Screening Experiments for Developing Dynamic Treatment Regimes

S. A. Murphy SAMURPHY@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109

D. Bingham DBINGHAM@STAT.SFU.CA
Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC, Canada
V5A 1S6

Abstract

Dynamic treatment regimes are time-varying treatments that individualize sequences of treatments to the patient. The construction of dynamic treatment regimes is challenging because a patient will be eligible for some treatment components only if he has not responded (or has responded) to other treatment components. In addition there are usually a number of potentially useful treatment components and combinations thereof. In this article, we propose new methodology for identifying promising components and screening out negligible ones. First, we define causal factorial effects for treatment components that may be applied sequentially to a patient. Second we propose experimental designs that can be used to study the treatment components. Surprisingly, modifications can be made to (fractional) factorial designs - more commonly found in the engineering statistics literature -for screening in this setting. Furthermore we provide an analysis model that can be used to screen the factorial effects. We demonstrate the proposed methodology using examples motivated in the literature and also via a simulation study.

Keywords: Multi-stage Decisions, Experimental Design, Causal Inference

1. Introduction

Dynamic treatment regimes (Robins, 1986, 1987; Lavori et al., 2000; Murphy et al., 2001) are time-varying treatments increasingly employed in the treatment and management of chronic, relapsing disorders such as drug dependence, HIV infection and depression (Brooner and Kidorf, 2002; Rush et al., 2003; Lisziewicz and Lori, 2002). These regimes individualize the treatment level and type via decision rules that input patient outcomes collected during treatment and output recommended treatment alterations. A two stage dynamic treatment regime specifies how to select the treatment combination by a sequence of decision rules, one per stage, $\{d_1,\ d_2\}$ where the decision rule d_1 takes the information available initially, say O_1 and outputs stage 1 treatments A_1 and the decision rule d_2 takes the information available at the beginning of the stage 2, say $\{O_1, A_1, O_2\}$, and recommends stage 2 treatments, A_2. The time order is O_1, A_1, O_2, A_2, O_3 where O_3 is observed at the end of stage 2. The primary outcome is $Y = f(O_1, A_1, O_2, A_2, O_3)$ for f known.

Dynamic treatment regimes are also often multi-component (i.e., multiple factor) treatments: components may include not only the treatments for the primary disorder but also adjunctive treatments for co-occurring problems, behavioral therapies to improve adherence and delivery mechanisms. Presently dynamic treatment regimes are constructed via expert opinion and then evaluated in randomized two group trials. Observational studies are typically used to assist experts in constructing the dynamic treatment regime. Observational analyses attempt to infer causal relations from non-randomized comparisons and are subject to bias when the compared groups differ in ways other than by type of treatment (Rubin, 1974, 1978). While there have been great advances in the epidemiological, statistical and other literatures in developing methods that precisely specify the assumptions under which this bias may be eliminated, randomized studies, when possible, remain the optimal approach to reducing bias (Rubin, 1974, 1978). The field of experimental design provides an approach to using randomization in the construction of multi-component treatments.

The focus of this article is on screening experiments for two stage dynamic treatment regimes. Screening experiments are used to whittle down the large number of treatment components resulting in a few promising dynamic treatment regimes. The formulation and analysis of screening experiments for multi-component dynamic treatment regimes requires the combination of ideas from causal inference and factorial experimental design.

Unfortunately, in settings such as substance abuse, mental health and HIV research, the classical design and analysis of screening experiments *cannot* be directly imported. This is because some stage 2 factors are only relevant for patients who responded (or did not respond) to prior stage 1 treatment factors. For this reason, in Section 2, we carefully define factorial effects so that these effects have a causal interpretation. In Section 3 we provide 2^k experimental designs and associated analysis methods that can be used to screen these effects. In Section 4, we consider 2^{k-m} designs. In this section we provide an approach to ascertain the aliasing of the factorial effects. Section 5 provides examples illustrating a variety of designs. In Section 6 a simulation study is used to evaluate the robustness of the proposed methods. Section 7 concludes with a summary and discussion of open problems.

2. Causal Effects of the Factors

Screening factors often occurs via the assessment of factorial effects in ANOVA decompositions (see Box, Hunter and Hunter, 1978; Wu and Hamada, 2000 or Montgomery and Jennings, 2006) for the conditional mean of the primary outcome. In our setting special care is required in defining effects. Consider a simple case where there is only one stage 1 factor and two stage 2 factors. Each factor takes on levels in $\{-1, 1\}$. Suppose N subjects are each randomized equally to the two levels of the stage one factor A_1. Then an indicator of early response is observed; R is 1 if the subject is responding following assignment of A_1 and 0 otherwise (there is no O_1 and $O_2 = R$). For simplicity, we refer to individuals with $R = 1, 0$ as responders, non-responders, respectively. At stage 2, responders are randomized equally between the $\{-1, 1\}$ levels of the factor $A_2^{(1)}$, and non-responders are randomized equally between the $\{-1, 1\}$ levels of the factor $A_2^{(0)}$ (the superscripts, (1), (0), indicate stage 2 factors for responders, non-responders, respectively). At the end of the study the primary outcome Y (coded so that high values are preferable) is observed. Assuming subject observations are independent and identically distributed, a "factorial" decomposition is

$$\begin{aligned}E[Y|A_1, R, A_2^{(1)}, A_2^{(0)}] &= \eta_0 + \eta_1 A_1 + \gamma_1 R + \gamma_2 R A_1 + \\ &\quad \beta_1 R A_2^{(1)} + \beta_2 R A_1 A_2^{(1)} + \alpha_1 (1-R) A_2^{(0)} + \alpha_2 (1-R) A_1 A_2^{(0)} \quad (1)\end{aligned}$$

with the interpretation that the coefficients $\{\eta_1, \beta_1, \beta_2, \alpha_1, \alpha_2\}$ represent factorial effects. R is included in the above decomposition since $A_2^{(1)}$ can only be assigned to subjects with $R = 1$ and similarly for $A_2^{(0)}$.

If we interpret the coefficients $\{\eta_1, \beta_1, \beta_2, \alpha_1, \alpha_2\}$ as factorial effects then to screen factors $A_1,\ A_2^{(1)}, A_2^{(0)}$ we will base our inference on these coefficients. However, this can lead to erroneous conclusions concerning the usefulness of A_1, via η_1, for at least two reasons. First in the medical/behavioral fields there are a plethora of both known and unknown common causes of R and Y (hence R is prognostic for Y). Consider the following example. Let U, a Bernoulli random variable with success probability $\frac{1}{2}$, represent an unknown common cause of both the outcome Y and early response R. U might be an unknown genetic factor. Suppose that $Y = \delta_0 + \delta_1 U + \epsilon$ where ϵ (mean zero, finite variance) is independent of $(U,\ R,\ A_1,\ A_2^{(0)},\ A_2^{(1)})$. Thus there is *no* effect of the stage 1 factor A_1 or of the stage 2 factors $(A_2^{(0)},\ A_2^{(1)})$ on the outcome Y. Next suppose that $P[R=1|U,\ A_1] = U\left[\frac{q_1+q_2}{2} + \frac{q_1-q_2}{2}A_1\right] + (1-U)\left[\frac{q_3+q_4}{2} + \frac{q_3-q_4}{2}A_1\right]$ where each $q_j \in [0,1]$. Note that A_1 may impact the early response ($q_1 - q_2 \neq 0$ or $q_3 - q_4 \neq 0$). We obtain

$$\begin{aligned} E[Y|A_1, R=0, A_2^{(0)},\ A_2^{(1)}] &= \delta_0 + \delta_1 E[U|A_1, R=0] \\ &= \delta_0 + \frac{\delta_1}{2}\left(\frac{1-q_1}{2-q_1-q_3} + \frac{1-q_2}{2-q_2-q_4}\right) \\ &\quad + \frac{\delta_1}{2}\left(\frac{1-q_1}{2-q_1-q_3} - \frac{1-q_2}{2-q_2-q_4}\right) A_1 \end{aligned}$$

Thus $\eta_1 = \frac{\delta_1}{2}\left(\frac{1-q_1}{2-q_1-q_3} - \frac{1-q_2}{2-q_2-q_4}\right)$ which may differ from the true effect of zero. That is, η_1 in (1) reflects biases that occur because we are conditioning on R which is both an outcome of A_1 and a prognostic variable for Y. Further discussion of this bias can be found in Robins and Greenland (1994) and Robins and Wasserman (1997).

Second even if there are no unknown common causes of R and Y, η_1 does not reflect the overall impact of A_1 as A_1 may impact Y at least partially by its effect on R (Parmigiani, 2002; Murphy, 2005). For both reasons, clinical trial guidelines (ICH E9, 1999) warn against conditioning on outcomes such as side effects or other post randomization covariates in comparing one treatment to another.

These two issues preclude the use of (1) for screening. To our knowledge the former issue does not arise in the traditional experimental design literature, even in clinical trials with multiple stages. For example, adaptive experimental designs (Hu and Rosenberger, 2006, Zacks, 1996, Patel, 1962) also involve multiple stages of (sequential) decision making, but each stage involves different subjects; that is "sequential" there means decisions on a sequence of subjects not sequential decisions on each subject. Since subjects can be generally assumed to respond independently (given membership in the same population of subjects), and the stages involve different subjects, the above issues do not occur.

2.1 Causal Effects in Terms of Potential Outcomes

In this section a model in terms of factorial effects for each fixed pattern of factor levels is developed; in the next subsection randomization of the factor levels is incorporated. This model will be in terms of potential outcomes (Rubin, 1978; Robins, 1986, 1987); the parameters in the model will be the causal effects. The idea is that each subject has multiple potential outcomes, one per factor level combination. However only one of the potential outcomes, that is, the outcome associated with the subject's assigned factor levels is observed; the remaining potential outcomes are missing. The potential outcome framework facilitates a precise definition of the causal factorial effects; as will be seen this is somewhat nuanced for stage 1 factorial effects. Also this framework facilitates the statement of the conditions under which estimators of the causal effects of a factor can be obtained.

Suppose there are p_1 stage 1 factors, p_{12} stage 2 factor levels for responders and p_{02} stage 2 factor for non-responders. As is generally the case in screening, we assume each factor has two levels (see Section 3 for discussion), coded by values ±1. Let a_1 denote a vector of stage 1 factor levels, $a_2^{(1)}$, $a_2^{(1)'}$ denote vectors of stage 2 factor levels for responders and $a_2^{(0)}$, $a_2^{(0)'}$ denote vectors of stage 2 factor levels for non-responders. Let $R_{a_1,a_2^{(0)},a_2^{(1)}} \in \{0,1\}$ and $Y_{a_1,a_2^{(0)},a_2^{(1)}}$ denote the stage 1 response and the primary outcome, respectively, that would be observed if the subject were assigned the dynamic treatment regime: "provide treatment a_1 in stage 1 and then if an early response occurs provide $a_2^{(1)}$ otherwise provide $a_2^{(0)}$." Note that there are $2^{p_1+p_{02}+p_{12}}$ different dynamic treatment regimes, thus, there are $2^{p_1+p_{02}+p_{12}}$ potential stage 1 outcomes and $2^{p_1+p_{02}+p_{12}}$ potential primary outcomes for each subject.

Throughout we make two assumptions that hold if the stage 2 treatment is not revealed to the subject/clinical staff until after the occurrence of the stage 1 outcome. These assumptions will enable us to reduce the number of groups of subjects in the experimental design (see the next section).

> Ignorability Assumption I: Assume that $R_{a_1,a_2^{(0)},a_2^{(1)}} = R_{a_1,a_2^{(0)'},a_2^{(1)'}}$ for all $\{a_1, a_2^{(0)}, a_2^{(0)'}, a_2^{(1)}, a_2^{(1)'}\}$.

In words, this is the assumption that a subject's stage 1 outcome remains the same regardless of the stage 2 treatment assignments. Below we use the notation, R_{a_1}, instead of $R_{a_1,a_2^{(0)},a_2^{(1)}}$.

> Ignorability Assumption II: Assume that $Y_{a_1,a_2^{(0)},a_2^{(1)}} R_{a_1} = Y_{a_1,a_2^{(0)'},a_2^{(1)}} R_{a_1}$ and $Y_{a_1,a_2^{(0)},a_2^{(1)}}(1-R_{a_1}) = Y_{a_1,a_2^{(0)},a_2^{(1)'}}(1-R_{a_1})$ for all $\{a_1, a_2^{(0)}, a_2^{(0)'}, a_2^{(1)}, a_2^{(1)'}\}$.

For example, this assumption states that the primary outcome of a subject who does not respond in stage 1 does not depend on what this subject would have been assigned had he/she responded in stage 1. These assumptions may be violated if, as part of the protocol, subjects are informed in stage 1 of which stage 2 treatment they will be offered if they do not respond at stage 1. It may happen that subjects, who know that they will be assigned a rather burdensome stage 2 treatment upon non-response, will try to proactively avoid this burdensome treatment by adhering more faithfully to their assigned stage 1 treatment than would otherwise be the case.

In the sequel, stage 1 effects refer to all effects involving *only* stage 1 factors and the stage 2 effects refer to all effects involving *at least* one stage 2 factor. The stage 2 causal effects are defined via a saturated linear regression model for

$$E[Y_{a_1,a_2^{(0)},a_2^{(1)}}|R_{a_1}] \tag{2}$$

in $a_1, a_2^{(1)}R_{a_1}, a_2^{(0)}(1-R_{a_1})$ and their higher order interactions. For example suppose that there is one stage 1 factor and two stage 2 factors ($p_1 = p_{02} = p_{12} = 1$) then similar to (1) we have

$$\begin{aligned}E[Y_{a_1,a_2^{(0)},a_2^{(1)}}|R_{a_1}] = \eta_0 + \eta_1 a_1 + \gamma_0 R_{a_1} + \gamma_1 R_{a_1} a_1 + \beta_1 R_{a_1} a_2^{(1)} + \beta_2 R_{a_1} a_1 a_2^{(1)} + \\ \alpha_1(1-R_{a_1})a_2^{(0)} + \alpha_2(1-R_{a_1})a_1 a_2^{(0)}.\end{aligned} \tag{3}$$

where $\alpha_j,\ \beta_j,\ j = 1, 2$ are the stage 2 causal effects; in particular α_1, β_1 are the stage 2 main effects and α_2, β_2 are the stage 2 interaction effects. If we solve for $\alpha_j, \beta_j,\ j = 1, 2$, we see that the causal effects are contrasts of conditional means involving potential outcomes; for example,

$$\begin{aligned}2\beta_1 = &\frac{1}{2}\left(E[Y_{1,a_2^{(0)},1}|R_1 = 1] - E[Y_{1,a_2^{(0)},-1}|R_1 = 1]\right) \\ &+ \frac{1}{2}\left(E[Y_{-1,a_2^{(0)},1}|R_{-1} = 1] - E[Y_{-1,a_2^{(0)},-1}|R_{-1} = 1]\right).\end{aligned}$$

Recall that under the ignorability assumption II, the quantities above do not depend on the value of $a_2^{(0)}$. The coefficients in the above linear regression are 1/2 the size of the usual factorial effects (Wu and Hamada, 2000, pg. 113). (Note there is some discrepancy in the literature in how factorial effects are defined; for example Byar and Piantadosi's (1985) definition of a main effect is one-half the size of Wu and Hamada's definition). Throughout this work we use the regression formulation and refer to the regression coefficients as factorial effects.

As discussed at the beginning of this section the usual definition of the stage 1 effect (for example, via η_1 in (3)) can lead to erroneous conclusions. To avoid this problem we do not condition on R_{a_1}; instead we marginalize over R_{a_1} to obtain causal effects. Consistent with ignorability assumption II, we use the notation $Y_{a_1,a_2^{(0)}}$ if $R_{a_1} = 0$ and $Y_{a_1,a_2^{(1)}}$ if $R_{a_1} = 1$. We define the stage 1 causal factorial effects via a saturated linear regression model for

$$E\left[\left(\frac{1}{2}\right)^{p_{12}} \sum_{a_2^{(1)}} Y_{a_1,a_2^{(1)}} R_{a_1} + \left(\frac{1}{2}\right)^{p_{02}} \sum_{a_2^{(0)}} Y_{a_1,a_2^{(0)}} (1 - R_{a_1})\right] \tag{4}$$

in a_1. Again suppose that $p_1 = p_{02} = p_{12} = 1$ then the linear regression is simple:

$$E\left[\frac{1}{2}\sum_{a_2^{(1)}} Y_{a_1,a_2^{(1)}} R_{a_1} + \frac{1}{2}\sum_{a_2^{(0)}} Y_{a_1,a_2^{(0)}} (1 - R_{a_1})\right] = \phi_1 + \phi_2 a_1.$$

Solving for ϕ_2 we obtain,

$$\begin{aligned} 2\phi_2 = {} & \frac{1}{2} E\left[\sum_{a_2^{(1)}} Y_{1,a_2^{(1)}} R_1 - \sum_{a_2^{(1)}} Y_{-1,a_2^{(1)}} R_{-1}\right] \\ & + \frac{1}{2} E\left[\sum_{a_2^{(0)}} Y_{1,a_2^{(0)}} (1 - R_1) - \sum_{a_2^{(0)}} Y_{-1,a_2^{(0)}} (1 - R_{-1})\right]; \end{aligned}$$

ϕ_2 is the causal main effect of A_1.

