Non-Inferiority and Equivalence Tests in A Sequential Multiple-Assignment Randomized Trial (SMART)

Palash Ghosh and Bibhas Chakraborty

Centre for Quantitative Medicine
Duke-NUS Medical School, National University of Singapore, Singapore
emails: palash.ghosh@duke-nus.edu.sg, bibhas.chakraborty@duke-nus.edu.sg

Abstract

In a sequential multiple-assignment randomized trial (SMART), a patient is given a sequence of treatments over the trial period. SMART designs are applicable in various branches of medical science like depression, behavioral science, cancer etc. These designs allow researchers to compare sequences of treatments over multiple stages rather than individual treatments in a single stage as in a traditional randomized controlled trial (RCT). As of now, only the traditional superiority testing framework has been employed in the context of SMARTs. However, there are instances where a superiority test may not be suitable to implement in a SMART design. To address this concern, we develop the non-inferiority and the equivalence test procedures to compare treatment sequences in a SMART. We argue that the developed tests can facilitate the uptake of SMART designs in the pharma-sponsored trials. We show that the trade-off between two competing outcomes (e.g., symptom severity and quality of life) from two treatment sequences can be handled well using a non-inferiority/equivalence test paradigm. Sample size and power issues have been discussed with supporting simulations. We show the applicability of the developed test procedures in the SMART context using the data from the well-known STAR*D trial.

Key words: Non-inferiority, Equivalence, Distinct-path, Shared-path, Inverse probability weighting, STAR*D.

1 Introduction

In clinical practice, a patient with a chronic condition is typically treated with a sequence of treatments rather than any stand-alone treatment; for example, a clinician may prescribe an initial treatment to a patient, and then after a prespecified period, s/he may choose to maintain the same treatment for a patient with a favorable proximal outcome (responder) and switch to another treatment for a patient with not-so-favorable proximal outcome (non-responder). In contrast, the framework of traditional two-arm randomized controlled trial (RCT) was developed to compare one treatment with another as opposed to comparing treatment sequences. Thus, at least in the context of chronic diseases, there is a disconnect between the clinical practice and the type of trial designs used to evaluate treatments. This issue is largely addressed by the modern sequential multiple-assignment randomized trial (SMART) design paradigm wherein one can compare multiple treatment sequences or treatment algorithms. A SMART is a multi-stage clinical trial that allows multiple randomizations of eligible patients at different stages of the trial as opposed to a single randomization in a traditional RCT; see Figure 1 for a schematic of such a trial design.
These trials allow researchers to compare multiple stage-specific treatment allocation rules embedded in the trial that factor in proximal response/non-response status; such multi-stage treatment allocation rules are often referred to as *dynamic treatment regimens* (DTRs)\[^{5,6,7,8}\]. Furthermore, SMARTs produce high-quality data to estimate, at the end of the trial, optimal DTRs that take into account the medical history of an individual patient in order to choose the best possible treatment for him/her. In other words, the DTR framework allows investigators to choose the optimized treatment sequence for individual patients.

SMARTs have gained considerable popularity in investigator-initiated “academic trials” in various chronic disease domains, particularly in psychology and behavioral sciences; see The Methodology Center website (https://methodology.psu.edu/ra/adap-inter/projects) for a list of recent SMARTs. While the early precursors of SMART, e.g., STAR*D\[^9\] and CATIE\[^10\], are trials involving multiple drugs approved by the US Food and Drug Administration (FDA), SMARTs have not yet become commonplace in trials sponsored by pharmaceutical companies. Generally speaking, any pharmaceutical company would feel reluctant to do a head-on comparison of a drug manufactured by them with another drug manufactured by a competitor using the traditional superiority testing framework even in an RCT set-up due to a legitimate concern about potential negative results adversely affecting their business. This fear of business loss may be somewhat elevated in case of a SMART design with traditional hypothesis testing because such a design involves more than one comparison of drugs possibly manufactured by different companies. To mitigate the above concerns and thereby to make SMARTs more attractive to pharmaceutical companies, Chuang-Stein et al.\[^11\] advocated for considering a non-inferiority testing framework in the context of a SMART design.

A non-inferiority test\[^12,13\] is a useful alternative to the traditional superiority test often employed in two-arm RCTs designed to confirm the superiority of one treatment over another (usually a
placebo). As is well-known, a placebo-controlled RCT is the gold standard approach to evaluate the efficacy of new medical interventions. However, such a trial is ethically justified only in absence of a proven effective intervention\textsuperscript{12,14} but not so when a proven effective alternative is available in the market\textsuperscript{15}. This is exactly the situation where a non-inferiority trial (i.e., a randomized trial with non-inferiority test) is appealing. In a non-inferiority trial, one aims to test whether a particular treatment is not inferior to another treatment, subject to a pre-specified non-inferiority margin. This testing paradigm is ethically sound in comparing an experimental intervention with an existing proven active control. This paradigm is also generally acceptable to pharmaceutical companies because in order to launch a new drug they only need to show its non-inferiority (but not superiority) to an existing competitor drug. A related testing framework in randomized trials is that of equivalence tests\textsuperscript{13}. In an equivalence trial, one aims to test the hypothesis that a particular drug (or other type of intervention) is equivalent (i.e., neither superior nor inferior) to another drug (or other type of intervention), subject to a pre-specified equivalence margin. An equivalence trial is used when the intention is to prove similarities rather than to establish a difference among treatments. Similar to a non-inferiority trial, an equivalence trial is also designed with the aim to compare a new treatment with an existing active control, in an effort to avoid the ethical issues associated with employing a placebo-controlled trial when active alternative treatments are already available in the market\textsuperscript{16}.

While Chuang-Stein et al.\textsuperscript{11} commented about the potential role of non-inferiority (and equivalence) testing in the context of a SMART design, to the best of our knowledge, there has been no follow up work on this topic in terms of developing the actual tests and associated power and sample size considerations. We view this as a critical gap in the literature. The current article aims to fill this gap by formally introducing non-inferiority and equivalence tests in a SMART design context; we believe that the option to consider the non-inferiority and equivalence tests within a SMART framework will greatly enhance the appeal of SMART designs in pharma-sponsored trials.

The above perspective of the pharmaceutical companies aside, non-inferiority and equivalence tests have other merits that can drive investigators to employ these tests. They offer a useful approach to handling more than one (competing) outcomes in making treatment decisions. Specifically, they can offer a way to consider the trade-off between a hard clinical outcome (e.g., symptom severity) and a patient-centric outcome (e.g., quality of life) when comparing two or more treatments in an RCT (or, treatment regimens in case of a SMART). Below we illustrate this point with the help of data from the well-known Sequenced Treatment Alternatives to Relieve Depression (STAR*D)\textsuperscript{9} study.

In the STAR*D study, a patient’s level of depression (symptom severity) was assessed using the quick inventory of depressive symptomatology (QIDS) score, with a lesser value of QIDS referring to a better patient outcome. For easier interpretation, we define a transformed score, $QIDS^* = \max\{QIDS\} - QIDS$, with a higher QIDS* score conforming to a better patient outcome in terms of symptom severity. The competing outcome considered is the quality of life (QOL) profile. As mentioned previously, the STAR*D study was an early precursor of a SMART design where more than one randomizations were administered on eligible patients. In this article, we consider only two randomized stages of STAR*D, even though there were more stages in this highly complex trial; see Appendix 8.4 for further details. The treatments given to the patients at each stage of STAR*D can fall into one of two broad categories: mono-therapy($M$) and combination therapy ($C$); more formally, the stage-specific treatment assignments $T_1, T_2 \in \{M, C\}$. Define two treatment regimens...
$d_{MC} \overset{\text{def}}{=} (M, M^R C^{1-R})$ and $d_{CC} \overset{\text{def}}{=} (C, C^R C^{1-R})$, respectively, where the notation $(M, M^R C^{1-R})$ stands for a treatment sequence experienced by a patient who starts with mono-therapy ($M$) at stage 1, continues with the same treatment if s/he is a responder ($R = 1$) at the first stage, and switches to a combination therapy ($C$) at the second stage if s/he is a non-responder ($R = 0$); likewise, $(C, C^R C^{1-R})$ corresponds to a patient who gets combination therapies ($C$) at both stages, irrespective of his/her response status. Now the key question is which of the two regimens is better; it may very well be the case that one regimen is better in terms of QIDS* whereas the other regimen is better in terms of QOL. Figure 2 shows two sunflower plots corresponding to the regimens $d_{MC}$ and $d_{CC}$ to compare their QIDS* and QOL profiles. In these plots, multiple points are depicted as sunflower with multiple leaves (petals) to facilitate the visual display of bivariate data when multiple observations appear at or near the same plotting location. In each panel, there are three ellipses containing 25%, 50% and 90% of the data points; the bright red dots in the middle represent the centers of these ellipses, and the lines going through the centers represent the regression lines. The ellipse containing 90% data points is wider for the regimen $d_{MC}$ than for the regimen $d_{CC}$. From Figure 2 it is difficult to conclude a particular regime as better in terms of either or both QIDS* and QOL. From Table 1 the mean value of QOL corresponding to the regimen $d_{MC}$ is higher than that of the regimen $d_{CC}$. A one-sided t-test with the null hypothesis of equality of the two regimes’ QOL-means with unequal variances is rejected with a p-value less than 0.0001. Considering the QOL profile alone, clearly $d_{MC}$ is a significantly better regimen. Thus, from a patient perspective, the regimen $d_{MC}$ may be a better choice if it also does reasonably well (but not necessarily better) in terms of QIDS*. In this scenario, a non-inferiority test can be used to check whether the regimen $d_{MC}$, having a better QOL profile, is non-inferior in efficacy (QIDS*) to the regimen $d_{CC}$. On the other hand, one could think of an equivalence test when the relative superiority/inferiority in efficacy of the regimen $d_{MC}$ is thought to be close to that of the regimen $d_{CC}$.

