Causal Inference in Hybrid Intervention Trials Involving Treatment Choice

Qi Long, Roderick J. Little, and Xihong Lin*

June 14, 2006

Abstract

Randomized allocation of treatments is a cornerstone of experimental design, but has drawbacks when a limited set of individuals are willing to be randomized, or the act of randomization undermines the success of the treatment. Choice-based experimental designs allow a subset of the participants to choose their treatments. We discuss here causal inferences for hybrid experimental designs where some participants are randomly allocated to treatments and others receive their treatment preference. This work was motivated by the "Women Take Pride" (WTP) study (Janevic et al., 2001), a doubly randomized preference trial (DRPT) to assess behavioral interventions for women with heart disease. We propose a model for estimating the causal effects in the subpopulations defined by treatment preferences, and hence preference effects. An EM algorithm is described for computing maximum likelihood estimates of the model parameters. The method is illustrated by analyzing sickness impact profile (SIP) scores and treatment adherence in the WTP data. Our results show 1) some evidence that SIP scores were improved when women received their prefered treatment; and 2) strong preference effects on program adherence; that is, women assigned to their prefered treatment were more likely to adhere to the program. We also provide a framework for assessing the DRPT and other hybrid trial designs, and discuss some alternative designs from the perspective of the strength of assumptions required to make causal inferences.

KEY WORDS: Clinical Trials; Doubly Randomized Preference Trials; EM algorithm; Partially Randomized Preference Trials; Randomization; Selection Bias.

*Qi Long is Rollins Assistant Professor, Department of Biostatistics, Emory University, Atlanta, GA, 30329; Roderick J. Little is Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109; Xihong Lin is Professor, Department of Biostatistics, Harvard University, Boston, MA, 02115. This research is supported by the National Cancer Institute grant R01CA76404. We would like to thank Noreen Clark for providing the WTP data and useful comments.

1 Introduction

Randomized assignment of subjects to treatments is a cornerstone of good experimental design. With full compliance and no missing data, randomization allows valid estimates of average treatment effects under minimal assumptions, by avoiding selection bias and ensuring that the observed mean for each treatment is an unbiased estimate of the overall mean if all individuals in the population had received that treatment (e.g., Little and Rubin, 2001).

However, it is also well known that randomization does not solve all problems in experiments involving alternative treatments. It is not always ethically feasible, as in medical trials when the principle of equipoise is not widely accepted, or in assessments of potentially harmful environmental effects that are unlikely to be beneficial. Inferences are only possible for the subset of subjects willing to be randomized, potentially excluding a significant fraction of the population, including subjects with strong treatment preferences. A behavioral treatment, in which strong motivation on behalf of the participant is required and treatment assignment cannot be blinded, may be more successful if subjects are allowed to choose, rather than are randomized to a treatment. Random assignment to a treatment perceived to be inferior may lead to issues of noncompliance and missing data that undermine the randomization and complicate causal inferences.

An alternative to randomization is to simply allow participants to choose their treatments. Participants that choose their treatment are more likely to participate and fully comply with the protocol, and the trial may more realistically measure the outcome of the treatment in the population of interest. On the other hand, this approach has all the well known problems of observational studies, in that treatment comparisons are obscured by the confounding effects of selection. Few experimentalists would accept the notion that the advantages of these designs compensate for their major weaknesses.

A natural question is whether there are hybrid designs between the extremes of randomization or choice that improve on either design. One candidate is Zelen's (1990) randomized consent design, which reverses the usual order of consent and randomization, by randomizing prior to consent and predicating the consent process on the randomized treatment. In a study with two alternative treatments (say A and B), participants are randomized to treatment (say A or B). Those randomized to A are asked if they are willing to receive A, after a discussion of the two treatments. If the participant agrees, A is given. If not, then B is given. The same procedure is followed in the other arm, with roles of A and B reversed. Zelen (1990) shows that this design allows for valid tests of the null hypothesis of no treatment effect, and can be more powerful than a randomized design restricted to participants willing to be randomized. However, the ethics of describing the treatments after the treatment assignment has already been made have been questioned (Ellenberg, 1992). See Altman et al. (1995) for more discussion of this design.

We consider here other hybrid designs that combine features of randomization and patient choice of treatments (Lambert and Wood, 2000). Perhaps the simplest approach is to ask the participants' treatment preferences in the context of a conventional fully randomized trial (Torgerson et al., 1996), and use that information as a covariate or effect-modifier; however, randomizing to a treatment other than the stated preference may be problematic. A more radical approach is the partially randomized preference trial (PRPT), where participants who are willing to be randomized to the treatments are randomized, and those who are not are assigned to their preferred treatment (Brewin and Bradley, 1989). A variation on this theme is the doubly-randomized preference trial (DRPT), where participants are randomized into a "randomization arm", within which treatments are randomized, and a "preference arm", within which participants get to choose their treatments. Versions of the DRPT are described by Rücker (1989), Wennberg et al. (1993) and Janevic et al. (2003). It seems plausible that in PRPT's and DRPT's, the additional information on participants who get to choose their treatments might usefully supplement the information from the participants who are randomized, although as discussed below this often requires modeling assumptions. Overall, hybrid designs enable us to estimate the preference effects and causal effects in subpopulations defined by treatment preferences, which cannot be estimated in a completely randomized trial.

We consider causal inference for these designs within the framework of potential outcomes, also known as Rubin's causal model (Holland, 1986). Originally formalized by Neyman (1923) in the context of randomized experiments, this framework was generalized and extended by Rubin (1974, 1977, 1978) to nonrandomized studies. The key underlying idea is that causal estimands are comparisons of the potential outcomes that would have been observed under different exposures of the same units to treatments at a particular place and time. Robins (1986, 1987) extended Rubin's "point treatment" potential outcome framework to evaluate direct and indirect effects of time-varying treatments in experimental and observational longitudinal studies. However, those methods are not directly applicable to the hybrid designs discussed in this paper.

The first objective of this paper is to propose a general model for assessing preference effects on the outcomes of interest. Our second objective is to propose a framework based on recent statistical ideas of causal inference for assessing the hybrid designs described above, and extensions. The basic idea is to classify individuals in the population into strata, which may or may not be observed for participants in the trial, and then assess assumptions required to identify the causal effects of treatments within these strata. Causal effects are defined as the difference in average outcome if all individuals within a stratum were assigned to treatment A and if all individuals within a stratum were assigned to treatment B (e.g. Rubin, 1974).

Our paper is motivated by the "Women Take Pride" study (Janevic et al., 2003), which utilized a DRPT to assess behavioral interventions for women with heart disease. In Section 2 we describe the design of that study, and show how to estimate causal effects in subpopulations defined by treatment preference, using a method of moments approach similar to that proposed by Rücker (1989). In Section 3, we propose a conceptual model for analyzing DRPTs and other hybrid trial designs. In Section 4, we propose a more general likelihood-based method of analysis that accommodates covariates and estimates causal parameters using the EM algorithm, and in Section 5 we apply these methods to the WTP data. In Section 6, we present a simulation study to evaluate the performance of the proposed method. In Section 7 we discuss alternative designs from the perspective of the strength of assumptions required to make causal inference. Section 8 presents some concluding remarks.

2 The "Women Take Pride" Study

The "Woman Take Pride" (WTP) intervention study (Janevic et al., 2001) concerns women aged 60 years and older with diagnosed cardiac disease being treated by daily heart medication. The interventions are behavioral programs aimed at enhancing the women's ability to manage their disease, based on principles of self-regulation from social cognitive theory. The comparison is between two versions of an intervention consisting of six weekly units: a Group treatment ($T_i = A$), where 6-8 women meet for 2-2 1/2 hours in a group setting; and a Self-directed treatment ($T_i = B$) where the participant studies at home following an initial orientation session. Motivation and support are provided through the social environment in the Group version, and through weekly telephone calls from the health educator and peer leader in the Self-directed treatment. These two versions of the intervention present the same material and only differ in format.

A DRPT design where some participants choose their treatment was seen as preferable to a completely randomized design in this setting, since the choice more accurately reflects a clinical situation. In a "real world" setting, patients may well have a preference for a group or self-directed format, and preference may well impact the motivation and adherence of the participants, and hence the success of the treatment. The design is summarized in Figure 1. At the first stage, a total of 3079 women with heart disease were randomized to a Random arm ($W_i = R$), where treatments were randomly assigned, and a Choice arm ($W_i = P$), where participants received their preferred treatment. Within the Random arm ($n = 1613$), 575 (35.6%) women agreed to participate; within the Choice arm ($n = 1466$), 496 (33.8%) agreed to participate. At the second stage, the women in the Random arm were randomized to three groups: control ($n = 184$) , the Group treatment A ($n = 190$) and the Self-Directed treatment B ($n = 201$); women in the Choice arm were asked to choose between Treatment A ($n = 321$) and Treatment B ($n = 175$). We do not analyze the data for the Control group here, since our focus is on comparisons of the Group and Self-Directed Treatments, and our general conceptual model is more simply presented for the case of two treatments. Extensions of our methods to more than two treatments are straightforward, however.