In general the causal main effect of one of the p_1 stage 1 factors is the average of contrasts between two means of Y one for each level of the stage 1 factor, marginal over the distribution of R (and all other intermediate outcomes in O_2 as well); the average is over the remaining factor levels (with all other factors taking levels ±1 with equal probability). The averaging/marginalization with respect to a uniform distribution over the other factor levels is consistent with the definition of factorial effects in experimental design (Wu and Hamada, 2000; Byar and Piantadosi, 1985). Note the definition of the stage 2 treatment effects also involves marginalization (except the stage 2 treatments are nested within the outcome R and thus the contrasts are between means conditional on R).

2.2 Effects in Terms of Conditional Expectations Involving Randomized Factors.

Suppose we have experimental data in which the factors are randomized according to some distribution on $\{-1, 1\}^{p_1+p_{02}+p_{12}}$. The following provides definitions of the factorial effects in terms of the resulting data distribution. Let A_1 denote the random vector of p_1 stage 1 factor levels, $A_2^{(1)}$ denote the random vector of p_{12} stage 2 factor levels for responders and $A_2^{(0)}$ denote the random vector of p_{02} stage 2 factor levels. Lower case letters denote realizations of these random vectors. In this case $\{A_1, A_2^{(0)}, A_2^{(1)}\}$ is clearly independent of the collection $\{R_{a_1},\ Y_{a_1,a_2^{(0)},a_2^{(1)}},\ a_1 \in \{-1,1\}^{p_1}, a_2^{(0)} \in \{-1,1\}^{p_{02}}, a_2^{(1)} \in \{-1,1\}^{p_{12}}\}$. On each subject we observe $A_1,\ R,\ A_2^{(0)}$, $A_2^{(1)},\ Y$ (when $R = 1$, factor levels given by $A_2^{(1)}$ are assigned and when $R = 0$, factor levels given by $A_2^{(0)}$ are assigned). We connect the potential outcomes to the observed data via the consistency assumption (Robins, 1997): assume that for all values of $\{a_1, a_2^{(0)}, a_2^{(1)}\}$, if $A_1 = a_1,\ A_2^{(0)} = a_2^{(0)},\ A_2^{(1)} = a_2^{(1)}$ then $Y = Y_{a_1,a_2^{(0)},a_2^{(1)}}$ and $R = R_{a_1}$.

The randomization of factor levels plus the consistency and ignorability assumptions imply that $E[R_{a_1}] = E[R_{a_1}|A_1 = a_1] = E[R|A_1 = a_1]$. Similarly for each $r \in \{0,1\}$, $P[R_{a_1} = r, A_1 = 1, A_2^{(r)} = a_2^{(r)}] = P[R = r, A_1 = 1, A_2^{(r)} = a_2^{(r)}]$. Assuming this probability is nonzero, $E[Y_{a_1,a_2^{(0)},a_2^{(1)}}|R_{a_1} = r]$ is equal to $E[Y_{a_1,a_2^{(0)},a_2^{(1)}}|A_1 = a_1, R_{a_1} = r, A_2^{(r)} = a_2^{(r)}]$, which by the consistency and ignorability assumptions is in turn equal to $E[Y|A_1 = a_1, R = r, A_2^{(r)} = a_2^{(r)}]$. As a result, the stage 2 causual effects in (2) are simply the coefficients of terms involving $a_2^{(0)}$ in a saturated linear model for conditional mean $E[Y|A_1 = a_1, R = 0, A_2^{(0)} = a_2^{(0)}]$ and the coefficients of terms involving $a_2^{(1)}$ in a saturated linear model for conditional mean $E[Y|A_1 = a_1, R = 1, A_2^{(1)} = a_2^{(1)}]$.

The discussion above implies that in the simple setting where there is only one stage 1 factor and two stage 2 factors the β (α) parameters in (1) represent the stage 2 causal effects for the responders (non-responders, respectively). A similar statement can not be made for the stage 1 effects. Instead recall that the stage 1 causal effects are given by a saturated linear model for the mean in (4). This mean can be rewritten as

$$\left(\frac{1}{2}\right)^{p_{12}} \sum_{a_2^{(1)}} E\left[Y_{a_1,a_2^{(1)}}|R_{a_1} = 1\right] E[R_{a_1}] + \left(\frac{1}{2}\right)^{p_{02}} \sum_{a_2^{(0)}} E\left[Y_{a_1,a_2^{(0)}}|R_{a_1} = 0\right] E[1 - R_{a_1}].$$

Again using randomization and the consistency assumption this mean can be reexpressed as

$$\begin{aligned} &\left(\frac{1}{2}\right)^{p_{12}} \sum_{a_2^{(1)}} E\left[Y|A_1 = a_1, R = 1, A_2^{(1)} = a_2^{(1)}\right] E[R|A_1 = a_1] \\ &+ \left(\frac{1}{2}\right)^{p_{02}} \sum_{a_2^{(0)}} E\left[Y|A_1 = a_1, R = 0, A_2^{(0)} = a_2^{(0)}\right] (1 - E[R|A_1 = a_1]). \qquad (5) \end{aligned}$$

Thus the stage 1 causal effects are given by the coefficients in a saturated linear model for the above sum. The above formula (5) is a version of Robins "G-computation" formula (Robins, 1986); this formula frequently arises in causal inference for time varying treatments. A more intuitive definition for the stage 1 effects obtains if the stage 2 factors are independently randomized according to discrete uniform distributions on $\{-1, 1\}$. Then (5) simplifies to $E[Y|A_1 = a_1]$; a saturated linear model for this conditional mean provides the definition of the stage 1 effects in this special setting.

Somewhat surprisingly the stage 1 and stage 2 causal effects are terms in one model for the conditional mean. Define

$$\begin{aligned} E_{A_2\sim U}\big[E[Y|A_1, R, A_2^{(R)}]|A_1, R\big] = \left(\frac{1}{2}\right)^{p_{12}} \sum_{a_2^{(1)}} E\left[Y|A_1, R=1, A_2^{(1)} = a_2^{(1)}\right] R \\ + \left(\frac{1}{2}\right)^{p_{02}} \sum_{a_2^{(0)}} E\left[Y|A_1, R=0, A_2^{(0)} = a_2^{(0)}\right] (1-R). \end{aligned}$$

This definition is used to make the next formulae concise; note (5) is equal to $E\left[\, E_{A_2\sim U}\big[E[Y|A_1, R, A_2^{(R)}]|A_1, R\big] \,\Big|A_1 = a_1\right]$. Consider the telescoping sum for the conditional mean, $E[Y|A_1, R, A_2^{(R)}]$:

$$\begin{aligned} &\Big(E[Y|A_1, R, A_2^{(R)}] - E_{A_2\sim U}\big[E[Y|A_1, R, A_2^{(R)}]|A_1, R\big]\Big) && (6) \\ &\quad + \Big(E_{A_2\sim U}\big[E[Y|A_1, R, A_2^{(R)}]|A_1, R\big] - E\big[E_{A_2\sim U}\big[E[Y|A_1, R, A_2^{(R)}]|A_1, R\big]|A_1\big]\Big) \\ &\qquad + E\left[E_{A_2\sim U}\big[E[Y|A_1, R, A_2^{(R)}]|A_1, R\big]|A_1\right]. && (7) \end{aligned}$$

The first term (6) contains the stage 2 causal effects. The last term (7) is just another way to write (5) and thus contains all stage 1 causal effects. The middle term is composed entirely of nuisance parameters and can be rewritten as $\big(R - E[R|A_1]\big)$ times

$$E_{A_2\sim U}\big[E[Y|A_1, R, A_2^{(R)}]|A_1, R = 1\big] - E_{A_2\sim U}\big[E[Y|A_1, R, A_2^{(R)}]|A_1, R = 0\big]. \qquad (8)$$

Thus in the example of one stage 1 factor and two stage 2 factors (one for responders, the other for non-responders), we replace the linear regression in (1) by

$$\begin{aligned} E[Y|A_1, R, A_2^{(R)}] \;=\; & \phi_1 + \phi_2 A_2 \\ & + (R - E[R|A_1])(\psi_1 + \psi_2 A_1) \\ & + \beta_1 (1-R) A_2^{(0)} + \beta_2 (1-R) A_2^{(0)} A_1 + \alpha_1 R A_2^{(1)} + \alpha_2 R A_2^{(1)} A_1, \end{aligned}$$

where the first row corresponds to (7), the second row to (8) times $(R - E[R|A_1])$ and the last row to (6). Note that even though we call the parameters, $\psi_1,\ \psi_2$ and the stage 1 response rate, $E[R|A_1]$, nuisance parameters, these parameters may be of independent interest. This is certainly the case for the stage 1 response rate; additionally $\psi_1,\ \psi_2$ reflect the ability of the stage 1 response to predict the primary outcome. Note that the nuisance parameters, $\psi_1,\ \psi_2$, would be absent from the above formula only if the response indicator, $R-E[R|A_1]$ were zero, that is, only if the stage 1 response, R, is a deterministic function of A_1.

3. 2^k Factorial Two-Stage Designs

Our goal is to screen out inactive factors. Usually we consider two levels for each factor. The two levels should be selected to be sufficiently disparate so that we can obtain an effect yet not be unethical. In settings where there is likely to be a downturn in the primary outcome at high levels of the factor, we select the higher level small enough so that the downturn is thought to be insufficient to eliminate the effect. If necessary, subsequent trials can be used to more fully explore the dose-response.

Suppose that there are two stage 1 factors $A_1 = \{A_{11},\ A_{12}\}$ and one stage 2 factor for responders $A_2^{(1)}$ and one for non-responders, $A_2^{(0)}$. Consider the experimental design in Table 1. Each row in the design of Table 1 provides the factor levels assigned to a group of subjects. The column labeled $A_2^{(1)} = A_2^{(0)}$ designates the identical factor level settings for $A_2^{(1)}$ and $A_2^{(0)}$. Note that once responder (non-responder) status is known for each subject, we can view the design as two 2^3 full factorial designs, one for responders to stage 1 treatment and the other for non-responders to stage 1 treatment.

Table 1: A 2^3 Factorial Two Stage Design[1]

A_{11}	A_{12}	$A_2^{(1)} = A_2^{(0)}$
+	+	+
+	+	-
+	-	+
+	-	-
-	+	+
-	+	-
-	-	+
-	-	-

[1] + denotes +1 and − denotes −1.

Each row of the design corresponds to a group of subjects, all of whom are assigned the same dynamic treatment regime. For example, the subjects in the first group are assigned the dynamic treatment regime: "Provide $A_{11} = +1,\ A_{12} = +1$, if they respond, assign $A_2^{(1)} = +1$ in stage 2 and if they do not respond, assign $A_2^{(0)} = +1$ in stage 2" (note that a dynamic treatment regime must specify the stage 2 treatments for both the responders and the non-responders). For convenience we use the term "stacked" to indicate situations in which the factor level settings of two factors are identical across all rows of the experimental design; factors $A_2^{(1)}$ and $A_2^{(0)}$ are stacked. The stacking is advantageous because investigators only have to implement a 2^3 design yet under the ignorability assumptions, we have information on all 2^4 dynamic treatment regimes. For example even though no row of the design corresponds to "assign $A_{11} = +1,\ A_{12} = +1$ and if response assign $A_2^{(1)} = +1$ but if non-response assign $A_2^{(0)} = -1$" we may combine the responders in row 1 and the non-responders in row 2 together to produce a new group of subjects assigned this dynamic treatment regime.

As in Table 1, all designs considered here possess the property that for each factor half of the rows in the design are set at the +1 level and half of the rows are set at the -1 level. This property will play a crucial role in ascertaining the aliasing between effects. Note however that due to the difficulties in recruiting subjects inherent in clinical trials, the groups of subjects may not be exactly equal in size, and second even if there are equal numbers of subjects in each group, the number of subjects assigned a particular level of $A_2^{(1)}$, $A_2^{(0)}$ depends on the response rate.

As before suppose there are p_1 stage 1 factors, p_{12} stage 2 factors for responders and p_{02} stage 2 factors for non-responders. Here we consider screening the stage 1 and stage 2 effects using data from a 2^k factorial two-stage experiment in which $k = p_1 + \max(p_{12}, p_{02})$. In contrast to classical analyses in experimental design (analyses that assume normality and are exact for small group sizes), we consider analyses that are justified by large sample theory (large group sizes) and thus are approximate for small group sizes. With this viewpoint we interpret a full factorial design as a design in which each subject is assigned with equal probability to one of the 2^k possible combinations of the factor levels. In large samples this randomization results in approximately equal sized groups assigned to each row of the design. Denote the collection of stage 1, stage 2 factors by A_1, respectively A_2. We have that $(A_1,\ A_2)$ has a discrete uniform distribution on $\{-1, 1\}^k$ where $A_2^{(1)}$ is the first p_{12} entries in A_2 (these will be the settings of the stage 2 factors for responders) and $A_2^{(0)}$ is the first p_{02} entries in A_2 (these will correspond to the settings of the stage 2 factors for non-responders). The data consists of N i.i.d. copies of $\{A_1, R, A_2, Y\}$. There are 2^k unique values of (A_1, A_2).

Define $\mathbf{X}_1$ as the random vector composed of a 1, all stage 1 factors and their two-way and higher order products (2^{p_1} terms), $\mathbf{X}_2^{(1)}$ as the random vector composed of all stage 2 factors for responders, their two-way and higher order products and the all products of these combinations with members of $\mathbf{X}_1$ ($2^{p_1+p_{12}} - 2^{p_1}$ terms). $\mathbf{X}_2^{(0)}$ is defined similarly. The decomposition in (6 - 8) can be written via a linear model

$$E[Y|A_1, R, A_2] = \mathbf{X}_1^T\phi + (R - p(\mathbf{X}_1))\mathbf{X}_1^T\psi + R\mathbf{X}_2^{(1)^T}\beta + (1-R)\mathbf{X}_2^{(0)^T}\alpha \tag{9}$$

where β and α are the vectors of all stage 2 causal effects for responders and non-responders, respectively, ψ is the vector of nuisance effects, ϕ is the vector of stage 1 causal effects and $p(\mathbf{X}_1) = E[R|A_1]$ (note $\mathbf{X}_1$ is a function of A_1).

Recall that a parameter is identifiable if different values of the parameter lead to different distributions of $\{A_1, R, A_2, Y\}$ (van der Vaart, 1998). Identifiability in this setting is straightforward.

Lemma 1 Assume that $P[0 < E[R|A_1] < 1] = 1$. The parameters, ϕ, ψ, β and α in (9) are identifiable.

The above lemma is a direct consequence of two facts. First the conditional means, $E[R|A_1 = a_1]$, $E[Y|A_1 = a_1, R = 1, A_2^{(1)} = a_2^{(1)}]$ and $E[Y|A_1 = a_1, R = 0, A_2^{(0)} = a_2^{(0)}]$ for all values of $\{a_1, a_2^{(1)}, a_2^{(0)}\}$, are identifiable. Second the expectation of $\left[\mathbf{X}_1^T,\ (R - p(\mathbf{X}_1))\mathbf{X}_1^T,\ R\mathbf{X}_2^{(1)^T},\ (1-R)\mathbf{X}_2^{(0)^T}\right]^T$ times its transpose is given by a block diagonal with blocks, $E\left[\mathbf{X}_1\mathbf{X}_1^T\right]$, $E\left[\mathbf{X}_1 p(\mathbf{X}_1)(1 - p(\mathbf{X}_1))\mathbf{X}_1^T\right]$, $E\left[\mathbf{X}_2^{(1)} p(\mathbf{X}_1)\mathbf{X}_2^{(1)^T}\right]$ and $E\left[\mathbf{X}_2^{(0)}(1 - p(\mathbf{X}_1))\mathbf{X}_2^{(0)^T}\right]$ (recall $p(\mathbf{X}_1) = E[R|A_1]$). These blocks are invertible under the conditions specified in Lemma 1. The assumption of Lemma 1 is not necessary for identifiability of ϕ and can be weakened to only $P[E[R|A_1] > 0] = 1$ for identifiability of β (with a similar statement for α). The details of the proof are omitted.

To utilize (9) in screening estimate $p(\mathbf{X}_1)$ by forming the average of R for each value of $\mathbf{X}_1$ to obtain $\hat{p}(\mathbf{X}_1)$. Conduct a linear regression of Y on $\{\mathbf{X}_1,\ (R - \hat{p}(\mathbf{X}_1))\mathbf{X}_1,\ R\mathbf{X}_2^{(1)},\ (1-R)\mathbf{X}_2^{(0)}\}$ to obtain $\hat{\phi}, \hat{\psi}, \hat{\beta}, \hat{\alpha}$. Note that in contrast to classical screening analyses, these estimators are not orthogonal. This is because homogenous variance is not assumed (e.g. the conditional variance of Y may vary by factor levels), the groups sizes may not be identical and because the responder rates will likely vary by the stage 1 factor levels.

