$Ht.NA.fit.C, type='S')
# abline(v=days, lty=2, col='darkgrey')
# abline(h=haz.C.df[tm.untl,3], lty=2, col='darkgrey')



###### trt data and MLE estimation
T.df.06 <- df06[which(df06$grp=='T'), vrrs]
T.df.03 <- df03[with(df03, which(grp=='T' & tm.g==tm.grp)), vrrs]
T.df.0  <- rbind.data.frame(T.df.06, T.df.03)
```

```
Z1.T <- with(T.df.0, ifelse(A0[nrow(T.df.0)]<Z0[1:(nrow(T.df.0)-1)],
A0[nrow(T.df.0)], Z0[1:(nrow(T.df.0)-1)]))
W0.T <- with(T.df.0, 1*(Z1.T==S1[1:(nrow(T.df.0)-1)]))
T0.T <- with(T.df.0, Z1.T-A0[1:(nrow(T.df.0)-1)])


tht.T        <- sum(T0.T)/sum(W0.T)
xii.T        <- xi.fn(tht.T, D, R)
haz.T        <- 1/tht.T
sd.tht.T     <- tht.T/sqrt(xii.T)


fitKM.T          <- survfit(Surv(T0.T, W0.T) ~ 1)
tm.untl.T        <- sum(fitKM.T$time<days)
ht.NA.fit.T      <- with(fitKM.T, n.event/n.risk)
Ht.NA.fit.T      <- cumsum(ht.NA.fit.T)


var.Ht.NA.fit.T  <- cumsum(with(fitKM.T, n.event/n.risk^2))
v.Ht.NA.fit.T    <- var.Ht.NA.fit.T*length(T0.T)
se.Ht.NA.fit.T   <- sqrt(v.Ht.NA.fit.T)


haz.T.df         <- data.frame(time=fitKM.T$time, ht.NA.fit.T, Ht.NA.fit.T,
var.Ht.NA.fit.T, v.Ht.NA.fit.T, se.Ht.NA.fit.T,
tm.untl.T=1*(fitKM.T$time<days))
tm.dat.T         <- haz.T.df[tm.untl.T,]
Ht.NA.T          <- tm.dat.T$Ht.NA.fit.T
```

```
se.Ht.NA.T       <- tm.dat.T$se.Ht.NA.fit.T


haz.exp.T.df     <- data.frame(time=fitKM.T$time, Ht.exp.T=fitKM.T$time/tht.T,
tm.untl.T=1*(fitKM.T$time<days))
tm.dat.T         <- haz.exp.T.df[tm.untl.T,]
Ht.exp.T         <- tm.dat.T$Ht.exp.T



###### putting estimated parms into overall data frame matrix
tmp5 <- paste0('df06$tht.',    grp.labb, '[tm.i-1] <- tht.',    grp.labb, '; ',
'df06$xii.',    grp.labb, '[tm.i-1] <- xii.',    grp.labb, '; ',
'df06$haz.',    grp.labb, '[tm.i-1] <- haz.',    grp.labb, '; ',
'df06$sd.tht.', grp.labb, '[tm.i-1] <- sd.tht.', grp.labb, '; ',
'df06$n.',      grp.labb, '[tm.i-1] <- nrow(',   grp.labb, '.df.06); ',
'df06$z1.sum.', grp.labb, '[tm.i-1] <- sum(Z1.', grp.labb, '); ',
'df06$w0.sum.', grp.labb, '[tm.i-1] <- sum(W0.', grp.labb, '); ',
'df06$t0.sum.', grp.labb, '[tm.i-1] <- sum(T0.', grp.labb, ')')
eval(parse(text=tmp5))


eval(parse(text=paste0('df06$Ht.NA.', grp.labb, '[tm.i-1] <- Ht.NA.', grp.labb, '; ',
'df06$se.Ht.NA.', grp.labb, '[tm.i-1] <- se.Ht.NA.', grp.labb, '; ',
'df06$Ht.exp.', grp.labb, '[tm.i-1] <- Ht.exp.', grp.labb)))
```