![Figure 2: Sunflower plot to compare the regimens $d_{MC}$ and $d_{CC}$ with respect to QIDS* and QOL](image-url)
Table 1: QOL profile of the regimens $d_{MC}$ and $d_{CC}$ from STAR*D

| Regimen | Mean  | SE    | No. of patients |
|---------|-------|-------|-----------------|
| $d_{MC}$ | 2.33  | 0.0584| 567             |
| $d_{CC}$ | 2.25  | 0.0599| 456             |

The rest of the article is organized as follows. In Section 2, we develop the general setup and introduce notations. In Section 3, methodologies for a non-inferiority SMART is established. We extend the methodologies to equivalence SMARTs in Section 4. Sections 5 and 6 present simulation studies and an analysis of the STAR*D data, respectively. The article ends with a discussion in Section 7. Additional details are relegated to the Appendix (Section 8).

2 General Set-up and Notations

Consider a SMART design where patients are randomized to one of the two treatments $A$ and $B$ at the first stage (see Figure 1). Depending on the proximal outcome after the first stage, the patients can be categorized as either responders ($R = 1$) or non-responders ($R = 0$). At the second stage, the responders are allowed to continue with the same treatment as in the first stage, whereas the non-responders are re-randomized to the second stage treatments $C$ or $D$. A dynamic treatment regimen (DTR) is a sequence of decision rules, one per stage of intervention, for adapting a treatment plan to the time-varying state of an individual patient. From Figure 1, we can consider four DTRs embedded in the trial: $d_1 = (A, A^{R_1-R}), d_2 = (A, A^{R_1-D}), d_3 = (B, B^{R_1-R}),$ and $d_4 = (B, B^{R_1-D})$, where for example, according to the regimen $d_1$, a patient will be given the treatment $A$ at the first stage, s/he will continue the same treatment if s/he is a responder at the end of stage 1, or switch to the treatment $C$ if s/he is a non-responder; other embedded regimens can be interpreted similarly. For clarity in comparing any two embedded regimens, we classify any pair of embedded regimens under comparison to be either (1) distinct-path (DP), i.e., those starting with different initial treatments (e.g., $\{d_1, d_3\}$ or $\{d_2, d_4\}$), or (2) shared-path (SP), i.e., those starting with the same initial treatment (e.g., $\{d_1, d_2\}$ or $\{d_3, d_4\}$).

In this work, we assume the primary outcome of interest, $Y_i$ for the $i^{th}$ patient, to be continuous; $i = 1, \cdots, N$, where $N$ is the total number of patients. Following the same notation as in Ghosh et al. denoted $N_d$ to be the number of patients whose treatment trajectories are consistent with the regimen $d \in \{d_1, \cdots, d_4\}$. For each individual, who obtained treatments $T_1$ and $T_2$ respectively at the two stages of a SMART, assume

$$E(Y_{i,T_1T_2}) = \mu_{T_1T_2} \quad \text{and} \quad Var(Y_{i,T_1T_2}) = \sigma^2$$

(1)

for all $i = 1, \cdots, N$, where $T_1 \in \{A, B\}, T_2 \in \{C, D\}$. For both the non-inferiority and the equivalence test procedures, the test statistic is based on the embedded regimen means. By design, a regimen mean is a weighted average of primary outcomes of patients having treatment trajectories consistent with that regimen. Note that, a weighted averaging is essential as there is a structural
imbalance between responders and non-responders. In the two-stage SMART considered here, the non-responders are re-randomized at the second stage, but the responders do continue with the same treatment. Define the first-stage randomization probability in favor of treatment $T_1$ as $\tau_{T_1}$, and the second-stage randomization probability for those who started with first-stage treatment $T_1$, in favor of treatment $T_2$ as $\tau_{T_1T_2}$ (often these probabilities are $1/2$). Note that, $\tau_B = 1 - \tau_A$, $\tau_{AD} = 1 - \tau_{AC}$ and $\tau_{BD} = 1 - \tau_{BC}$. A responder, due to a single randomization at stage 1, receives the treatment sequence $(T_1, T_1)$ with probability $\tau_{T_1}$. On the other hand, due to two stages of randomization, a non-responder receives a treatment sequence $(T_1, T_2)$ with probability $\tau_{T_1T_2}$. Using the principles of inverse probability weighting, the weight used for the $i^{th}$ patient in regimen $d$ is $W_i = \frac{1}{\tau_{T_1}}$ for a responder and $W_i = \frac{1}{\tau_{T_1} \times \tau_{T_1T_2}}$ for a non-responder. In general, the weight can be expressed as $W_i = 1/(\tau_{T_1} \times \tau_{T_1T_2}^{1-R_i})$, where $R_i$ is the responder indicator: 1 for responders and 0 for non-responders. The observed mean corresponding to the regimen $d$ can be written as

$$\bar{Y}_d = \sum_{i=1}^{N_d} W_i Y_{i,T_1T_2} / \sum_{i=1}^{N_d} W_i. \tag{2}$$

Define $\gamma_{T_1} \in \{\gamma_A, \gamma_B\}$ as the response rate, with respect to some pre-specified definition of response, to the initial treatment $T_1$. In the SMART (see figure [1], at the first stage, all the $N$ patients are randomized among the two interventions $A$ and $B$ with randomization probability in favor of treatment $T_1$ as $\tau_{T_1}$. Then the expected number of responders in a regimen that starts with treatment $T_1$ is $N \times \tau_{T_1} \times \gamma_{T_1}$. So, the number of non-responders that start with treatment $T_1$ is $N \times \tau_{T_1} \times (1 - \gamma_{T_1})$. Note that, only $N \times \tau_{T_1} \times (1 - \gamma_{T_1}) \times \tau_{T_1T_2}$ non-responders get re-randomized to the treatment $T_2$ at the second stage. As a consequence, the expected number of patients whose treatment trajectories are consistent with the regimen $d$ is $N \times \tau_{T_1} \times \gamma_{T_1} + N \times \tau_{T_1} \times (1 - \gamma_{T_1}) \times \tau_{T_1T_2} = N\tau_{T_1} \times \gamma_{T_1} + (1 - \gamma_{T_1}) \tau_{T_1T_2}$. Considering $W_i = \frac{1}{\tau_{T_1}}$ for responders and $W_i = \frac{1}{\tau_{T_1} \times \tau_{T_1T_2}}$ for non-responders and for a fixed $N_d$, we have $\sum_{i=1}^{N_d} W_i = \frac{1}{\tau_{T_1}} \times (N \times \tau_{T_1} \times \gamma_{T_1}) + \frac{1}{\tau_{T_1} \times \tau_{T_1T_2}} \times N \times \tau_{T_1} \times (1 - \gamma_{T_1}) \times \tau_{T_1T_2} = N \tau_{T_1} \times \gamma_{T_1} + (1 - \gamma_{T_1}) \tau_{T_1T_2}$.

Now we can write the mean of regimen $d$ as

$$\mu_d = E(\bar{Y}_d) = \frac{1}{N} \sum_{i=1}^{N_d} W_i E(Y_{i,T_1T_1})$$

$$= \frac{1}{N} \left[ \sum_{i:R_i=1} W_i E(Y_{i,T_1T_1}) + \sum_{i:R_i=0} W_i E(Y_{i,T_1T_2}) \right]$$

$$= \gamma_{T_1} \mu_{T_1T_1} + (1 - \gamma_{T_1}) \mu_{T_1T_2}, \tag{3}$$

and the variance,

$$Var(\bar{Y}_d) = \sigma^2 \sum_{i=1}^{N_d} W_i^2 / \left( \sum_{i=1}^{N_d} W_i \right)^2 = \frac{1 - \gamma_{T_1} \gamma_{T_1T_2}}{\tau_{T_1} \tau_{T_1T_2}} \times \frac{\sigma^2}{N}. \tag{4}$$

When $\tau_{T_1} = \tau_{T_1T_2} = \frac{1}{2}$, then $Var(\bar{Y}_d) = \frac{2(2-\gamma_{T_1})}{N} \sigma^2$. For simplicity, to develop the non-inferiority and equivalence test procedures in the context of SMART, we consider $\tau_A = \tau_B = \tau_{AC} = \tau_{BC} = \frac{1}{2}$. In other words, the randomization procedures at the both stages 1 and 2 are even.

3 Non-Inferiority Test

The objective of a classical non-inferiority trial is to show that the efficacy of the experimental treatment, when compared to an active treatment, is not below a pre-specified non-inferiority
margin. In practice, a non-inferiority margin is the maximum clinically acceptable distance from active control on average that researchers agree to accept in exchange of the secondary benefits of the new treatment. A non-inferiority test becomes a traditional one-sided superiority test when the non-inferiority margin is set to zero. In this section, we describe the non-inferiority test procedure in the context of a SMART. For completeness, we consider the non-inferiority test for both distinct-path and shared-path comparison of DTRs. In general, for any two regimens \( d_1 \) and \( d_j \) with corresponding regimen means \( \mu_i \) and \( \mu_j \) respectively, the null and alternative hypotheses for the non-inferiority test can be written as

\[
H_0 : \mu_j \leq \mu_i - \theta \text{ vs. } H_1 : \mu_j > \mu_i - \theta,
\]

where \( \theta \) is a pre-specified non-inferiority margin (\( \theta > 0 \)); \( j \neq i \in \{1, \ldots, 4\} \). This hypothesis tests that the average efficacy of regimen \( d_j \) is not inferior to that of the regimen \( d_i \), with the non-inferiority margin \( \theta \). The unscaled test statistic is \( \bar{Y}_{d_j} - \bar{Y}_{d_i} \). Note that, the choice of the non-inferiority margin depends on both statistical reasoning and clinical judgment.

