Published in final edited form as:

J R Stat Soc Series B Stat Methodol. 2007 November 1; 69(5): 879–901. doi:10.1111/j.
1467-9868.2007.00615.x.

Semiparametric estimation of treatment effects given base-line covariates on an outcome measured after a post-randomization event occurs

Yannis Jemai,
Cytel Inc., Cambridge, USA

Andrea Rotnitzky,
DiTella University, Buenos Aires, Argentina, and Harvard School of Public Health, Boston, USA

Bryan E. Shepherd, and
Vanderbilt University, Nashville, USA

Peter B. Gilbert
University of Washington, Seattle, USA

Summary

We consider estimation, from a double-blind randomized trial, of treatment effect within levels of base-line covariates on an outcome that is measured after a post-treatment event $\mathbf{E}$ has occurred in the subpopulation $\mathcal{P}_{\mathbf{E},\mathbf{E}}$ that would experience event $\mathbf{E}$ regardless of treatment. Specifically, we consider estimation of the parameters γ indexing models for the outcome mean conditional on treatment and base-line covariates in the subpopulation $\mathcal{P}_{\mathbf{E},\mathbf{E}}$. Such parameters are not identified from randomized trial data but become identified if additionally it is assumed that the subpopulation $\mathcal{P}_{\mathbf{E},\mathbf{E}}$ of subjects that would experience event $\mathbf{E}$ under the second treatment but not under the first is empty and a parametric model for the conditional probability that a subject experiences event $\mathbf{E}$ if assigned to the first treatment given that the subject would experience the event if assigned to the second treatment, his or her outcome under the second treatment and his or her pretreatment covariates. We develop a class of estimating equations whose solutions comprise, up to asymptotic equivalence, all consistent and asymptotically normal estimators of γ under these two assumptions. In addition, we derive a locally semiparametric efficient estimator of γ . We apply our methods to estimate the effect on mean viral load of vaccine *versus* placebo after infection with human immunodeficiency virus (the event $\mathbf{E}$) in a placebo-controlled randomized acquired immune deficiency syndrome vaccine trial.

Keywords

Counterfactuals; Missing data; Potential outcomes; Principal stratification; Structural model; Vaccine trials

1. Introduction

In this paper we consider the problem of estimating, from double-blind randomized trials, the effect of a treatment on an outcome that is measured after a certain post-treatment event $\mathbf{E}$ has occurred. Our work is motivated by the need to develop adequate methodology for addressing the important question in human immunodeficiency virus (HIV) vaccine research of whether exposure to an imperfect vaccine (i.e. one that can prevent infection in some but not all recipients) has an effect on the progression of disease after infection. In this context, the relevant question is whether exposure to vaccine prevents or delays the onset of acquired immune deficiency syndrome in those in which it fails to prevent infection. In particular, the goal is to compare post-infection outcomes in the subpopulation $\mathcal{P}_{\mathbf{EE}}$ of individuals for whom, in a placebo-controlled vaccine trial, infection (the event $\mathbf{E}$ in this case) would occur during the course of the trial regardless of treatment assignment. This differs from the goal of comparing the direct effects of two regimens controlling for the occurrence of the event $\mathbf{E}$. Direct effect contrasts measure the effect of treatment in the hypothetical scenario in which the event $\mathbf{E}$ is controlled for, i.e. it is forced to occur, in all subjects. In the context of vaccine studies and with $\mathbf{E}$ denoting infection, such a hypothetical scenario is of no scientific interest.

Counterfactual, also referred to as potential, outcomes (Rubin, 1978; Robins, 1986) are useful tools for constructing causal contrasts measuring treatment effects. Our problem is peculiar in that, although in principle it may be possible to conceptualize counterfactual outcomes for all subjects in the population, such outcomes do not hold any relevant meaning except in the subpopulation $\mathcal{P}_{\mathbf{EE}}$. For example, counterfactual markers of disease progression after infection, such as viral load 3 months after infection, could in principle be defined for everyone (by envisioning that infection could somehow be forced by intervention in everyone). However, they are irrelevant variables except for the subpopulation in which infection cannot be prevented as this is the pertinent analysis subpopulation. Another setting in which a similar situation arises is when interest lies in causal contrasts that are defined in terms of outcomes that can be censored by death. For example, in a cancer trial, we may be interested in determining which of two competing chemotherapy treatments would result in better quality of life in those who would survive a fixed period of time after the end of treatment.

The importance of estimation of quantities that are defined only in the subpopulation in which outcomes hold meaning was first discussed in the context of outcomes that are censored by death by Kalbfleisch and Prentice (1980). As far as we know, Robins (1986), remark 12.2, is the earliest reference that considers inference about causal effects in the $\mathcal{P}_{\mathbf{EE}}$ -subpopulation of subjects who would survive under either treatment. Robins (1986) considered time-to-event outcomes but its formulation applies likewise to outcomes that are measured at a prespecified time point. Later Robins (1995) provided a set of strong untestable assumptions under which the sharp null hypothesis of no causal effect in the $\mathcal{P}_{\mathbf{EE}}$ -subpopulation is testable. Rubin (1998), section 6, and Rubin (2006) also considered censoring by death and proposed focusing the inference on the causal contrast in the $\mathcal{P}_{\mathbf{EE}}$ -subpopulation. This idea was revisited in the discussion of Scharfstein *et al.* (1999) by Little and Rubin and in the authors' rejoinder. Later, Robins and Greenland (2000) and Rubin (2000), in their separate discussions of a paper by Dawid (2000), reiterated the idea, using counterfactual notation to define the causal contrasts. Finally, Frangakis and Rubin (2002) coined the term 'principal strata' for the subpopulations that are defined by the occurrence or non-occurrence of one or many general post-treatment events and discussed various settings in which causal contrasts in principal strata would be the relevant estimands of scientific interest.

Because causal contrasts in the subpopulation $\mathcal{P}_{\mathbf{E},\mathbf{E}}$ are not identified even from randomized trial data, two approaches have been considered in the literature to address the non-identifiability problem. The first approach considers inference about sharp bounds for the

causal contrasts (Hudgens *et al.*, 2003; Jemiai and Rotnitzky, 2003; Zhang and Rubin, 2003). Sharp bounds determine the range of all possible values of the causal contrast that are compatible with the observed data distribution. This approach has the advantage of providing a most objective assessment of the treatment effect in the light of the available information. However, it is uninformative about what conditions might give rise to each value in the range.