Primary outcomes Y of the study are measures of improved physical and psychosocial functioning, frequency and severity of symptoms, and health-care utilization measured at baseline, 4, 12 and 18 months. In this paper, we analyze physical, psychological and total scores of the sickness impact profile (SIP) at month 12. These are measures of physical, psychological and total functional health status, scored between 0 and 100, with higher scores indicating greater impairment due to illness (Bergner et al., 1981). To improve normality assumptions we analyze all three SIP scores on the log-transformed scale after adding a constant 0.05. We conduct an intent-to-treat analysis of the SIP score outcomes. However, we also analyze a measure of treatment adherence as a secondary outcome measure, specifically a binary variable for whether a woman completed at least one unit of materials. Table 1a summarizes the observed data for those outcomes.

Table 1a shows that in the Random arm ($W_i = R$), women assigned the Self-Directed treatment ($T_i = A$) have higher average SIP scores than women assigned the Group treatment ($T_i = B$); the same trend is observed in the Choice arm ($W_i = P$), with a greater mean difference. Adherence rates are 76% for both treatments in the Random arm, while in the Choice arm the adherence rate is higher for the Group treatment (93%) than for the Self-Directed treatment (77%).

The treatment comparisons in the Random arm are valid because of the random allocation. On the other hand, the treatment comparisons in the Choice arm are potentially biased by the effects of self-selection. Let C_i denote treatment preference, with $C_i = A$ if an individual prefers treatment A and $C_i = B$ if an individual prefers treatment B. The mean outcome for treatment A are for individuals with $C_i = A$, and the mean outcome for treatment B is for individuals with $C_i = B$. The comparison of these two means does not estimate a causal effect in a particular population, which is the key requirement for a causal effect in Rubin's (1974) sense. A direct comparison of these two means requires the very debatable assumption that the subpopulation with $C_i = A$ and the subpopulation with $C_i = B$ are equivalent with respect to treatment outcomes. This assumption might be improved by regression adjustment for known characteristics of participants in the two groups, but as in any observational setting, such adjustments do not necessarily remove the bias. A direct comparison of individuals with $C_i = A(B)$ in the choice arm and individuals with $T_i = A(B)$ in the random arm as was done in Janevic et al. (2003), is problematic for similar reasons.

A causal analysis that addresses the issue of choice is to construct estimates of mean outcomes of treatments A and B within each of the two preference subpopulations $C_i = A$ and $C_i = B$. This is not possible from data in the Choice arm alone, because it requires outcome data for participants who do not receive their treatment of choice. However, it can be addressed with a DRPT, since some participants in the Random arm do not receive their treatment of choice, and treatment assignment remains random within the two preference subpopulations, $C_i = A$ and $C_i = B$. Specifically, define

- $\mu(A)$=overall mean outcome if assigned to treatment A in the whole population
- $\mu_A(A)$=mean outcome if assigned to treatment A in the subpopulation that prefers A ($C_i = A$)
- $\mu_B(A)$=mean outcome if assigned to A in the subpopulation that prefers B ($C_i = B$),

and define $\mu(B)$, $\mu_A(B)$ and $\mu_B(B)$ as the corresponding mean outcomes if assigned to the Group treatment B, in the overall population and the two subpopulations respectively. Let π_B be the proportion of the population that prefers the Group treatment ($C_i = B$). Then

$$\begin{aligned}\mu(A) &= \pi_B \mu_B(A) + (1 - \pi_B)\mu_A(A) \\ \mu(B) &= \pi_B \mu_B(B) + (1 - \pi_B)\mu_A(B).\end{aligned}$$

From the choice arm, we can estimate $\hat{\mu}_A(A)$, $\hat{\mu}_B(B)$, and $\hat{\pi}_B = 321/496 = 0.65$ (Fig. 1). From the random arm, we can estimate $\hat{\mu}(A)$ and $\hat{\mu}(B)$. Thus, $\hat{\mu}_B(A)$ and $\hat{\mu}_A(B)$ can be estimated by solving a set of linear equations.

Let $\theta_B = \mu_B(B) - \mu_B(A)$, the difference in outcome means for treatments B and A in the subpopulation that prefers B. Then θ_B can be estimated as

$$\hat{\theta}_B = \hat{\mu}_B(B) - \hat{\mu}_B(A),$$

Similarly in the subpopulation that prefers A, we have $\theta_A = \mu_B(B) - \mu_B(A)$, which is estimated by

$$\hat{\theta}_A = \hat{\mu}_A(B) - \hat{\mu}_A(A).$$

The difference $\theta_B - \theta_A$ is defined as the preference effect for the two treatments, and measures the extent to which treatment preference modifies the treatment effect.

We first apply this method to our health outcome measures in the WTP study. Our results (Table 1b) show that for women who preferred the Group format, average SIP physical scores at month 12 were lower (-0.370) when they were assigned to the Group format than when they were assigned to the Self-Directed format; for women who preferred the SD format, average SIP physical scores at month 12 were higher (+0.080) when they were assigned to the Group format than when they were assigned to the Self-Directed format. These results, though not statistically significant, are in the direction of women having better physical functional health status when assigned their treatment of choice. Our results also show a similar trend for the other two SIP scores (results not included). The intervention effects as well as the preference effects, however, are not statistically significant.

We now apply our method to study intervention adherence. The results show that for women who prefer the Group format, the adherence rate when assigned to the Group format is estimated to be 18% ($P < 0.001$) higher than the adherence rate when assigned to the Self-Directed format; for women who prefer the Self-Directed format, the adherence rate when assigned to the Self-Directed format is estimated to be 33% ($P < 0.001$) higher than the adherence rate when assigned to the Group format. These results indicate that women are more likely to adhere to the program they prefer. The preference effect, defined as the treatment effect difference between the two subpopulations, is thus estimated as $\hat{\theta}_B - \hat{\theta}_A = 0.51$ ($P < 0.001$). It follows that the treatment effects on adherence are highly significant in the two subpopulations and they are significantly different. The results suggest that the very similar adherence rates for the two intervention groups in the Random arm mask strong preference effects, with much higher adherence rates for the prefered interventions.

This analysis method is a more formal description of the method of Rücker (1989). It does not require a distributional assumption and is applicable to estimating causal effects of an arbitrary outcome, e.g., the causal effect of a health outcome at a post-intervention time point such as 12 month. We now describe a framework that elucidates the implicit assumptions in this analysis, and then generalize the analysis to include covariates.

3 A Conceptual Model for Analyzing Hybrid Trial Designs

We present a general conceptual model for assessing hybrid intervention trials like the WTP study, which clarifies assumptions that are implicit in the above analysis. This framework will also be applied to assess other designs in Section 7. For simplicity we focus on designs involving just two treatments A and B, although the framework extends in an obvious way to designs with more than two treatments. We first stratify the target population into five groups (Figure 2(a)):

1. The set of individuals unwilling to participate even if given their choice of treatments ($\overline{P}$). Clearly we cannot learn anything empirically about treatment effects for this group without

making assumptions that relate it to a group we can study. We do not consider this group further here.

2. The set of individuals willing to participate if given the choice of treatment (P). We stratify this group into four subpopulations:

   (a) Individuals that prefer A and will not participate unless allowed to choose A ($P\overline{R}A$).

   (b) Individuals that prefer A but are willing to participate in a randomized trial (PRA).

   (c) Individuals that prefer B but are willing to participate in a randomized trial (PRB).

   (d) Individuals that prefer B and will not participate unless allowed to choose B ($P\overline{R}B$).

We consider two versions of each treatment, a version where the treatment is chosen by the participant ($\mathcal{A}_\mathcal{C}$ and $\mathcal{B}_\mathcal{C}$) and a version where the treatment is assigned by randomization ($\mathcal{A}_\mathcal{R}$ and $\mathcal{B}_\mathcal{R}$). We allow for the possibility that outcomes under these two versions of each treatment might differ. Throughout we assume the potential outcomes for each individual do not depend on the treatment status of other individuals in the sample, the so-called Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1978).

The table in Figure 2(b) results from crossing subpopulation stratum with treatment. Mean outcomes in the empty cells can be estimated from the data, but mean outcomes in the cells labeled $\overline{F}$ are inestimable, since they are a-priori counterfactuals (Angrist, Imbens and Rubin, 1996) (AIR): we cannot observe outcomes in these cells under any design. For example, we do not get to see the effect of randomizing to A ($\mathcal{A}_\mathcal{R}$) in the subpopulation of individuals who prefer A but will not participate in a randomized trial ($P\overline{R}A$). Opinions differ on the extent to which it is meaningful to consider treatment outcomes in such cells, and this needs to be considered in the specific context of each trial. In our discussion we follow AIR and focus attention on outcomes that are measurable under some design, that is, the cells without $\overline{F}$'s in Figure 2(b). Causal comparisons (in Rubin's sense) are comparisons of column means within rows of Figure 2(b). Comparisons between means in different rows are not causal since they concern different subpopulations.