### 3.1 Distinct-path DTRs

Let \( \mu_{d_1} \) and \( \mu_{d_3} \) be the regimen means of the two distinct-path DTRs \( d_1 \) and \( d_3 \), respectively. Here, we assume the regimen \( d_3 \) as the standard (control) treatment regimen. We want to test whether the regimen \( d_1 \) is non-inferior in efficacy to the regimen \( d_3 \). From the general hypothesis, the hypothesis of interest is \( H_0 : \mu_{d_1} \leq \mu_{d_3} - \theta \text{ vs. } H_1 : \mu_{d_1} > \mu_{d_3} - \theta \). The unscaled test statistic is \( \bar{Y}_{d_1} - \bar{Y}_{d_3} \) with mean \( \mu_{d_1} - \mu_{d_3} \) and variance \( \nu_{d_1 d_3}^{DP} = 2\sigma^2(4 - \gamma_A - \gamma_B)/N \), where \( \sigma^2 = \text{var}(Y_i) \), \( \gamma_A, \gamma_B \) are the response rates corresponding to \( d_1 \) and \( d_3 \) respectively, and \( N \) is the total sample size of the trial. Thus, under \( H_0 \), the large-sample distribution of the test statistic

\[
Z_{d_1 d_3} = \frac{(\bar{Y}_{d_1} - \bar{Y}_{d_3}) - (\mu_{d_1} - \mu_{d_3})}{\sqrt{\text{var}(\bar{Y}_{d_1} - \bar{Y}_{d_3})}} = \frac{\bar{Y}_{d_1} - \bar{Y}_{d_3} + \theta}{\sqrt{\nu_{d_1 d_3}^{DP}}} \rightarrow N(0, 1).
\]

So we reject \( H_0 \) and conclude non-inferiority if \( Z_{d_1 d_3} > z_{\alpha} \) (one-sided), where \( \alpha \) is the type-I error and \( z_{\alpha} \) is the \( (1 - \alpha)^{th} \) quantile of the standard normal distribution. Under a specified effect size \( \delta \) such that \( \mu_{d_3} - \mu_{d_1} = \delta \), the required sample size is

\[
N = \frac{2(z_{\alpha} + z_{\beta})^2}{\delta^2} \frac{\sigma^2}{(4 - \gamma_A - \gamma_B)(\theta - \delta)^2},
\]

where \( \beta \) is the type-II error. When \( \gamma_A = \gamma_B = \gamma \), then \( N = 4(z_{\alpha} + z_{\beta})^2(2 - \gamma)/(\theta - \delta)^2 \). See Appendix \ref{appendix} for the derivation. The sample size \( N \) is a decreasing function of both the response rates \( \gamma_A \) and \( \gamma_B \). Note that, \( N \) depends on the difference between the non-inferiority margin \( \theta \) and the effect size \( \delta \) rather than their individual values.

### 3.2 Shared-path DTRs

Next we consider two shared-path DTRs \( d_1 \) and \( d_2 \) with corresponding regimen means \( \mu_{d_1} \) and \( \mu_{d_2} \), respectively. Here, we consider the regimen \( d_2 \) as the standard (control) treatment regimen. We want to test whether the regimen \( d_1 \) is non-inferior in efficacy to the regimen \( d_2 \). Similar to the distinct-path, the unscaled test statistic is \( \bar{Y}_{d_1} - \bar{Y}_{d_2} \) with mean \( \mu_{d_1} - \mu_{d_2} \). However, in this case, the variance of the unscaled test statistic will be different from that of distinct-path regimens due to the shared path, in that both the regimens \( d_1 \) and \( d_2 \) have the same set of patients as responders.
The variance of $\bar{Y}_{d_1} - \bar{Y}_{d_2}$ is $\nu^{SP}_{d_1,d_2} = \frac{8(1-\gamma_A)^2}{N}$ (see Appendix 8.2 for the derivation). Therefore, under $H_0$, the large-sample distribution of the test statistic

$$Z_{d_1,d_2} = \frac{(\bar{Y}_{d_1} - \bar{Y}_{d_2}) - (\mu_{d_2} - \mu_{d_1})}{\sqrt{\text{var}(\bar{Y}_{d_1} - \bar{Y}_{d_2})}} = \frac{\bar{Y}_{d_1} - \bar{Y}_{d_2} + \theta}{\sqrt{\nu^{SP}_{d_1,d_2}}} \rightarrow N(0,1).$$

So we reject $H_0$ and conclude non-inferiority if $Z_{d_1,d_2} > z_\alpha$ (one-sided). Similar to the distinct-path, under a specified effect size $\delta$ such that $\mu_{d_2} - \mu_{d_1} = \delta$, the required sample size for the trial is

$$N = 8(z_\alpha + z_\beta)^2(1 - \gamma_A)\frac{\sigma^2}{(\theta - \delta)^2}. \quad (7)$$

Note that, in the shared-path comparison of two regimens, only one response rate is involved. The sample size $N$ is a decreasing function of that single response rate. In case $\gamma_A = \gamma_B = 0$, there is no responder in the SMART, resulting in $\nu^{DP}_{d_1,d_2} = \nu^{SP}_{d_1,d_2}$. Furthermore, in this scenario, from (6) and (7), the required sample size $N$ becomes identical in the distinct-path and the shared-path comparison of regimens. However, $\gamma_A = 1$ makes all the patients initially treated with $A$ as responders after the first stage and $\nu^{SP}_{d_1,d_2} = 0$. In this case, there is no difference between the two regimens $d_1$ and $d_2$. On the contrary, the distinct-path comparison of regimens remains valid even when both the response rates $\gamma_A = \gamma_B = 1$.

4 Equivalence Test

As mentioned earlier, the term equivalence means that the efficacies of the two considered interventions are close to each other in a way that cannot be considered superior or inferior. Similar to the non-inferiority margin, here we consider an equivalence margin to define the closeness of two interventions. An equivalence trial becomes a traditional two-sided superiority test when the equivalence margin is set to zero. An equivalence margin $(-\theta, \theta)$ is defined at the beginning of a trial such that any value in that range is clinically unimportant. After the trial, an equivalence is achieved when the estimated confidence interval of the treatment difference falls in $(-\theta, \theta)$. In general, for any two regimens $d_i$ and $d_j$ with corresponding regimen means $\mu_i$ and $\mu_j$ respectively, the null and alternative hypotheses for the equivalence test are specified as

$$H_0: \mu_j \leq \mu_i - \theta \text{ or } \mu_j \geq \mu_i + \theta \text{ vs. } H_1: \mu_i - \theta < \mu_j < \mu_i + \theta, \quad (8)$$

where $\theta$ is a pre-specified equivalence margin ($\theta > 0$). The above test actually contains two separate tests: (1) a non-inferiority test ($H_{01}: \mu_j \leq \mu_i - \theta$ vs. $H_{11}: \mu_j > \mu_i - \theta$), and (2) a non-superiority test ($H_{02}: \mu_j \geq \mu_i + \theta$ vs. $H_{12}: \mu_j < \mu_i + \theta$). In traditional clinical trials, due to ease of use and recommendation from the US Food and Drug Administration (FDA), two one-sided tests (TOST) procedure is widely used for equivalence test. Note that, in a TOST, we consider two one-sided tests, each of them having type-I error $\alpha$ resulting in the overall type-I error of $2\alpha$. In this section, we develop an equivalence test procedure for SMART, separately for the distinct-path and the shared-path DTRs.

4.1 Distinct-path DTRs

First we consider the non-inferiority test procedure to conduct an equivalence test for distinct-path DTRs. For the two distinct-path DTRs $d_1$ and $d_3$ with means $\mu_{d_1}$ and $\mu_{d_3}$ respectively, we repeat the non-inferiority test procedure described in section 3.1. The same test statistic $Z_{d_1,d_3}$ is used.
to test $H_{01} : \mu_{d1} \leq \mu_{d3} - \theta$ vs. $H_{11} : \mu_{d1} > \mu_{d3} - \theta$. The null hypothesis $H_{01}$ will be rejected if $Z_{d1,d3} > z_{a}$ (one-sided). For the non-superiority test $H_{02} : \mu_{d1} \geq \mu_{d3} + \theta$ vs. $H_{12} : \mu_{d1} < \mu_{d3} + \theta$, the large-sample distribution of the test statistic

$$Z_{d1,d3}^* = \frac{(\bar{Y}_{d1} - \bar{Y}_{d3}) - (\mu_{d1} - \mu_{d3})}{\sqrt{\text{var}(Y_{d1} - Y_{d3})}} = \frac{\bar{Y}_{d1} - \bar{Y}_{d3} - \theta}{\sqrt{\nu_{d1,d3}}} \to N(0, 1),$$

under $H_{02}$. Note that, $Z_{d1,d3}^*$ is different from $Z_{d1,d3}$ with respect to sign of $\theta$ in their expression. Reject $H_{02}$ and conclude non-superiority if $Z_{d1,d3}^* < -z_{a}$ (one-sided). We conclude equivalence when both tests reject the corresponding null hypotheses. Given an effect size $\delta = \mu_{d3} - \mu_{d1} \neq 0$, there is no closed-form solution for the minimum required sample size, $N$. However, the power of the equivalence test is given by

$$\Phi \left( -z_{a} + \frac{\theta - \delta}{\nu_{d1,d3}} \right) - \Phi \left( z_{a} - \frac{\theta + \delta}{\nu_{d1,d3}} \right).$$

(see Appendix 8.3 for the derivation). The formula in (9) can be used to estimate the power corresponding to a particular sample size ($N$) when all the involved parameters are known. Repeating the procedure to estimate the power corresponding to different $N$ gives the required sample size corresponding to a specified power. However, in the special case when the effect size $\delta = 0$, the minimum sample size has a closed form, and is given by

$$N = 2 \left( z_{a} + \frac{z_{\gamma}}{\tau} \right)^2 (4 - \gamma_A - \gamma_B) \frac{\sigma^2}{\theta^2}.$$  

(10)

Note that, unlike the distinct-path comparison of regimens in the non-inferiority trial setting, here the sample size $N$ based on (9) depends on the individual values of $\theta$ and $\delta$ as opposed to their difference $\theta - \delta$ alone.