The second approach considers assumptions under which the causal contrasts are identified from randomized trial data. These include postulating values for easily interpretable features of the distribution of the counterfactuals. The causal contrasts are then estimated regarding the features' value as fixed and known, and the estimation procedure is repeated under various such values within a plausible range in a form of sensitivity analysis (Gilbert *et al.*, 2003; Hayden *et al.*, 2005). In particular, Gilbert *et al.* (2003) described an estimator of the mean difference of the treatment-specific counterfactual outcomes in the subpopulation $\mathcal{P}_{E,E}$ under the following identifying assumptions:

- a. the subpopulation $\mathcal{P}_{E,E}$ of subjects who would experience the event under the second treatment, say **B**, but not under the first treatment, say **A**, is empty (this assumption is often referred to as monotonicity; Angrist *et al.* (1996), Gilbert *et al.* (2003) and Zhang and Rubin (2003))
- b. the distributions of the treatment A counterfactual outcome in $\mathcal{P}_{E,E}$ and in the subpopulation $\mathcal{P}_{E,\bar{E}}$ of subjects that would experience the event under A but not under B are related by a specified tilting transformation, which is indexed by a function g .

Gilbert *et al.* (2003) recommended repeating the estimation each time regarding a different tilting transformation as known, as a form of sensitivity analysis.

One attractive feature of the approach of Gilbert *et al.* (2003) is that inference about treatment effects is conducted under a flexible semiparametric model that makes no distributional shape assumptions and, as such, is protected from the possibility of distributional shape misspecifications. However, Gilbert *et al.* (2003) did not consider the estimation of treatment effects conditional on high dimensional base-line covariates. Yet, these conditional effects are important to understand whether and how the effect of treatment varies across levels of base-line covariates. In this paper, we develop extensions of the semiparametric approach of Gilbert *et al.* (2003) that allow the estimation of treatment effects conditional on covariates. For this, we introduce a new model, the $\mathcal{P}_{E,E}$ -marginal structural mean model, which postulates that the conditional mean of the treatment-specific counterfactual outcome given base-line covariates X in the subpopulation $\mathcal{P}_{E,E}$ is known up to a finite dimensional parameter γ . We consider inference about γ assuming that monotonicity and that condition (b) above hold conditionally on covariates. We prove that, even though γ is identified under our assumptions, it cannot be well estimated with the moderate sample sizes that are found in practice when X is high dimensional, except in the trivial case in which the distributions of the counterfactual outcomes in the populations $\mathcal{P}_{E,E}$ and $\mathcal{P}_{E,\bar{E}}$ are assumed to be the same (under this assumption, ordinary regression restricted to those experiencing the event E yields valid inferences). The source of this difficulty is that consistent estimation of γ requires the additional estimation of an unknown nuisance function of X , which cannot be well estimated under the model owing to the curse of dimensionality. We show that well-behaved estimators of γ do exist, if we additionally assume that

- (c) a parametric model for the conditional probability that a subject experiences the event **E** if assigned to treatment A given that the subject would experience the event if assigned to treatment B, his or her outcome under B and his or her pretreatment covariates.

We derive a locally efficient estimator of γ in the semiparametric model that is defined by the $\mathcal{P}_{\mathbf{E}, \mathbf{E}}$ -marginal structural mean model and assumptions (a)–(c), which is guaranteed to achieve the semiparametric variance bound whenever a working parametric model for the distribution of the treatment $\mathbf{B}$ counterfactual outcome given the covariates in the population $\mathcal{P}_{\mathbf{E}, \mathbf{E}}$ is correctly specified. This work is a sequel to Shepherd *et al.* (2006) who considered estimation under the more restrictive fully parametric model, that on top of assumptions (a)–(c) assumes parametric models for the conditional distributions of the treatment-specific counterfactual outcomes given covariates in the subpopulation $\mathcal{P}_{\mathbf{E}, \mathbf{E}}$.

2. The set-up

Consider an experiment which randomizes n subjects, who are independently selected from a given population of interest, to one of two treatments. A vector $X_i = (X_{i1}, \dots, X_{ik_i})'$ of pre-randomization covariates and the treatment indicator Z_i are recorded on each individual i . In addition, a continuous or discrete outcome of interest Y_i is to be measured after occurrence of the event $\mathbf{E}$. Thus, Y_i is observable only if an event $\mathbf{E}$ has occurred for subject i . Let S_i denote the binary indicator that the event $\mathbf{E}$ has occurred for subject i (i.e. $S_i = 1$ if the event has occurred and $S_i = 0$ otherwise). We assume that the observed data $O_i = (Z_i, X_i, S_i, Y_i)$, $i = 1, \dots, n$, are n independent and identically distributed copies of the random vector $O \equiv (Z, X, S, Y)$, where by notational convention we define $Y_i \equiv *$ if $S_i = 0$. In the context of the HIV vaccine trial that was described in Section 1, $Z = 1$ and $Z = 0$ indicate assignment to vaccine and placebo respectively, $\mathbf{E}$ is HIV infection, $S = 1$ and $S = 0$ indicate infection and non-infection respectively during the trial, Y is viral load measured at a prespecified post-infection time and X includes base-line covariates such as host genetics, behavioural risk factors and age.

To define the treatment effect of interest and our model, we use counterfactual random variables (Neyman, 1990; Rubin, 1978; Robins, 1986). Specifically, for each $z \in \{0, 1\}$, define $S_i(z)$ and $Y_i(z)$ to be the event indicator and outcome for subject i if, possibly contrary to fact, he or she had been assigned to treatment z , and where, if $S_i(z) = 0$, we define $Y_i(z) \equiv *$. Here, we have implicitly made the stable unit treatment value assumption (Rubin (1978, 1986) and Cox (1958), page 19), which states that a subject i 's counterfactuals are independent of the other subjects' treatment assignments. The stable unit treatment value assumption implies the consistency assumption that the subject's observed outcome is equal to the counterfactual outcome under the treatment that is actually assigned to him or her, i.e.

$$(S_i(z), Y_i(z)) = (S_i, Y_i) \quad \text{if } Z_i = z. \quad (1)$$

We assume that the vectors $W_i = (Z_i, X_i, S_i(0), S_i(1), Y_i(0), Y_i(1))$, $i = 1, \dots, n$, are n independent and identically distributed copies of the random vector $W \equiv (Z, X, S(0), S(1), Y(0), Y(1))$. Randomization, possibly depending on the base-line covariates X , guarantees that

$$Z \perp\!\!\!\perp (S(0), S(1), Y(0), Y(1)) \mid X \quad \text{with probability 1,} \quad (2)$$

because counterfactual random variables are unobserved base-line characteristics of each individual. Here, $A \perp\!\!\!\perp B \mid C$ indicates conditional independence of A and B given C (Dawid, 1979).