The WTP study described above is an example of a particular hybrid trial design, the Doubly Randomized Preference Trial (DRPT), which has the generic form of Figure 3(a). People willing to participate (P) are first randomized to a choice arm and a random arm. Within the choice arm, people receive their treatment of choice (PA or PB). Within the random arm, individuals willing to be randomized (R) are randomized to $\mathcal{A}_\mathcal{R}$ or $\mathcal{B}_\mathcal{R}$, and those not willing to be randomized do not participate. Figure 3(b) indicates that mean outcomes can be estimated directly for four pooled subpopulations: the mean for $\mathcal{A}_{\mathcal{C}}$ in the subpopulation that prefers A, namely $PRA \bigcup P\overline{R}A$; the means for $\mathcal{A}_{\mathcal{R}}$ and $\mathcal{B}_{\mathcal{R}}$ in the subpopulation that is willing to be randomized, namely $PRA \bigcup PRB$; and the mean for $\mathcal{B}_{\mathcal{C}}$ in the subpopulation that prefers B, namely $PRB \bigcup P\overline{R}B$. The only causal effect that is estimable directly without additional assumptions is the comparison of $\mathcal{A}_{\mathcal{R}}$ and $\mathcal{B}_{\mathcal{R}}$ in the combined population $PR = PRA \cup PRB$; this is the treatment comparison from the random arm of the study.

If the outcome for an individual randomly assigned to a treatment is the same as if that individual had chosen that treatment, then $\mathcal{A}_{\mathcal{R}} = \mathcal{A}_{\mathcal{C}} = \mathcal{A}$ and $\mathcal{B}_{\mathcal{R}} = \mathcal{B}_{\mathcal{C}} = \mathcal{B}$. We follow other authors by calling this assumption the "exclusion restriction" (ER), since it is an example of an exclusion restriction in the sense that the term is used in econometrics (Angrist and Rubin, 1996). Under the ER assumption, the four columns in Figure 3(b) reduce to two. Additional assumptions are still needed to estimate the individual cells in the table, since there are eight cells (two a-priori counterfactual) and only four means can be directly estimated from the data. Suppose now we also assume that the random arm and choice arm participants are random samples from the same population, that is $PRA = PA$ and $PRB = PB$. We call this assumption "no selection bias from randomization" (NSBR), which allows us to combine the information from the Random and Choice arms of the study. Under ER and NSBR, the table in Figure 3(b) collapses to Figure 3(c) with just four cells, and the mean outcomes of A and B in the PA and PB subpopulations are then identified. The method described in Section 2 estimates these means, using information from the random and choice arms.

In particular, the analysis of the WTP data in the previous section implicitly makes the ER and NSBR assumptions. The mean outcomes in the two diagonal cells of Figure 3(c) are estimated from the Choice arm, and the column marginal mean outcomes are estimated from the Random arm. The remaining off-diagonal cells can then be estimated, with the proportion of participants that prefer A estimated from the choice arm. In support of the NSBR assumption, we note that the proportion of screened individuals agreeing to participate is comparable in the randomization (575/1613) and choice (496/1466) arms. If a sizeable proportion of the population only participated if given their treatment of choice, we would expect the participation rate to be higher in the choice arm than in the randomization arm. We discuss designs under which the ER and NSBR assumptions can be relaxed in Section 7.

4 A General Model for a DRPT with covariates

Janevic, et al. (2003) found that the probability of choosing each treatment was affected by demographic variables and disease severity. The outcomes within the two subpopulations ($C_i = A$ and

$C_i = B$) are also likely to be affected by baseline covariates other than the treatments. We hence extend the analysis of the previous section to accommodate covariates. As before, we consider a DRPT with two treatments, generically denoted as A and B. In this section, we assume ER and NSBR within the subpopulations defined by values of the covariates.

4.1 The Model

Suppose the data are comprised of n subjects. For subject i, let Y_i denote the observed outcome of interest, T_i denote the treatment assignment (A or B), C_i denote the treatment preference (A or B), $W_i = R$ if subject i is randomized to the random arm and $W_i = P$ if subject i is randomized to the choice arm, X_{1i} be a set of covariates associated with Y_i, and X_{2i} be a set of covariates associated with C_i, where X_{1i} and X_{2i} may overlap. In the application to the WTP data, the observed outcome of interest is the SIP physical score, the covariates associated with the SIP physical score are age, employment status and some baseline measures, and the covariates associated with treatment preference C_i are employment status, baseline total symptom impact and baseline SIP physical score.

Let $Y_i(A)$ and $Y_i(B)$ be the potential outcomes of Y for subject i when $T_i = A$, and $T_i = B$, respectively. The average causal effect of treatment assignment for the whole population is

$$\theta = E\left\{Y_i(B)\right\} - E\left\{Y_i(A)\right\},$$

The average causal effect of treatment assignment for the subpopulation preferring treatment m, ($m = A, B$) and having covariates X_1 is

$$\theta_m(X_1) = E\left\{Y_i(B)|C_i = m, X_1\right\} - E\left\{Y_i(A)|C_i = m, X_1\right\}.$$

Averaging over the distribution of X_1, the average causal effect of treatment assignment for the subpopulation preferring treatment m equals to

$$\theta_m = E_{X_1|C_i=m}\left[E\left\{Y_i(B)|C_i = m, X_1\right\} - E\left\{Y_i(A)|C_i = m, X_1\right\}\right].$$

The causal parameters $\theta_m(X_1)$ and θ_m can be related to estimable quantities under the ER and NSBR assumptions. Specifically, since subjects are randomized to the Choice or Random arms, we have

$$E\left\{Y_i(j)|C_i = m, X_1\right\} = E\left\{Y_i(j)|C_i = m, X_1, W_i\right\},$$

where m and j take values A, B. In the Random arm, since subjects are randomized to treatment groups, we have

$$E\left\{Y_i(j)|C_i = m, X_1, W_i = R\right\} \quad = \quad E\left\{Y_i(j)|T_i = j, C_i = m, X_1, W_i = R\right\}$$

$$= \quad E(Y_i|T_i = j, C_i = m, X_1, W_i = R).$$

In the Choice arm, subjects are assigned to treatments they prefer, that is, $T_i = C_i$. Thus for $j = A, B$,

$$\begin{aligned} E\left\{Y_i(j)|C_i = j, X_1, W_i = P\right\} &= E\left\{Y_i(j)|T_i = C_i, C_i = j, X_1, W_i = P\right\} \\ &= E(Y_i|T_i = j, C_i = j, X_1, W_i = C). \end{aligned}$$

We can not estimate $E\{Y_i(A)|C_i = B, X_1, W_i = P\}$ and $E\{Y_i(B)|C_i = A, X_1, W_i = P\}$ from data from the Choice arm alone. However, under the ER and NSBR assumptions, we have

$$\begin{aligned} E\{Y_i(A)|C_i = B, X_1, W_i = P\} &= E\{Y_i(A)|T_i = A, C_i = B, X_1, W_i = R\} = E\{Y_i|T_i = A, C_i = B, X_1, W_i = R\} \\ E\{Y_i(B)|C_i = A, X_1, W_i = P\} &= E\{Y_i(B)|T_i = B, C_i = A, X_1, W_i = R\} = E\{Y_i|T_i = B, C_i = A, X_1, W_i = R\} \end{aligned}$$

Hence, we can then use data in the random arm in conjunction with the data in the choice arm to estimate these quantities, by viewing each group in the random arm as a mixture of the two preference subpopulations.

For notational simplicity, we recode the values of T_i and C_i by replacing A by 1 and B by 0. We assume that the distribution of Y_i given T_i, C_i, X_{1i} and W_i belongs to the exponential family

$$f(Y_i|T_i, C_i, X_{1i}, W_i) = exp\left\{\frac{Y_i\gamma_i - b(\gamma_i)}{\phi a_i^{-1}} + c(Y_i, \phi)\right\},$$

where a_i is a known constant, ϕ is a scale parameter, γ_i is the canonical parameter, $b(\cdot)$ and $c(\cdot)$ are known functions. The mean of Y_i is $\mu_i = E(Y_i|T_i, C_i, X_{1i}, W_i) = b'(\gamma_i)$ and is assumed to have the form

$$g(\mu_i) = \beta_0 + X_{1i}^T\beta_{X_1} + T_i\beta_T + C_i\beta_C + T_iC_i\beta_{TC}, \tag{1}$$

where $g(\cdot)$ is a monotonic link function (McCullagh and Nelder, 1989). The model is completed by assuming the treatment preference C_i given X_{2i} follows a logistic model with $\pi_i = \Pr(C_i = 1|X_{2i})$ satisfying

$$\text{logit}(\pi_i) = \alpha_0 + X_{2i}^T\alpha_{X_2}. \tag{2}$$

The causal effect of treatment assignment for the subpopulation preferring treatment m ($m = 1, 0$) given covariates X_1 is

$$\begin{aligned} \theta_m(X_1) &= E\left\{Y(1)|C = m, X_1\right\} - E\left\{Y(0)|C = m, X_1\right\} \\ &= g^{-1}\left\{\beta_0 + X_1\beta_{X_1} + \beta_T + m(\beta_C + \beta_{TC})\right\} - g^{-1}(\beta_0 + X_1\beta_{X_1} + m\beta_C) \end{aligned}$$

The marginal causal effects, θ_m have the form

$$\theta_m = \int \theta_m(X_1 = x_1) f(x_1|C = m) dx_1, \tag{3}$$

where $f(x_1|C = m)$ can be empirically estimated from the Choice arm, that is, $\hat{f}(x_1|C = m) = \sum_i I(C_i = m, X_{1i} = x_1)/\sum_i I(C_i = m)$, where $I(\cdot)$ is an indicator function.