```
###### randomization

df06$rho.hat[tm.i] <- with(df06[tm.i-1,], rand(Ht.NA.T, Ht.NA.C, se.Ht.NA.T, se.Ht.NA.C))

df06$grp[tm.i]     <- with(df06[tm.i,], ifelse(U<rho.hat, 'T', 'C'))

if(df06$grp[tm.i]=='C'){

df.add <- C.df.03

} else {

df.add <- T.df.03

}

eval(parse(text=paste0('df06$', vrrs, '[tm.i] <- df.add$', vrrs)))




###### testing for power

zz.stat      <- (Ht.NA.C - Ht.NA.T) / sqrt(se.Ht.NA.C^2 + se.Ht.NA.T^2)

pvalu        <- 2*pnorm(-abs(zz.stat))

rejj         <- 1*(pvalu<0.05)

hr.obs.na    <- Ht.NA.C/Ht.NA.T

hr.obs.exp   <- haz.C/haz.T

eval(parse(text=paste0('df06$', pwr.vrs.df, '[tm.i] <- ', pwr.vrs.est)))

}


Rh[,j]               <- df06$rho.hat

Gp[,j]               <- df06$grp


CumHaz.NA.C[,j]      <- df06$Ht.NA.C
```

```
CumHaz.NA.T[,j]     <- df06$Ht.NA.T

SE.CumHaz.NA.C[,j]  <- df06$se.Ht.NA.C

SE.CumHaz.NA.T[,j]  <- df06$se.Ht.NA.T

CumHaz.Exp.C[,j]    <- df06$Ht.exp.C

CumHaz.Exp.T[,j]    <- df06$Ht.exp.T


N.T[,j]             <- df06$n.T

N.C[,j]             <- df06$n.C

Xi.T[,j]            <- df06$xii.T

Xi.C[,j]            <- df06$xii.C


Z.Stat[,j]          <- df06$z.stat

Pval[,j]            <- df06$pval

Rej[,j]             <- df06$rej


HR.Obs.NA[,j]       <- df06$HR.obs.NA

HR.Obs.Exp[,j]      <- df06$HR.obs.exp


Haz.T[,j]           <- df06$haz.T

Haz.C[,j]           <- df06$haz.C

Tht.T[,j]           <- df06$tht.T

Tht.C[,j]           <- df06$tht.C

SD.Tht.T[,j]        <- df06$sd.tht.T

SD.Tht.C[,j]        <- df06$sd.tht.C
```

```
A0.obs[,j]          <- df06$A0

C0.Obs[,j]          <- df06$C0

C1.Obs[,j]          <- df06$C1

S0.Obs[,j]          <- df06$S0

S1.Obs[,j]          <- df06$S1

Z0.Obs[,j]          <- df06$Z0


Z1.Sums.C[,j]       <- df06$z1.sum.C

W0.Sums.C[,j]       <- df06$w0.sum.C

T0.Sums.C[,j]       <- df06$t0.sum.C

Z1.Sums.T[,j]       <- df06$z1.sum.T

W0.Sums.T[,j]       <- df06$w0.sum.T

T0.Sums.T[,j]       <- df06$t0.sum.T


print(j)