In the causal inference literature, a causal effect measure on the outcome of interest (say, viral load) is defined as some measure of discrepancy between the distributions of $Y(0)$ and $Y(1)$ in the target population. However, this measure is irrelevant when, as in our context, both $Y(0)$

and $Y(1)$ only hold meaning in the subpopulation $\mathcal{P}_{\mathbf{E},\mathbf{E}}$ of subjects for whom $S(1) = 1$ and $S(0) = 1$. Gilbert *et al.* (2003), Hudgens *et al.* (2003) and Hayden *et al.* (2005), following Robins (1995), Rubin (2000), Robins and Greenland (2000) and Frangakis and Rubin (2002), suggested considering such measures of discrepancy in the subpopulation $\mathcal{P}_{\mathbf{E},\mathbf{E}}$ assuming that this population exists. In particular, Gilbert *et al.* (2003) described estimators of $E\{Y(z)|S(1) = S(0) = 1\}$, $z = 0, 1$, and their difference. In this paper we extend their methods to allow estimation of $m(z, X) \equiv E\{Y(z)|S(1) = S(0) = 1, X\}$, $z = 0, 1$, i.e. of the mean of $Y(z)$ conditional on covariates X in the subpopulation $\mathcal{P}_{\mathbf{E},\mathbf{E}}$.

3. Identification

Randomization is insufficient to identify $m(z, X)$ from the observed data O_i , $i = 1, \dots, n$, because subjects with $S(0) = S(1) = 1$ are never detectable. Specifically, in each arm z , $z = 0, 1$, of a randomized study, we can observe subjects for whom $S(z) = 1$, but we can never observe which of these also have $S(1-z) = 1$. To identify $m(z, X)$ we shall make the following further assumptions.

Assumption 1 (monotonicity). $\Pr\{S(1) = 1, S(0) = 0|X\} = 0$ with probability 1.

Assumption 2 (positivity assumption 1). $\Pr\{S(1) = 0, S(0) = 1|X\} > 0$ with probability 1.

Assumption 3 (positivity assumption 2). $\Pr\{S(1) = 1, S(0) = 1|X\} > 0$ with probability 1.

Assumption 4 (pattern mixture assumption).

$$f\{Y(0)|S(1)=S(0)=1, X\} = \frac{f\{Y(0)|S(1)=0, S(0)=1, X\}}{c(X)} \exp[g\{X, Y(0)\}] \quad (3)$$

where $g(\cdot, \cdot)$ is a known function and, for each $X = x$, $c(x)$ is a normalizing constant.

Assumption 1 implies that the subpopulation $\mathcal{P}_{\mathbf{E},\mathbf{E}}$ is empty. In contrast, assumptions 2 and 3 imply that $\mathcal{P}_{\mathbf{E},\mathbf{E}}$ and $\mathcal{P}_{\mathbf{E},\mathbf{E}}$ respectively are non-empty. In the context of placebo-controlled vaccine trials, $\mathcal{P}_{\mathbf{E},\mathbf{E}}$, $\mathcal{P}_{\mathbf{E},\mathbf{E}}$ and $\mathcal{P}_{\mathbf{E},\mathbf{E}}$ correspond to the subpopulations of *harmful*, *protected* and *always infected* individuals, as defined by Gilbert *et al.* (2003). Under randomization (2), assumption 1 implies the testable restriction $\Pr(S=1|Z=0, X) \geq \Pr(S=1|Z=1, X)$. In the analysis of vaccine trials it might be reasonable to assume *a priori* assumption 1 if the treatments being compared are vaccine ($z=1$) and placebo ($z=0$). However, this assumption would usually not be defensible if the treatments are two experimental vaccines. Under randomization (2) and assumption 1, assumption 2 is also testable as it implies the observed data law restriction $\Pr(S=1|Z=0, X) > \Pr(S=1|Z=1, X)$. In assumption 4, the function g calibrates the discrepancy between the conditional distributions of the outcome under treatment $Z=0$ given covariates X in the subpopulations $\mathcal{P}_{\mathbf{E},\mathbf{E}}$ and $\mathcal{P}_{\mathbf{E},\mathbf{E}}$. In particular $g=0$ establishes that these distributions are identical. In the placebo-controlled vaccine trial setting and with viral load being the outcome, the assumption $g=0$ would hold if, among subjects who are infected if they receive placebo, the covariates X are all the simultaneous predictors of

- a. the potential outcome after infection under placebo and
- b. the potential infection state under vaccine

Note also that assumption 4 carries the implicit assumption that, within levels of X , the support of $Y(0)$ is the same in the subpopulations $\mathcal{P}_{\mathbf{E},\mathbf{E}}$ and $\mathcal{P}_{\mathbf{E},\mathbf{E}}$. In the context of placebo-controlled vaccine trials, this equal support assumption establishes that, if there is an always infected

subject who would attain a given viral load value following infection after receiving placebo, then there also is a protected individual who, under the same circumstances, would attain the same viral load value.

A straightforward application of Bayes rule shows that assumptions 2–4 are equivalent to the existence of a real-valued function $r(\cdot)$ such that, with $\omega(u) \equiv \{1 + \exp(-u)\}^{-1}$

$$\begin{aligned} \Pr\{S(1)=1|S(0)=1, Y(0) \\ , X\} = \omega[r(X) \\ + g\{Y(0), X\}]. \end{aligned} \quad (4)$$

We let $\mathcal{A}(g)$ denote the model for the distribution of W defined by randomization (2) and assumptions 1–4. Shepherd *et al.* (2006) noted that $m(z, X)$ is identified under condition (1) and $\mathcal{A}(g)$ and satisfies

$$\begin{aligned} m(1, X) &= E(Y|S=1, Z=1, X), \\ m(0, X) &= \frac{E[\omega\{r(X)+g(Y, X)\}Y|S=1, Z=0, X]}{E[\omega\{r(X)+g(Y, X)\}|S=1, Z=0, X]} \end{aligned} \quad (5)$$

where, for each x , $r(x)$ is the unique solution of the equation

$$\begin{aligned} E[\omega\{r(X)+g(Y, X)\}|S \\ =1, Z=0, X=x] &= \frac{\Pr(S=1|Z=1, X=x)}{\Pr(S=1|Z=1, X=x)}. \end{aligned} \quad (6)$$

Theorem 1 below establishes that, under condition (1), model $\mathcal{A}(g)$ imposes on the distribution of the observed data the unique restriction

$$\begin{aligned} \Pr(S=1|Z=0, X) &> \Pr(S \\ &=1|Z \\ &=1, X) > 0 \quad \text{with probability 1.} \end{aligned} \quad (7)$$

An important consequence is that, for any g and g' models $\mathcal{A}(g)$ and $\mathcal{A}(g')$ impose the same restrictions on the observed data distribution. Thus, the observed data cannot help to discriminate g from g' , i.e. g is not identified. In settings where it can be *a priori* assumed that assumptions 1–3 hold, these results provide the mathematical justification for an analytical strategy in which inference about $E\{Y(z)|X, S(1)=S(0)=1\}$ is drawn repeatedly under model $\mathcal{A}(g)$, each time regarding a different function g as fixed and known in a form of sensitivity analysis. Expression (4) facilitates the interpretation of the function g and thus may help the investigator to determine its plausible range in each particular application. For example, in the HIV vaccine trial setting, identity (4) implies that, for any fixed x , the function $g(x, \cdot)$ determines whether and how the rate of infection under vaccine, among subjects who would also be infected under placebo and have covariate values x , changes with viral load level under placebo. Interestingly, model $\mathcal{A}(g)$ is like the biased sampling model of Bickel *et al.* (1993), page 113, when their selection probabilities are all equal except that, in their model, the stratum weights ω are known, whereas, under model $\mathcal{A}(g)$ they depend on the unknown function $r(\cdot)$ and must be estimated.