Lemma 2 Assume that the variance of Y is finite and that $P[0 < E[R|A_1] < 1] = 1$. Then as $N \to \infty$, $\{\sqrt{N}(\hat{\phi} - \phi),\ \sqrt{N}(\hat{\beta} - \beta),\ \sqrt{N}(\hat{\alpha} - \alpha)\}$ converges in distribution to a multivariate normal. If $P\left[Var(Y|A_1, R, A_2) = Var(Y|A_1, R, A_2^{(R)})\right] = 1$, then the estimators $\hat{\beta}$, $\hat{\alpha}$ and $\hat{\phi}$ are locally semiparametric efficient.

In many settings all responders will be provided the same treatment. In this case there are no $\hat{\beta}$, β and $P[0 < E[R|A_1] < 1] = 1$ can be replaced by $P[E[R|A_1] > 0] = 1$ (similar statements can be made if all non-responders are provided the same treatment). See Tsiatis (Chapter 4, 2006) for an introduction to and definition of locally semiparametric efficiency. Note the assumption on the conditional variances holds under the consistency and ignorability assumptions. This lemma is a special case of Lemma 4 below. The formula for the variance-covariance matrix is provided in Appendix A.1 along with an asymptotically consistent estimator and the proof of Lemma 4.

4. 2^{k-m} Factorial Two Stage Design

A better use of resources for screening is use a 2^{k-m} fractional factorial design, which uses only a $1/2^m$ fraction of the groups in a 2^k design. An example of a 2^{3-1} design for the setting in which there are two stage 1 factors and one stage 2 factor each for responders/non-responders is

Table 2: A 2^{3-1} Fractional Factorial Two Stage Design

A_{11}	A_{12}	$A_2^{(1)} = A_2^{(0)}$
+	+	+
+	-	-
-	+	-
-	-	+

This design has one-half as many rows (groups of subjects) as the full factorial design in Table 1; see the examples in Section 6 for realistic designs. Usually a 2^{k-m} design is selected by first ascertaining plausible working assumptions concerning the factorial effects. Wu and Hamada (2000) provide principles that can be used to guide these working assumptions in the absence of scientific knowledge (e.g. often it is plausible that three way and higher order effects are negligible). Next, the aliasing associated with candidate 2^{k-m} designs is ascertained; parameters, that is, effects, are aliased when only their sum can be identified. Note that the aliasing occurring here is different from the confounding that occurs in causal inference as the latter is unplanned and due to potentially unknown variables whereas the former is the result of the planned study design and is known. Lastly the design with the aliasing structure that is most consistent with working assumptions is selected. See Wu and Hamada, (2000) for simple approaches to creating 2^{k-m} designs.

As was the case for full factorial designs, we interpret a fractional factorial design as a design in which each subject is assigned with equal probability to one of the 2^{k-m} possible combinations of the factor levels. Thus $(A_1,\ A_2)$ has a discrete uniform distribution across the factor levels given by the 2^{k-m} rows of the design (recall that $A_2^{(1)}$ ($A_2^{(0)}$) is set to the first p_{12} (p_{02}, respectively) entries in A_2). The experimental data consists of N i.i.d. copies of $\{A_1, R, A_2, Y\}$.

4.1 Identification of Effects

We use a large sample notion of aliasing. Effects will be aliased if only their sum can be identified (operationalized here to mean that only asymptotically consistent estimators of their sum are available). This is a weaker concept than the finite sample concept of aliasing as discussed, for example, in Wu and Hamada (2000); in that setting, effects are aliased if we are only able to obtain an unbiased estimator of their sum.

Defining words are generally used to ascertain finite sample aliasing. Consider the design in Table 2. The defining word for this design is $1 = A_{11}A_{12}A_2^{(1)}$ (the product of the ±1's across the three columns is equal to 1). Equivalently $1 = A_{11}A_{12}A_2^{(0)}$ since the levels of $A_2^{(1)}$ are equal to the levels of $A_2^{(0)}$. Similarly $A_2^{(1)} = A_{11}A_{12}$ (the product of the ±1's in columns 1 and 2 is equal to the ±1's in column 3). In a standard 2^{3-1} design, the defining word $1 = A_{11}A_{12}A_2^{(0)}$ means that each main effect is aliased with a two way interaction (e.g. A_{11} is aliased with $A_{12}A_2^{(0)}$, etc.). *It turns out that even though the screening analysis model for a dynamic treatment regime is based on a nonstandard model (9), we will, nonetheless, be able to use the defining words to ascertain the large sample aliasing.* This means that we can use commonly available designs such as those provided by Wu and Hamada (2000) to design screening studies for dynamic treatment regimes. Lemma 3, below, provides conditions under which the defining words can be used to determine the aliasing of effects.

As in Section 3 we construct the vector $\left[\mathbf{X}_1^T, \left(\mathbf{X}_2^{(1)}\right)^T, \left(\mathbf{X}_2^{(0)}\right)^T\right]$ from $(A_1,\ A_2)$. This vector takes on 2^{k-m} equally likely vector values. Construct the matrix $\left[\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2^{(1)}, \tilde{\mathbf{X}}_2^{(0)}\right]$ with each row corresponding to one of these vectors values. The defining words identify the identical columns in $\left[\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2^{(1)}\right]$ and in $\left[\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2^{(0)}\right]$. Consider the design in Table 2; here

$$\tilde{\mathbf{X}}_1 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \ \tilde{\mathbf{X}}_2^{(1)} = \tilde{\mathbf{X}}_2^{(0)} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix}. \tag{10}$$

Labeling the columns of $\tilde{\mathbf{X}}_1$ as $\{1,\ A_{11},\ A_{12},\ A_{11}A_{12}\}$ and the columns of $\tilde{\mathbf{X}}_2^{(j)}$ by $\{A_2^{(j)},\ A_{11}A_2^{(j)},\ A_{12}A_2^{(j)},\ A_{11}A_{12}A_2^{(j)}\}$ we see that the defining word, $1 = A_{11}A_{12}A_2^{(j)}$ $(j = 0, 1)$ identifies the identical columns.

Lemma 3 Assume that $P[0 < E[R|A_1] < 1] = 1$. Make the *Formal* assumption:

> For each identical column in both $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2^{(1)}$ and for each identical column in both $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2^{(0)}$ either
>
> (3a) the associated nuisance effect(s) (ψ parameters) in (9) are zero or
>
> (3b) the associated stage 2 effects (β, α parameters) in (9) are zero.

Then the defining words can be used to ascertain aliasing.

To make Lemma 3 concrete, consider the analog of (9) for the design in Table 2:

$$\begin{aligned} E[Y|A_1, R, A_2] \quad = \quad & \phi_1 + \phi_2 A_{11} + \phi_3 A_{12} + \phi_4 A_{11}A_{12} + \\ & (R - p(\mathbf{X}_1))(\psi_1 + \psi_2 A_{11} + \psi_3 A_{12} + \psi_4 A_{11}A_{12}) + \\ & R(\beta_1 A_2^{(1)} + \beta_2 A_{11}A_2^{(1)} + \beta_3 A_{12}A_2^{(1)} + \beta_4 A_{11}A_{12}A_2^{(1)}) + \\ & (1 - R)(\alpha_1 A_2^{(0)} + \alpha_2 A_{11}A_2^{(0)} + \alpha_3 A_{12}A_2^{(0)} + \alpha_4 A_{11}A_{12}A_2^{(0)}) \end{aligned} \tag{11}$$

The matrices in (10) permit a variety of formal assumptions. For example, we might make the formal assumption that $\psi_2 = \psi_3 = \psi_4 = 0$ and $\beta_4 = \alpha_4 = 0$. Then Lemma 3 allows us to read off the large sample aliasing from the defining word, $1 = A_{11}A_{12}A_2^{(j)},\ j = 0, 1$. That is each main effect is aliased with a two-way interaction between the remaining predictors (e.g. A_{11} is aliased with $A_{12}A_2^{(0)}$ and so on). This design is most useful if, according to the working assumptions, there are no two-way interactions.

See the Appendix A.1 for a proof of Lemma 3. As before if there are only stage 2 factors for responders (non-responders) then we need only assume $P[E[R|A_1] > 0] = 1$ (respectively $P[E[R|A_1] < 1] = 1$). In the next section we provide an algorithm based on this proof for determining the aliasing and constructing the predictors in the screening analysis. We only use 2^{k-m} designs for which either (3a) or (3b) holds for any common stage 1 and stage 2 columns in $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$ of Lemma 3. The use of designs in which this formal assumption is violated produces aliasing that is not easily discernable by the defining words (see Section A.2 in the Appendix for an example). Note that we can choose a design in which $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$ do not share a column (see example 1 in Section 5) and thus formal assumptions are not required. The formal assumptions, in addition to the working assumptions, are used to select the appropriate 2^{k-m} design. The design is selected so that according to the working assumptions all but one of the effects in each set of aliased effects is thought to be negligible.

4.2 Screening Effects

Of course we cannot fit the model (9), or in the case of our example, (11), since the 2^{k-m} design does not provide Y outcomes for all 2^k values of $\{A_1, A_2\}$ (the omitted outcomes correspond to the omitted rows in the design). To permit a regression model we rewrite (9) in terms of aliased effects. Using the results of Lemma 3, we reduce the number of columns in $\tilde{\mathbf{X}}_1$, $\tilde{\mathbf{X}}_2^{(1)}$ and $\tilde{\mathbf{X}}_2^{(0)}$. Note removing columns from these matrices is equivalent to eliminating predictors in $\mathbf{X}_1$, $\mathbf{X}_2^{(1)}$ and $\mathbf{X}_2^{(0)}$. The construction of the regressors for the screening analysis is straightforward except when formal assumption (3b) is made; see step 4 below.

Algorithm for Constructing the Regressors

1. Construct $\tilde{\mathbf{X}}_1$, $\tilde{\mathbf{X}}_2^{(1)}$ and $\tilde{\mathbf{X}}_2^{(0)}$. Eliminate duplicate columns in each of $\tilde{\mathbf{X}}_1$, $\tilde{\mathbf{X}}_2^{(1)}$ and $\tilde{\mathbf{X}}_2^{(0)}$ and associated predictors from $\mathbf{X}_1$ ($\mathbf{X}_2^{(1)}$, $\mathbf{X}_2^{(0)}$), retaining only the predictor that, according to the working assumptions, may have a non-negligible effect. Name the resulting vector $\mathbf{U}_1$ ($\mathbf{U}_2^{(1)}$, $\mathbf{U}_2^{(0)}$).

2. If any β (α) parameters are assumed zero (via assumption 3b) delete the predictors associated with these parameters in $\mathbf{U}_2^{(1)}$ (respectively $\mathbf{U}_2^{(0)}$); similarly delete the associated columns in $\tilde{\mathbf{X}}_2^{(1)}$, $\tilde{\mathbf{X}}_2^{(0)}$.

3. If some of the nuisance, ψ, parameters are assumed to be zero (via assumption 3a) then create $\mathbf{Z}_3$ from $\mathbf{U}_1$ by eliminating predictors in $\mathbf{U}_1$ with an assumed zero valued ψ parameter. Otherwise $\mathbf{Z}_3 = \mathbf{U}_1$. At this point we can rewrite (9) as

$$E[Y|A_1, R, A_2] = \mathbf{U}_1^T \phi'' + (R - p(\mathbf{U}_1))\mathbf{Z}_3^T \psi'' + R{\mathbf{U}_2^{(1)}}^T \beta'' + (1-R){\mathbf{U}_2^{(0)}}^T \alpha''$$

where the primes on the regression coefficients indicate each entry in ψ'', ψ'', β'', α'' may be a sum of stage 1 effects, nuisance effects, stage 2 effects for responders and stage 2 effects for non-responders, respectively. Here the response rate, $E[R|A_1]$ is written as $p(\mathbf{U}_1)$.

4. If assumption 3b is made (e.g. some stage 2 effects, β, α, parameters are assumed to be zero) then the defining words have identified columns common to at least two of the three matrices $\tilde{\mathbf{X}}_1$, $\tilde{\mathbf{X}}_2^{(0)}$,$\tilde{\mathbf{X}}_2^{(0)}$. If there are one or more columns common to all three matrices then for each common column delete an associated predictor from one of either $\mathbf{U}_1$, $\mathbf{U}_2^{(0)}$ or $\mathbf{U}_2^{(0)}$ resulting in $\mathbf{Z}_1$, $\mathbf{Z}_2^{(0)}$ or $\mathbf{Z}_2^{(0)}$. The choice of which predictor to eliminate determines the aliasing as follows. If the omitted predictor is a stage 1 variable then the coefficient of the stage 2 predictor for responders (non-responders) in the regression (12) is an estimator of the sum of the stage 1 and stage 2 effects for responders $\beta' = \phi'' + \beta''$ ($\alpha' = \phi'' + \alpha''$, respectively). If the omitted predictor is a stage 2 variable for responders (non-responders) then the coefficient of the stage 1 predictor in the regression (12) is an estimator of sum $\phi' = \phi'' + \beta''$ ($\phi' = \phi'' + \alpha''$, respectively) and the coefficient of the stage 2 predictor for non-responders is an estimator of the difference between the stage 1 and stage 2 effects $\alpha' = \alpha'' - \beta''$ ($\beta'' = \beta'' - \alpha''$, respectively). We obtain

$$E[Y|A_1, R, A_2] = \mathbf{Z}_1^T\phi' + (R - p(\mathbf{U}_1))\mathbf{Z}_3^T\psi' + R{\mathbf{Z}_2^{(1)}}^T\beta' + (1-R){\mathbf{Z}_2^{(0)}}^T\alpha'. \qquad (12)$$

Using the working assumptions the regression coefficients can be labeled as corresponding to particular stage 1 effects, nuisance effects, a stage 2 effects for responders and stage 2 effects for non-responders.

Consider once again the example from Table 2 with the formal assumptions that $\psi_2 = \psi_3 = \psi_4 = 0$ and $\beta_4 = \alpha_4 = 0$ (see (11)). After step 3, we have $\mathbf{U}_1 = \{1, A_{11}, A_{12}, A_{11}A_{12}\}$, $\mathbf{Z}_3 = \{1\}$, $\mathbf{U}_2^{(j)} = \{A_2^{(j)}, A_{11}A_2^{(j)}, A_{12}A_2^{(j)}\}$, $j = 0, 1$. In step 4 we select one of several analysis models (each corresponding to a different form of aliasing). Suppose we eliminate $A_{11}A_{12}$ in $\mathbf{U}_1$ to form $\mathbf{Z}_1$ and eliminate $A_{11}A_2^{(1)}, A_{12}A_2^{(1)}$ from $\mathbf{U}_2^{(1)}$ to form $\mathbf{Z}_2^{(1)}$ ($\mathbf{Z}_2^{(0)} = \mathbf{U}_2^{(0)}$), then according to step 4 the conditional mean becomes

$$\begin{aligned} E[Y|A_1, R, A_2] &= \phi_1' + \phi_2' A_{11} + \phi_3' A_{12} + (R - p(\mathbf{U}_1))\psi_1' + \\ &\quad R\beta_1' A_2^{(1)} + (1-R)(\alpha_1' A_2^{(0)} + \alpha_2' A_{11} A_2^{(0)} + \alpha_3' A_{12} A_2^{(0)}) \end{aligned}$$

where $\phi_1' = \phi_1$, $\phi_2' = \phi_2 + \beta_3$, $\phi_3' = \phi_3 + \beta_2$, $\psi_1' = \psi_1$, $\beta_1' = \beta_1 + \phi_4$, $\alpha_1' = \alpha_1 + \phi_4$, $\alpha_2' = \alpha_2 - \beta_3$ and $\alpha_3' = \alpha_3 - \beta_2$ from (11).

The steps in the algorithm follow the steps in the proof of identifiability for Lemma 3; see the proof in Appendix A.1 for a detailed explanation. At this point the total number of entries in $\{\mathbf{Z}_1, \mathbf{Z}_3, \mathbf{Z}_2^{(1)}, \mathbf{Z}_2^{(0)}\}$ is at most $2\left(2^{k-m}\right)$ and the total number of predictors in $\mathbf{U}_1$ is at most 2^{k-m}. To conduct the screening analysis we estimate $p(\mathbf{U}_1)$ for each value of $\mathbf{U}_1$ in the design by the proportion of responders assigned that value to obtain $\hat{p}(\mathbf{U}_1)$ ($\hat{p}(\mathbf{U}_1)$ can be formed from a saturated regression of R on $\mathbf{U}_1$). The screening analysis is a regression of Y on $\left\{\mathbf{Z}_1, (R - \hat{p}(\mathbf{U}_1))\mathbf{Z}_3, R\mathbf{Z}_2^{(1)}, (1-R)\mathbf{Z}_2^{(0)}\right\}$ thus obtaining $\hat{\phi}', \hat{\psi}', \hat{\beta}', \hat{\alpha}'$. See Section 5 for further concrete examples.

Lemma 4 Assume that the variance of Y is finite and that $P[0 < E[R|A_1] < 1] = 1$. Then as $N \to \infty$, the multivariate distribution of $\{\sqrt{N}(\hat{\phi}' - \phi'),\ \sqrt{N}(\hat{\beta}' - \beta'),\ \sqrt{N}(\hat{\alpha}' - \alpha')\}$ converges to an multivariate normal distribution. Furthermore if the model in (12) is saturated then the estimators $\hat{\phi}',\ \hat{\beta}', \hat{\alpha}'$ are semiparametric efficient.