In the random arm (W_i, T_i, Y_i) are observed but C_i is not observed. In the choice arm, (W_i, T_i, Y_i, C_i) are observed and $T_i = C_i$, since subjects are assigned to their preferred treatment. Denote by $\delta = (\alpha, \beta, \phi)$ the parameter vector. Define $Y = (Y_1, Y_2, \ldots, Y_n)^T$, and T, C, X_1, X_2, W similarly. The observed data loglikelihood is given by

$$\begin{aligned} \ell(Y, C_{obs}|T, X_1, X_2, W; \delta) &= \sum_{W_i = P} \left[\log\{f(Y_i|T_i, X_{1i}, C_i)\} + C_i \log(\pi_i) + (1 - C_i)\log(1 - \pi_i)\right] \qquad (4) \\ &+ \sum_{W_i = R} \left[\log\{\pi_i f(Y_i|T_i, X_{1i}, C_i = 1) + (1 - \pi_i) f(Y_i|T_i, X_{1i}, C_i = 0)\}\right], \end{aligned}$$

where C_{obs} denote the observed C values for the Choice arm, and $f(Y_i|T_i, X_{1i}, C_i)$ follows the generalized linear model (1) and π_i follows the logistic model (2).

4.2 Estimation Using the EM Algorithm

An EM algorithm (Dempster, Laird and Rubin 1977) can be used to calculate the maximum likelihood (ML) estimate of δ for the above model. The complete data $(Y_i, C_i, T_i, X_{1i}, X_{2i}, W_i)$ have loglikelihood

$$\begin{aligned} \ell(Y, C|T, X_1, X_2, W; \delta) &= \sum_i \ell(Y_i, C_i|T_i, X_{1i}, X_{2i}, W_i; \delta) \\ &= \sum_i \left[\log\{f(Y_i|T_i, X_i, C_i, W_i)\} + C_i \log(\pi_i) + (1 - C_i)\log(1 - \pi_i)\right]. \qquad (5) \end{aligned}$$

The EM algorithm iterates between an E step, which replaces missing values of C_i in (5) by their conditional expectations given the observed data, and an M step, which maximizes the expected complete-data loglikelihood (5) to yield updated parameter estimates.

1. *E Step at the kth iteration.* We calculate the expected complete-data loglikelihood given Y, C_{obs}, T, X, Z, W and the current parameter estimates $\delta^{(k)}$, namely

$$\begin{aligned} & E\{\ell(Y, C|T, X_1, X_2, W)|Y, C_{obs}, T, X_1, X_2, W; \delta^{(k)}\} \\ = & \sum_{W_i = R,} \sum_{m = 0,1} w_{i,m}^{(k)} \ell(Y_i, C_i = m|T_i, X_{1i}, X_{2i}; \delta^{(k)}) + \sum_{W_i = P} \ell(Y_i, C_i|T_i, X_{1i}, X_{2i}; \delta^{(k)}), \qquad (6) \end{aligned}$$

where for participants in the random arm ($W_i = R$),

$$\begin{aligned} w_{i,m}^{(k)} &= p(C_i = m|Y_i, T_i, X_i; \delta^{(k)}) \\ &= \frac{f(Y_i|T_i, X_{1i}, C_i = m; \delta^{(k)}) p(C_i = m|T_i, X_{2i}; \delta^{(k)})}{\pi_i^{(k)} f(Y_i|T_i, X_{1i}, C_i = 1; \delta^{(k)}) + (1 - \pi_i^{(k)}) f(Y_i|T_i, X_{1i}, C_i = 0; \delta^{(k)})}. \end{aligned}$$

where $\pi_i^{(k)} = p(C_i = 1|T_i, X_{2i}; \delta^{(k)})$. The E-step estimates the weights $w_{i,m}^{(k)}$ in the weighted complete-data log-likelihood (5) for the random arm.

2. *M Step at the kth iteration.* This step updates the parameter estimates $\delta^{(k+1)}$ by maximizing the expected complete-data loglikelihood (5). We first construct an augmented data set as follows. Each observation in the random arm, with C_i missing, is replaced by two "filled-in" observations, in which the missing treatment preference indicator C_i is replaced by 0 and 1 respectively and the corresponding weights $w_{i,0}$ and $w_{i,1}$ are computed using the current estimates of the parameters. The observations in the choice arm are unchanged, with weights are set to one. Using this augmented data set, we fit a weighted generalized linear model for $\beta^{(k+1)}$ and $\phi^{(k+1)}$, and a weighted logistic model for $\alpha^{(k+1)}$.

The resulting estimates from the EM algorithm at convergence give the ML estimates of δ. For the special case of no covariates X_{1i}, X_{2i} and a binary outcome, the ML estimates of β transformed to the mean scale reduce to the method of moments estimates in Section 2.2. A similar EM algorithm was proposed by Ibrahim (1990) for missing categorical covariates. Our EM algorithm extends his algorithm to the hybrid randomized-preference design by estimating the weights in the random arm and fixing the weights to be one in the choice arm. We consider estimating standard errors of the ML estimates using bootstrap, the observed information and the approximation proposed in Ibrahim (1990). The observed information is obtained by directly computing the second derivative of the observed likelihood using a symbolic differentiation algorithm.

5 Analysis of the WTP data

We now apply the methods of Section 4 to the WTP data, to estimate preference effects adjusting for covariates. Group format is coded as 1 and Self-Directed format is coded as 0. Based on the previous analysis in Janevic, et al. (2003), we consider the following covariates in models (1) and (2): employment status, age, total symptom impact at baseline, which is a measure of symptom severity scored between 0 and 70 (Clark et al., 1997), and SIP scores at baseline.

Table 2 presents the estimates of the regression parameters in (1) and (2) and their estimated standard errors for the outcome SIP physical score. The results from model (2) show that employed women ($OR = 1.84$, $P = 0.046$), and women with greater physical limitations at baseline ($OR = 1.03$, $P = 0.012$) are more likely to choose the Self-Directed format , suggesting that these women tend to opt for the more flexible scheduling that the Self-Directed format provides. Women with a higher total symptom impact score at baseline are less likely to choose the Self-Directed format ($OR = 0.97$,

$P = 0.013$), suggesting that these women may be interested in opportunites to meet women in a similar situation, which the Group format provides. The magnitudes of these effects on treatment preference are comparable to those in Janevic et al. (2003), which includes more detailed discussion on predictors of program format preference. In terms of the covariate effects on SIP physical score, women with higher baseline SIP score or total symptom impact score have higher SIP physical score at month 12; other baseline covariates are not significant ($P > 0.05$). The significant interaction between treatment preference and treatment assignment ($P = 0.039$) suggests an effect of treatment preference for this outcome, with women having a better SIP physical score when assigned their treatment of choice.

Table 3 summarizes the treatment and preference effects on the subpopulation means for SIP physical score and the other two SIP outcomes. These adjusted means are obtained by integrating over the distributions of the covariates using equation (3). The treatment and preference effects for the SIP psychological and SIP Total scores have a similar pattern to that for SIP physical score, but are not statistically significant. Overall, our analysis shows limited evidence of a benefit of women getting the types of treatment they prefer. This finding addresses one of the hypotheses proposed by the investigators in the WTP study, that is, making both treatment formats available to women may be advantageous.

The last row of Table 3 shows strong preferences effects on treatment adherence. The marginal causal treatment effect on adherence on the probability scale for the subpopulation preferring the Group treatment is $\hat{\theta}_1 = 0.208$ ($P < 0.001$), and the marginal causal treatment effect on adherence on the probability scale for the subpopulation preferring the Self-Directed treatment is $\hat{\theta}_0 = -0.336$ ($P < 0.001$). These results show that women who prefer the Group treatment are 20.8% more likely to adhere to the treatment if assigned to the Group format than if assigned to the Self-Directed form, whereas women who prefer the Self-Directed treatment are 33.6% more likely to adhere if assigned to the Self-Directed treatment than if assigned to the Group treatment. These results are consistent with the findings in Section 2 that women are more likely to adhere to the program they prefer. The covariate-adjusted treatment effects are slightly stronger than those in Section 2.2 without covariate adjustments.