Theorem 1

Suppose that condition (1) holds. Then, for any function g , the observed data laws that are allowed by model $\mathcal{A}(g)$ are those that satisfy restriction (7).

Remark 1

(identification in the absence of covariates). Suppose that assumptions 1–4 are restated without conditioning on X . Then theorem 1 remains valid when X is removed from all conditional statements and $r(\cdot)$ is replaced by a constant α . Indeed, without covariates, conditions 1 and 3 and equation (4) are precisely the assumptions that were made by Gilbert *et al.* (2003), to identify $m(z)$. However, without explicitly stating it, they actually conducted maximum likelihood (ML) estimation of $m(z)$ under a model that allowed for the possibility that α could be ∞ , i.e. that assumption 2 did not hold. Although Gilbert *et al.* (2003) considered only functions $g(y) = \beta y$ for some user-specified β , their ML estimator equally applies to any arbitrary user-specified function $g(\cdot)$

4. Estimation of $m(z, X)$

When X has a finite discrete sample space, a consistent estimator of $m(z, X)$ under model $\mathcal{A}(g)$ may be obtained by using the estimator of $m(z, x)$ that was given by Gilbert *et al.* (2003), within each level x of X . However, if X is continuous and/or high dimensional, these estimators are infeasible, because the data are too sparse to conduct stratified estimation. Furthermore, estimating $m(z, x)$ by using smoothing methods will not be useful in practice because of the curse of dimensionality. For this reason, in this paper we consider estimation of $m(z, X)$ by assuming that

$$m(z, X) = m(z, X; \gamma^*), \gamma^* \in \{0, 1\}, \quad (8)$$

where, for each z and x , $m(z, x; \gamma)$ is a smooth function of a $q \times 1$ vector γ and its true value γ^* is unknown. We call expression (8) a $\mathcal{P}_{\mathbf{E}, \mathbf{E}}$ -marginal structural mean model, because it is a model for the marginal mean of a counterfactual random variable in the subpopulation $\mathcal{P}_{\mathbf{E}, \mathbf{E}}$. In addition, we denote by $\mathcal{B}(g)$ the model that is defined like $\mathcal{A}(g)$ but with the additional restriction (8).

Although γ^* is identified under model $\mathcal{B}(g)$, its estimation remains infeasible when X is high dimensional owing to the curse of dimensionality. Specifically, as implied by theorem 2 below, influence functions of regular asymptotically linear (RAL) estimators of γ^* in model $\mathcal{B}(g)$ typically depend on the unknown function $r(\cdot)$ that is defined by equation (6) and do not have mean 0 when evaluated at a misspecified function $r(\cdot)$ (The only exceptions are models $\mathcal{B}(g)$ where $m(z = 1, X)$ is determined by the parameters indexing the model for $m(z = 0, X)$, but these are of little or no interest in studies that are aimed at investigating treatment effects, because they *a priori* assume the relationship between $m(z = 1, X)$ and $m(z = 0, X)$.) The implication is that estimation of γ^* under model $\mathcal{B}(g)$ would require the preliminary consistent estimation of the function $r(\cdot)$ by using smoothing techniques. However, when X is a vector with two or more continuous components, this function would not be well estimated given the moderate sample sizes that are found in practice, essentially because no two subjects would have values of X that are sufficiently near to allow the borrowing of information that is necessary for smoothing. Thus, in practice, we are forced to conduct inference under a reduced model for $r(\cdot)$. In this paper, we consider inference under the assumption that $r(\cdot)$ follows a parametric model

$$r(x) = r(x; \alpha^*) \quad (9)$$

where, for each x , $r(x; \alpha)$ is a smooth function of a $P \times 1$ vector α and its true value α^* is unknown. Denote by $\mathcal{C}(g)$ the model that is defined like $\mathcal{B}(g)$ but with the additional restriction (9). The following theorem establishes the restrictions that are placed on the observed data law by $\mathcal{C}(g)$. We subsequently discuss its implications for inference about γ^* .

Theorem 2

Under condition (1), the observed data laws that are allowed by $\mathcal{C}(g)$ are those satisfying the restrictions $E\{q_1(O; \alpha^*, \gamma^*)|X\} = 0$ (restriction 1) and $E[S \omega\{r(X; \alpha^*) + g(Y, X)\}^{1-Z}|Z, X]$ (restriction 2) does not depend on Z , where

$$q_1(O; \alpha, \gamma) \equiv \begin{pmatrix} SZ\{Y - m(Z, X; \gamma)\} \\ S(1 - Z)\omega\{r(X; \alpha) + g(Y, X)\}\{Y - m(Z, X; \gamma)\} \end{pmatrix}.$$

Remark 2

Consider model $\mathcal{H}$, which is defined like model $\mathcal{C}(g)$, except that g is assumed to follow a parametric model that is indexed by an unknown parameter δ . It is easily shown that theorem 2 remains valid if $\mathcal{C}(g)$, $g(X, Y)$ and $q_1(O; \alpha, \gamma)$ are replaced by $\mathcal{H}$, $g(X, Y; \delta)$ and $q_1(O; \alpha, \gamma, \delta)$ respectively. This, in turn, implies that, if the dimensions of (δ, α, γ) are not too high, then δ will be identified. Consequently, one could estimate δ consistently. However, since δ is only identified because of models that are imposed to alleviate the curse of dimensionality, we take the philosophical stance that model $\mathcal{H}$ should not be considered for inference. Rather, we should adopt model $\mathcal{C}(g)$ (or, better yet, model $\mathcal{D}(g)$ which is defined below) and conduct a sensitivity analysis over various choices of g (for more discussion on this point, see Scharfstein *et al.* (1999)).