The proof, and an asymptotically consistent estimator for the variance-covariance matrix, is provided in Appendix A.1. In general the model (12) will be saturated. Exceptions occur when more formal assumptions are made than is necessary (see example 3 in Section 5 for an illustration). Or when according to the design, there are one or more stage 2 factors (only for responders or only for non-responders) with levels completely crossed with the levels of all other factors (e.g. these stage 2 factors are randomized independently of the remaining factors and they are not stacked with other factors in the design). For completeness Appendix A.1 also supplies an estimating function yielding semiparametric efficient estimators for use in the case when (12) is not saturated. As was the case for Lemmas 2 and 3, if all responders are provided the same treatment then there are no $\hat{\beta}',\ \beta'$ and $P[0 < E[R|A_1] < 1] = 1$ can be replaced by $P[E[R|A_1] < 1] = 1$ (similar statements can be made if all non-responders are provided the same treatment).

4.3 Sample Size Considerations

The primary goal is to assess the activity of all stage 1 and stage 2 main effects, that is to test if each main effect is zero. To choose the sample size N, we make the following rough approximations. We assume that our formal assumptions, if any, are correct and that the residual variance is equal across the 2^{k-m} rows in the design, say σ^2. If the the stage 1 responder rates were equal (say p) then the asymptotic variance-covariance matrix of the estimated effects would be a diagonal matrix with diagonal elements equal to $\frac{\sigma^2}{Np}$ for the stage 2 effects for responders, $\frac{\sigma^2}{N(1-p)}$ for the stage 2 effects for non-responders and $\frac{\sigma^2}{N}$ for the stage 1 effects. We act as if this is roughly the case. If there are stage 2 factors for both responders and non-responders then the sample size is determined by the smaller of $\frac{\sigma^2}{p}$ and $\frac{\sigma^2}{(1-p)}$. The sample size to detect a main effect of size Δ with power $1-\beta$ using a two-sided test with Type one error rate α is

$$N = \frac{2^{k-m}(z_\beta + z_{\alpha/2})^2}{\min(p_{min}, 1-p_{max})(\Delta/\sigma)^2}$$

where p_{min}, p_{max} are minimal and maximal response rates (across the 2^{k-m} rows) and z_u is the $(1-u)$th standard normal percentile.

Suppose there are stage 2 components for both responders and non-responders. If the signal-to-noise ratio ($SNR = \Delta/\sigma$) is very large then the recommended row size will be so small that the chance of a group (corresponding to a row in the design) occurring with no responders or no non-responders is too high. Given the response rate p and group size $n = N/2^{k-m}$, the chance that a group of this type will occur is $1-(1-p^n-(1-p)^n)^{2^{k-m}}$. To reduce the chance of such groups we find the smallest n for which the above probability is smaller than a cutoff (in the simulations we use .01). To set the group size we use the maximum of the above number times 2^{k-m} and N. Note that if there are no stage 2 components for responders (respectively non-responders) then $1-p^n-(1-p)^n$ in the above formula may be replaced by $1-p^n$ (respectively by $1-(1-p)^n$).

5. Examples

In this section we illustrate several 2^{k-m} designs and associated working and formal assumptions. The examples below are 2^{k-m} generalizations of the clinical studies STAR*D (Fava et al., 2003) and ExTENd (Oslin, personal communication) in which the stage 2 factors vary by whether the patient exhibits an early response. Three different designs, associated with different formal assumptions are illustrated.

Consider a study with four stage 1 factors: S (speciality or general practice clinic), B (adjunctive therapy to improve adherence: yes/no), C (counseling: intensive/less intensive) and T (level of staff training: high/low). All individuals are offered medication. There is one stage 2 factor for early non-responders F_2 (switch or augment medication) and one stage 2 factor for early responders G_2 (telephone disease management: yes/no). Suppose that the working assumptions are that the main effects of the factors and the two way interactions CG_2, TG_2, BF_2, TF_2 and BC may be active with the remaining effects assumed negligible. Recall, these working assumptions are used to guide the choice of the 2^{k-m} design, but are not necessarily assumed true in the screening analysis.

In Design 1, a 2^{4-1} design is constructed for the stage 1 factors and then the levels of the stage 2 factor are crossed with this design. In Design 2 the stage 2 factor is aliased with the four way interaction between the stage 1 factors. Both of these designs are 2^{5-1} designs. In the Design 3 below we allow for an additional stage 2 factor which is applicable only for the non-responders. Here a 2^{6-2} design is discussed.

Design 1: This design crosses a 2^{4-1} design with defining word $SBCT = 1$ with the stage 2 factor levels to produce a 2^{5-1} design; no formal assumptions are made. The defining word indicates that the stage 1 and stage 2 matrices $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2^{(1)}$ (or $\tilde{\mathbf{X}}_2^{(0)}$) do not share columns and thus stage 2 effects are not aliased with nuisance effects (or stage 1 effects). As a result the aliasing follows directly from the defining word (e.g. the two way interaction BC is aliased with the two way interaction ST, and since $SBCTG_2 = G_2$, the two interaction TG_2 is aliased with the interaction $SBCG_2$, and so on). Note that this design is consistent with the working assumptions in that all of the interesting effects are aliased with effects that are thought to be negligible.

The screening analysis uses the model in (12) with $\mathbf{Z}_1 = [1, S, B, C, T, BC, SB, SC]$, $\mathbf{Z}_3 = \mathbf{Z}_1$, $\mathbf{Z}_2^{(1)} = [G_2, SG_2, BG_2, CG_2, TG_2, SBG_2, SCG_2, BCG_2]$ and similarly $\mathbf{Z}_2^{(0)} = [F_2, SF_2, BF_2, CF_2, TF_2, SBF_2, SCF_2, BCF_2]$. This is a saturated model. Recall that due to the nonorthogonality of the predictors, omitting predictors from this model is equivalent to making formal assumptions. An advantage of this design is that no negligibility assumptions on the nuisance parameters (ψ's) need be made. However, this design aliases BC with ST; a better design would avoid aliasing two way interactions.

Design 2: Suppose that we are willing to make the formal assumption that there are no three way or higher order stage 2 interactions (α and β regression coefficients of interactions between a stage 2 factor and two or more stage 1 factors are 0), and that there are no four way or higher nuisance interactions involving R and stage 1 factors (ψ regression coefficients of interactions between R and three or more stage 1 factors are 0). Consider a 2^{5-1} design with defining word $SBCTF_2 = 1$ with F_2 and G_2 stacked (so another expression for the defining word is $SBCTG_2 = 1$). The design is provided in Table 3. The defining words indicate that under the formal assumptions, none of the stage 1 main effects or two-way interactions are aliased. Potentially active stage 2 effects, such as CG_2, can also be estimated since the nuisance effects associated with the same column in the stage 1 matrix $\tilde{\mathbf{X}}_1$, (here the four way interaction between R and SBT) are negligible. And potentially active nuisance effects such as the three way interaction between R and SB can be estimated since these nuisance effects are associated with the same column in the stage 2 matrices $\tilde{\mathbf{X}}_2^{(1)}$ (or $\tilde{\mathbf{X}}_2^{(0)}$) as negligible stage 2 effects (here CTG_2 and CTF_2).

Table 3. Design 2:

S	B	C	T	$G_2 = F_2$
-	-	-	-	+
-	-	-	+	-
-	-	+	-	-
-	-	+	+	+
-	+	-	-	-
-	+	-	+	+
-	+	+	-	+
-	+	+	+	-
+	-	-	-	-
+	-	-	+	+
+	-	+	-	+
+	-	+	+	-
+	+	-	-	+
+	+	-	+	-
+	+	+	-	-
+	+	+	+	+

To screen the factors using data from design 2, we use the model in (12) with with $\mathbf{Z}_1 = [1, S, B, C, T, BC, SB, SC, ST, BT, CT]$, $\mathbf{Z}_3 = \mathbf{Z}_1$, $\mathbf{Z}_2^{(1)} = [G_2, SG_2, BG_2, CG_2, TG_2]$ and similarly $\mathbf{Z}_2^{(0)} = [F_2, SF_2, BF_2, CF_2, TF_2]$. This is a saturated model. Note that the defining word indicates the aliasing, for example the $\hat{\beta}'$ coefficient of G_2 is actually estimating the sum of main effect of G_2 and the four-way interaction $SCBT$.

Design 3: In order to illustrate some of the more subtle considerations, consider the inclusion of an additional stage 2 factor H_2 (additional behavioral contingency to improve long term medication adherence) that is only assigned to non-responders ($R = 0$). Further suppose the formal assumptions are that there are no three way or higher nuisance interactions involving R and stage 1 factors and that three way and higher stage 2 causal effects are negligible. The working assumptions are that all effects except the main effects of the factors and the two way interactions TG_2, TF_2, CG_2, BF_2, F_2H_2 and BC are expected to be negligible. Consider the 2^{6-2} design, with defining words $SCTF_2 = 1$ and $BTF_2H_2 = 1$ and F_2 and G_2 stacked (so we could express the above with a G_2 instead of F_2; note the product of these two defining words yields $1 = SBCH_2$.)

In contrast to designs 1 and 2, not only the experimental design but also the choice of the regression model determines the aliasing (we use step 4 of the algorithm in Section 4.2 to determine the aliasing). Indeed we have a choice of several regression models that can be used to screen the factors each corresponding to different aliasing.

Using the steps in Section 4.2 we see that one possibility is to use the model in (12) with $\mathbf{Z}_1 = [1, S, B, C, T, SB, BC, BT, SBC, SBT, BCT]$, $\mathbf{Z}_3 = [1, S, B, C, T]$, $\mathbf{Z}_2^{(1)} = [G_2, SG_2, BG_2, CG_2, TG_2]$ and $\mathbf{Z}_2^{(0)} = [F_2, SF_2, BF_2, CF_2, TF_2, H_2, SH_2, CH_2, F_2H_2]$. Using the defining words, we can deduce the aliasing. For example since the defining words indicate that $G_2 = F_2 = SCT$ and SCT has been omitted from $\mathbf{X}_1$, we have that the β' coefficient of G_2 is the sum of the main effect of G_2 and the three way interaction SCT; similarly the α' coefficient of F_2 is the sum of the main effect of F_2 and the three way interaction SCT. Note that the defining words indicate that $BF_2 = BG_2 = TH_2 = SBCT$ and we included only BF_2, BG_2 in the regression thus the α' coefficient of BF_2 is the sum of the two way interactions BF_2, TH_2 and the four way interaction $SBCT$ whereas the β' coefficient of BG_2 is estimating the sum of the two way interaction BG_2 and the four way interaction $SBCT$. In general, given the formal assumptions, the defining words along with the choice of $\mathbf{Z}_1$, $\mathbf{Z}_2^{(1)}$ and $\mathbf{Z}_2^{(0)}$ result in the aliasing: $\beta'_{G_2} = \beta_{G_2} + \phi_{SCT}$, $\beta'_{SG_2} = \beta_{SG_2} + \phi_{CT}$, $\beta'_{BG_2} = \beta_{BG_2} + \phi_{SBCT}$, $\beta'_{CG_2} = \beta_{CG_2} + \phi_{ST}$, $\beta'_{TG_2} = \beta_{TG_2} + \phi_{SC}$, $\beta'_{F_2} = \beta_{F_2} + \phi_{SCT}$, $\beta'_{SF_2} = \beta_{SF_2} + \phi_{CT}$, $\beta'_{BF_2} = \beta_{BF_2} + \beta_{TH_2} + \phi_{SBCT}$, $\beta'_{CF_2} = \beta_{CF_2} + \phi_{ST}$, $\beta'_{TF_2} = \beta_{TF_2} + \beta_{BH_2} + \phi_{SC}$ (for clarity the associated predictor is given as the subscript on the parameter). The remaining parameters are unaliased (e.g. the (ϕ', β', α') parameters are equal to the corresponding (ϕ, β, α) parameters).

Another possible screening analysis would use (12) with $\mathbf{Z}_1 = [1, S, B, C, T, SB, BC, BT, CT, SBC, SBT, BCT]$ and $\mathbf{Z}_2^{(0)} = [F_2, BF_2, CF_2, TF_2, H_2, SH_2, CH_2]$ but leave $\mathbf{Z}_3$ and $\mathbf{Z}_2^{(1)}$ as is; that is the predictor CT is added to $\mathbf{Z}_1$ and the predictor SF_2 is removed from $\mathbf{Z}_2^{(0)}$. In this case the aliasing is the same as before except $\phi'_{CT} = \phi_{CT} - \alpha_{SF_2}$, $\beta'_{SG_2} = \beta_{SG_2} - \alpha_{SF_2}$ (there is no longer a α'_{SF_2} parameter).

Both of the two screening analysis models discussed above have 30 parameters and hence are not saturated models. If desired we could fit a saturated model by adding two nuisance interactions to $\mathbf{Z}_3$, SBT and BCT. These two nuisance interactions were needlessly assumed to be negligible in the formal assumptions. For example, from the defining words, note that the column associated with SBT is the same as the column associated with the three way stage 2 interactions SF_2H_2, TCH_2 and BCF_2 all of which were assumed to be negligible. That is we made both assumption 3a and 3b in Lemma 3 instead of one or the other. Estimation of these two effects acts as a check on the formal assumptions since under the formal assumptions these two nuisance effects are zero.

6. Simulation Results

Extensive simulations were conducted so as to evaluate the proposed analysis, examine the impact of violations of the formal assumptions and examine the impact of rows with no responders (or no non-responders) on the analysis. We are particularly interested in mental health settings such as the treatment of major depression and substance abuse in which response (absence of symptoms) rates to initial treatment are around 50 to 70%; as a result the simulations below use initial response rates in this range. Note that in actual practice, when response (non-response) rates are low, usually only factors for non-responders (respectively, responders) are investigated.

The findings were as follows. First when the formal assumptions hold, the standard errors performed well and the Type 1 error is as planned. In general the sample size calculations depend on the group with the smallest proportion of non-responders (or responders). As a result the sample size calculations are conservative and the power to detect active stage 1 main effects and stage 2 main effects is higher than the nominal value. Simulations with both normal and non-normal error distributions such as a t-distribution with 3 degrees of freedom were also conducted; this did not substantially alter the results. When the formal assumptions are violated (e.g. effects assumed negligible in the formal assumptions are not negligible) there is a surprising degree of robustness; that is, the bias is of smaller magnitude than would be expected. Also when one or more rows contain only responders, fitting an analysis model that omitted some active effects led to relatively robust estimators of the remaining effects. This robustness is discussed at greater length following simulation 2. We present three simulations that exemplify the above findings.

The simulations below use Design 2 (see Table 3). In this example the formal assumptions are that there are no three way or higher order stage 2 interactions and there are no four way or higher order nuisance interactions involving R and stage 1 factors. In the simulations provided below, R is generated using a response rate given by $logit\big(E[R|F_1]\big) = .6 + .1S + .1B + .1C + .1T$. This results in response rates varying from .55 to .73 across the 16 rows; this wide range of early response rates is extreme for the mental health/substance abuse fields however it allows us to illustrate the issues. Y is normally distributed with residual variance, σ^2, set to 1. We present results with the signal-to-noise ratio (SNR= effect size/residual variance) equal to .25 or .35 units per standard deviation. In the simulations, active main effects are equal to the SNR$\times\sigma$, active two way interactions are equal to SNR$\times .5 \times \sigma$ and active three way interactions are set equal to SNR$\times .25 \times \sigma$. These settings are consistent with the Hierarchical Ordering Principle (Wu and Hamada, 2000, pg. 112) which states that the lower order effects are more likely to be important than higher order effects and effects of the same order are equally likely to be important (this principle is used in the absence of domain knowledge indicating otherwise). Throughout all main effects and the interactions SB, SC, SG_2, TG_2 and TF_2 are active; thus the working assumptions are incorrect (the working assumptions are that only the main effects of the factors and the two way interactions TG_2, TF_2, CG_2, BF_2 and BC are likely to be active).

We used the formula in Section 4.3 to determine the sample size required to detect a given SNR at 90% power with a 10% Type 1 error rate for $p_{min} = .55$ and $p_{max} = .73$. Following the recommendations in Collins et. al. (2007) interesting effects are tested using a Type 1 error rate of .1 (no correction for multiple tests) and the remaining effects are tested using a overall Type 1 error rate of .1 (using a Bonferroni correction). In all simulations we estimate the variance-covariance matrix of the effects using the formula given in Appendix A.1 (near the beginning of the proof of Lemma 4 in Appendix A.1). Table 4 provides the results from 1000 data sets; the formal assumptions hold for these data sets.

Note that there are 6 effects that according to the working model should be negligible and are actually negligible. Given an overall error rate of .1, the empirical error rate should be around .011 and this is the case as can be seen by the rows labeled ST, BT, CT, BG_2, SF_2 and CF_2. Similarly the empirical Type 1 error rate for the interesting effects should be around .1 and this can be seen by rows labeled BC, CG_2 and BF_2. The high power in detecting main effects is due to the conservatism used in selecting the group size. Recall that the response rates per group range from a low of .55 to an high of .73. The group sizes are chosen then to achieve the power .9 to detect stage 2 effects for non-responders when the non-response rate is .27. Since only one of the groups exhibits this low non-response rate, the group sizes are conservative, resulting in higher power. In simulations with a constant response rate across all 16 groups (not shown here) the power to detect a main effect for stage 2 non-responders varied from around .88 to .94 depending on the simulation.