}
```

References

[1] L Friedman, C Furberg, and D DeMets. *Fundamentals of Clinical Trials*. Springer-Verlag, New York, NY, 4 edition, 2010.

[2] U.S. Department of Health and Human Services. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, April 1979.

[3] A Atkinson and A Biswas. *Randomised Response-Adaptive Designs in Clinical Trials*. CRC Press, Boca Raton, FL, 2014.

[4] A Biswas and R Bhattacharya. Optimal response-adaptive designs for normal responses. *Biometrical J*, 51:193–202, 2009.

[5] A Biswas and S Mandal. Optimal adaptive designs in phase III clinical trials for continuous responses with covariates. In A D Bucchianico, H Lauter, and H P Wynn, editors, *mODa7-Advances in Model-Oriented Design and Analysis*, pages 51–58. Heidelber: Physica-Verlag, 2004.

[6] L Zhang and W F Rosenberger. Response-adaptive randomization for clinical trials with continuous outcomes. *Biometrics*, 62:562–569, 2006.

[7] A Biswas, R Bhattacharya, and L Zhang. Optimal response-adaptive designs for continuous responses in phase III trials. *Biometrical J*, 49:928–940, 2007.

[8] F Hu and L Zhang. Asymptotic properties of doubly adaptive biased coin designs for multi-treatment clinical trials. *Ann Stat*, 32:268–301, 2004.

[9] J Eisele. The doubly adaptive biased coin design for sequential clinical trials. *J Stat Plan Inference*, 38:249–262, 1994.

[10] J Eisele and M Woodroofe. Central limit theorems for doubly adaptive biased coin design. *Ann Stat*, 23:234–254, 1995.

[11] A Ivanova, A Biswas, and H Lurie. Response-adaptive designs for continuous outcomes. *J Stat Plan Inference*, 136:1845–1852, 2006.

[12] A Ivanova. A play-the-winner-type urn design with reduced variability. *Metrika*, 58:1–13, 2003.

[13] Q Yao and L Wei. Play-the-winner for phase II/III clinical trials. *Stat Med*, 15:2413–2423, 1996.

[14] W Rosenberger. Asymptotic inference with response-adaptive treatment allocation designs. *Ann Stat*, 21:2098–2107, 1993.

[15] W Rosenberger and P Seshaiyer. Adaptive survival trials. *J Biopharm Stat*, 7:617–624, 1997.

[16] A Hallstrom, M Brooks, and M Peckova. Logrank, play the winner, power and ethics. *Stat Med*, 15:2135–2142, 1996.

[17] U Bandyopadhyay and A Biswas. An adaptive allocation for continuous response using Wilcoxon-Mann-Whitney score. *J Stat Plan Inference*, 123:207–224, 2004.

[18] F Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

[19] H Mann and D Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat*, 18:50–60, 1947.

[20] D Freedman. Bernard freedman's urn. *Ann Math Stat*, 65:965–970, 1965.

[21] L Wei. A class of designs for sequential clinical trials. *J Am Stat Assoc*, pages 382–386, 1977.

[22] L Wei. On the random allocation design for the control of selection bias in sequential experiments. *Biometrika*, 65:79–90, 1978.

[23] C Jennison and B W Turnbill. Group sequential tests with outcome-dependent treatment assignment. *Sequential Analysis*, 20:209–234, 2001.

[24] G Casella and R L Berger. *Statistical Inference*. Thomson Learning, Pacific Grove, CA, 2 edition, 2002.

[25] L Zhang and W F Rosenberger. Response-adaptive randomization for survival trials: the parametric approach. *Appl. Statist.*, 56:153–165, 2007.

[26] Mathworks Support Documentation. Nonparametric estimates of cumulative distribution functions and their inverses. Web page, 2017.

[27] J Banks, J Carson, B Nelson, and D Nichol. *Discrete-Event System Simulation*. Prentice Hall, Upper Saddle River, NJ, 3 edition, 2001.

[28] L Leemis and S Park. *Discrete-Event Simulation: A First Course*. Prentice Hall, Upper Saddle River, NJ, 2006.

[29] P Muldowney, K Ostaszewski, and W Wojdowski. The Darth Vader Rule. *Tatra Mt Math Publ*, 52:53–63, 2012.

[30] ML Moeschberger and JP Klein. A Comparison of Several Methods of Estimating the Survival Function When There is Extreme Right Censoring. *Biometrics*, 41:253–259, 1985.