Theorem 2 has the following consequences for inference about (γ^*, α^*) . If restriction 1 were the only restriction defining the model, then results from Chamberlain (1987) concerning models that are defined by zero conditional mean restrictions would imply that RAL estimators of (α^*, γ^*) have influence functions of the form $d_1(X) q_1(O; \alpha^*, \gamma^*)$ for some $(q + p) \times 2$ function $d_1(X)$. Likewise, if restriction 2 were the only restriction defining the model, then results in Chamberlain (1987) about models that are defined by conditional mean independence restrictions would imply that, if an RAL estimator of α^* existed, its influence function would have to equal $d_2(X) q_2(O; \alpha^*)$ for some $p \times 1$ function $d_2(X)$, where

$$q_2(O; \alpha) \equiv (S \omega\{r(X; \alpha) + g(Y, X)\}^{1-Z} - E[S \omega\{r(X; \alpha) + g(Y, X)\}^{1-Z}|X])\{Z - E(Z|X)\}.$$

(Note that, if restriction 2 were the only restriction defining the model, then γ^* would not be defined.) In the light of this discussion, we would expect that RAL estimators of (α^*, γ^*) in model $\mathcal{C}(g)$ would have influence functions of the form $d_1(X) q_1(O; \alpha^*, \gamma^*) + d_2(X) q_2(O; \alpha^*)$ for some $(q + p) \times 2$ functions $d_1(X)$ and $d_2(X)$. This is indeed so, as stated below in part (c) of theorem 3. However, results in Klaassen (1987) indicate that influence functions of RAL estimators must be estimable consistently. Since $q_2(O; \alpha^*)$ depends on the unknown functions $\Pr(Z = 1|X)$ and $E[S \omega\{r(X; \alpha^*) + g(Y, X)\}^{1-Z}|X]$ and these are not estimable consistently without further assumptions, we conclude that RAL estimators in model $\mathcal{C}(g)$ have influence functions of the form $d_1(X) q_1(O; \alpha^*, \gamma^*)$, i.e. with $d_2(X) = 0$. In essence, this result implies that under model $\mathcal{C}(g)$ we shall never be able to exploit restriction 2 for estimation of α^* . This in turn implies that, if the dimension of α^* is large, then RAL estimators of α^* and γ^* in model $\mathcal{C}(g)$ will generally have large sampling variability or may not even exist because α^* and γ^* will be weakly identified or may even be non-identified by restriction 1. To see this consider the

following extreme situation. Suppose that X is discrete with, say, k levels, and we pose models for $m(z = 1, X)$ and $m(z = 0, X)$ that are indexed by variation-independent parameters γ_0 and γ_1 , and a fully saturated model for $r(\cdot)$. Then, restriction 1 determines a set of k population equations, which are insufficient to identify the k components of α and the additional parameter γ_0 .

Fortunately, in the context of randomized studies, the difficulty can be remedied, because the relevant model under which to conduct inference is model $\mathcal{D}(g)$, which is defined like $\mathcal{C}(g)$, but with the additional restriction that the randomization probabilities are known, i.e. $\Pr(Z = 1|X)$ is a known function of X .

Part (a) of theorem 3 below establishes that RAL estimators of (α^*, γ^*) under model $\mathcal{D}(g)$ have influence functions that can be written as $h(O; d^*, \alpha^*, \gamma^*) \equiv d^*(X) q(O; \alpha^*, \gamma^*)$ for some $(q + p) \times 4$ vector function $d^*(X)$, where $q(O; \alpha, \gamma) \equiv (q_1(O; \alpha, \gamma)', q_2(O; \alpha, \gamma), Z - E(Z|X))'$ for any (α, γ) . Subsequently, in remark 3, we explain why the functional form of $h(O; d^*, \alpha^*, \gamma^*)$ implies that estimators of (α, γ) that exploit restriction 2 can be constructed under model $\mathcal{D}(g)$. Define

$$d_{\text{eff}}(X) = E \left\{ \frac{\partial}{\partial(\alpha', \gamma')} q(O; \alpha, \gamma) \Big|_{(\alpha, \gamma) = (\alpha^*, \gamma^*)} \Big| X \right\}' \text{var}\{q(O; \alpha^*, \gamma^*)|X\}^{-1}. \quad (10)$$

Throughout, for any pair of matrices T_1 and T_2 , $T_1 \leq T_2$ means that $T_2 - T_1$ is positive semi-definite and $T_1^{\otimes 2}$ denotes $T_1 T_1^T$.

Theorem 3

- a. If $(\hat{\alpha}, \hat{\gamma})$ is an RAL estimator of (α, γ) in model $\mathcal{D}(g)$, then there is a $(q + p) \times 4$ vector function $d(X)$ such that $(\hat{\alpha}, \hat{\gamma})$ has influence function $h(O; d^*, \alpha^*, \gamma^*)$, where $d^*(X) = [E\{h(O; d, \alpha^*, \gamma^*)h(O; d_{\text{eff}}, \alpha^*, \gamma^*)^T\}]^{-1}d(X)$. Equivalently,

$$n^{1/2} \left\{ \begin{pmatrix} \hat{\alpha} \\ \hat{\gamma} \end{pmatrix} - \begin{pmatrix} \alpha^* \\ \gamma^* \end{pmatrix} \right\} \xrightarrow{n} N(0, \Sigma) \quad \text{where } \Sigma \equiv E\{h(O; d^*, \alpha^*, \gamma^*)^{\otimes 2}\}.$$

- b. The function $h(O; d_{\text{eff}}, \alpha^*, \gamma^*)$ is the efficient influence function for RAL estimators of (α, γ) in model $\mathcal{D}(g)$. Thus, for any $h(O; d^*, \alpha^*, \gamma^*)$

$$E\{h(O; d_{\text{eff}}, \alpha^*, \gamma^*)^{\otimes 2}\} \leq E\{h(O; d^*, \alpha^*, \gamma^*)^{\otimes 2}\}.$$

- c. If $(\tilde{\alpha}, \tilde{\gamma})$ is an RAL estimator of (α, γ) in model $\mathcal{C}(g)$ then there is a $(q + p) \times 4$ function $d(X)$ with fourth column equal to $\mathbf{0}$ such that $(\tilde{\alpha}, \tilde{\gamma})$ has influence function $h(O; d^*, \alpha^*, \gamma^*)$.

4.1. A class of estimating equations

Consider the estimating equation

$$\sum_i d(X_i) q(O_i; \alpha, \gamma) = 0 \quad (11)$$

where $d(X)$ is an arbitrary $(q + p) \times 4$ vector function. Using standard Taylor series expansion arguments, it can be shown that if equation (11) has a solution $(\hat{\alpha}, \hat{\gamma})$ that is RAL then the influence function of $(\hat{\alpha}, \hat{\gamma})$ must be equal to $h(O; d^*, \alpha^*, \gamma^*)$ with d^* defined in theorem 3. Thus,

part (a) of theorem 3 implies that if RAL estimators $(\hat{\alpha}, \hat{\gamma})$ of (α^*, γ^*) in model $\mathcal{D}(g)$ exist we can indeed obtain them all (up to asymptotic equivalence) by solving estimating equations of the form (11) for adequate choices of $d(X)$. In addition, a consistent estimator of the asymptotic variance Σ can be constructed as $n^{-1} \sum_i h(O_i; d^{\wedge*}, \hat{\alpha}, \hat{\gamma})^{\otimes 2}$ where

$$\widehat{d}^*(X) = \{n^{-1} \sum_i d(X_i) \partial q(O; \alpha, \gamma) / \partial(\alpha', \gamma')|_{(\alpha, \gamma) = (\hat{\alpha}, \hat{\gamma})}\}^{-1} d(X).$$