6 A Simulation Study

We conducted a simulation study to evaluate the finite sample performance of the proposed method. The design of the simulation study was similar to that of the WTP study. Each data set consisted of 1000 observations. Independent binary observations C_i of the treatment preference indicator were generated using the logistic model

$$logit\{\Pr(C_i = 1)\} = \alpha_0 + X_{2i}\alpha_1,$$

where $\alpha_1 = -1$, $\alpha_1 = 2$, and the X_{2i} were generated from a uniform distribution on the interval $[0, 1]$. The outcome variable Y_i was assumed to be binary and generated independently using the logistic model

$$logit\{E(y_i)\} = \beta_0 + X_{1i}\beta_1 + T_i\beta_T + C_i\beta_C + T_iC_i\beta_{TC},$$

where $\beta_0 = -2$, $\beta_1 = \beta_T = \beta_C = 2$, $\beta_{TC} = -2$, the X_{1i} were generated from a uniform distribution on the interval $[0, 1]$. Values of W_i denoting random versus choice arm were generated from a Bernoulli distribution with $P(W_i = R) = 0.5$. The treatment assignment indicators T_i were set equal to C_i in the choice arm ($W_i = P$), and were generated from a Bernoulli distribution with $P(T_i = 1) = 0.5$ in the random arm ($W_i = R$). The preference indicators C_i were set to missing in the random arm ($W_i = R$).

A total of 125 simulated data sets were generated and analyzed. Table 4 presents the simulation results. The point estimates are very close to the true values and there is no evidence of bias. We compared three methods for estimating the standard errors: the bootstrap method, the observed information, and the approximation given by Ibrahim (1990). These estimated standard errors can be compared with the empirical standard errors. Our results show that the observed information based standard errors are closest to the empirical standard errors, the bootstrap standard errors perform similarly but are slight overestimates for the coefficients of treatment preference and the interaction between treatment preference and treatment assignment. The approximation given by Ibrahim (1990) seems to slightly underestimate the standard errors. This might be due to the fact that the estimators of the regression parameters β and α in model (1) and (2) are assumed independent in the approximation.

7 Alternative Hybrid Trial Designs

We now consider some other hybrid designs using the framework developed in Section 3. A simple alternative to the DRPT is the partially randomized preference trial (PRPT) mentioned in Section 1. In this design, individuals willing to be randomized are assigned to the random arm and receive $\mathcal{A}_\mathcal{R}$ or $\mathcal{B}_\mathcal{R}$, and individuals not willing to be randomized are allowed to choose the treatment, $\mathcal{A}_\mathcal{C}$ or $\mathcal{B}_\mathcal{C}$ (Figure 4(a)). Thus the PR individuals are randomized to $\mathcal{A}_\mathcal{R}$ and $\mathcal{B}_\mathcal{R}$, the $P\overline{R}A$ individuals receive $\mathcal{A}_\mathcal{C}$ and the $P\overline{R}B$ individuals receive $\mathcal{B}_\mathcal{C}$. Figure 4(b) shows which of the cell means in Figure 2(b) can be directly estimated under this design (E); the cells denoted by $\overline{E}$ are not a-priori counterfactual

($\overline{F}$), but cannot be estimated without modeling assumptions, since outcomes for individuals in these cells are not directly observed. The randomized part of the design allows comparisons of $\mathcal{A}_{\mathcal{R}}$ and $\mathcal{B}_{\mathcal{R}}$ for the population PR willing to be randomized, but does not distinguish preference effects for this population. The choice arm provides estimates for the outcome under $\mathcal{A}_{\mathcal{C}}$ for $P\overline{R}A$ and the outcome under $\mathcal{B}_{\mathcal{C}}$ for $P\overline{R}B$, but additional modeling assumptions are needed for this information to yield estimates of causal effects within the $P\overline{R}A$ and $P\overline{R}B$ rows of the table. Under ER, $\mathcal{A}_{\mathcal{C}} = \mathcal{A}_{\mathcal{R}}$ and $\mathcal{B}_{\mathcal{C}} = \mathcal{B}_{\mathcal{R}}$, and we can compare the outcomes of A and B in $PRA \bigcup PRB$, and estimate the outcomes of A in $P\overline{R}A$ and B in $P\overline{R}B$. With the additional assumptions $P\overline{R}A = PRA = PA$ and $P\overline{R}B = PRB = PB$, the model is identified; the estimation approach of the previous section can be applied to this design. These assumptions imply the NSBR assumption $PRA = PA, PRB = PB$ that we used to identify the DRPT design, but the latter also apply when $P\overline{R}A$ and $P\overline{R}B$ are null sets, that is there is no one who would participate if assigned their preference but would not participate if randomized. In that case the DRPT can estimate preference effects, as we have seen, but the PRPT cannot, because everyone is assigned to the randomization arm where preference is not elicited. The framework helps to clarify the quite subtle differences in these two designs.

Suppose, however, that after adjustment for covariates, the average outcomes for treatment A were similar in $P\overline{R}A$ and $PRA \bigcup PRB$ (e.g. a test of the null hypothesis that the means are equal is not significant). Also, the average outcomes for treatment B were similar in $P\overline{R}B$ and $PRA \bigcup PRB$. Then one could argue that after covariate adjustment, $P\overline{R}A$ and $P\overline{R}B$ are homogeneous with respect to the outcomes of interest, and therefore the comparison of A in $P\overline{R}A$ with B in $P\overline{R}B$ is plausibly causal. The strata might be collapsed to provide a more precise estimate of the treatment effect. This kind of analysis would not be possible in a strictly randomized design that does not include the choice arm.

A simple way of estimating preference effects is to ask a direct question about treatment preference, and use the answers to that question to classify people directly into preference subpopulations. An important issue with this approach is whether it is feasible to combine such a question with randomization of treatments, and in particular whether asking the participant to answer a question about preference undermines their willingness to agree to randomization. However, this approach has been found feasible in some randomized trials (Torgerson et al., 1996). Figure 5 describes a DRPT design with a question about treatment preference, which we label DRPTQ, and shows the corresponding set of directly estimable means. The random arm allows us to estimate causal effects of treatments within each treatment preference group; the choice arm now provides additional information over DRPT. Specifically if NSBR is assumed then the ER can be checked, since the treatment effects can be estimated for both the randomized and chosen versions of the treatments. On the other hand if the ER is assumed, then the NSBR assumption can be assessed, since treatment effects can be estimated in the strata both willing and unwilling to be randomized. Even with no NSBR or ER assumption, useful bounding information might be derived by assuming that the outcome under $\mathcal{A}_{\mathcal{C}}(\mathcal{B}_{\mathcal{C}})$ is at least as good as the outcome under $\mathcal{A}_{\mathcal{R}}(\mathcal{B}_{\mathcal{R}})$. A direct question about treatment preference can also be added to the PRPT design.

Figure 6 shows a design that includes both a preliminary question about preferences and combines the features of the PRPT and DRPT designs. We call this a combined randomized preference trial with a question about preference (CRPTQ). The corresponding set of estimable effects, shown in Figure 6(b), includes all the effects that are not a-priori counterfactual. In this case all four versions of the treatment ($\mathcal{A}_{\mathcal{R}}$, $\mathcal{A}_{\mathcal{C}}$, $\mathcal{B}_{\mathcal{R}}$, $\mathcal{B}_{\mathcal{C}}$) can be compared for the subpopulations who prefer A and prefer B, and are willing to be randomized; the outcomes of the treatment can also be compared in the three subpopulations where they are not a-priori counterfactual. The NSBR and ER assumptions can be checked using this design. A further elaboration of the initial question is to ask the question using a Likert scale: SA = strongly prefer A; A = prefer A; N = neutral; B = prefer B, SB = strongly prefer B. This more detailed classification may provide more information for modeling treatment effects in the different strata.

8 Discussion

Although classical randomized trials enjoy many advantages, it is well recognized that they have limitations, particularly for behavioral interventions like the WTP trial. Hybrid designs like the DRPT provide an alternative to classical randomized trials to examine how treatment preference can influence treatment effects. However, the naive approach of comparing outcomes under the two treatments is flawed because of selection bias in the choice arm. We have proposed a two stage model for estimating the causal treatment effects and the preference effect in a DRPT for discrete and continuous outcomes, treating the preference indicator as missing in the random arm. An EM algorithm can be used for ML estimation of parameters. The analysis can be readily generalized to model other types of outcomes. For outcomes other than adherence our analysis is based on Intention-To-Treat (ITT), that is, it does not consider information about compliance. In future we plan to develop methods to estimate complier average causal effects, as discussed in Angrist, Imbens and Rubin (1996).

Our analyses of WTP data show limited evidence of a benefit to women of receiving their preferred treatment, so making both treatment formats available to women may be advantageous. It is also shown that there is strong preferences effects on treatment adherence, that is, allowing patients to choose their treatments improves treatment adherence. This is likely to lead to improved outcomes, to the extent that treatments are effective.