### 4.2 Shared-path DTRs

In an equivalence test procedure to compare two shared-path DTRs $d_1$ and $d_2$, first we use the non-inferiority test described in section 3.2. The test statistic $Z_{d1,d2}$ is used to test $H_{01} : \mu_{d1} \leq \mu_{d2} - \theta$ vs. $H_{11} : \mu_{d1} > \mu_{d2} - \theta$. The null hypothesis $H_{01}$ will be rejected if $Z_{d1,d2} > z_{a}$ (one-sided). For the non-superiority test $H_{02} : \mu_{d1} \geq \mu_{d2} + \theta$ vs. $H_{12} : \mu_{d1} < \mu_{d2} + \theta$, the large-sample distribution of the test statistic

$$Z_{d1,d2} = \frac{(\bar{Y}_{d1} - \bar{Y}_{d2}) - (\mu_{d1} - \mu_{d2})}{\sqrt{\text{var}(Y_{d1} - Y_{d2})}} = \frac{\bar{Y}_{d1} - \bar{Y}_{d2} - \theta}{\sqrt{\nu_{d1,d2}}} \to N(0, 1),$$

under $H_{02}$. We reject $H_{02}$ and conclude non-superiority if $Z_{d1,d2}^* < -z_{a}$ (one-sided). Similar to the distinct-path scenario, we conclude equivalence when both tests reject the corresponding null hypotheses. To estimate the minimum sample size when the effect size $\delta \neq 0$, we use the expression in (9) with $\nu_{d1,d3}^D$ replaced by $\nu_{d1,d2}^S$. In case of $\delta = 0$, the minimum sample size

$$N = 8(\alpha + \frac{\beta}{2})^2(1 - \gamma_A) \frac{\sigma^2}{\theta^2}.$$ 

(11)

Similar to the non-inferiority test, here also, $\gamma_A = \gamma_B = 0$ implies that there is no responder in the SMART resulting in $\nu_{d1,d3}^D = \nu_{d1,d2}^S$, and then the sample size $N$ is the same for both the distinct-path and the shared-path comparison of regimens. In case $\gamma_A = \gamma_B = 1$, the conclusions made in the context of the non-inferiority test remain valid in the equivalence test context as well.
5 Simulation Studies

In this section, we consider simulation studies to show the applicability of the procedures developed above. Our main objective in this section is to check how the estimated (Monte Carlo) power varies with respect to the effect size ($\delta$) and the sample size ($N$) when the non-inferiority/equivalence margin ($\Theta$) is fixed. We also investigate the impact of the response rates ($\gamma_A$ and/or $\gamma_B$) on the estimated sample size in both non-inferiority and equivalence tests.

The data generation process for a SMART is complex. So here we outline an algorithm that generates datasets from a SMART (as depicted in Figure 1) to conduct simulation studies. The algorithm is:

1. Set the values of: $\text{Var}(Y) = \sigma^2$, $\gamma_A$, $\gamma_B$, $\alpha$, $\beta$.
2. Fix the value $N$ for power curve, otherwise calculate the required minimum value of $N$.
3. Set the mean and variance of a latent variable $L$ for the patient with treatment $A$ at first stage as $\mu_{LA}, \sigma_L^2$, respectively, and generate $N_A = \text{ceiling}(N/2)$ random draws from $N(\mu_{LA}, \sigma_L^2)$. The latent variable can be viewed as a proximal outcome which determine responder/non-responder status after the first stage.
4. Define the cut-off value $\eta$ as that value of the latent variable $L$ above which a patient is considered a responder. So
   \[
   \eta = \mu_L + \sigma_L \times \Phi^{-1}(1 - \gamma_A),
   \]
   where $\Phi$ is the cumulative distribution function of the standard normal distribution.
5. The cut-off $\eta$ should be fixed for a SMART irrespective of the choice of initial treatment ($A$ or $B$). So we define
   \[
   \mu_{LB} = \eta - \sigma_L \times \Phi^{-1}(1 - \gamma_B),
   \]
   and generate $N_B = N - N_A$ random draws from $N(\mu_{LB}, \sigma_L^2)$.
6. Define the truncated normal means of the non-responders (NR) who obtained the first-stage treatment $A$ or $B$ as
   \[
   \mu_{LA,NR} = \mu_L - \sigma_L \times \frac{\phi(w^*)}{\Phi(w^*)}, \quad \text{where} \quad w^* = \frac{\eta - \mu_{LA}}{\sigma_L}, \quad \text{and}
   \]
   \[
   \mu_{LB,NR} = \mu_L - \sigma_L \times \frac{\phi(w^*)}{\Phi(w^*)}, \quad \text{where} \quad w^* = \frac{\eta - \mu_{LB}}{\sigma_L}, \quad \text{respectively.}
   \]
   Here, $\phi$ is the probability density function of the standard normal distribution.
7. The regimen means corresponding to the regimens $d_1 : (A, A^R C^{1-R})$, $d_2 : (A, A^R D^{1-R})$, $d_3 : (B, B^R C^{1-R})$ and $d_4 : (B, B^R D^{1-R})$ can be defined as
   \[
   \mu_{d_1} = (\zeta_0 + \zeta_{1A} \mu_{LA}) \times \gamma_A + (\xi_0 + \xi_{1A} \mu_{LA} + \xi_{2AC} \mu_{LA,NR}) \times (1 - \gamma_A),
   \]
   \[
   \mu_{d_2} = (\zeta_0 + \zeta_{1A} \mu_{LA}) \times \gamma_A + (\xi_0 + \xi_{1A} \mu_{LA} + \xi_{2AD} \mu_{LA,NR}) \times (1 - \gamma_A),
   \]
   \[
   \mu_{d_3} = (\zeta_0 + \zeta_{1B} \mu_{LB}) \times \gamma_B + (\xi_0 + \xi_{1B} \mu_{LB} + \xi_{2BC} \mu_{LB,NR}) \times (1 - \gamma_B),
   \]
   \[
   \mu_{d_4} = (\zeta_0 + \zeta_{1B} \mu_{LB}) \times \gamma_B + (\xi_0 + \xi_{1B} \mu_{LB} + \xi_{2BD} \mu_{LB,NR}) \times (1 - \gamma_B),
   \]
   respectively. Here, $\{\zeta_0, \zeta_{1A}, \zeta_{1B}, \zeta_0, \xi_1A, \xi_1B, \xi_{2AC}, \xi_{2AD}, \xi_{2BC}, \xi_{2BD}\}$ are (fixed) parameters to be supplied by the user.
8. Define the effect size between the two distinct-path regimens $d_1$ and $d_3$ as $\delta_{DP} = \mu_{d_3} - \mu_{d_1}$.
Similarly, define the effect size between the two shared-path regimens $d_1$ and $d_2$ as $\delta_{SP} = \mu_{d_2} - \mu_{d_1}$.

9. Finally generate $N_{T_1T_2}$ number of primary outcomes $Y$ from $N(\mu_{T_1T_2}, \sigma^2)$, where

$$
\mu_{T_1T_2} = \begin{cases} 
\xi_0 + \xi_1 A \mu_{L_A} & \text{if } T_1 = T_2 = A \\
\xi_0 + \xi_1 A \mu_{L_A} + \xi_2 A \mu_{L_A.NR} & \text{if } T_1 = A, T_2 = C \\
\xi_0 + \xi_1 A \mu_{L_A} + \xi_2 A \mu_{L_A.NR} & \text{if } T_1 = A, T_2 = D \\
\xi_0 + \xi_1 B \mu_{L_B} & \text{if } T_1 = T_2 = B \\
\xi_0 + \xi_1 B \mu_{L_B} + \xi_2 B \mu_{L_B.NR} & \text{if } T_1 = B, T_2 = C \\
\xi_0 + \xi_1 B \mu_{L_B} + \xi_2 B \mu_{L_B.NR} & \text{if } T_1 = B, T_2 = D
\end{cases}
$$

and

$$
N_{T_1T_2} = \begin{cases} 
N_A \gamma_A & \text{if } T_1 = T_2 = A \\
N_A (1 - \gamma_A) \times \frac{1}{2} & \text{if } T_1 = A, T_2 \in \{C, D\} \\
N_B \gamma_B & \text{if } T_1 = T_2 = B \\
N_B (1 - \gamma_B) \times \frac{1}{2} & \text{if } T_1 = B, T_2 \in \{C, D\}.
\end{cases}
$$

If $N_{T_1T_2}$ contains decimal part, then $N_{T_1T_2} = \text{ceiling}(N_{T_1T_2})$.