Remark 3

We noted earlier that under model $\mathcal{C}(g)$ restriction 2 of theorem 2 cannot be effectively exploited for estimation of (α^*, γ^*) because $q_2(O; \alpha)$ depends on the unknowns $\Pr(Z = 1|X)$ and $E[S \omega\{r(X; \alpha^*) + g(Y)\}^{1-Z} | X]$. In contrast, under $\mathcal{D}(g)$, we can use estimating equations with summands $d_2(X_i) q_2(O_i; \alpha) + d_3(X_i) \{Z_i - E(Z|X_i)\} = 0$ which do exploit this restriction. For instance, the equation

$$\sum_i d_2(X_i) S_i \omega\{r(X_i; \alpha) + g(Y_i)\}^{1-Z_i} \{Z_i - E(Z|X_i)\} = 0$$

that is obtained after setting $d_3(X)$ equal to $E[S \omega\{r(X; \alpha^*) + g(Y)\}^{1-Z} | X]$ has this form and does not depend on unknown nuisance functions because, under model $\mathcal{D}(g)$, $E(Z|X)$ is known.

Remark 4

The estimating equations (11) may fail to have a solution. This is best understood in the absence of covariates. Then, $r(x; \alpha) = \alpha$ and $\gamma = (\gamma_0, \gamma_1)$ with $\gamma_0^* = m(z=0)$ and $\gamma_1^* = m(z=1)$. Suppose that $g(Y) = \beta Y$ and $d = d(X)$ is the constant 4×3 matrix with last row equal to $\mathbf{0}$ and first three rows defining the 3×3 identity matrix. Then, solving the third equation in equations (11) is equivalent to solving

$$\begin{aligned} & \left\{ \sum_i S_i (1 - Z_i) \right\}^{-1} \sum_i S_i (1 - Z_i) \omega(\alpha + \beta Y_i) \\ &= \widehat{p}_1 / \widehat{p}_0, \end{aligned}$$

where

$$\widehat{p}_j \equiv \sum_i S_i I(Z_i = j) / \sum_i I(Z_i = j)$$

is an unbiased estimator of $\Pr\{S(j) = 1\}$. The resulting estimator of α is precisely the estimator of Gilbert *et al.* (2003). Since $\omega(\alpha + \beta Y_i)$ is between 0 and 1, the left-hand side of this equation, being an average of values of ω , is also between 0 and 1 for any value of α . However, although $\Pr\{S(1) = 1\} < \Pr\{S(0) = 1\}$ under model $\mathcal{D}(g)$, it may happen that $\widehat{p}_1 > \widehat{p}_0$, in which case the right-hand side is a number which is greater than 1. In such circumstances, the equation has no solution α . Such situations, of course, happen with higher probability the closer $\Pr\{S(1) = 1\}$ is to $\Pr\{S(0) = 1\}$, or, equivalently, under assumption 1, the closer $\Pr\{S(1) = 0 | S(0) = 1\}$ is to 0 and hence the closer we are to violating assumption 2. In their simulation studies, Gilbert *et al.* (2003) resolved this problem by estimating $m(0)$ with $\sum_i S_i (1 - Z_i) Y_i / \sum_i S_i (1 - Z_i)$ whenever $\widehat{p}_1 > \widehat{p}_0$, the justification being that, under assumption 1, violation of assumption 2 implies that $m(0) = E(Y|S = 1, Z = 0)$. However, Jemiai (2005) showed that Wald confidence intervals

that are centred at such estimators of $m(0)$ will have coverage probability that is smaller than the prescribed level, especially for large values of $|\beta|$. In our context, the strategy of Gilbert *et al.* (2003) when equation (11) does not have a solution would correspond to estimating γ^* as the solution to equation (11) with $\omega\{r(x, \alpha) + \beta Y_i\}$ equal to 1 for all i in the second entry of $q_1(O; \alpha, \gamma)$. In light of the difficulties that were mentioned concerning Wald interval estimation, we recommend following this strategy but with one *caveat*: that a warning about the possible inadequacy of inference for large values of $|\beta|$ be added to the conclusions of the analysis whenever such replacements are performed. The investigation of valid confidence intervals when assumption 2 is almost violated certainly warrants further study but is beyond the scope of this paper.

4.2. Locally efficient estimation

Part (b) of theorem 3 implies that solving equation (11) with $d_{\text{eff}}(X)$ instead of $d(X)$ yields a globally efficient estimator of (α^*, γ^*) in model $\mathcal{D}(g)$, i.e. an estimator whose limiting distribution, after centring at (α^*, γ^*) and standardizing by $\sqrt{n}$, is normal with mean 0 and variance equal to $E\{h(O; d_{\text{eff}}^*, \alpha^*, \gamma^*)^{\otimes 2}\}$ under any law that is allowed by $\mathcal{D}(g)$. Unfortunately, such an estimator is infeasible, because $d_{\text{eff}}(X)$ depends on components of the unknown observed data distribution F_O . However, we can obtain an estimator $(\hat{\alpha}_{\text{loc,eff}}, \hat{\gamma}_{\text{loc,eff}})$ that is locally semipara-metric efficient in model $\mathcal{D}(g)$ at a ‘working’ parametric submodel $\mathcal{D}_{\text{work}}(g)$, i.e. an estimator that is consistent and asymptotically normal under model $\mathcal{D}(g)$ with asymptotic variance equal to the semiparametric variance bound under any law that is allowed by $\mathcal{D}_{\text{work}}(g)$. To construct such an estimator, we can proceed as follows.

Step 1: specify a working parametric submodel $\mathcal{D}_{\text{work}}(g)$ that is indexed by (α, γ) and finite dimensional nuisance parameters, say η . This model determines a parametric model $F_O(\cdot; \alpha, \gamma, \eta)$ for the distribution $F_O(\cdot)$ of the observed data O .

Step 2: compute the ML estimator $(\hat{\alpha}^-, \hat{\gamma}^-, \hat{\eta}^-)$ of (α, γ, η) under model $\mathcal{D}_{\text{work}}(g)$.

Step 3: compute $d^{\text{eff}}(X)$, which is defined as $d_{\text{eff}}(X)$ in equation (10), but with expectations calculated under the law $F_O(\cdot; \hat{\alpha}^-, \hat{\gamma}^-, \hat{\eta}^-)$:

Step 4: solve equation (11) by using $d^{\text{eff}}(X)$ for $d(X)$.