Table 4. Simulation 1: Formal Assumptions Hold[1]

Effect	True effect	Avg. Est. Effect	Std. Error	Avg. Est. Std. Error	Power[2]	Bonf. Corrected Power[3]
Stage 1						
S	.25	.248	.044	.046	1.00	
B	.25	.253	.044	.046	1.00	
C	.25	.250	.045	.046	1.00	
T	.25	.250	.045	.046	1.00	
SB	.125	.124	.045	.046		.54
SC	.125	.124	.046	.046		.55
ST	.0	.002	.046	.046		.01
BC	.0	.000	.045	.046	.09	
BT	.0	.000	.043	.046		.01
CT	.0	.000	.046	.046		.01
Stage 2-Responders						
G_2	.25	.247	.057	.058	1.00	
SG_2	.125	.126	.056	.058		.34
BG_2	.0	-.002	.055	.058		.01
CG_2	.0	-.001	.055	.058	.10	
TG_2	.125	.124	.058	.058	.71	
Stage 2 Non-responders						
F_2	.25	.251	.074	.076	.96	
SF_2	.0	.001	.077	.076		.01
BF_2	.0	-.002	.077	.076	.11	
CF_2	.0	-.001	.079	.076		.02
TF_2	.125	.125	.077	.076	.52	

[1] SNR= .25, Nuisance term in (12) is $(R - E[R|A_1])(.25 + .125S + .125B + .125T + .0625SB + .0625SC + .0625ST + .0625BC + .0625BT + .0625CT)$, $N = 512$.

[2] Individual Type 1 error rate is .1

[3] Overall Type 1 error rate is .1

In the next simulation, also of 1000 simulated data sets, the formal assumptions are violated; there are active three way stage 2 effects for responders. The results in Table 5 are surprisingly good; the results in this table would be expected *if* the interactions involving R were zero (no nuisance interactions). To see this, recall the defining words for this design are $1 = SBCTG_2$ (equivalently $1 = SBCTF_2$). Thus, for example the column in the stage 2 matrix $\tilde{\mathbf{X}}_2^{(0)}$ matrix associated with CTF_2 is the same as the column associated with SB in $\tilde{\mathbf{X}}_1$. If the coefficient of the interaction $(R - E[R|A_1])SB$ were zero (it is not) then we would expect that estimators of the stage 1 interactions such as SB will estimate this stage 1 interaction plus the product of the response rate times the effect of CTG_2 plus the non-response rate times the effect of CTF_2. In this case, this is .175+ response rate(.0875)+ non-response rate(0).

If we use the average non-response rate (here .64), this yields .231 which is close to the average estimated effect for SB. Similar statements can be made for remaining stage 1 two way interactions. This robustness is explained by the fact that the response rates vary only from .55 to .73 across the groups. As discussed in Section 4, the less the response rates vary, the closer the predictors are to being orthogonal. Thus the estimators of the effects are approximately uncorrelated (the off diagonal elements of the correlation matrix are small–in this simulation the maximum correlation in absolute value is .12 and the average of the absolute value of the correlations is .03). Thus as long as the response rates do not vary greatly (as is the case in many areas of mental health and substance abuse) we can expect this robustness. Similar results hold when the formal assumptions concerning the nuisance effects are violated (not shown here).

Table 5. Simulation 2: Formal Assumptions are Violated[1]

Effect	True effect	Avg. Est. Effect	Std. Error	Avg. Est. Std. Error	Power[2]	Bonf. Corrected Power[3]
Stage 1						
S	.35	.359	.054	.056	1.00	
B	.35	.353	.050	.056	1.00	
C	.35	.355	.053	.056	1.00	
T	.35	.356	.053	.056	1.00	
SB	.175	.228	.054	.056		.95
SC	.175	.230	.054	.056		.94
ST	.0	.056	.054	.056		.06
BC	.0	.057	.053	.056	.26	
BT	.0	.054	.054	.056		.05
CT	.0	.056	.054	.056		.06
Stage 2-Responders[4]						
G_2	.35	.349	.065	.069	1.00	
SG_2	.175	.181	.064	.069		.52
BG_2	.0	.006	.063	.069		.01
CG_2	.0	.004	.064	.069	.08	
TG_2	.175	.182	.067	.069	.86	
Stage 2 Non-responders						
F_2	.35	.348	.086	.090	.99	
SF_2	.0	.001	.092	.090		.01
BF_2	.0	.005	.089	.090	.09	
CF_2	.0	-.001	.090	.090		.01
TF_2	.175	.182	.090	.090	.65	

[1] SNR= .35, Nuisance term in (12) is $(R - E[R|A_1])(.35 + .175S + .175B + .175T + .0875SB + .0875SC + .0875ST + .0875BC + .0875BT + .0875CT)$, $N = 384$.

[2] Individual Type 1 error rate is .1

[3] Overall Type 1 error rate is .1

[4] All six three way stage 2 interactions for responders are equal to .0875 but are not included in analysis model.

Also this simulation, by chance, did not result in any samples with one or more groups containing only responders. However this can happen and did happen in other simulations (not shown). Indeed when SNR= .35 the group sizes had to be adjusted upwards to ensure that the probability that all groups have some non-responders is not too low (low is arbitrarily chosen to be .01; see Section 4.3). Even with this adjustment, a simulation size of 1000 samples will sometimes include some samples in which groups with no responders occur. If in a given data set some groups (e.g. corresponding to rows in the design) contain no responders we fit a model using the working assumptions, that is, $\mathbf{Z}_1 = [1, S, B, C, T, BC]$ $\mathbf{Z}_3 = \mathbf{Z}_1$, $\mathbf{Z}_2^{(1)} = [F_2, BF_2, TF_2]$ and $\mathbf{Z}_2^{(0)} = [G_2, CG_2, TG_2]$.

The working assumptions are incorrect thus the omission of active effects biases the estimated effects. To examine the degree of bias we conducted a simulation in which the groups corresponding to the rows of the design were sufficiently small so that the chance of one or more groups with all responders was likely; in effect we simulated many samples saving only those that had one or more groups with all responders. Table 6 reports the results for 729 such samples.

Table 6. Simulation 3: One or More Cells with No Non-responders[1]

Effect	True effect	Avg. Est. Effect	Std. Error	Avg. Est. Std. Error	Power[2]
Stage 1[3]					
S	.50	.509	.083	.088	1.00
B	.50	.512	.083	.088	1.00
C	.50	.504	.085	.088	1.00
T	.50	.506	.086	.088	1.00
BC	.0	.017	.086	.088	.09
Stage 2 Responders[3]					
G_2	.50	.506	.115	.117	1.00
CG_2	.0	.017	.117	.117	.11
TG_2	.25	.289	.123	.118	.78
Stage 2 Non-responders					
F_2	.50	.484	.193	.175	.86
BF_2	.0	-.028	.191	.174	.14
TF_2	.125	.188	.192	.173	.32

[1] SNR= .5; 729 samples; nuisance term is $(R - E[R|F_1])(.5 + .25S + .25B + .25T + .125SB + .125SC + .125ST + .125BC + .125BT + .125CT)$; $N = 176$.

[2] Individual Type 1 error rate is .1

[3] SB, SC, SG_2 have coefficients equal to .25 but are omitted in the regression analysis.

As with simulation 2, this simulation demonstrates robustness of the analysis method to the omission of active effects. Note that both the bias of the estimators and the quality of the standard errors is poorest for the stage 2 effects for non-responders. This is not surprising as many rows will have few non-responders.

7. Discussion

An attractive alternative to the use of randomized clinical trials in constructing dynamic treatment regimes is the use of observational data in which the components are not randomized. Here the variation in treatment is usually due to patient/clinician preference, patient adherence, component availability, etc. Observational data has the great advantage in that it often already exists and/or is less expensively obtained than data from the screening designs discussed here. However note that observational data suffers from two substantial drawbacks. First inferences regarding causal relations in the absence of randomization requires untestable assumptions (Rubin, 1978; Robins, 1992) concerning why individuals receive, or are offered, differing treatments. Second, even if we are comfortable with the required assumptions, then because we do not control the treatment levels in the observational study we have uncontrolled aliasing. That is, when our modeling assumptions do not hold, the form of the aliasing may be complex and difficult to ascertain. In contrast the screening designs considered here not only improve our ability to make causal inferences but also provide a greater understanding of the aliasing that occurs due to incorrect modeling assumptions. However there are many open problems. Four such problems follow.

First there is the question of how to construct the optimal experimental designs. A first thought might be to utilize the maximum resolution criterion or its refinement, the minimum aberration criterion (see Wu and Hamada, 2000). Note that the design in example 1 has only resolution IV whereas the design in example 2 has higher resolution (V). It is not necessarily true that the example 2 design is to be preferred since it requires formal assumptions whereas the example 1 design does not. More work is required to appropriately incorporate the roll of the formal assumptions into methods for comparing these designs.

Second as discussed in Step 4 of the Algorithm in Section 4.2 some designs permit multiple analysis models; because of the non-orthogonality of the predictors, different models result in tests with different powers. A strategy for choosing among these different analysis models is needed.

Third, designs with a smaller number of groups (e.g. treatment combinations) might be considered. To do so, we would begin by eliminating negligible effects from a potential model via formal and working assumptions. Next, we could attempt to find, say, a D-optimal design (e.g., see Wu and Hamada, 2000) that would permit estimation of the remaining effects. This would yield designs that have fewer than 2^{k-m} groups. However, such designs rarely have the nice aliasing structure found with 2^{k-m} factorial designs. Furthermore, the designs proposed here frequently permit us to assess the working assumptions, unlike most D-optimal designs (see design 3 in Section 5). We view this as an important advantage.

Lastly, this paper has not discussed potential secondary analyses that might be used with data from a 2^{k-m} design. Such analyses might consider how best to utilize time-varying patient covariates in the construction of the decision rules. The methods of Murphy (2003) and Robins (2004) require generalization as these methods require randomization or, at a minimum, stochastic variation in assigned stage 2 factors.

Acknowledgments

We acknowledge support for this project from National Institute of Health grants RO1 MH080015 and P50 DA10075.

References

1. Box GEP, Hunter WG & JS Hunter (1978). Statistics for Experimenter: An Introduction to Design, Data Analysis, and Model Building. New York: Wiley & Sons.

2. Byar DP, Piantadosi S (1985). Factorial designs for randomized clinical trials. *Cancer Treat Rep.* Oct;69(10):1055-63.

3. Brooner, RK and M. Kidorf (2002). Using behavioral reinforcement to improve methadone treatment participation. *Science and Practice Perspectives* **1**:38-48.

4. Chamberlain, G (1986). Asymptotic efficiency in semi-parametric models with censoring, *Journal of Econometrics.* **32**:189-218.

5. Collins LM, Murphy SA and V. Strecher (2007). The Multiphase Optimization Strategy (MOST) and the Sequential Multiple Assignment Randomized Trial (SMART): New Methods for More Potent e-Health Interventions. *American Journal of Preventive Medicine*, **32(5S)**:S112-118.

6. Fava M, Rush AJ, Trivedi MH, Nierenberg AA, Thase ME, Sackeim HA, Quitkin FM, Wisniewski S, Lavori PW, Rosenbaum JF, Kupfer DJ. (2003) Background and Rationale for the Sequenced Treatment Alternative to Relieve Depression (STAR*D) Study. *Psychiatric Clinics of North America* **26(3)**:457-494.

7. Fries, A & WG Hunter (1980). Minimum Aberration 2^{k-p} Designs, *Technometrics*, **22**:601-608.

8. Hu, F & WF Rosenberger (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials.* Hoboken, NJ:John Wiley & Sons, Inc.

9. ICH E9 (1999). ICH Harmonised Tripartite Guideline, Statistical Principles for Clinical Trials. *Statistics in Medicine*, **18**:1905-1942.

10. Lavori, PW, R Dawson, & AJ Rush, (2000). Flexible treatment strategies in chronic disease: Clinical and research implications. *Biological Psychiatry*, **48**: 605-614.

11. Montgomery DC & CL Jennings (2006). An Overview of Industrial Screening Experiments. In A. Dean and S. Lewis, eds. *Screening Methods for Experimentation in Industry, Drug Discovery, and Genetics.* New York: Springer.

12. Murphy S.A. (2003), Optimal Dynamic Treatment Regimes. *Journal of the Royal Statistical Society, Series B (with discussion)* **65**(2):331-366.

13. Murphy SA (2005). An Experimental Design for the Development of Adaptive Treatment Strategies., *Statistics in Medicine.* **24**:1455-1481.

14. Murphy, SA, van der Laan MJ, Robins JM and CPPRG (2001), marginal mean models for dynamic regimes *JASA*, **96** 1410-1423.

15. Newey WK (1990), Semiparametric efficiency bounds *Journal of Applied Econometrics*, **5(2)** 99-135.

16. Parmigiani, G. (2002). Modeling in Medical Decision Making: A Bayesian Approach. Wiley & Sons, Ltd: England.

    Patel MS (1962). Group-Screening with More Than Two Stages *Technometrics*, **4**(2) 209-217.

17. 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. *Computers and Mathematics with Applications* **14**, 1393–1512.

18. 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." *Computers and Mathematics with Applications* **14**, 923–945.

19. Robins JM. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, **79**, 321-34.

20. Robins, J.M. (1997). Causal Inference from complex longitudinal data. *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics (120)*, 69–117, (eds: M. Berkane). Springer-Verlag, Inc, New York.

21. Robins, J.M. and Greenland S. (1994). Adjusting for differential rates of PCP prophylaxis in high- versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association*, **89**, 737-749.

22. Robins, JM and Wasserman, L (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, 409–42, (eds: D. Geiger, P. Shenoy). Morgan Kaufmann, San Francisco.

23. Robins JM (2004). Optimal structural nested models for optimal sequential decisions. In DY Lin and P Heagerty (Eds.), Proceedings of the Second Seattle Symposium on Biostatistics , New York. Springer.

24. Rubin DB, (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, **66**(5), 688-701.

25. Rubin, DB, (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, **6**, 34-58.

26. Tsiatis AA (2006). *Semiparametric Theory and Missing Data.* New York: Springer.

27. Wu JCF & M Hamada (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization.* New York: John Wiley & Sons, Inc.

28. Zacks S (1996). Adaptive Designs for Parametric Models in S. Ghosh and C.R. Rao, eds,. *Handbook of Statistics*, Vol. 13 Elsevier Science B.V. Amsterdam

Appendix

A.1 Lemmas

Proof of Lemma 3

First assume there are factors for both responders and non-responders. We have for $R = 0, 1$,

$$E[\mathbf{Y}|R] = \tilde{\mathbf{X}}_1\phi + (R - \mathbf{D_p})\tilde{\mathbf{X}}_1\psi + R\tilde{\mathbf{X}}_2^{(1)}\beta + (1-R)\tilde{\mathbf{X}}_2^{(0)}\alpha \tag{13}$$

where the defining words specify the common columns in $\tilde{\mathbf{X}}_1$, $\tilde{\mathbf{X}}_2^{(1)}$ and $\tilde{\mathbf{X}}_2^{(0)}$ ($\mathbf{D_p}$ is a diagonal matrix with $\mathbf{p}$ on the diagonal). Note that the unique columns in $\{\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2^{(1)}\}$ and in $\{\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2^{(0)}\}$ each number at most 2^{k-m}. We group terms in the above display so that only unique columns remain. First if a column occurs k times in $\tilde{\mathbf{X}}_1$ then delete $k-1$ of the columns and sum all terms in ψ and ϕ associated with the column. Each sum becomes the coefficient of the remaining column; all of the remaining entries in ψ, ϕ remain the same. This process is repeated until all columns are unique resulting in a matrix $\tilde{\mathbf{Z}}_1$ and two vectors of coefficients ψ', ϕ' for which $\tilde{\mathbf{X}}_1\psi = \tilde{\mathbf{Z}}_1\psi'$ and

$\tilde{\mathbf{X}}_1\phi = \tilde{\mathbf{Z}}_1\phi'$. We follow a similar procedure for the two stage 2 matrices resulting in $\tilde{\mathbf{X}}_2^{(1)}\beta = \tilde{\mathbf{Z}}_2^{(1)}\beta'$ and $\tilde{\mathbf{X}}_2^{(0)}\alpha = \tilde{\mathbf{Z}}_2^{(0)}\alpha'$.

Next if assumption 3a was used then some of the entries in ψ' are known to be zero. Delete the associated columns from $\tilde{\mathbf{Z}}_1$ to form $\tilde{\mathbf{Z}}_3$ and delete these entries from ψ'. If assumption 3b was made then follow the same procedure for $\tilde{\mathbf{Z}}_2^{(1)}$ and β' and similarly for $\tilde{\mathbf{Z}}_2^{(0)}$ and α'.