5.1 Power Curves

Now we draw the power curves with respect to the effect size ($\delta$) and the sample size ($N$) when the non-inferiority/equivalence margin ($\theta$) is fixed. First, we generate data by the above mentioned algorithm setting $\sigma = 2, \gamma_A = 0.3, \gamma_B = 0.4, \alpha = 0.05$ and $\beta = 0.2$. We consider four different values of the fixed sample size $N$ as 100, 200, 300 and 500. These values of $N$ will illustrate the effect of sample size on the power of the test. Note that, the non-inferiority/equivalence margin $\theta$ is not needed in the data generation process; however it is essential to conduct the non-inferiority/equivalence test. For illustration purpose, we consider $\theta = 2$. Values of other parameters mentioned in the above algorithm are set as: $\mu_{L_A} = 2, \sigma_L = 0.2, \zeta_0 = 0.02, \xi_1 A = 0.5, \xi_1 B = 0.8, \xi_0 = 0.03, \xi_2 AD = 1, \xi_1 B = 0.5, \xi_2 BC = 1$ and $\xi_2 BD = 1$. We consider different values of the parameter $\xi_{2AC} \in \{3.4, 3.2, 3.0, 2.8, 2.6, 2.4, 2.2, 2.0, 1.8, 1.5, 1.3, 1.0, 0.8, 0.6, 0.4, 0.2, 0, -0.2, -0.4, -0.6, -0.8\}$ corresponding to different values of the effect sizes $\delta_{DP}$ and $\delta_{SP}$. In Figure 3, we have shown the power curves for both the distinct-path and the shared-path settings in non-inferiority and equivalence tests. In the non-inferiority test for comparison of distinct-path regimen means, we want to test whether the regimen $d_1$ is non-inferior in efficacy to the regimen $d_3$ with a non-inferiority margin $\theta$. Similarly, in the equivalence test for comparing distinct-path regimen means, we want to test whether the regimen $d_1$ is equivalent in efficacy to the regimen $d_3$ with an equivalence margin $\theta$. On the other hand, when comparing shared-path regimen means, we want to test whether the regimen $d_1$ is non-inferior (equivalent) in efficacy to the regimen $d_2$ with a non-inferiority (equivalence) margin $\theta$. It is evident from Figure 3 that the power increases with the increase in the sample size $N$ when other parameters are fixed. Note that, we have considered the effect size $\delta$ as $\delta_{DP}$ and $\delta_{SP}$ for the distinct-path and the shared-path comparison of regimen
means, respectively. However, in the horizontal axis of Figure 3 we plotted \((\theta - \delta)\), rather than \(\delta\) itself \((\delta \in \{\delta_{DP}, \delta_{SP}\})\), since \((\theta - \delta)\) controls the power of the test. A higher value of \((\theta - \delta)\) means the mean of the regimen \(d_1\) is close to (from left side) or higher than the that of the regimen \(d_2\) or \(d_3\) depending on \(\delta = \delta_{SP}\) or \(\delta_{DP}\), respectively. In case of non-inferiority test for the distinct-path (shared-path) regimen comparison, \(\theta - \delta > 0\) implies that the regimen mean of \(d_1\) is within the non-inferiority margin \(\theta\) corresponding to the regimen mean of \(d_3\) \((d_2)\), i.e, \(\mu_{d_1} > \mu_{d_3} - \theta\) (or, \(\mu_{d_1} > \mu_{d_2} - \theta\)). In Figure 3, for \(\theta - \delta = 2\) (i.e., \(\delta = 0\)), all the power curves achieve (or very close to) the power 1. In other words, \(\delta = 0\) indicates equality of corresponding regimen means. Note that, for the equivalence tests, the power curves are symmetric around \(\theta - \delta = 2\) (i.e., \(\delta = 0\)). In this case, a large value (positive or negative) of the effect size \(\delta\) indicates non-equivalence.

![Figure 3: Power Curves for non-inferiority and equivalence tests with distinct-path and shared-path regimens.](image)

### 5.2 Effect of Response Rates on Sample Size

Next we investigate the effect of response rates \((\gamma_A, \gamma_B)\) on the sample size \(N\) in the non-inferiority and equivalence tests. We set the generative model parameters as: \(\sigma = 3, \alpha = 0.05, \beta = 0.2, \theta = 2, \mu_{LA} = 2, \sigma_L = 0.2, \zeta_0 = 0.02, \zeta_{1A} = 0.5, \zeta_{1B} = 0.8, \xi_0 = 0.03, \xi_{1A} = 0.25, \xi_{2AD} = 1, \xi_{1B} = 0.5, \xi_{2BC} = 1\) and \(\xi_{2BD} = 1\). We consider five different values of \(\xi_{2AC} \in \{0.8, 0.6, 0.4, 0.2, 0\}\) corresponding to five different panels (top to bottom) of Table 2. Note that, in each of the five panels of Table 2, the only parameter that changes its value is \(\gamma_B\) for the distinct-path scenario and
\(\gamma_A\) for the shared-path scenario; \(\delta\) is a function of other parameters. In Table 2, for the distinct-path scenario, we set \(\gamma_A = 0.30\), whereas \(\gamma_B\) can take values in \(\{0.35, 0.40, 0.50, 0.60, 0.70\}\); similarly, for the shared-path scenario, \(\gamma_B = 0.50\) and \(\gamma_A\) can take values in \(\{0.10, 0.20, 0.30, 0.40, 0.50\}\). Note that, in case of shared-path regimens, we are comparing the regimen \(d_1\) with the regimen \(d_2\). Thus, in this case, there is no role of the parameter \(\gamma_B\) in the non-inferiority or equivalence tests. It is evident from the left hand side (distinct-path) of the table that as \(\gamma_B\) increases or deviates from \(\gamma_A\) the value of \(N\) decreases. In other words, when response rates \(\gamma_A\) and \(\gamma_B\) are wide apart from each other, a smaller sample size \((N)\) is required to take a decision in the non-inferiority test. Similarly, in case of shared-path regimens, we need a smaller sample size \((N)\) when the value of \(\gamma_A\) is high.

Note that, the required minimum sample size \(N\) is a decreasing function of both \(\gamma_A\) and \(\gamma_B\) for a distinct-path comparison of regimens, whereas \(N\) is a decreasing function of either \(\gamma_A\) or \(\gamma_B\) for a shared-path comparison. In general, the estimated power based on 5000 Monte Carlo simulations is close to the nominal power of 0.80 in case of distinct-path regimens; however, for the shared-path regimens, the estimated power is slightly less than the nominal power.

As mentioned in Section 4.1, for an equivalence test, there is no closed-form expression of the sample size \(N\) in general; however, in the special case of effect size \(\delta = 0\), \(N\) can be written in a closed form. In Table 3 we present results from a simulation study to check the effect of response rates \(\gamma \in \{\gamma_A, \gamma_B\}\) on the sample size \(N\). Here we assume \(\delta = \delta_{DP} = \delta_{DP} = 0\) and \(\gamma_A = \gamma_B = \gamma\). It is evident from Table 2 that the estimated minimum sample size is a decreasing function of \(\gamma\) for both the distinct-path and the shared-path comparisons of regimens. The estimated power in all the cases is close to the nominal power of 0.80.

6 STAR*D Data Analysis

In this section, we consider both the non-inferiority and the equivalence tests in the context of a SMART using data from the STAR*D trial mentioned in Section 1. The STAR*D was a multi-site, multi-level randomized trial to compare the effects of different treatment regimens for patients with major depressive disorder. A part of data analysis has been discussed in the Section 1. Here we demonstrate the non-inferiority/equivalence tests to compare some of the embedded regimens in the STAR*D trial. A brief description about the actual treatments given at different stages of STAR*D is presented in the Appendix 8.4. For illustration purpose, as done in Section 1 we categorize the various treatments as either mono-therapy \((M)\) or combination therapy \((C)\).

There are outcomes at different stages for various types of patients within the STAR*D study. For example, the end-of-first-stage outcomes are the last recorded outcomes for the patients who are either a responder or a dropout at the end of stage 1; on the other hand, patients who remain in the study have two observed outcomes at the end of stages 1 and 2, respectively. To overcome the discrepancy in the observed outcome pattern, we define the overall primary outcome of interest for our current study as:

\[
Y = R_1Y_1 + (1 - R_1)\left(D_1Y_1 + (1 - D_1)\left(\frac{Y_1 + Y_2}{2}\right)\right),
\]

where \(R_1\) represents the responder/non-responder status: 1 for responder and 0 for non-responder, after stage 1; and \(D_1\) denotes the dropout status: 1 for dropout and 0 for non-dropout, after stage 1. Once the overall primary outcome is available for all the patients, we define the inverse probability
Table 2: Simulation for non-inferiority test