The framework in Section 3 distinguishes effects that cannot be estimated by any design without modeling assumptions (comparisons involving a-priori counterfactuals) and effects that are estimable under some designs but not under others. In particular, we show in Section 7 how designs that ask directly about treatment preference can avoid the need for the NSBR and ER assumptions. This advantage needs to be weighed against the potential drawbacks of asking about treatment preference in the context of randomization, and additional administrative overhead. Pilot studies might be helpful in casting light on the feasibility of such designs in a particular context. These more complex designs follow W.G. Cochran's principle of providing the opportunity to learn more by making hypotheses more complex (Rubin, 1984).

In conclusion, designs that combine randomization and choice seem to us a fruitful area for future research and application, and the ideas of causal inference and population stratification are helpful for guiding the design and resulting analysis.

References

Altman, D. G., Whitehead, J., Parmar, M. K. B., Stenning, S. P., Fayers, P. M., and Machin, D. (1995), "Randomised Consent Designs in Cancer Clinical Trials," *European Journal of Cancer*, 31A (12), 1934-1944.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables (Disc: p456-472)", *Journal of the American Statistical Association*, 91, 444-455.

Bergner, M., Bobbitt, R. A., Carter, W. B., and Gilson, B. S. (1981), "The Sickness Impact Profile: Development and Final Revision of a Health Status Measure," *Medical Care*, 19,787-805.

Brewin, C. R., and Bradley, C. (1989), "Patient Preferences and Randomised Clinical Trials," *British Medical Journal*, 299, 313-315.

Clark, N. M., Janz, N. K., Dodge, J. A., Schork, M. A., Wheeler, J. R. C., Liang, J., Keteyian, S. J., and Santinga, J. T. (1997), "Self-management of Heart Disease by Older Adults: Assessment of an Intervention Based on Social Cognitive Theory," *Research on Aging*, 19, 362-382.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society Series B*, 39, 1-38.

M., and Schoenfeld, D. A. (1992), "Statistical Issues Arising in AIDS Clinical Trials," *Journal of the American Statistical Association*, 87, 562-569.

Ellenberg, S. S., Finkelstein, D. M., and Schoenfeld, D. A. (1992), "Statistical Issues Arising in AIDS Clinical Trials," *Journal of the American Statistical Association*, 87, 562-569.

Holland, P. W. (1986), "Statistics and Causal Inference(with discussion)," *Journal of the American Statistical Association,* 81, 945-970.

Ibrahim, J. G. (1990), "Incomplete Data in Generalized Linear Models," *Journal of the American Statistical Association*, 85, 765-769.

Janevic, M. R., Janz, N. K., Lin, X., Pan, W., Sinco, B. R., and Clark, N. M. (2003), "The Role of Choice in Health Education Interventional Trials: A Review and Case Study," *Social Science and Medicine*, 56(7), 1581-1594.

Lambert, M. F., and Wood, J. (2000), "Incorporating Patient Preferences into Randomized Trials," *Journal of Clinical Epidemiology*, 53, 163-166.

Little, R. J., and Rubin, D. B. (2001), "Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches," *Annual Review of Public Health*, 21, 121-145.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*. Chapman and Hall, London.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," translated in *Statistical Science, (with discussion),* 5, 465-480, 1990.

Robins, J.M. (1986), "A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods - Application to Control of the Healthy Worker Survivor Effect," *Mathematical Modeling,* 7, 1393-1512.

Robins, J.M. (1987), Addendum to "A new Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods - Application to Control of the Healthy Worker Survivor Effect," it Computers and Mathematics with Applications, 14, 923-945.

Rubin, D.B. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology,* 66, 688-701.

Rubin, D.B. (1977), “Assignment to a Treatment Group On the Basis of a Covariate,” it Journal of Educational Statistics, 2, 1-26.

Rubin, D.B. (1978), “Bayesian Inference for Causal Effects: The Role of Randomization,” *The Annals of Statistics*, 6, 34-58.

Rubin, D.B. (1984), “William G. Cochran’s Contributions to the Design, Analysis and Evaluation of Observational Studies,” In *W. G. Cochran’s Impact on Statistics*, ed. PSRS Rao, J Sedransk, 37-69. New York: Wiley.

Rücker, G. (1989), “A Two-stage Trial Design for Testing Treatment, Self-selection, and Treatment Preference Effects,” *Statistics in Medicine*, 8, 477-485.

Torgerson, D. J., Klaber-Moffett, J., and Russell, I. T. (1996), “Patient Preferences in Randomized Trials: Threat or Opportunity?” *Journal of Health Services Research and Policy*, 1(4), 194-197.

Wennberg, J. E., Barry, M. J., Fowler, F. J., and Mulley, A. (1993), “Outcomes Research, PORTs, and Health Care Reform,” *Annals of the New York Academy of Sciences*, 703, 52-62.

Zelen, M. (1990), “Randomized Consent Designs for Clinical Trials: An Update,” *Statistics in Medicine*, 9, 645-656.

Table 1. Summary Data from the WTP Study

(a) Observed Sample Means

Outcome	Random Arm		Choice Arm	
	Group	Self-Directed	Group	Self-Directed
12-month SIP Physical Score[1]	0.770	0.982	0.557	1.087
12-month SIP Psychological Score[1]	0.083	0.277	-0.179	0.407
12-month SIP Total Score[1]	1.160	1.384	1.096	1.507
Adherence Rate	0.76	0.76	0.93	0.77

(b) Estimated Subpopulation Means Using Method of Moments

Outcome	Preference	Treatment		Treatment Effect
	Subpopulation	Group	Self-Directed	(P-value)
12-month SIP Physical Score[1]	Group	0.557	0.927	-0.370 (0.279)
	Self-Directed	1.177	1.087	0.090 (0.900)
Adherence Rate	Group	0.93	0.75	-0.18 (<0.001)
	Self-Directed	0.44	0.77	0.33 (<0.001)

[1] on log-transformed scale.

Table 2. Parameter Estimates of Models (1) and (2), and Estimated Standard Errors, for SIP physical score

Parameter	Estimate	SE[1]	P-value
Outcome Model (1)			
Intercept	-0.606	0.757	0.402
Age	0.010	0.010	0.309
Employment[2]	-0.171	0.222	0.447
Baseline Total Symptom Impact	0.025	0.006	< 0.001
Baseline SIP Physical Score	0.628	0.035	< 0.001
Treatment Assignment	0.587	0.335	0.071
Treatment Preference	-0.230	0.273	0.404
Preference*Assignment	-0.918	0.434	0.039
Model (2)			
Intercept	0.620	0.181	0.0008
Employment	-0.612	0.308	0.046
Baseline Total Symptom Impact	0.027	0.012	0.013
Baseline SIP Physical Score	-0.033	0.013	0.012

[1] Computed using the bootstrap method based on 2500 bootstrapped data sets.

[2] 1=Employed, 0=Unemployed.

Table 3. WTP Data: Estimates of Mean Outcomes Adjusted for Covariates

Outcome	Preference Subpopulation	Treatment G	 SD	Trt Effect (SE) $\mu(G) - \mu(SD)$	P-value[1]	Pref Effect(SE)	P-value[1]
SIP Physical[2]	Group	0.566	0.897	-0.331 (0.221)	0.141	-0.918 (0.434)	0.039
	Self-Directed	1.725	1.138	0.587 (0.335)	0.071		
SIP Psychological[2]	Group	-0.197	-0.024	-0.175 (0.299)	0.586	-0.618 (0.718)	0.405
	Self-Directed	0.961	0.518	0.443 (0.589)	0.443		
SIP Total[2]	Group	1.082	1.214	-0.132 (0.238)	0.649	-0.307 (0.704)	0.591
	Self-Directed	1.757	1.582	0.174 (0.608)	0.624		
Adherence Rate	Group	0.930	0.722	0.208 (0.054)	< 0.001	-0.544 (0.127)	< 0.001
	Self-Directed	0.433	0.770	-0.336 (0.105)	< 0.001		

[1] Computed using the bootstrap method based on 2500 bootstrapped data sets.

[2] On log-transformed scale.

Table 4. Simulation Results Based on 125 Replications

Parameters	True Value	Estimate	Est. SE[1]	Est SE[2]	Est. SE[3]	Emp. SE[4]
			Outcome Model (1)			
Intercept	-2.00	-2.01	0.22	0.22	0.22	0.21
X_1	2.00	2.00	0.28	0.28	0.31	0.27
A	2.00	2.02	0.40	0.35	0.32	0.39
C	2.00	2.04	0.50	0.44	0.35	0.40
$A * C$	-2.00	-2.05	0.70	0.60	0.48	0.60
			Preference Model (2)			
Intercept	-1.00	-0.98	0.18	0.18	0.16	0.19
X_2	2.00	1.97	0.32	0.32	0.27	0.31

[1] Estimated SE using the bootstrap method based on 300 bootstrapped data sets.
[2] Estimated SE using the observed information.
[3] Estimated SE using the approximation in Ibrahim (1990).
[4] Empirical SE.