[31] ET Lee and OT Go. Survival analysis in public health research. *Annu Rev Public Health*, 18:105–134, 1997.

[32] T M Therneau and P M Grambsch. *Modeling survival data: Extending the Cox model*. 2000.

[33] W Kaczynski, L Leemis, N Loehr, and J McQueston. Nonparametric random variate generation using a piecewise-linear cumulative distribution function. *Commun Stat Simul Comput*, 41:449–468, 2012.

[34] C Kooperberg and C J Stone. Hazard Estimation with Flexible Tails, Technical Report No. 388. Technical report, University of California, Department of Statistics, 1993.

[35] M Tai. A Mathematical Model for the Determination of Total Area Under Glucose Tolerance and Other Metabolic Curves. *Diabetes Care*, 17:152–154, 1994.

[36] M H Hansen, J Z Huang, C Kooperberg, C J Stone, and Y K Truong. *Statistical Modeling with Spline Functions: Methodology and Theory*. http://kooperberg.fhcrc.org/monopdf/mono.html, 2006.

[37] C Kooperberg, C J Stone, and Y K Truong. Hazard Regression, Technical Report No. 389. Technical report, University of California, Berkeley, Department of Statistics, 1993.

[38] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[39] C Kooperberg. *polspline: Polynomial Spline Routines*, 2018. R package version 1.1.13.

[40] I Walker. Statistics for Psychology: Making sense of our world through analysis. Web page, 2007.

[41] G M Sullivan and R Feinn. Using Effect Size - or Why the P Value Is Not Enough. *J Grad Med Educ*, 4:279–282, 2012.

[42] T M Fragoso and F L Neto. Bayesian model averaging: A systematic review and conceptual classification. *Stat Sci*, 2015.

[43] L M Haines and H Sadiq. Start-up designs for response-adaptive randomization procedures with sequential estimation. *Stat Med*, 34:2958–2970, 2015.

[44] J A Hoeting, D Madigan, A E Raftery, and C T Volinsky. Bayesian Model Averaging: A Tutorial. *Stat Sci*, 14:382–417, 1999.

[45] K Burham and D Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York, NY, 2002.

[46] S Kullback. *Information Theory and Statistics*. John Wiley and Sons, New York, NY, 1959.

[47] S Kullback and R Leibler. On information and sufficiency. *Ann Math Stat*, 22:79–86, 1951.

[48] M Stephens. Ecf statistics for goodness of fit and some comparisons. *J Am Stat Assoc*, 69(347):730–737, 1974.

[49] T Anderson. On the distribution of the two-sample cramer-von mises criterion. *Ann Math Stat*, 33(3):1148–1159, 1962.

[50] N Kuiper. Tests concerning random points on a circle. *Proc K Ned Akad Wet A*, 63:38–47, 1960.

[51] O Sverdlov, Y Ryeznik, and WK Wong. Efficient and ethical response-adaptive randomization designs for multi-arm clinical trials with Weibull time-to-event outcomes. *J Biopharm Stat*, 24:732–754, 2014.

[52] O Sverdlov, Y Ryeznik, and WK Wong. Optimal response-adaptive randomized designs for multi-armed survival trials. *Stat Med*, 30:2890–2910, 2011.

[53] F Hu and W F Rosenberger. Optimality, variability, and power: evaluating response-adaptive randomization procedures for treatment comparisons. *J Am Stat Ass*, 98:671–678, 2003.

[54] W Liu and D S Coad. Group-sequential response-adaptive designs for censored survival outcomes. *J Am Stat Ass*, 205:293–305, 2020.

[55] C Hochheimer. *cpsurvsim: Simulating Survival Data from Change-Point Hazard Distributions*, 2019.

[56] SE Jones, J Erban, B Overmoyer, GT Budd, L Hutchins, E Lower, L Laufman, S Sundaram, WJ Urba, KI Pritchard, R Mennel, D Richards, S Olsen, ML Meyers, and PM Ravdin. Randomized phase III study of docetaxel compared with paclitaxel in metastatic breast cancer. *J Clin Oncol*, 23:5542–51, 2005.

[57] S P Hey and J Kimmelman. Are outcome-adaptive allocation trials ethical? *Clin Trials*, 12:102–106, 2015.

[58] T Laage, J W Loewy, S Menson, E R Miller, E Pulkstenis, N Kan-Dobrosky, and C Coffey. Ethical considerations in adaptive design clinical trials. *Ther Innov Regul Sci*, 51:190–199, 2017.