| Distinct-path: \{d_1, d_3\} | Shared-path: \{d_1, d_4\} |
|------------------|------------------|
| (\gamma_A, \gamma_B) | (\gamma_A, \gamma_B) |
| \delta_{DP} | \delta_{SP} | N | \hat{\text{power}} | N | \hat{\text{power}} |
| (0.30, 0.35) | 0.76 | 243 | 0.80 | (0.10, 0.50) | 0.35 | 148 | 0.77 |
| (0.30, 0.40) | 0.71 | 221 | 0.81 | (0.20, 0.50) | 0.31 | 125 | 0.78 |
| (0.30, 0.50) | 0.62 | 187 | 0.81 | (0.30, 0.50) | 0.27 | 105 | 0.79 |
| (0.30, 0.60) | 0.53 | 160 | 0.80 | (0.40, 0.50) | 0.22 | 85 | 0.77 |
| (0.30, 0.70) | 0.44 | 138 | 0.80 | (0.50, 0.50) | 0.18 | 68 | 0.77 |
| (0.30, 0.35) | 1.02 | 389 | 0.80 | (0.10, 0.50) | 0.71 | 241 | 0.81 |
| (0.30, 0.40) | 0.98 | 353 | 0.81 | (0.20, 0.50) | 0.62 | 187 | 0.79 |
| (0.30, 0.50) | 0.89 | 290 | 0.79 | (0.30, 0.50) | 0.53 | 145 | 0.79 |
| (0.30, 0.60) | 0.80 | 240 | 0.79 | (0.40, 0.50) | 0.45 | 112 | 0.78 |
| (0.30, 0.70) | 0.71 | 201 | 0.79 | (0.50, 0.50) | 0.37 | 84 | 0.78 |
| (0.30, 0.35) | 1.29 | 740 | 0.80 | (0.10, 0.50) | 1.06 | 454 | 0.79 |
| (0.30, 0.40) | 1.25 | 653 | 0.80 | (0.20, 0.50) | 0.93 | 312 | 0.81 |
| (0.30, 0.50) | 1.16 | 505 | 0.81 | (0.30, 0.50) | 0.80 | 217 | 0.78 |
| (0.30, 0.60) | 1.07 | 399 | 0.80 | (0.40, 0.50) | 0.67 | 151 | 0.78 |
| (0.30, 0.70) | 0.98 | 321 | 0.81 | (0.50, 0.50) | 0.55 | 106 | 0.77 |
| (0.30, 0.35) | 1.55 | 1842 | 0.79 | (0.10, 0.50) | 1.41 | 1151 | 0.78 |
| (0.30, 0.40) | 1.51 | 1530 | 0.80 | (0.20, 0.50) | 1.24 | 617 | 0.80 |
| (0.30, 0.50) | 1.42 | 1059 | 0.78 | (0.30, 0.50) | 1.06 | 353 | 0.78 |
| (0.30, 0.60) | 1.33 | 769 | 0.80 | (0.40, 0.50) | 0.90 | 221 | 0.78 |
| (0.30, 0.70) | 1.24 | 579 | 0.81 | (0.50, 0.50) | 0.74 | 141 | 0.77 |
| (0.30, 0.35) | 1.82 | 11507 | 0.78 | (0.10, 0.50) | 1.76 | 6956 | 0.79 |
| (0.30, 0.40) | 1.78 | 7588 | 0.78 | (0.20, 0.50) | 1.54 | 1683 | 0.80 |
| (0.30, 0.50) | 1.69 | 3706 | 0.79 | (0.30, 0.50) | 1.33 | 695 | 0.80 |
| (0.30, 0.60) | 1.60 | 2157 | 0.80 | (0.40, 0.50) | 1.12 | 345 | 0.77 |
| (0.30, 0.70) | 1.51 | 1391 | 0.80 | (0.50, 0.50) | 0.92 | 191 | 0.77 |

weighted regimen mean for the regimen \(d\) as

\[
\bar{Y}_d = \frac{\sum_{i=1}^{N_d} W_i Y_i}{\sum_{i=1}^{N_d} W_i}, \text{ where}
\]

\[
W_i = \begin{cases} 
\frac{1}{\tau_{T_1}} & \text{if } R_1 = 1 \\
\frac{1}{\tau_{T_1}} \times \frac{1}{P(D_1 = 1 | T_1)} & \text{if } (1 - R_1)D_1 = 1 \\
\frac{1}{\tau_{T_1}} \times \frac{1}{1 - P(D_1 = 1 | T_1)} \times \frac{1}{\tau_{T_2}} & \text{if } (1 - R_1)(1 - D_1) = 1,
\end{cases}
\]

(13)

where \(\tau_{T_1}\) and \(\tau_{T_1T_2}\) are the first-stage and the second-stage randomization probabilities for \(T_1, T_2 \in \{M, C\}\), respectively. In case of even randomization at both stages 1 and 2, \(\tau_{T_1} = \tau_{T_1T_2} = 0.5\). Note that, a responder \((R_1 = 1)\) who started with mono-therapy in STAR*D, obtain the corresponding treatment sequence with probability \(\tau_M\); a non-responder who started with mono-therapy in the study and who is also a dropout after stage 1 obtain the corresponding treatment sequence with probability \(\tau_M \times P(D_1 = 1 | T_1 = M)\); and a non-dropout and non-responder patient obtain mono-therapies at both stages with probability \(\tau_M \times (1 - P(D_1 = 1 | T_1 = M)) \times \tau_{MM}\). Similarly probabilities for various trajectories can be computed for a patient starting with a combination therapy. The inverse probability weight \(W_i\) has been calculated accordingly. Note that, in (13), we have \(\tau_M = 1 - \tau_C\), \(\tau_{MM} = 1 - \tau_{MC}\) and \(\tau_{CM} = 1 - \tau_{CC}\).
Here we consider QIDS*, as defined in Section 1, to measure the efficacy of a treatment sequence. Note that the original QIDS score can take values in $[0, 27]$ (lesser is better). Hence, QIDS* also takes values in the same range but in opposite direction (higher is better). In the current article, for illustration purpose, we consider both the non-inferiority and equivalence test procedures to show that: (1) the regimen $d_{MC}$, having a significantly better QOL (see Table 1), is non-inferior in efficacy to the regimen $d_{CC}$, and (2) the relative superiority/inferiority in efficacy of the regimen $d_{MC}$ is equivalent to that of the regimen $d_{CC}$, respectively. In Table 4, the values of t-statistics corresponding to the non-inferiority and the non-superiority tests along with the rejection of null hypothesis decisions are shown against the different values of non-inferiority/equivalence margin $\theta$. The last column of Table 4 shows that all the non-superiority tests corresponding to different positive values of $\theta$ are rejected; this is due to high negative values of the corresponding t-statistics. Thus, for all the considered values of $\theta$, the weighted regimen mean of $d_{MC}$ is non-superior than that of the regimen $d_{CC}$. On the other hand, in case of non-inferiority test, the null hypotheses have been rejected when $\theta \geq 1$. Note that equivalence is established when both the non-inferiority and non-superiority tests are rejected. Based on the above data analysis, we can conclude that the regimen $d_{MC}$, having a (statistically) significantly better QOL profile than the regimen $d_{CC}$, is non-inferior (equivalent) in term of efficacy based on QIDS* with non-inferiority (equivalence) margin $\theta = 1$.

The above result is consistent with clinical intuition. According to the results, if a patient starts with the mono-therapy (M) at the first stage and then switches to the combination therapy (C) at the second stage, s/he will experience a better quality of life due to likely less side-effects from the “less aggressive” mono-therapy, and the efficacy of the treatment sequence is comparable (non-inferior/equivalent) to the treatment sequence consisting of combination therapies at both the stages. So it seems better to start with a less aggressive mono-therapy at the first stage to experience better QOL without compromising the efficacy (up to a pre-specified margin).
Table 4: STAR*D data analysis: t-statistic for different values of the non-inferiority/equivalence margin $\theta$, along with the results of both non-inferiority and non-superiority tests (reject the null hypothesis if t-statistic is greater than 1.645 for non-inferiority test and less than -1.645 for non-superiority test).

| $\theta$ | t-statistic    | Equation Value | Reject the null hypothesis | Decision |
|----------|----------------|----------------|----------------------------|----------|
|          | non-inferiority | non-superiority| non-inferiority            | non-superiority | Equation Value |
| 0.20     | -30.05         | -49.57         | N                          | Y        |
| 0.40     | -20.30         | -59.33         | N                          | Y        |
| 0.60     | -10.54         | -69.09         | N                          | Y        |
| 0.80     | -0.78          | -78.85         | N                          | Y        |
| 1.00     | 8.98           | -88.60         | Y                          | Y        |
| 1.20     | 18.74          | -98.36         | Y                          | Y        |
| 1.40     | 28.49          | -108.12        | Y                          | Y        |
| 1.60     | 38.25          | -117.88        | Y                          | Y        |
| 1.80     | 48.01          | -127.64        | Y                          | Y        |
| 2.00     | 57.77          | -137.39        | Y                          | Y        |

7 Discussion

In the current era of increased emphasis on ongoing, patient-centric care for chronic disorders, SMART is a very relevant, cutting-edge clinical trial design that is getting increasing visibility in the relevant scientific community. However, to the best of our knowledge, all SMARTs designed and/or conducted so far have employed the traditional (superiority) testing framework, which to some extent, has limited the wide uptake of SMARTs in clinical research. In the present article, we have developed non-inferiority and equivalence tests in the SMART design context, which we believe will pave the way for designing non-inferiority SMARTs or equivalence SMARTs in future. As such, SMARTs have not yet gained much ground in trials sponsored by pharmaceutical companies, part of the reason being fear of business loss from comparing (potentially multiple times) drugs manufactured by competing companies. We have argued in the current article that by employing a non-inferiority or equivalence testing framework, the concerns of pharmaceutical companies about SMARTs can be largely alleviated.

Another potential application area of the proposed methodology is mobile health (mHealth). In the current era of fast technological advancements, many health interventions are being delivered via smartphone or other mobile devices; these interventions are generally referred to as mHealth interventions. Time-varying mHealth interventions are special instances of dynamic treatment regimens, albeit with the technology-enabled platform and potentially a larger number of treatment stages compared to traditional drugs-based treatment regimens. The use of SMART designs in the mHealth domain has already been advocated for\[26\]. With the availability of non-inferiority and equivalence testing framework, the appeal of SMART in mHealth is expected to increase, because in order to launch a new mHealth intervention, the associated medical technology company (just like a pharmaceutical company, in case of drug trials) only needs to establish its non-inferiority compared to an existing mHealth intervention.