Figure 1. Design of the WTP Study

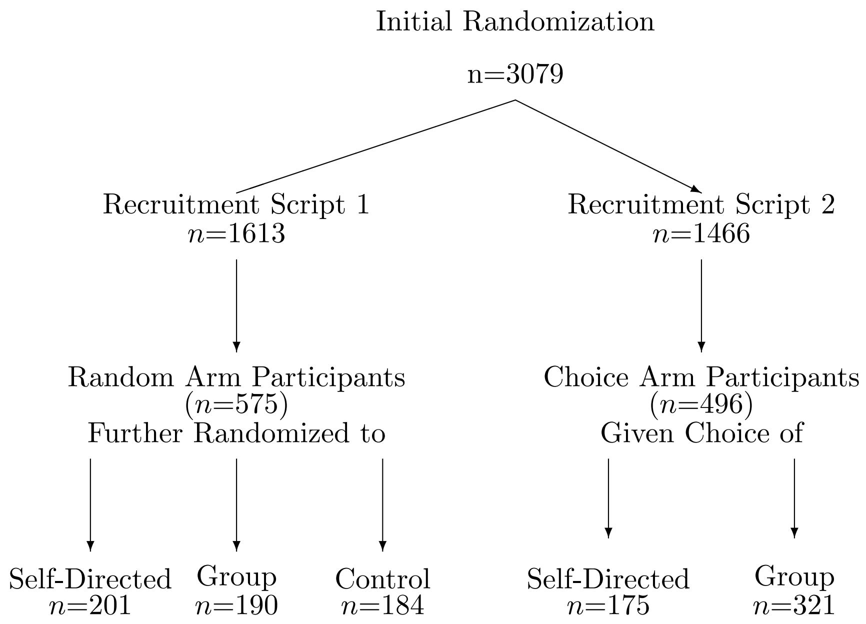
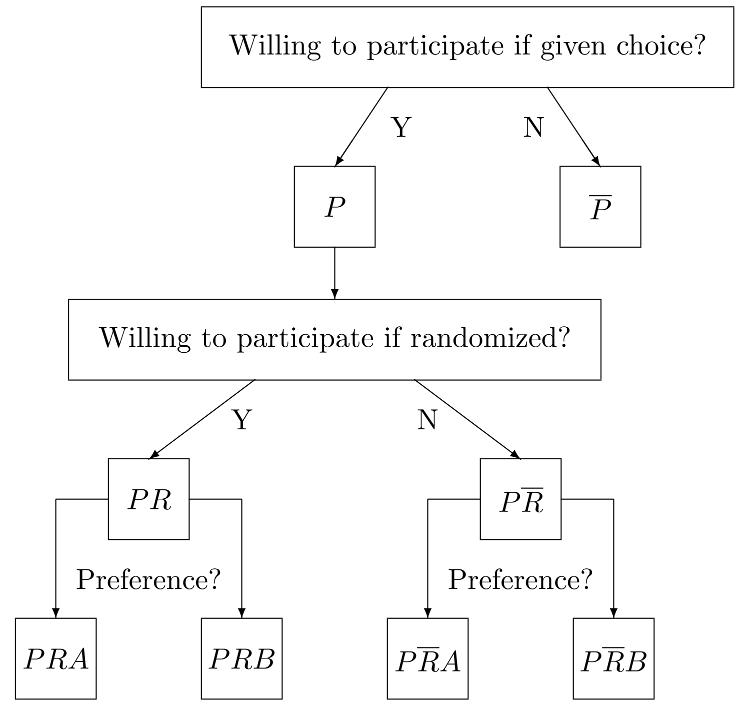


Figure 2. Population Stratification for Hybrid Trials with Two Treatments

(a) Population Stratification



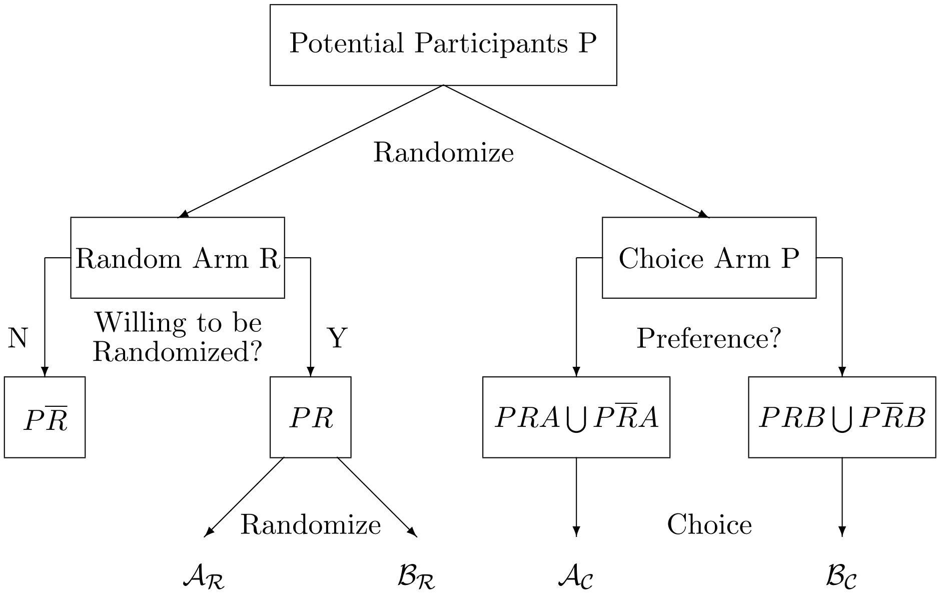
(b) Combinations of Population Stratum and Treatment

	Treatment			
Stratum	$\mathcal{A}_{\mathcal{C}}$	$\mathcal{A}_{\mathcal{R}}$	$\mathcal{B}_{\mathcal{R}}$	$\mathcal{B}_{\mathcal{C}}$
$P\overline{R}A$		$\overline{F}$	$\overline{F}$	$\overline{F}$
PRA				$\overline{F}$
PRB	$\overline{F}$			
$P\overline{R}B$	$\overline{F}$	$\overline{F}$	$\overline{F}$	

$\overline{F}$ = A-priori counterfactual; cannot be observed from data.
Empty cells indicate quantities may be estimated from the data.

Figure 3: The Doubly Randomized Preference Trial (DRPT)

(a) Design of the DRPT


Potential Participants P
Randomize
Random Arm R
Choice Arm P
N
Willing to be Randomized?
Y
Preference?
$P\overline{R}$
PR
$PRA \bigcup P\overline{R}A$
$PRB \bigcup P\overline{R}B$
Randomize
Choice
$\mathcal{A}_{\mathcal{R}}$
$\mathcal{B}_{\mathcal{R}}$
$\mathcal{A}_{\mathcal{C}}$
$\mathcal{B}_{\mathcal{C}}$

(b) Estimable Effects from the DRPT

	Treatment			
Stratum	$\mathcal{A}_{\mathcal{C}}$	$\mathcal{A}_{\mathcal{R}}$	$\mathcal{B}_{\mathcal{R}}$	$\mathcal{B}_{\mathcal{C}}$
$P\overline{R}A$	E	$\overline{F}$	$\overline{F}$	$\overline{F}$
PRA		E	E	$\overline{F}$
PRB	$\overline{F}$			E
$P\overline{R}B$	$\overline{F}$	$\overline{F}$	$\overline{F}$	

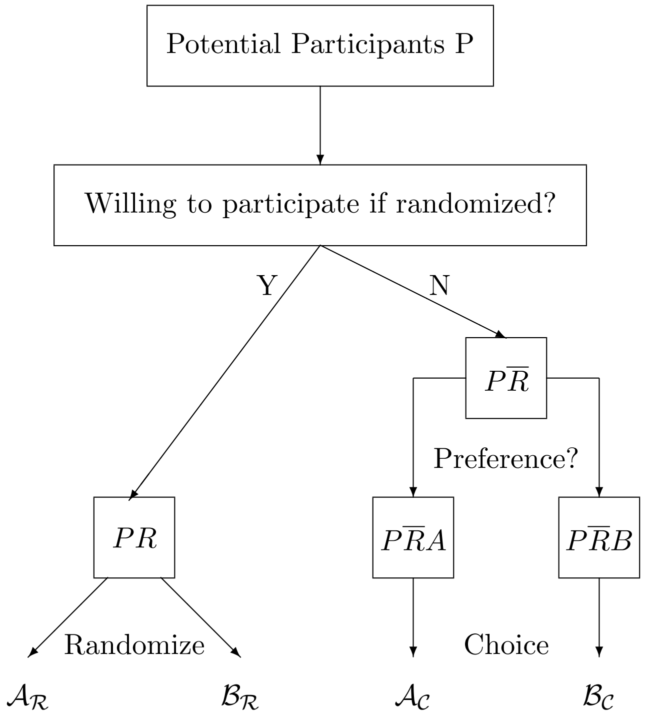
(c) Estimable Effects under NSBR and ER assumptions from the DRPT

	Treatment	
Stratum	A	B
PA	E	E
PB	E	E

$\overline{F}$ = A-priori counterfactual; cannot be observed from data.
$\overline{E}$ = not an A-priori counterfactual; not estimable from this design.
E = estimable directly from this design.