It is standard to show (see, for example, Robins *et al.* (1992)) that, under regularity conditions, the estimating equations that are solved in step 4 have a solution $(\hat{\alpha}_{\text{loc,eff}}, \hat{\gamma}_{\text{loc,eff}})$ that is locally semiparametric efficient in model $\mathcal{D}(g)$, at the submodel $\mathcal{D}_{\text{work}}(g)$. We can ensure that the working parametric model is indeed a submodel of $\mathcal{D}(g)$ by arguing as follows. Suppose that we postulate parametric models that are indexed by η_1 and η_0 for the conditional distributions of $e_1 \equiv Y(1) - m(z=1, X)$ and $e_0 \equiv Y(0) - m(z=0, X)$ given X , $S(0)=1$ and $S(1)=1$ respectively, and a working parametric submodel $\Pr\{S(1)=1|X; \eta_2\}$ that is indexed by η_2 for $\Pr\{S(1)=1|X\}$. These submodels are necessarily compatible with model $\mathcal{D}(g)$ because model $\mathcal{D}(g)$ does not impose restrictions on the conditional distributions of e_1 and e_0 given X , $S(0)=1$ and $S(1)=1$ nor on $\Pr\{S(1)=1|X\}$. To construct the likelihood $L_n(\alpha, \gamma, \eta)$ under such a model, we marginalize the joint distribution of the counterfactuals over the missing information as in Frangakis and Rubin (2002). We effectively implement this marginalization by first deriving the restrictions that are implied by our model for the counterfactuals on the observed data distribution and then constructing the likelihood based on the derived model for the observed data. Specifically, under assumption (1), the submodels for the conditional distributions of e_1 and e_0 , together with the $\mathcal{D}_{\mathbf{E}, \mathbf{E}}$ marginal mean model, determine fully parametric models $f(Y|S=1, Z=1, X; \gamma, \eta_1)$ for the distribution $f(Y|S=1, Z=1, X)$, and $f^*(Y|X; \gamma, \eta_0)$ for the distribution

$$f^*(Y|X) \equiv \frac{\omega\{r(X;\alpha^*)+g(X, Y)\} f(Y|S=1, Z=0, X)}{E[\omega\{r(X;\alpha^*)+g(X, Y)\} |S=1, Z=0, X]}. \quad (12)$$

In turn, model $f^*(Y|X; \gamma, \eta_0)$ determines a parametric model $f(Y|S=1, Z=0, X; \gamma, \alpha, \eta_0)$ for $f(Y|S=1, Z=0, X)$. Also, restriction 2 of theorem 2 is equivalent to the restriction

$$\begin{aligned} \Pr(S=1|Z=1, X) &= E[\omega\{r(X;\alpha^*)+g(X, Y)\} |S \\ &= 1, Z=0, X] \Pr(S=1|Z=0, X). \end{aligned} \quad (13)$$

Thus, under condition (1), the model $\Pr\{S(1)=1|X; \eta_2\}$ for $\Pr\{S(1)=1|X\}$ determines a parametric model for $\Pr(S=1|Z=1, X)$ that is indexed by η_2 and a parametric model for $\Pr(S=1|Z=0, X)$ that is indexed by $(\gamma, \alpha, \eta_0, \eta_2)$. Under the working model, the likelihood $\mathcal{L}_n(\alpha, \gamma, \eta)$ is equal to

$$\prod_{i=1}^n f\{\varepsilon_{0,i}(\gamma) | S_i=1, Z_i=0, X_i; \eta_0\}^{S_i(1-Z_i)} f\{\varepsilon_{1,i}(\gamma) | S_i=1, Z_i=1, X_i; \eta_1\}^{S_i Z_i} Q_i(\alpha, \gamma, \eta_0, \eta_2)$$

where $\eta = (\eta_0, \eta_1, \eta_2)$,

$$Q_i(\alpha, \gamma, \eta_0, \eta_2) = \{\Gamma_i(\eta_2) R_i(\alpha, \gamma, \eta_0)^{Z_i-1}\}^{S_i} \{1 - \Gamma_i(\eta_2) R_i(\alpha, \gamma, \eta_0)^{Z_i-1}\}^{1-S_i},$$

$\Gamma_i(\eta_2) = P(S_i=1|Z_i=1, X_i; \eta_2)$, $\varepsilon_{z,i}(\gamma) = Y_i - m(z, X_i; \gamma)$, $z \in \{0, 1\}$ and

$$\begin{aligned} R_i(\alpha, \gamma, \eta_0) &= E[\omega\{X; \alpha\} + g(X, Y) | S_i \\ &= 1, Z_i=0, X; \gamma, \eta_0]. \end{aligned}$$

Remark 5

In steps 2 and 3 we can replace the ML estimator $(\hat{\alpha}^-, \hat{\gamma}^-, \hat{\eta}^-)$ with the easier-to-compute estimator $(\tilde{\alpha}^-, \tilde{\gamma}^-, \tilde{\eta}^-)$ where $(\tilde{\alpha}^-, \tilde{\gamma}^-)$ solves equation (11) for an arbitrary choice of $d(X)$ and $\tilde{\eta}^-$ is the value of η that maximizes $\mathcal{L}_n(\tilde{\alpha}^-, \tilde{\gamma}^-, \eta)$. The output of the thus-modified algorithm still returns a locally efficient estimator of (α, γ) .

Remark 6

The ML estimator $(\tilde{\alpha}^-, \tilde{\gamma}^-)$ is one of the estimators that was proposed by Shepherd *et al.* (2005). Note, however, that in contrast with $(\tilde{\alpha}^-, \tilde{\gamma}^-)$ our proposed estimator $(\hat{\alpha}_{\text{loc,eff}}, \hat{\gamma}_{\text{loc,eff}})$ is consistent, regardless of whether or not the models for $f\{\varepsilon_1|S(0)=1, S(1)=1, X\}$, $f\{\varepsilon_0|S(0)=1, S(1)=1, X\}$ and $\Pr(S=1|Z=0, X)$ are correctly specified.

5. A simulation study

A simulation study was carried out to evaluate the finite sample properties of our proposed estimator $(\hat{\alpha}_{\text{loc,eff}}, \hat{\gamma}_{\text{loc,eff}})$. We conducted seven experiments, each consisting of 1000 repetitions so that 95% confidence intervals were expected to cover the true parameters with roughly a 1:3% margin of error. We generated independent vectors $W_i = (Z_i, X_i, S_i(0), S_i(1))$,

$Y_i(0), Y_i(1)), i = 1, \dots, n$, where $n = 2000$ or 20000 depending on the experiment, so that conditions (1), (2), 1–4, (9) and (8) would hold with

$$\left. \begin{aligned} r(X; \alpha) &= \alpha_0 + \alpha_1 X, \\ m(z; X; \gamma) &= \gamma_0 + \gamma_1 X + \gamma_2 z + \gamma_3 z X, \\ g(X, Y) &= \beta Y \end{aligned} \right\} \quad (14)$$

where the values of $(\alpha_0^*, \alpha_1^*)(\gamma_0^*, \gamma_1^*)$ and β are given below, and $\gamma_2^* = \gamma_3^* = 0$. This set-up implies that treatment Z has indeed no effect on the conditional mean of the outcome in the subpopulation $\mathcal{P}_{\mathbf{E}, \mathbf{E}}$. In addition, under our data-generating process the distributions of $Y(0)$ and $Y(1)$ given X in the subpopulation $\mathcal{P}_{\mathbf{E}, \mathbf{E}}$ are the same and equal to $N\{m(z, X; \gamma^*), 1\}$. The distribution of $Y(0)$ in the subpopulation $\mathcal{P}_{\mathbf{E}, \mathbf{E}}$ is normal with mean $m(z, X; \gamma^*) - \beta$ and variance 1. Our parameter values were chosen so that roughly 70% of the subjects with $S(0) = 1$ and $X = 38$ (with 38 being the mean of X , as indicated next) would also have $S(1) = 1$, i.e. so that $\Pr\{S(1) = 1 | S(0) = 1, X = 38\} \approx 0.7$.