At this point there are no duplicate columns in $\tilde{\mathbf{Z}}_1$ or in $\tilde{\mathbf{Z}}_3$ or in $\tilde{\mathbf{Z}}_2^{(1)}$ or in $\tilde{\mathbf{Z}}_2^{(0)}$. Furthermore there is no column common to both $\tilde{\mathbf{Z}}_3$ and $\tilde{\mathbf{Z}}_2^{(1)}$ and no column common to both $\tilde{\mathbf{Z}}_3$ and $\tilde{\mathbf{Z}}_2^{(0)}$. This is because such a column would have been assumed to either have a ψ coefficient or α, β coefficients equal to zero by assumptions 3a), 3b) respectively. There may, however, be one or more columns which are common to all three, $\tilde{\mathbf{Z}}_1$, $\tilde{\mathbf{Z}}_2^{(1)}$, $\tilde{\mathbf{Z}}_2^{(0)}$, matrices. Note the above discussion implies these columns are not in $\tilde{\mathbf{Z}}_3$. To obtain regression coefficients that are identifiable, we need to eliminate each common column from one of these three matrices. To see this consider one such common column, $\mathbf{z}$ which is, say, the ith column of $\tilde{\mathbf{Z}}_1$, the jth column of $\tilde{\mathbf{Z}}_2^{(1)}$ and the kth column of $\tilde{\mathbf{Z}}_2^{(0)}$. This column contributes $\mathbf{z}\phi'_i + R\mathbf{z}\beta'_j + (1-R)\mathbf{z}\alpha'_k$ to $E[\mathbf{Y}|R]$. We can write this sum in three ways corresponding to how we delete this column from one of the three matrices:

$$\begin{aligned}\mathbf{z}\phi'_i + R\mathbf{z}\beta'_j + (1-R)\mathbf{z}\alpha'_k &= R\mathbf{z}(\beta'_j + \phi'_i) + (1-R)\mathbf{z}(\alpha'_k + \phi'_i)\\ &= \mathbf{z}(\phi'_i + \beta'_j) + (1-R)\mathbf{z}(\alpha'_k - \beta'_j)\\ &= \mathbf{z}(\phi'_i + \alpha'_k) + R\mathbf{z}(\beta'_j - \alpha'_k).\end{aligned}$$

If we delete $\mathbf{z}$ from $\tilde{\mathbf{Z}}_1$ then we will be able to identify the sum of the associated stage 1 and stage 2 effects; if we delete $\mathbf{z}$ from $\tilde{\mathbf{Z}}_2^{(1)}$ then we will be able to identify the sum of the associated stage 1 effect and the stage 2 effect for responders and also the difference between the stage 2 effect for non-responders and the stage 2 effect for responders; if we delete $\mathbf{z}$ from $\tilde{\mathbf{Z}}_2^{(0)}$ then we will be able to identify the sum of the associated stage 1 effect and the stage 2 effect for non-responders and also the difference between the stage 2 effect for responders and the stage 2 effect for non-responders.

Note the assumption that $P[E[R|A_1] > 0] = 1$ ($P[E[R|A_1] < 1] = 1$) implies that the $2^{k-m} \times 1$ vector $E[\mathbf{Y}|R = 1]$, respectively $E[\mathbf{Y}|R = 0]$, is identifiable. Furthermore the vector of response rates $\mathbf{p}$ is identifiable. We have

$$\begin{aligned}E[\mathbf{Y}|R=1] &= \tilde{\mathbf{Z}}_1\phi' + \mathbf{D}_{1-\mathbf{p}}\tilde{\mathbf{Z}}_3\psi' + \tilde{\mathbf{Z}}_2^{(1)}\beta'\\ E[\mathbf{Y}|R=0] &= \tilde{\mathbf{Z}}_1\phi' - \mathbf{D}_{\mathbf{p}}\tilde{\mathbf{Z}}_3\psi' + \tilde{\mathbf{Z}}_2^{(0)}\alpha'.\end{aligned}$$

Multiply all terms in both of these equations by $\tilde{\mathbf{Z}}_3^T$ and subtract the lower equation from the top. Note that $\tilde{\mathbf{Z}}_3$ does not share a column with $\tilde{\mathbf{Z}}_2^{(i)}$ so $\tilde{\mathbf{Z}}_3^T\tilde{\mathbf{Z}}_2^{(i)} = 0$ for $i = 0, 1$ and that $\tilde{\mathbf{Z}}_3^T\tilde{\mathbf{Z}}_3$ is the identity matrix. We obtain $\psi' = \tilde{\mathbf{Z}}_3^T\left(E[\mathbf{Y}|R=1] - E[\mathbf{Y}|R=0]\right)$. So ψ' is identifiable. A little more matrix algebra yields

$$\begin{aligned}\tilde{\mathbf{Z}}_1^T\left(E[\mathbf{Y}|R=1] - \mathbf{D}_{1-\mathbf{p}}\tilde{\mathbf{Z}}_3\psi'\right) &= \phi' + \tilde{\mathbf{Z}}_1^T\tilde{\mathbf{Z}}_2^{(1)}\beta'\\ \tilde{\mathbf{Z}}_1^T\left(E[\mathbf{Y}|R=0] + \mathbf{D}_{\mathbf{p}}\tilde{\mathbf{Z}}_3\psi'\right) &= \phi' + \tilde{\mathbf{Z}}_1^T\tilde{\mathbf{Z}}_2^{(0)}\alpha'.\end{aligned}$$

Identifiability of ϕ', α' and β' follows from the fact that there are no columns common to all three of the matrices.

In there are no factors for responders (non-responders) remove $\tilde{\mathbf{X}}_2^{(1)}\beta$ (respectively $\tilde{\mathbf{X}}_2^{(0)}\alpha$) from (13). Then follow the above arguments.

Proof of Lemma 4

We assume there are factors for both responders and non-responders. Similar arguments can be used when there are only factors for one or the other. Recall a 2^{k-m} design is used; that is $\{A_1, A_2\}$ have a discrete uniform distribution over the rows of the 2^{k-m} design. Recall that $\hat{p}(\mathbf{U}_1)$ denotes the observed response proportion for each unique value of $\mathbf{U}_1$. Let η be the parameter in the bernoulli distribution of R given A_1 parameterized as $p(\mathbf{U}_1) = \mathbf{U}_1^T\eta$. Let $\gamma' = [(\psi')^T,\ (\phi')^T,\ (\beta')^T,\ (\alpha')^T]^T$ (placing ψ' first is for convenience in the invertibility proof below). Note that $\{\mathbf{Z}_3,\ \mathbf{Z}_1,\ \mathbf{Z}_2^{(1)},\ \mathbf{Z}_2^{(0)}\}$ in (12) are simply functions of A_1, A_2. $\mathbf{Z}_1$, $\mathbf{Z}_3$ are subvectors of $\mathbf{U}_1$.

Define $\mathbf{Z}(\eta) = \left[(R - \mathbf{U}_1^T\eta)\mathbf{Z}_3^T, \mathbf{Z}_1^T, R\mathbf{Z}_2^{(1)^T}, (1-R)\mathbf{Z}_2^{(0)^T}\right]$. Also let $\mathbf{E}_N$ be expectation with respect to the empirical distribution of the data. Then solve

$$\begin{aligned} 0 &= \mathbf{E}_N\left[\mathbf{Z}(\eta)(Y - \mathbf{Z}(\eta)^T\gamma')\right] \\ 0 &= \mathbf{E}_N\left[\mathbf{U}_1(R - \mathbf{U}_1^T\eta)\right] \end{aligned} \tag{14}$$

for γ', η to obtain $\hat{\gamma}'$, $\hat{\eta}$. The second equation is the normal equation for a regression of R on $\mathbf{U}_1$. Recall the model used for the conditional distribution of R is saturated so weighting by the variance of R would not alter the results; in fact the value of $u_1^T\hat{\eta}$ will be equal to the observed combined response rate for the rows consistent with u_1. The asymptotic arguments are standard exercises. As a result we provide only an outline. First, standard asymptotic arguments are sufficient to show convergence in probability of $\hat{\gamma}'$ to γ' and $\hat{\eta}$ to η. Furthermore define $\Sigma_\gamma = E\left[\mathbf{Z}(\eta)\mathbf{Z}(\eta)^T\right]$, $\Sigma_{\gamma,\eta} = E\left[\mathbf{Z}(\eta)\left(\mathbf{Z}_3^T\psi'\right)\mathbf{U}_1^T\right]$ and $\Sigma_{\eta,\eta} = E\left[\mathbf{U}_1\mathbf{U}_1^T\right]$ and the corresponding estimators by $\hat{\Sigma}_\gamma = \mathbf{E}_N\left[\mathbf{Z}(\hat{\eta})\mathbf{Z}(\hat{\eta})^T\right]$, $\hat{\Sigma}_{\gamma,\eta} = \mathbf{E}_N\left[\mathbf{Z}(\hat{\eta})(\mathbf{Z}_3^T\hat{\psi})\mathbf{U}_1^T\right]$ and $\hat{\Sigma}_{\eta,\eta} = \mathbf{E}_N\left[\mathbf{U}_1\mathbf{U}_1^T\right]$. Note that $\Sigma_{\eta,\eta}$ is the identity matrix due to the orthogonality of design ($\hat{\Sigma}_{\eta,\eta}$ may not be the identity since in finite samples, the groups sizes may be unequal). It is easily shown that $\hat{\Sigma}_\gamma$, $\hat{\Sigma}_{\gamma,\eta}$ and $\hat{\Sigma}_{\eta,\eta}$ converge in probability (as the sample size $N \to \infty$) Σ_γ, $\Sigma_{\gamma,\eta}$ and $\Sigma_{\eta,\eta}$.

Below we show that Σ_γ is invertible. Again standard arguments can be used to show that

$$\Sigma_\gamma\sqrt{N}\left(\hat{\gamma}' - \gamma'\right) = \sqrt{N}(\mathbf{E}_N - E)\left[\ell(\gamma', \eta, \Sigma_{\gamma,\eta}, \Sigma_{\eta,\eta})\right] + o_P(1).$$

where $\ell(\gamma', \eta, \Sigma_{\gamma,\eta}, \Sigma_{\eta,\eta}) = \mathbf{Z}(\eta)(Y - \mathbf{Z}(\eta)^T\gamma') - \Sigma_{\gamma,\eta}\Sigma_{\eta,\eta}^{-1}\mathbf{U}_1(R - \mathbf{U}_1^T\eta)$. Thus the asymptotic variance-covariance matrix of $\sqrt{N}\left(\hat{\gamma}' - \gamma'\right)$ is

$$\Sigma_\gamma^{-1}E\left[\left[\ell(\gamma', \eta, \Sigma_{\gamma,\eta}, \Sigma_{\eta,\eta})\right]\left[\ell(\gamma', \eta, \Sigma_{\gamma,\eta}, \Sigma_{\eta,\eta})\right]^T\right]\Sigma_\gamma^{-1}.$$

A consistent estimator of the asymptotic variance-covariance matrix is provided by replacing Σ_γ by $\hat{\Sigma}_\gamma$, $\Sigma_{\gamma,\eta}$ by $\hat{\Sigma}_{\gamma,\eta}$, $\Sigma_{\eta,\eta}$ by $\hat{\Sigma}_{\eta,\eta}$, η by $\hat{\eta}$, γ' by $\hat{\gamma}'$ and the operator E by $\mathbf{E}_N$ in the above. To improve the accuracy of the variance estimators when samples are small, we recommend adjusting the formula for the number of estimated parameters.

That is multiply the estimated variance-covariance matrix by $N/(N - q_\gamma - q_\eta)$ where q_γ is the dimension of γ' and q_η is the dimension of η.

Proof that Σ_γ is invertible: Construct the matrices $\tilde{\mathbf{Z}}_1$, $\tilde{\mathbf{Z}}_3$, $\tilde{\mathbf{Z}}_2^{(1)}$, $\tilde{\mathbf{Z}}_2^{(0)}$ as in the proof of Lemma 3 above. Note each row is equal to one of the 2^{k-m} equally probable realizations of $\left[\mathbf{Z}_1^T\ \mathbf{Z}_3^T,\ \left(\mathbf{Z}_2^{(1)}\right)^T,\ \left(\mathbf{Z}_2^{(0)}\right)^T\right]$. Define $\mathbf{p}$ to be a vector of the 2^{k-m} response probabilities (e.g. the ith entry is given by $u_{i1}^T\eta$ where u_{i1} is the value of $\mathbf{U}_1$ corresponding to the ith row of the 2^{k-m} design). Define $\mathbf{D}_{-\mathbf{p}}$ to be a diagonal matrix with diagonal elements equal to $-\mathbf{p}$. Define

$$\tilde{\mathbf{V}} = \begin{bmatrix} \mathbf{D}_{1-\mathbf{p}}\tilde{\mathbf{Z}}_3 & \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_2^{(1)} & 0 \\ \mathbf{D}_{-\mathbf{p}}\tilde{\mathbf{Z}}_3 & \tilde{\mathbf{Z}}_1 & 0 & \tilde{\mathbf{Z}}_2^{(0)} \end{bmatrix}$$

where the 0's denote conforming matrices with all entries equal to zero. Then $2^{k-m}\Sigma_\gamma$ can be written as

$$\tilde{\mathbf{V}}^T \begin{bmatrix} \mathbf{D}_{\mathbf{p}} & 0 \\ 0 & \mathbf{D}_{1-\mathbf{p}} \end{bmatrix} \tilde{\mathbf{V}}.$$

We show that $\tilde{\mathbf{V}}$ is of full rank. This combined with the assumption that the response probabilities are bounded away from both 0 and 1 will imply that Σ_γ is invertible.

Note that $\tilde{\mathbf{V}}$ will have less than $2(2^{k-m})$ columns (it has $2(2^{k-m})$ rows) if there are unequal numbers of factors for responders and non-responders and/or unnecessary formal assumptions have been made. We can add columns to $\tilde{\mathbf{V}}$ by adding back in columns to $\tilde{\mathbf{Z}}_3$ associated with nuisance effects that did not necessarily need to be assumed zero (see example 3 for an illustration). Suppose there are fewer stage 2 factors for responders than for non-responders; this means that some of the stage 2 factors for non-responders are not stacked with a stage 2 factor for responders. Add the columns in $\tilde{\mathbf{Z}}_2^{(0)}$ associated with these non-stacked factors to $\tilde{\mathbf{Z}}_2^{(1)}$. A similar procedure can be followed in there are fewer stage 2 factors for non-responders than responders. If we prove this augmented $2(2^{k-m}) \times 2(2^{k-m})$ matrix $\tilde{\mathbf{V}}$ is of full rank then certainly any subset of columns is also full rank. Hence it suffices to prove that $\tilde{\mathbf{V}}$ is of full rank in the case when it is square (has $2(2^{k-m})$ columns).

Next $\tilde{\mathbf{V}} = \tilde{\mathbf{V}}_1 + \tilde{\mathbf{V}}_2$ where

$$\tilde{\mathbf{V}}_1 = \begin{bmatrix} \tilde{\mathbf{Z}}_3 & \tilde{\mathbf{Z}}_1 & 0 & \tilde{\mathbf{Z}}_2^{(0)} \\ 0 & \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_2^{(1)} & 0 \end{bmatrix}, \ \tilde{\mathbf{V}}_2 = \begin{bmatrix} \mathbf{D}_{-\mathbf{p}}\tilde{\mathbf{Z}}_3 & 0 & 0 & 0 \\ \mathbf{D}_{-\mathbf{p}}\tilde{\mathbf{Z}}_3 & 0 & 0 & 0 \end{bmatrix}$$

$\tilde{\mathbf{V}}_1$ is easily shown to be of full rank, once one identifies common columns in $\tilde{\mathbf{Z}}_3$, $\tilde{\mathbf{Z}}_1$, $\tilde{\mathbf{Z}}_2^{(1)}$, $\tilde{\mathbf{Z}}_2^{(0)}$. Note by assumption there are no columns common to $\tilde{\mathbf{Z}}_3$ and $\tilde{\mathbf{Z}}_2^{(1)}$ or $\tilde{\mathbf{Z}}_3$ and $\tilde{\mathbf{Z}}_2^{(0)}$. And there are no columns common to all three $\tilde{\mathbf{Z}}_1$, $\tilde{\mathbf{Z}}_2^{(1)}$, $\tilde{\mathbf{Z}}_2^{(0)}$. Thus $\tilde{\mathbf{Z}}_3$ is composed of columns common with $\tilde{\mathbf{Z}}_1$ and columns unique to $\tilde{\mathbf{Z}}_3$. Similarly $\tilde{\mathbf{Z}}_1$ is composed of columns that are common with $\tilde{\mathbf{Z}}_3$, columns common with $\tilde{\mathbf{Z}}_2^{(1)}$ and columns common with $\tilde{\mathbf{Z}}_2^{(0)}$. $\tilde{\mathbf{Z}}_1$ has no unique columns(if it did then this column should have been added to $\tilde{\mathbf{Z}}_3$ when $\tilde{\mathbf{V}}$ was augmented). The inner product of a column with itself is 2^{k-m} and the inner product of any other two columns is zero. These facts imply that the inverse of $\tilde{\mathbf{V}}_1^T\tilde{\mathbf{V}}_1$ is easily found. Let q_3 be the number of columns in $\tilde{\mathbf{Z}}_3$. For the purposes of this proof we need only note that the first q_3 rows of $\left(\tilde{\mathbf{V}}_1^T\tilde{\mathbf{V}}_1\right)^{-1}$ are proportional to the matrix:

$$\begin{bmatrix} 2I_{q_{3c}\times q_{3c}} & 0 & -I_{q_{3c}\times q_{3c}} & 0 \\ 0 & I_{q_{3u}\times q_{3u}} & 0 & 0 \end{bmatrix}$$

where q_{3c} is the number of common columns between $\tilde{\mathbf{Z}}_3$ and $\tilde{\mathbf{Z}}_1$ and q_{3u} is the number of columns unique to $\tilde{\mathbf{Z}}_3$. In this we reordered the columns of $\tilde{\mathbf{Z}}_3$ and $\tilde{\mathbf{Z}}_1$ so that the common columns appear first.