As illustrated in detail through an analysis of data from the well-known STAR*D trial, a first-generation SMART, the non-inferiority/equivalence testing framework offers a novel way to handle the trade-off between two competing outcomes, one of which can be a hard clinical outcome (e.g.,
QIDS, a measure of depression severity considered in STAR*D) and the other can be a more patient-centric outcome (e.g., quality of life). Previously, Lizotte et al.\(^\text{27}\) developed a procedure to combine multiple outcomes via a convex combination of them in the context of estimating optimal dynamic treatment regimens from SMART data. However, that methodology is only relevant for post hoc analysis of SMART data in order to find the optimal regimen for individual patients, but not relevant for comparing regimens that are already embedded in the SMART, and thus not pertinent to designing and sizing a SMART \textit{per se}. Our methodology offers a new approach to sizing a SMART; the topic of power and sample size has been dealt with in great detail.

In the current article, we have developed the non-inferiority and equivalence testing framework with respect to continuous outcomes only. However, the same methodology can be easily extended to binary outcomes, following the work of Ghosh et al.\(^\text{20}\). For survival outcomes, some additional work may be necessary to develop non-inferiority/equivalence versions of the weighted log-rank tests considered by Kidwell et al.\(^\text{19}\).

Statistically it is impossible to show that two treatments are identical, because an infinite sample size would be required for doing so.\(^\text{28}\) Thus, one can aim to establish similarity of two treatments under consideration in an equivalence trial, or to establish that one treatment is no worse than the other in a non-inferiority trial when one (new) treatment is “sufficiently close” to the other (standard or placebo). In this regard, the non-inferiority/equivalence margin \(\theta\) is a key concept that formalizes the notion of “sufficiently close” in non-inferiority/equivalence trials. However, setting an appropriate value of \(\theta\) remains a challenge. A too narrow margin may lead to unnecessarily large sample size requirements, whereas a too wide margin may end up establishing an equivalence for two interventions which are in fact substantially different from each other. Information from recent clinical trials involving the standard treatment can provide useful guidelines to choosing an appropriate value of \(\theta\).\(^\text{12}\) Further discussion on selection of non-inferiority margin in an RCT can be found in Chow and Song.\(^\text{29}\) In the context of SMARTs, if the treatment sequences consist of marketed drugs, data from the corresponding phase-III trials can be used to choose a value of \(\theta\). More in-depth thought about how to optimally set the value of \(\theta\) is necessary in the SMART design context; we view this as an important future work.
8 Appendix

8.1 Non-inferiority Test: Sample Size for Comparing Distinct-path Regimens

Here, based on Section 3.1, we calculate the sample size formula. We can write the power of the test as

\[ P_{H_1 \mid \text{reject } H_0} = P_{\mu_{d_3} - \mu_{d_1} = \delta} \left[ \frac{\bar{Y}_{d_1} - \bar{Y}_{d_3}}{\sqrt{\nu_{d_1 d_3}}} > z_\alpha \right] = P_{\mu_{d_3} - \mu_{d_1} = \delta} \left[ \frac{\bar{Y}_{d_1} - \bar{Y}_{d_3} + \delta}{\sqrt{\nu_{d_1 d_3}}} + \frac{\theta - \delta}{\sqrt{\nu_{d_1 d_3}}} > z_\alpha \right] = P \left[ Z > z_\alpha - \frac{\theta - \delta}{\sqrt{\nu_{d_1 d_3}}} \right], \]

since, \( Z = \frac{\bar{Y}_{d_3} - \bar{Y}_{d_1} + \delta}{\sqrt{\nu_{d_1 d_3}}} \rightarrow N(0,1) \) under \( H_1 \)

\[ 1 - \Phi \left( z_\alpha - \frac{\theta - \delta}{\sqrt{\nu_{d_1 d_3}}} \right), \]

or, \( 1 - \beta = 1 - \Phi \left( z_\alpha - \frac{\theta - \delta}{\sqrt{\nu_{d_1 d_3}}} \right), \)

\[ \Rightarrow \Phi \left( z_\alpha - \frac{\theta - \delta}{\sqrt{\nu_{d_1 d_3}}} \right) = \beta = \Phi(z_{1-\beta}) \]

\[ \Rightarrow z_\alpha - \frac{\theta - \delta}{\sqrt{\nu_{d_1 d_3}}} = z_{1-\beta} = -z_\beta \]

\[ \Rightarrow z_\alpha + z_\beta = \frac{\theta - \delta}{\sqrt{\nu_{d_1 d_3}}} = \frac{\theta - \delta}{\sqrt{2\sigma^2(4 - \gamma_A - \gamma_B)/N}} \]

\[ \Rightarrow (z_\alpha + z_\beta)^2 = \frac{N(\theta - \delta)^2}{2\sigma^2(4 - \gamma_A - \gamma_B)} \]

\[ \Rightarrow N = 2(z_\alpha + z_\beta)^2(4 - \gamma_A - \gamma_B)\frac{\sigma^2}{(\theta - \delta)^2}. \]

When \( \gamma_A = \gamma_B = \gamma \), then \( N = 4(z_\alpha + z_\beta)^2(2 - \gamma)\frac{\sigma^2}{(\theta - \delta)^2} \). Note that, \( \theta = \delta \) means \( H_0 \) is true, then \( N = \infty \). In all other cases, the above formula will give reasonable sample sizes. If \( 0 < \delta < \theta \), \( N \) will be high; if \( \delta < 0 \), \( N \) will be relatively small.

8.2 Non-inferiority Test: Variance Formula for Comparing Shared-path Regimens

Based on Section 3.2, we can write the variance as

\[ V(\bar{Y}_{d_1} - \bar{Y}_{d_2}) = V(\bar{Y}_{d_1}) + V(\bar{Y}_{d_2}) - 2Cov(\bar{Y}_{d_1}, \bar{Y}_{d_2}) = \frac{4(2 - \gamma_A)\sigma^2}{N} - 2Cov(\bar{Y}_{d_1}, \bar{Y}_{d_2}). \]
Denote \( R \): responder and \( NR \): non-responder. Now,

\[
\text{Cov}(\bar{Y}_{d_1}, \bar{Y}_{d_2}) = \text{Cov} \left( \frac{\sum_{i=1}^{N_{d_1}} W_i Y_i}{\sum_{i=1}^{N_{d_1}} W_i}, \frac{\sum_{i=1}^{N_{d_2}} W_i Y_i}{\sum_{i=1}^{N_{d_2}} W_i} \right)
\]

\[
= \frac{1}{(\sum_{i=1}^{N_{d_1}} W_i)(\sum_{i=1}^{N_{d_2}} W_i)} \text{Cov} \left( \sum_{i=1}^{N_{d_1}} W_i Y_i, \sum_{i=1}^{N_{d_2}} W_i Y_i \right)
\]

\[
= \frac{1}{N_2} \text{Cov} \left( \sum_{R \in d_1} W_i Y_i + \sum_{NR \in d_1} W_i Y_i, \sum_{R \in d_2} W_i Y_i + \sum_{NR \in d_2} W_i Y_i \right)
\]

\[
= \frac{1}{N_2} \text{Cov} \left( \sum_{R \in d_1} W_i Y_i, \sum_{R \in d_2} W_i Y_i \right)
\]

\[
= \frac{1}{N_2} V \left( \sum_{R \in d_1} W_i Y_i \right)
\]

\[
= \frac{2\gamma_A g^2}{N}.
\]

Hence, using [14], \( V(\bar{Y}_{d_2} - \bar{Y}_{d_1}) = \frac{8(1 - \gamma_A)\sigma^2}{N} \).

### 8.3 Equivalence Test: Power of Comparing Distinct-path Regimens

Here, based on Section 4.1, we provide the formula which can be used to obtain the sample size in the equivalence trial. We can write the power of the test as

\[
P_{H_1} \left[ \frac{\bar{Y}_{d_1} - \bar{Y}_{d_3} + \theta}{\sqrt{\nu_{d_1}^{DP}}} > z_\alpha \text{ and } \frac{\bar{Y}_{d_3} - \bar{Y}_{d_1} - \theta}{\sqrt{\nu_{d_1}^{DP}}} < -z_\alpha \right],
\]

where \( \mu_{d_3} - \mu_{d_1} \in (-\theta, \theta) \) under \( H_1 \). When \( \mu_{d_3} - \mu_{d_1} = \delta \in (-\theta, \theta) \), then the power

\[
P_{\mu_{d_3} - \mu_{d_1} = \delta} \left[ \frac{\bar{Y}_{d_1} - \bar{Y}_{d_3} + \theta}{\sqrt{\nu_{d_1}^{DP}}} > z_\alpha \text{ and } \frac{\bar{Y}_{d_3} - \bar{Y}_{d_1} - \theta}{\sqrt{\nu_{d_1}^{DP}}} < -z_\alpha \right]
\]

\[
= P_{\mu_{d_3} - \mu_{d_1} = \delta} \left[ \frac{\bar{Y}_{d_1} - \bar{Y}_{d_3} - \delta}{\sqrt{\nu_{d_1}^{DP}}} + \frac{\theta + \delta}{\sqrt{\nu_{d_1}^{DP}}} > z_\alpha \text{ and } \frac{\bar{Y}_{d_3} - \bar{Y}_{d_1} - \delta}{\sqrt{\nu_{d_1}^{DP}}} - \frac{\theta - \delta}{\sqrt{\nu_{d_1}^{DP}}} < -z_\alpha \right]
\]

\[
= P \left[ Z > z_\alpha - \frac{\theta + \delta}{\sqrt{\nu_{d_1}^{DP}}} \text{ and } Z < -z_\alpha + \frac{\theta - \delta}{\sqrt{\nu_{d_1}^{DP}}} \right]
\]

\[
= P \left[ z_\alpha - \frac{\theta + \delta}{\sqrt{\nu_{d_1}^{DP}}} < Z < -z_\alpha + \frac{\theta - \delta}{\sqrt{\nu_{d_1}^{DP}}} \right]
\]

\[
= \Phi \left( -z_\alpha + \frac{\theta - \delta}{\sqrt{\nu_{d_1}^{DP}}} \right) - \Phi \left( z_\alpha - \frac{\theta + \delta}{\sqrt{\nu_{d_1}^{DP}}} \right),
\]