Figure 4. The Partially Randomized Preference Trial (PRPT)

(a) Design of the PRPT


Potential Participants P
Willing to participate if randomized?
Y
N
$P\overline{R}$
Preference?
PR
$P\overline{R}A$
$P\overline{R}B$
Randomize
Choice
$\mathcal{A}_{\mathcal{R}}$
$\mathcal{B}_{\mathcal{R}}$
$\mathcal{A}_{\mathcal{C}}$
$\mathcal{B}_{\mathcal{C}}$

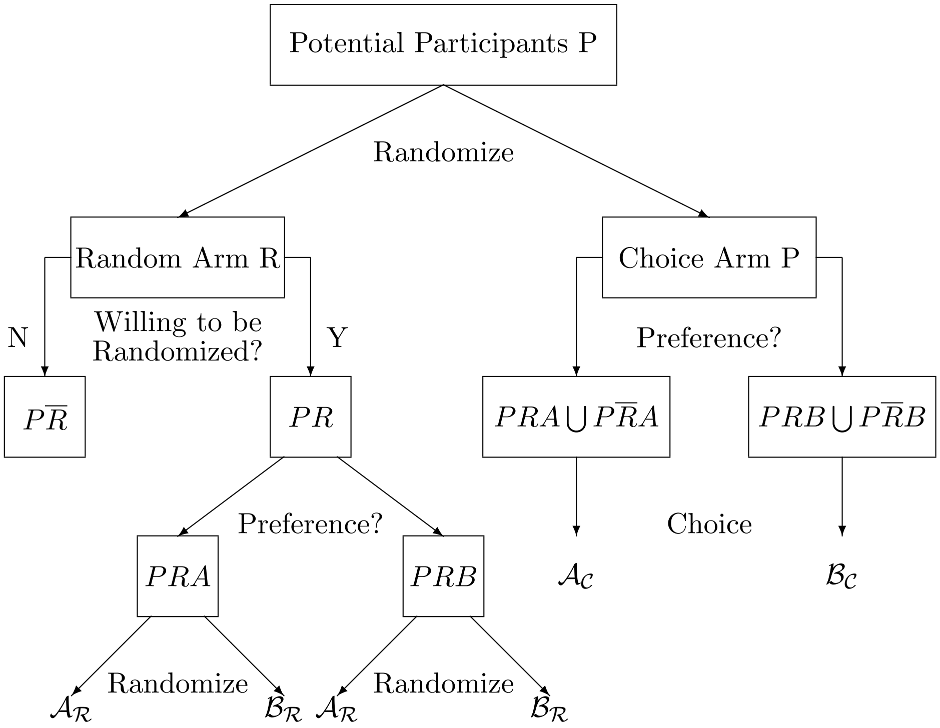
(b) Estimable Effects from the PRPT

	Treatment			
Stratum	$\mathcal{A}_{\mathcal{C}}$	$\mathcal{A}_{\mathcal{R}}$	$\mathcal{B}_{\mathcal{R}}$	$\mathcal{B}_{\mathcal{C}}$
$P\overline{R}A$	E	$\overline{F}$	$\overline{F}$	$\overline{F}$
$PRA \bigcup PRB$	$\overline{E}$	E	E	$\overline{E}$
$P\overline{R}B$	$\overline{F}$	$\overline{F}$	$\overline{F}$	E

$\overline{F}$ = A-priori counterfactual; cannot be observed from data.
$\overline{E}$ = not an A-priori counterfactual; not estimable from this design.
E = estimable directly from this design.

Figure 5. The Doubly Randomized Preference Trial with the Preference Question (DPRPTQ)

(a) Design of the DPRPTQ

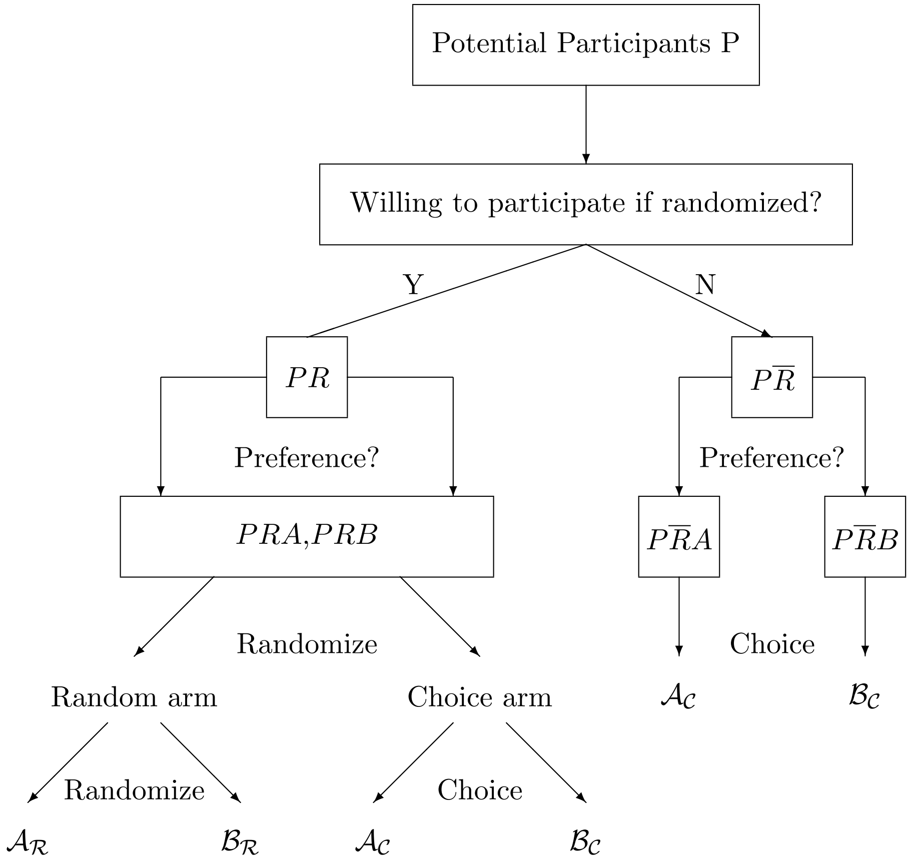

Potential Participants P
Randomize
Random Arm R
Choice Arm P
N
Willing to be Randomized?
Y
Preference?
$P\overline{R}$
PR
$PRA \bigcup P\overline{R}A$
$PRB \bigcup P\overline{R}B$
Preference?
Choice
PRA
PRB
$\mathcal{A}_\mathcal{C}$
$\mathcal{B}_\mathcal{C}$
Randomize
Randomize
$\mathcal{A}_\mathcal{R}$
$\mathcal{B}_\mathcal{R}$
$\mathcal{A}_\mathcal{R}$
$\mathcal{B}_\mathcal{R}$

(b) Estimable Effects from the DRPTQ

	Treatment			
Stratum	$\mathcal{A}_\mathcal{C}$	$\mathcal{A}_\mathcal{R}$	$\mathcal{B}_\mathcal{R}$	$\mathcal{B}_\mathcal{C}$
$P\overline{R}A$	E	$\overline{F}$	$\overline{F}$	$\overline{F}$
PRA		E	E	$\overline{F}$
PRB	$\overline{F}$	E	E	E
$P\overline{R}B$	$\overline{F}$	$\overline{F}$	$\overline{F}$	

Figure 6. The Combined PRPT/DRPT with the Preference Question(CRPTQ)

(a) Design of the CRPTQ


Potential Participants P
Willing to participate if randomized?
Y
N
PR
$P\overline{R}$
Preference?
Preference?
PRA,PRB
$P\overline{R}A$
$P\overline{R}B$
Randomize
Choice
Random arm
Choice arm
$\mathcal{A}_{\mathcal{C}}$
$\mathcal{B}_{\mathcal{C}}$
Randomize
Choice
$\mathcal{A}_{\mathcal{R}}$
$\mathcal{B}_{\mathcal{R}}$
$\mathcal{A}_{\mathcal{C}}$
$\mathcal{B}_{\mathcal{C}}$

(b) Estimable Effects from the CRPTQ

	Treatment			
Stratum	$\mathcal{A}_{\mathcal{C}}$	$\mathcal{A}_{\mathcal{R}}$	$\mathcal{B}_{\mathcal{R}}$	$\mathcal{B}_{\mathcal{C}}$
$P\overline{R}A$	E	$\overline{F}$	$\overline{F}$	$\overline{F}$
PRA	E	E	E	$\overline{F}$
PRB	$\overline{F}$	E	E	E
$P\overline{R}B$	$\overline{F}$	$\overline{F}$	$\overline{F}$	E