Next because $\tilde{\mathbf{V}}_1$ is of full rank and square (thus invertible) we can write,

$$\tilde{\mathbf{V}} = \tilde{\mathbf{V}}_1\left(I_{2(2^{k-m})\times 2(2^{k-m})} + \left(\tilde{\mathbf{V}}_1^T\tilde{\mathbf{V}}_1\right)^{-1}\tilde{\mathbf{V}}_1^T\tilde{\mathbf{V}}_2\right).$$

Using the formula for the first q_3 rows of $\left(\tilde{\mathbf{V}}_1^T\tilde{\mathbf{V}}_1\right)^{-1}$ we find that the term in parentheses is proportional to a matrix of the form

$$\begin{bmatrix} I_{q_3\times q_3} & 0 \\ \mathbf{U} & I_{2(2^{k-m})-q_3\times 2(2^{k-m})-q_3} \end{bmatrix}$$

where $\mathbf{U}$ is of dimension $2(2^{k-m})-q_3\times q_3$. This square matrix has nonzero determinant and is thus invertible. $\tilde{\mathbf{V}}$ is invertible.

Efficiency: First we derive the semiparametric efficient score. Then we discuss how this score can be used to produce a locally semiparametric efficient estimator of γ'. Lastly we prove that in the saturated model the estimators are semiparametric efficient and that under an additional assumption on the variances (this assumption is provided in Lemma 2) certain unsaturated models will also yield semiparametric efficient estimators.

Semiparametric Efficient Score: The semiparametric model is given by

$$Y = \mathbf{Z}(\eta)^T\gamma' + \epsilon$$

where $E[\epsilon|A_1, R, A_2] = 0$, R has a Bernoulli distribution with success rate, $E[R|A_1, A_2] = E[R|A_1] = \mathbf{U}_1^T\eta$ and γ' belongs R^{q_γ} ($q_\gamma \leq 2\left(2^{k-m}\right)$) and η belongs to R^{q_η} (q_η is equal to the number of rows in the 2^{k-m} design corresponding to unique levels of the stage 1 factors). Put $q = q_\gamma + q_\eta$. Recall that $\mathbf{Z}(\eta) = \left[(R - \mathbf{U}_1^T\eta)\mathbf{Z}_3^T, \mathbf{Z}_1^T, R\mathbf{Z}_2^{(1)^T}, (1-R)\mathbf{Z}_2^{(0)^T}\right]$ and $\gamma' = [(\psi')^T,\ (\phi')^T,\ (\beta')^T,\ (\alpha')^T]^T$. The density for a single observation is given by

$$f_\epsilon(Y - \mathbf{Z}(\eta)^T\gamma'|A_1, R, A_2)(\mathbf{U}_1^T\eta)^R(1-\mathbf{U}_1^T\eta)^{1-R}\prod_{i=1}^{2^{k-m}} p_i^{1\{(A_1,A_2)=(a_{1i},a_{2i})\}}.$$

The nuisance parameters are the densities $f_\epsilon(\cdot|a_1, r, a_2)$ where each is a mean zero density with respect to Lebesgue measure and the p_i's. Actually we know that $p_i = 1/2^{k-m}$, however the efficient score is the same whether we know the p_i's or not. This is because $\{A_1, A_2\}$ is ancillary (Newey, 1990). All of the expectations in this subsection are with respect to this density. As is customary, to derive the efficient score, we make regularity and smoothness conditions on the likelihood of the semiparametric model; these conditions are given in definition A.1 of the appendix in Newey (1990). This model is only slightly different from the semiparametric restricted moment model treated at great length by Tsiatis (Ch. 4, 2006); in particular it differs since a parameter in the Bernoulli distribution of R (e.g. η) also appears in $\mathbf{Z}(\eta)$. In the following we use Tsiatis's (2006) terminology and where possible use his results.

To derive the efficient score function for $\{\gamma', \eta\}$ we calculate the score vector for $\{\gamma', \eta\}$ and then form the projection of this score vector on the nuisance tangent space (see Theorem 4.1 of Tsiatis (2006)). The residual is the efficient score. Let $\mathcal{H}$ is the Hilbert space of q-dimensional, mean zero, finite variance functions of $\{A_1, R, A_2, Y\}$ equipped with the inner product, $< h_1, h_2 >= E[h_1^T h_2]$. The tangent space for a particular parameter is the subset of $\mathcal{H}$ formed by mean square closure of all functions that are score functions for the parameter in a parametric submodel containing the true value of that parameter. Define

$$\begin{aligned}
\Lambda_1 &= \{h_1(\epsilon, a_1, r, a_2) \in \mathcal{H} : E[h_1(\epsilon, A_1, R, A_2)|A_1, R, A_2] = 0\} \\
\Lambda_2 &= \{h_2(\epsilon, a_1, r, a_2) \in \mathcal{H} : E[\epsilon h_2(\epsilon, A_1, R, A_2)|A_1, R, A_2] = 0\} \\
\Lambda_3 &= \{h_3(a_1, r, a_2) \in \mathcal{H} : E[h_3(A_1, R, A_2)|A_1, A_2] = 0\} \\
\Lambda_4 &= \{h_4(a_1, a_2) \in \mathcal{H} : E[h_3(A_1, A_2)] = 0\}
\end{aligned}$$

From Tsiatis (Section 4.5, 2006) we have that the tangent space for the densities $f_\epsilon(\epsilon, a_1, r, a_2)$ is $\Lambda_1 \cap \Lambda_2$ and that the tangent space for the distribution of $\{A_1, A_2\}$ (e.g. the p_i's) is given by Λ_4. Because the p_i's are variationally independent of each of the $f_\epsilon(\epsilon|a_1, r, a_2)$ densities the nuisance tangent space is given by the direct sum of two orthogonal spaces,

$$\Lambda = (\Lambda_1 \cap \Lambda_2) \oplus \Lambda_4.$$

Second, Tsiatis (Section 4.5, 2006) proves that

$$(\Lambda_1 \cap \Lambda_2) \oplus \Lambda_3 \oplus \Lambda_4 = \Lambda_2.$$

Since $\mathcal{H} = \Lambda_2 \oplus \Lambda_2^\perp$, we have that the nuisance tangent space, $\Lambda = (\Lambda_1 \cap \Lambda_2) \oplus \Lambda_4$ is the subset of Λ_2 which is orthogonal to Λ_3, that is, $\Lambda = \Lambda_2 \cap \Lambda_3^\perp$. It follows that $\Lambda^\perp$ is the direct sum of $\Lambda_2^\perp$ and the part of Λ_2 that is contained in Λ_3. Since $\Lambda_3 \subset \Lambda_2$ we have $\Lambda^\perp = \Lambda_2^\perp \oplus \Lambda_3$.

Recall the efficient score is the residual of the projection of the score vector for $\{\gamma', \eta\}$ onto the nuisance tangent space, Λ. Equivalently the efficient score is the projection of the score vector for $\{\gamma', \eta\}$ onto $\Lambda^\perp$. In this case the latter is easier to compute. Tsiatis (Section 4.5, 2006) proves that

$$\Lambda_2^\perp = \{\epsilon g(a_1, r, a_2)\ \forall\ g \text{ arbitrary } q\text{-dimensional functions of } \{a_1, r, a_2\}\}.$$

Because R is a binary random variable, members of Λ_3 also have a simple form: $h_3(A_1, R, A_2) = g'(A_1, A_2)(R - E[R|A_1, A_2])$ for g', a q-dimension function of $\{A_1, A_2\}$.

We have that the efficient score must be of the form, $\epsilon g(A_1, R, A_2) + g'(A_1, A_2)(R - E[R|A_1, A_2])$ where both g and g' are q-dimensional. Assume that $E[\epsilon^2|A_1, R, A_2]$ is positive. Then it is easy to see that the projection of an $h(\epsilon, A_1, R, A_2) \in \mathcal{H}$ onto $\Lambda_2^\perp$ is

$$\Pi\left(h(\epsilon, A_1, R, A_2)|\Lambda_2^\perp\right) = E[h\epsilon|A_1, R, A_2]\frac{\epsilon}{E[\epsilon^2|A_1, R, A_2]}.$$

Similarly, assuming that $E[R|A_1, A_2] \in (0,1)$,

$$\Pi\left(h(\epsilon, A_1, R, A_2)|\Lambda_3\right) = E[h(R - E[R|A_1, A_2])|A_1, A_2] \frac{R - E[R|A_1, A_2]}{E[R|A_1, A_2](1 - E[R|A_1, A_2])}.$$

If we denote the scores for γ', η, by $S_{\gamma'}$, S_η, respectively we see need only calculate $E[S_{\gamma'}\epsilon|A_1, R, A_2]$, $E[S_{\gamma'}(R - E[R|A_1, A_2])|A_1, A_2]$ and similar quantities for S_η. Following Tsiatis (2006) we do this indirectly without explicitly calculating $S_{\gamma'}$ and S_η. We have

$$\begin{aligned}
1 &= \int f_\epsilon(y - z(\eta)^T\gamma'|a_1, r, a_2)\, dy \\
0 &= \int (y - z(\eta)^T\gamma') f_\epsilon(y - z(\eta)^T\gamma'|a_1, r, a_2)\, dy \\
0 &= \sum_{r=0,1} (r - u_1^T\eta) \left(u_1^T\eta\right)^r \left(1 - u_1^T\eta\right)^{1-r}
\end{aligned}$$

for all values of $a_1, r, a_2, \eta, \gamma'$ (recall u and z are functions of these). Differentiating both sides of the first two equations with respect to γ' we obtain

$$\begin{aligned}
0 &= \int S_{\gamma'}(y - z(\eta)^T\gamma', a_1, r, a_2) f_\epsilon(y - z(\eta)^T\gamma'|a_1, r, a_2)\, dy \\
0 &= -z(\eta) + \int (y - z(\eta)^T\gamma') S_{\gamma'}(y - z(\eta)^T\gamma', a_1, r, a_2) f_\epsilon(y - z(\eta)^T\gamma'|a_1, r, a_2)\, dy.
\end{aligned}$$

That is, $0 = E[S_{\gamma'}(\epsilon, A_1, R, A_2)|A_1, R, A_2]$ and $\mathbf{Z}(\eta) = E[\epsilon S_{\gamma'}(\epsilon, A_1, R, A_2)|A_1, R, A_2]$. The former implies that $E[S_{\gamma'}(\epsilon, A_1, R, A_2)(R - \eta^T\mathbf{U}_1)|A_1, A_2] = 0$. Differentiating both sides of each of the last two equations with respect to η results in:

$$\begin{aligned}
0 &= (z_3^T\psi')u_1 + \int (y - z(\eta)^T\gamma') S_\eta(y - z(\eta)^T\gamma', a_1, r, a_2) f_\epsilon(y - z(\eta)^T\gamma'|a_1, r, a_2)\, dy \\
0 &= -u_1 + \sum_{r=0,1} (r - u_1^T\eta) S_\eta(a_1, r, a_2) \left(u_1^T\eta\right)^r \left(1 - u_1^T\eta\right)^{1-r}.
\end{aligned}$$

That is, $-(\mathbf{Z}_3^T\psi')\mathbf{U}_1 = E[\epsilon S_\eta(\epsilon, A_1, R, A_2)|A_1, R, A_2]$ and
$\mathbf{U}_1 = E[(R - \mathbf{U}_1^T\eta)S_\eta(A_1, R, A_2)|A_1, A_2]$. Putting these results together we obtain the efficient score for (γ', η):

$$S_{eff}(Y, A_1, R, A_2; \gamma', \eta, \sigma) = \begin{pmatrix} \mathbf{Z}(\eta)\frac{Y - \mathbf{Z}(\eta)^T\gamma'}{\sigma^2_{A_1,R,A_2}} \\ -(\mathbf{Z}_3^T\psi')\mathbf{U}_1\frac{Y - \mathbf{Z}(\eta)^T\gamma'}{\sigma^2_{A_1,R,A_2}} + \mathbf{U}_1\frac{R - \mathbf{U}_1^T\eta}{\mathbf{U}_1^T\eta(1 - \mathbf{U}_1^T\eta)} \end{pmatrix}$$

where $\sigma^2_{A_1,R,A_2} = E[(Y - \mathbf{Z}(\eta)^T\gamma')^2|A_1, R, A_2]$. The variance-covariance matrix of S_{eff} ($\Sigma_{eff} = E\left[S_{eff}(Y, A_1, R, A_2; \gamma', \eta, \sigma)S_{eff}(Y, A_1, R, A_2; \gamma', \eta, \sigma)\right]$) is given by

$$\begin{pmatrix} E\left[\mathbf{Z}(\eta)\sigma^{-2}_{A_1,R,A_2}\mathbf{Z}(\eta)^T\right] & -E\left[\mathbf{Z}(\eta)\sigma^{-2}_{A_1,R,A_2}(\mathbf{Z}_3^T\psi')\mathbf{U}_1^T\right] \\ -E\left[\mathbf{U}_1(\mathbf{Z}_3^T\psi')\sigma^{-2}_{A_1,R,A_2}\mathbf{Z}(\eta)^T\right] & E\left[\mathbf{U}_1\left((\mathbf{U}_1^T\eta(1 - \mathbf{U}_1^T\eta))^{-1} + (\mathbf{Z}_3^T\psi')^2\sigma^{-2}_{A_1,R,A_2}\right)\mathbf{U}_1^T\right] \end{pmatrix}$$

Although we did not show this above, the smoothness assumptions imply that the estimators $\{\hat{\gamma}', \hat{\eta}\}$ are regular, asymptotically linear estimators (see Tsiatis (2006) for definitions). If Σ_{eff} were singular then no regular, asymptotically linear estimators can exist (Chamberlain, 1986). Thus Σ_{eff} must be invertible. Furthermore for any regular and asymptotically linear estimator, say $((\tilde{\gamma}')^T, \tilde{\eta}^T)$, and a q-dimensional vector, a, we have that a lower bound on the asymptotic variance of $\sqrt{N}\left(((\tilde{\gamma}')^T, \tilde{\eta}^T) - ((\gamma')^T, \eta^T)\right) a$ is $a^T \Sigma_{eff}^{-1} a$.

A Semiparametric Efficient Estimator: To construct a semiparametric efficient estimator we employ the efficient score, S_{eff} as follows. As before we assume that $P[0 < E[R|A_1] < 1] = 1$. We also assume that $P[Var(Y|A_1, R, A_2) > 0] = 1$. First use the method described in the body of the paper (and above in this Appendix) to produce consistent estimators $\hat{\gamma}'$, $\hat{\eta}$. Next set $\hat{\sigma}^2_{A_1,1,A_2}$ ($\hat{\sigma}^2_{A_1,0,A_2}$) to be the sample variance of the residual $(Y - \mathbf{Z}(\hat{\eta})^T \hat{\gamma}')$ for responders (respectively, non-responders) for each group (each row) of the design. Solve

$$\begin{aligned} 0 &= \mathbf{E}_N \left[\mathbf{Z}(\hat{\eta}) \frac{Y - \mathbf{Z}(\eta)^T \gamma'}{\hat{\sigma}^2_{A_1,R,A_2}} \right] \\ 0 &= \mathbf{E}_N \left[-(\mathbf{Z}_3^T \hat{\psi}') \mathbf{U}_1 \frac{Y - \mathbf{Z}(\eta)^T \hat{\gamma}'}{\hat{\sigma}^2_{A_1,R,A_2}} + \mathbf{U}_1 \frac{R - \mathbf{U}_1^T \eta}{\mathbf{U}_1^T \hat{\eta}(1 - \mathbf{U}_1^T \hat{\eta})} \right] \qquad (15) \end{aligned}$$

for γ', η to obtain $\hat{\gamma}'_{eff}$, $\hat{\eta}_{eff}$. Standard arguments can be used to show that

$$\begin{aligned} &\Sigma_{eff} \sqrt{N} \left(((\hat{\gamma}'_{eff})^T, \hat{\eta}^T_{eff})^T - ((\gamma')^T, \hat{\eta}^T) \right) \\ &\qquad = \sqrt{N} (\mathbf{E}_N - E) \left[S_{eff}(Y, A_1, R, A_2; \gamma', \eta, \sigma) \right] + o_P(1). \end{aligned}$$

as $N \to \infty$.

Semiparametric Efficiency of $\{\hat{\gamma}', \hat{\eta}\}$*:* In general the model(12) should be saturated; this is because each predictor associated with an effect (or one of its aliases) will occur in two vectors, either in $\mathbf{Z}_1$ and $\mathbf{Z}_3$ or in $\mathbf{Z}_1$ and $\mathbf{Z}_2^{(1)}$ or in $\mathbf{Z}_1$ and $\mathbf{Z}_2^{(0)}$ or in $\mathbf{Z}_2^{(1)}$ and $\mathbf{Z}_2^{(0)}$ (Similarly (9) will generally be saturated as well). An exception to this pattern (and thus the model will not be saturated) occurs if there are unnecessary formal assumptions; that is both assumptions 3a) and 3b) are made concerning the effects associated with a particular column in Lemma 3 rather than just one of these. In this case the "Algorithm for Constructing the Regressors for Use in the Screening Analysis" of Section 4.2 will remove too many predictors from $\mathbf{Z}_3$.