(15)
which is equal to $1 - \beta$. In a special case, when $\mu_{d3} - \mu_{d1} = \delta = 0$, \[15\] becomes

$$1 - \beta = \Phi \left( -z_\alpha + \frac{\theta}{\sqrt{\nu_{d1d3}}} \right) - \Phi \left( z_\alpha - \frac{\theta}{\sqrt{\nu_{d1d3}}} \right) = 1 - 2\Phi \left( z_\alpha - \frac{\theta}{\sqrt{\nu_{d1d3}}} \right)$$

$$\Rightarrow \Phi \left( z_\alpha - \frac{\theta}{\sqrt{\nu_{d1d3}}} \right) = \frac{\beta}{2} = \Phi \left( z_\frac{1-\beta}{2} \right)$$

$$\Rightarrow z_\alpha - \frac{\theta}{\sqrt{\nu_{d1d3}}} = z_\frac{1-\beta}{2} = -z_\frac{\beta}{2}$$

$$\Rightarrow z_\alpha + z_\frac{\beta}{2} = \frac{\theta}{\sqrt{\nu_{d1d3}}} = \frac{\theta}{\sqrt{2\sigma^2(4 - \gamma_A - \gamma_B)/N}}$$

$$\Rightarrow \left( z_\alpha + z_\frac{\beta}{2} \right)^2 = \frac{N\theta^2}{2}\sigma^2(4 - \gamma_A - \gamma_B)$$

$$\Rightarrow N = 2 \left( z_\alpha + z_\frac{\beta}{2} \right)^2 \left( 4 - \gamma_A - \gamma_B \right)\frac{\sigma^2}{\theta^2}$$

8.4 STAR*D Description

In this section, we describe the different treatment options given to patients in the STAR*D study. In this article, among the four levels of the STAR*D study we only consider up to level three for illustration. We do not consider the level 1 in our analysis as all the patients in this level obtained the same treatment citalopram (CIT) (no randomization). The non-responders from level 1 went to level 2 where patients received one of up to seven different treatment options, which we categorize as: (i) mono therapy: bupropion sustained release (BUP-SR) or sertraline (SER) or venlafaxine extended release (VEN-XR) or cognitive psychotherapy (CT); and (ii) combination therapy: CIT + BUP-SR, or CIT + buspirone (BUS), or CIT + CT. Note that, the patients with non-satisfactory response from the randomized treatments either CT or CIT+CT at level 2 could enter a supplementary level 2A with randomized treatment options VEN-XR or BUP-SR. The non-responders from level 2 (and level 2A, if applicable) went to level 3. At the level 3, all the patients were randomized to (i) mono therapy options: mirtazapine (Mirt) or nortriptyline (NTP), or (ii) combination therapy options: lithium (Li) or thyroid hormone (Thy) combined with treatments given at the previous level; see supplementary material of Chakraborty et al.\[17\] for additional details.

8.5 Acknowledgements

Bibhas Chakraborty acknowledges support from the Duke-NUS Medical School, National University of Singapore, and an AcRF Tier 2 grant (MOE2015-T2-2-056) from the Ministry of Education, Singapore. We used the limited access datasets distributed from the National Institute of Mental Health (NIMH), USA. The STAR*D study (ClinicalTrials.gov identifier NCT00021528) was supported by NIMH Contract # N01MH90003 to the University of Texas Southwestern Medical Center. This manuscript reflects the views of the authors and may not reflect the views of the STAR*D Investigators or the NIH.
References

1 Lavori PW, Dawson R. A design for testing clinical strategies: Biased adaptive within-subject randomization. Journal of the Royal Statistical Society, Series A. 2000;163:29 – 38.

2 Lavori PW, Dawson R. Dynamic treatment regimes: Practical design considerations. Clinical Trials. 2004;1:9 – 20.

3 Thall PF, Millikan RE, Sung HG. Evaluating multiple treatment courses in clinical trials. Statistics in Medicine. 2000;30:1011 – 1128.

4 Murphy SA. An experimental design for the development of adaptive treatment strategies. Statistics in Medicine. 2005:24:1455 – 1481.

5 Chakraborty B, Moodie EEM. Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine. New York: Springer; 2013.

6 Kosorok MR, Moodie EEM, editors. Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine. SIAM; 2016.

7 Murphy SA. Optimal dynamic treatment regimes (with discussions). Journal of the Royal Statistical Society, Series B. 2003:65:331 – 366.

8 Robins JM. Optimal structural nested models for optimal sequential decisions. In: Lin DY, Heagerty P, editors. Proceedings of the Second Seattle Symposium on Biostatistics. New York: Springer; 2004. p. 189 – 326.

9 Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA, et al. Sequenced treatment alternatives to relieve depression (STAR*D): Rationale and design. Controlled Clinical Trials. 2004;25:119 – 142.

10 Stroup TS, McEvoy JP, Swartz MS, Byerly MJ, Glick ID, Canive JM, et al. The National Institute of Mental Health clinical antipschotic trials of intervention effectiveness (CATIE) project: schizophrenia trial design and protocol development. Schizophrenia Bulletin. 2003;29:15 – 31.

11 Chuang-Stein C, Follman D, Chappell R. University of Pennsylvania 6th annual conference on statistical issues in clinical trials: Dynamic treatment regimes (afternoon session). Clinical Trials. 2014;11(4):457–466. Available from: http://ctj.sagepub.com/content/11/4/457.

12 D´Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues-the encounters of academic consultants in statistics. Statistics in Medicine. 2003;22.

13 Walker E, Nowacki AS. Understanding Equivalence and Noninferiority Testing. Journal of General Internal Medicine. 2011;26(2):192–196.

14 Harrington DP. The randomized clinical trial. Journal of the American Statistical Association. 2000;95:312–315.

15 Association WM. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. Journal of the American Medical Association. 2013;310(20):2191–2194. Available from: http://dx.doi.org/10.1001/jama.2013.281053.
16 Ebbutt AF, Frith L. Practical issues in equivalence trials. Statistics in medicine. 1998;17(15-16):1691–1701.

17 Chakraborty B, Ghosh P, Moodie EEM, Rush AJ. Estimating optimal shared-parameter dynamic regimens with application to a multistage depression clinical trial. Biometrics. 2016;72(3):865–876. Available from: http://dx.doi.org/10.1111/biom.12493.

18 Schilling MF, Watkins AE. A Suggestion for Sunflower Plots. The American Statistician. 1994;48(4):303–305. Available from: http://www.jstor.org/stable/2684839.

19 Kidwell KL, Wahed AS. Weighted log-rank statistic to compare shared-path adaptive treatment strategies. Biostatistics. 2013;14(2):299–312. Available from: http://biostatistics.oxfordjournals.org/content/14/2/299.abstract.

20 Ghosh P, Cheung YK, Chakraborty B. Sample size calculations for clustered SMART designs. In: Kosorok MR, Moodie EEM, editors. Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine. SIAM; 2016.

21 Oetting AI, Levy JA, Weiss RD, Murphy SA. In: Shrout PE, Shrout PE, Keyes KM, Ornstein K, editors. Statistical Methodology for a SMART Design in the Development of Adaptive Treatment Strategies. Arlington, VA: American Psychiatric Publishing, Inc.; 2011. p. 179 – 205.

22 Nahum-Shani I, Qian M, Almiral D, Pelham W, Gnagy B, Fabiano G, et al. Experimental Design and Primary Data Analysis Methods for Comparing Adaptive Interventions. Psychological Methods. 2012;17:457 – 477.

23 Lauzon C, Caffo B. Easy Multiplicity Control in Equivalence Testing Using Two One-Sided Tests. The American Statistician. 2009;63(2):147–154.

24 Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics. 1987;15(6):657–680.

25 Fava M, Rush AJ, Trivedi MH, Nierenberg AA, Thase ME, Sackeim HA, et al. Background and rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study. Psychiatric Clinics of North America. 2003;26:457 – 494.

26 Kumar S, Nilsen WJ, Abernethy A, Atienza A, Patrick K, Pavel M, et al. Mobile health technology evaluation: the mHealth evidence workshop. American Journal of Preventive Medicine. 2013;45:228–236.

27 Lizotte DJ, Bowling M, Murphy SA. Linear Fitted-Q Iteration with Multiple Reward Functions. Journal of Machine Learning Research. 2012;13:3253 – 3295.

28 Friedman LM, Furberg CD, DeMets D, Reboussin DM, Granger CB. Basic Study Design. In: Fundamentals of Clinical Trials. Springer; 2015.

29 Chow SC, Song F. On Selection of Margin in Non-Inferiority Trials. Journal of Biometrics & Biostatistics. 2016;7:1–6.