Assuming that no unnecessary formal assumptions are made, the only other reason why (12) (or 9) will not be a saturated model is if there is a stage 2 factor for only responders (or only non-responders) and the design specifies that this factor's levels are crossed with the levels of the remaining factors (e.g. this factor is independent of the remaining factors). In this latter case effects involving this stage 2 factor can only be aliased with other effects involving this stage 2 factor. As a result each predictor associated with effects of this stage 2 factor (or one of its aliases) can only be included once and only in $\mathbf{Z}_2^{(1)}$. However under the assumption that $P[Var(Y|A_1, R, A_2) = RVar(Y|A_1, R, A_2^{(R)})] = 1$ we show that in this non-saturated case, $\{\hat{\gamma}', \hat{\eta}\}$ is locally semiparametric efficient. The adjective, "locally" is used because the semiparametric efficiency holds if this assumption holds and otherwise not.

First we prove that $\{\hat{\gamma}', \hat{\eta}\}$ is semiparametric efficient when (12) is a saturated model. For simplicity suppose there are stage 2 factors for both responders and nonresponders. Assume $P[0 < E[R|A_1] < 1] = 1$ and $P[Var(Y|A_1, R, A_2) > 0] = 1$. This means that the probability of $P[\{\min_{a_1,r,a_2} \hat{\sigma}_{a_1,r,a_2} > 0\} \cup \{\min_{u_1} |u_1^T\hat{\eta}|\,|1 - u_1^T\hat{\eta}| > 0\}] \to 1$ as $N \to \infty$. Thus we assume in the sample the proportion of responders in each of the 2^{k-m} groups is not equal to one or zero. In this case we prove that $\{\hat{\gamma}', \hat{\eta}\} = \{\hat{\gamma}'_{eff}, \hat{\eta}_{eff}\}$.

First since $\mathbf{E}_N\left[\mathbf{U}_1(R - \mathbf{U}_1^T\hat{\eta})\right] = 0$ we have that $\mathbf{E}_N\left[\mathbf{U}_1\frac{R-\mathbf{U}_1^T\hat{\eta}}{\mathbf{U}_1^T\hat{\eta}(1-\mathbf{U}_1^T\hat{\eta})}\right]$ is zero as well. To see this suppose that there are ν unique values of $\mathbf{U}_1$ in the design ($\nu \leq 2^{k-m}$). Then,

$$\begin{aligned}
\mathbf{E}_N\left[\mathbf{U}_1(R - \mathbf{U}_1^T\hat{\eta})\right] &= \sum_{i=1}^{\nu}\mathbf{E}_N\left[1_{\mathbf{U}_1=\mathbf{u}_{1i}}\mathbf{u}_{1i}(R - \mathbf{u}_{1i}^T\hat{\eta})\right] \\
&= \sum_{i=1}^{\nu}\mathbf{E}_N\left[1_{\mathbf{U}_1=\mathbf{u}_{1i}}\right]\mathbf{u}_{1i}\left(\frac{\mathbf{E}_N\left[R1_{\mathbf{U}_1=\mathbf{u}_{1i}}\right]}{\mathbf{E}_N\left[1_{\mathbf{U}_1=\mathbf{u}_{1i}}\right]} - \mathbf{u}_{1i}^T\hat{\eta}\right) \\
&= \tilde{\mathbf{U}}_1^T\mathbf{D}\left(\bar{\mathbf{R}} - \tilde{\mathbf{U}}_1\hat{\eta}\right)
\end{aligned}$$

where $\tilde{\mathbf{U}}_1$ is a $\nu \times \nu$ matrix with columns corresponding to the ν realizations of $\mathbf{U}_1$, $\mathbf{D}$ is a diagonal matrix with ith entry equal to $\mathbf{E}_N\left[R1_{\mathbf{U}_1=\mathbf{u}_{1i}}\right]$ and $\bar{\mathbf{R}}$ is the ν dimensional vector of observed response proportions. Since $\mathbf{E}_N\left[\mathbf{U}_1(R - \mathbf{U}_1^T\hat{\eta})\right] = 0$, $\tilde{\mathbf{U}}_1^T$ and $\mathbf{D}$ are invertible, we have that $\hat{\eta}$ satisfies $\bar{\mathbf{R}} = \tilde{\mathbf{U}}_1\hat{\eta}$. Since we can write $\mathbf{E}_N\left[\mathbf{U}_1\frac{R-\mathbf{U}_1^T\hat{\eta}}{\mathbf{U}_1^T\hat{\eta}(1-\mathbf{U}_1^T\hat{\eta})}\right]$ in a similar matrix formulation (only the diagonal matrix is altered) we see that this formula is equal to zero as well.

Next since $\mathbf{E}_N\left[\mathbf{Z}(\hat{\eta})\left(Y - \mathbf{Z}(\hat{\eta})^T\hat{\gamma}'\right)\right] = 0$, both $\mathbf{E}_N\left[\mathbf{Z}(\hat{\eta})\frac{Y-\mathbf{Z}(\hat{\eta})^T\hat{\gamma}'}{\hat{\sigma}^2_{A_1,R,A_2}}\right]$ and $\mathbf{E}_N\left[-(\mathbf{Z}_3^T\hat{\psi}')\mathbf{U}_1\frac{Y-\mathbf{Z}(\hat{\eta})^T\hat{\gamma}'}{\hat{\sigma}^2_{A_1,R,A_2}}\right]$ will be zero as well. To see this we use a matrix formulation. First let $\mathbf{E}_N[\mathbf{Y}_1]$ be a 2^{k-m} dimensional vector in which the ith entry corresponds to the sample average of YR for the ith group (row in design) divided by the sample average of R in the ith group. Similarly define $\mathbf{E}_N[\mathbf{Y}_0]$ be a 2^{k-m} dimensional vector in which the ith entry is the sample average of $Y(1-R)$ for the ith group (row in design) divided by the sample average of $1 - R$ in the ith group. Define a vector of response proportions of the same dimension, say $\hat{\mathbf{p}}$ (the ith entry corresponds to a $u1i^T\hat{\eta}$. With these definitions we construct an empirical version of $\tilde{\mathbf{V}}$ defined in the proof of Lemma 4. Define

$$\hat{\mathbf{V}} = \begin{bmatrix} \mathbf{D}_{1-\hat{\mathbf{p}}}\tilde{\mathbf{Z}}_3 & \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_2^{(1)} & 0 \\ \mathbf{D}_{-\hat{\mathbf{p}}}\tilde{\mathbf{Z}}_3 & \tilde{\mathbf{Z}}_1 & 0 & \tilde{\mathbf{Z}}_2^{(0)} \end{bmatrix}$$

where the 0's denote conforming matrices with all entries equal to zero and $\tilde{\mathbf{Z}}_1$, $\tilde{\mathbf{Z}}_3$, $\tilde{\mathbf{Z}}_2^{(1)}$, $\tilde{\mathbf{Z}}_2^{(0)}$ were defined in the proof of Lemma 4 above. Since the model is saturated, $\hat{\mathbf{V}}$ is a square matrix of dimension $2\left(2^{k-m}\right)$. Using the same proof as used in Lemma 4 to show that $\tilde{\mathbf{V}}$ is full rank, we can show that $\hat{\mathbf{V}}$ is full rank as well (and hence invertible). Define $\mathbf{D}_1$ to be a diagonal matrix with ith diagonal entry equal to the proportion of responders in the ith group multiplied by the group size. Similarly define $\mathbf{D}_0$ to be a diagonal matrix with ith diagonal entry equal to the proportion of non-responders in the ith group multiplied by the group size.

We can write $0 = \mathbf{E}_N \left[\mathbf{Z}(\hat{\eta}) \left(Y - \mathbf{Z}(\hat{\eta})^T \hat{\gamma}'\right)\right]$ as

$$0 = \hat{\mathbf{V}}^T \begin{pmatrix} \mathbf{D}_1 & 0 \\ 0 & \mathbf{D}_0 \end{pmatrix} \left(\begin{pmatrix} \mathbf{E}_N[\mathbf{Y}_1] \\ \mathbf{E}_N[\mathbf{Y}_0] \end{pmatrix} - \hat{\mathbf{V}}\hat{\gamma}' \right). \tag{16}$$

Thus in the saturated model, $\hat{\gamma}'$

$$0 = \begin{pmatrix} \mathbf{E}_N[\mathbf{Y}_1] \\ \mathbf{E}_N[\mathbf{Y}_0] \end{pmatrix} - \hat{\mathbf{V}}\hat{\gamma}'.$$

Since we can write $\mathbf{E}_N \left[\mathbf{Z}(\hat{\eta}) \frac{Y - \mathbf{Z}(\hat{\eta})^T \hat{\gamma}'}{\hat{\sigma}^2_{A_1,R,A_2}}\right]$ and $\mathbf{E}_N \left[-(\mathbf{Z}_3^T \hat{\psi}')\mathbf{U}_1 \frac{Y - \mathbf{Z}(\hat{\eta})^T \hat{\gamma}'}{\hat{\sigma}^2_{A_1,R,A_2}}\right]$ as matrices times the above difference we have these two terms are zero as well. Combining the above with the results for $\hat{\eta}$, we have that when the model is saturated and the sample is sufficiently large so that the response rates in each of the 2^{k-m} groups are neither zero or one, $\hat{\gamma}' = \hat{\gamma}'_{eff}$ and $\hat{\eta} = \hat{\eta}_{eff}$.

Next we prove that in the non-saturated model, $\{\hat{\gamma}', \hat{\eta}\}$ will be locally semiparametric efficient if 1) no unnecessary formal assumptions are made (assumptions 3a and 3b are made concerning the effects associated with a particular column in Lemma 3 rather than just one of these) and 2) there is a stage 2 factor for responders (or non-responders) that is independent of the remaining factors. Assume that $Var(Y|A_1, R = r, A_2) = Var(Y|A_1, R = r, A_2^{(r)})$ for $r = 0, 1$. Under these assumptions on the residual variance we estimate $\sigma_{A_1,1,A_2}$ by the sample variance of the residual $(Y - \mathbf{Z}(\hat{\eta})^T \hat{\gamma}')$ for responders for each combination of groups (rows) in the design with a unique value of $\{A_1, A_2^{(1)}\}$ and similarly for non-responders. Denote the estimators by $\hat{\sigma}_{A_1,r,A_2^{(r)}}$ for $r = 0, 1$. A locally semiparametric estimator can be found by solving (15) (with $\hat{\sigma}_{A_1,R,A^{(R)}}$ instead of $\hat{\sigma}_{A_1,R,A_2}$).

As in the above both $\mathbf{E}_N \left[\mathbf{U}_1 \frac{R - \mathbf{U}_1^T \hat{\eta}}{\mathbf{U}_1^T \hat{\eta}(1 - \mathbf{U}_1^T \hat{\eta})}\right]$ and $\mathbf{E}_N \left[\mathbf{U}_1(R - \mathbf{U}_1^T \hat{\eta})\right]$ are identically zero. Next consider the first equation in (14) and (15) (in the latter $\hat{\sigma}_{A_1,R,A_2}$ is replaced by $\hat{\sigma}_{A_1,R,A^{(R)}}$).

For clarity, assume there is only one more stage 2 factor for responders than non-responders and that this stage 2 factor is randomized independently of all other factors. This means that while there are 2^{k-m} unique realizations of $\{A_1, A_2^{(1)}\}$ there are only 2^{k-m-1} unique realizations of $\{A_1, A_2^{(0)}\}$. Now (16) continues to hold but $\hat{\mathbf{V}}$ although a full rank matrix, is no longer square. It's dimensions are $2\left(2^{k-m}\right) \times \left(2^{k-m} + 2^{k-m-1}\right)$. A closer inspection of $\hat{\mathbf{V}}$ reveals that in the lower 2^{k-m} rows there are only 2^{k-m-1} unique rows, each has a one double. Thus if we remove these 2^{k-m-1} doubles from $\hat{\mathbf{V}}$ (say to form $\hat{\mathbf{V}}'$) we now have a square matrix which is full rank. We can re-express (16) as follows. Let $\mathbf{D}_0$ and $\mathbf{E}_N[\mathbf{Y}_0]$ be defined as before. The latter vector is of dimension 2^{k-m}. Define $\mathbf{E}_N[\mathbf{Y}_0]'$ to be 2^{k-m-1} dimension with the ith entry corresponds to the sample average of $Y(1 - R)$ for the ith unique realization of $\{A_1, A_2^{(0)}\}$ divided by the sample average of R in the ith realization. Define $\mathbf{D}_0'$ to be a diagonal matrix with ith diagonal entry equal to proportion of responders in the group formed by the ith unique realization of $\{A_1, A_2^{(0)}\}$ multiplied by the group size. Now we can re-express (16) as

$$0 = \left(\hat{\mathbf{V}}'\right)^T \begin{pmatrix} \mathbf{D}_1 & 0 \\ 0 & \mathbf{D}_0' \end{pmatrix} \left(\begin{pmatrix} \mathbf{E}_N[\mathbf{Y}_1] \\ \mathbf{E}_N[\mathbf{Y}_0]' \end{pmatrix} - \hat{\mathbf{V}}'\hat{\gamma}' \right).$$

As before the invertibility of $\hat{\mathbf{V}}'$ implies that the last term in parentheses is equal to zero. Because we can express both

$$\mathbf{E}_N\left[\mathbf{Z}(\hat{\eta})\frac{Y-\mathbf{Z}(\hat{\eta})^T\hat{\gamma}'}{\hat{\sigma}^2_{A_1,R,A_2^{(R)}}}\right] \text{ and } \mathbf{E}_N\left[-(\mathbf{Z}_3^T\hat{\psi}')\mathbf{U}_1\frac{Y-\mathbf{Z}(\hat{\eta})^T\hat{\gamma}'}{\hat{\sigma}^2_{A_1,R,A_2^{(R)}}}\right]$$

as matrices times the above difference we have these two terms are zero as well. Note we would not have been able to do this had we used the estimator $\hat{\sigma}^2_{A_1,R,A_2^{(R)}}$ instead of $\hat{\sigma}^2_{A_1,R,A_2}$ in (15). Combining the above with the results for $\hat{\eta}$, we have that when the sample is sufficiently large so that the response rates in each of the groups defined by unique values of $\{A_1, A_2^{(1)}\}$ and $\{A_1, A_2^{(0)}\}$ are neither zero or one then $\hat{\gamma}' = \hat{\gamma}'_{eff}$ and $\hat{\eta} = \hat{\eta}_{eff}$.

A.2 Aliasing and The Formal Assumptions

Here we provide a simple example of how the aliasing becomes difficult to interpret when the formal assumptions do not hold. Suppose there are two factors at stage 1 and one stage 2 factor for responders. We are interested in the main effects of these factors and we are willing to make the working assumption that there are no interactions. Suppose we are willing to make the formal assumption that all effects of R (including interactions between R and the stage 1 factors) are negligible. We use the experimental design in Table 2 with defining word $1 = A_{11}A_{12}A_2$ (A_2 is $A_2^{(0)}$ here).

The conditional mean of Y given A_1, R, A_2 is given by (12). Suppose that it is known that $\psi_j = 0$, $j = 2, 3, 4$. However we do not know whether any of the remaining parameters are zero or not. The experimental design in Table 2 permits the identification of the η parameters in $p(A_1) = \eta_1 + \eta_2 A_{11} + \eta_3 A_{12} + \eta_4 A_{11}A_{12}$ and the eight conditional means $E[Y|A_{11} = a_{11}, A_{12} = a_{12}, R = r, A_2 = a_{11}a_{12}]$ (recall that $A_2 = A_{11}A_{12}$ in this design). The conditional means are

$$\begin{aligned} E[Y|A_{11} = a_{11}, A_{12} = a_{12}, R = r, A_2 = a_{11}a_{12}] = \\ (\phi_1+\psi_1) + (\phi_2+\eta_2\psi_1)a_{11} \\ +(\phi_3+\eta_3\psi_1)a_{12} + (\phi_4+\eta_4\psi_1)a_{11}a_{12} \\ +(\psi_1+\beta_1 a_{11}a_{12}+\beta_2 a_{12}+\beta_3 a_{11}+\beta_4)r \end{aligned}$$

for $a_{11} \in \{-1, 1\}$, $a_{12} \in \{-1, 1\}$ and $r \in \{0, 1\}$. So from data resulting from the experimental design in Table 2 we can identify the eight quantities such as: $\phi_1 + \psi_1$, $\phi_2 + \eta_2\psi_1$, $\phi_3 + \eta_3\psi_1$, $\phi_4 + \eta_4\psi_1$, $\psi_1 + \beta_4$, β_1, β_2 and β_3. The above nonlinearity in parameters makes it difficult to interpret the identifiable quantities involving the ϕ_j's. More complex designs result in similarly uninterpretable quantities. In summary, to maintain interpretability, we consider experimental designs that *do not* alias a potentially active stage 2 effect with a potentially active nuisance effect.