For each $i = 1, \dots, n$, Z_i , X_i and $S_i(0)$ were generated independently as $Z_i \sim \text{Bernoulli}(0.5)$, $X_i \sim N(38, 36)$ and $S_i(0) \sim \text{Bernoulli}(0.25)$ (the distribution of X was chosen to resemble that of age in the study that is described in Section 6). For units with $S_i(0) = 1$, $S_i(1)$ was generated from a Bernoulli distribution with success probability equal to

$$1 - [1 + \exp\{\alpha_0^* + \alpha_1^* X_i + \beta(\gamma_0^* + \gamma_1^* X_i - 0.5\beta)\}]^{-1}$$

, where $\gamma_0 = 2.3$, $\gamma_1 = 0.05$, $\alpha_1^* = \log(2)/10$ and, depending on the experiment, $(\alpha_0^*, \beta) = (-2.2, 0.1)$, $(-5.5, 1)$, $(-9.9, 3)$. For reasons that are explained below, we also conducted an additional experiment with $n = 2000$ and $(\alpha_0^*, \beta) = (-8.8, 2)$. The values of β were chosen to reflect little, moderate and serious differences in the distributions of the outcome $Y(0)$ in the subpopulations $\mathcal{P}_{\mathbf{E}, \mathbf{E}}$ and $\mathcal{P}_{\mathbf{E}, \mathbf{E}}$. We avoided the value $\beta = 0$ since, under such value, our proposal reduces to standard analysis based on weighted least squares among subjects with $S = 1$. For units with $S_i(0) = 0$, $S_i(1)$ was set to 0. Given $S_i(0)$, $S_i(1)$, X_i and Z_i , $(Y_i(0), Y_i(1))$ were generated as follows. First we generated $Y_i^* \sim N[\gamma_0^* + \gamma_1^* X_i - \{1 - S_i(1)\}\beta, 1]$. Then, we set $Y_i(0) = Y_i(1) = Y_i^*$ if $S_i(0) = S_i(1) = 1$, $Y_i(0) = Y_i^*$ and $Y_i(1) = *$ if $S_i(0) = 1$ and $S_i(1) = 0$, and $Y_i(0) = Y_i(1) = *$ if $S_i(0) = 0$. Finally, we set $S_i = S_i(Z_i)$ and $Y_i = Y_i(Z_i)$. It can be easily seen that the data-generating process in our simulation satisfies the conditions that were described in the preceding paragraph.

Table 1 and Table 2 report the results for inference about the parameters $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ and $\alpha = (\alpha_0, \alpha_1)$ respectively of the models (14) for $n = 20000$ and $n = 2000$. Each table reports the Monte Carlo mean (labelled ‘Mean’) and median (labelled ‘Median’) of the estimators, the Monte Carlo coverage probability of nominal 95% Wald confidence intervals (labelled CP) and their median length (labelled ‘Length’). Table 1 and Table 2 report results for the following estimators: the naïve, OLS, estimator of Y on Z and X based on observations with $S = 1$ (labelled OLS), the inefficient estimator of (α, γ) (labelled INE) solving equation (11) that uses

$$d(X) = \begin{pmatrix} 1 & X & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & X & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & X \\ 0 & 0 & 0 & 0 & b(X) & X b(X) \end{pmatrix}, \quad (15)$$

where $b(X)$ was chosen so that

$$\begin{aligned}
& q_2(O;\alpha)+b(X)\{Z \\
& \quad - E(Z|X)\} \\
& = S \omega\{r(X;\alpha)+g(Y)\}^{1-Z}\{Z - E(Z|X)\},
\end{aligned}$$

and the locally efficient estimator under (correctly specified) working models that assume a logistic model for $\Pr(S = 1|Z = 0, X)$ and normal distributions with variance equal to 1 for $f^*(Y|X)$ and $f(Y|S = 1, Z = 1)$ (labelled EFF). All estimators were computed under the true value of β . The starting values of γ and α for the algorithm solving the inefficient estimating equation were set at the OLS estimator of γ and at $\alpha = (0, 0)$. The resulting estimates were used as starting values for the algorithm solving the locally efficient estimating equation. When $n = 2000$, the inefficient estimation algorithm did not converge in 0.4–14.4% of runs, depending on the value of β (more frequently for the larger values of β), because the algorithm failed to find a root of the estimating equation. These runs were discarded and replaced with new runs. In each of the faulty runs we examined whether the algorithm solving the locally efficient estimating equation converged when started at values of γ and α near the true values. In all the runs that were investigated, the algorithm converged. We therefore attribute the lack of convergence of the algorithm solving the inefficient equation to the poor choice of function $d(X)$. We do not regard this failure as serious, since, faced with it, an investigator would try different $d(X)$ and different starting values until the algorithm converged.

Our results for $n = 20000$ confirm that the properties that are established by theorem 3 hold. The naïve OLS estimator was biased, more so as the value of β departed from 0. When $\beta = 3$, this bias was sufficiently severe to reverse the sign of $\hat{\gamma}_0$. OLS estimators of the mean shift parameter γ_2 were significantly far from 0 even when $\beta = 1$. Also, as predicted by the theory, both the efficient and the inefficient estimators were unbiased. Coverage probabilities of the 95% confidence intervals were close to the nominal value, and efficiency, as measured by the median interval length, was somewhat better when the locally efficient estimator was used to centre the intervals, the gains in efficiency being more pronounced for estimation of α . Curiously, intervals that are centred at the inefficient estimator had poor coverage when $\beta = 3$ whereas the coverage was substantially improved when the intervals were centred at the efficient estimator. No significant gains in efficiency were obtained from the locally efficient estimator when β was 0.1. This came as no surprise since this value, accounting for the variance in Y , is very close to 0. At $\beta = 0$, it can be easily shown that $d_{\text{eff}}(X) = d(X)$ given in equation (15), and ω does not depend on Y , thus resulting in algebraic identity between the locally efficient and inefficient estimators. Results for $n = 2000$ were qualitatively similar to those for $n = 20000$, except that for $\beta = 3$ both the inefficient and the efficient estimators were biased and the coverage probability of the interval was poor. This demonstrates that the asymptotic distribution $(\hat{\gamma}, \hat{\alpha})$ is not a good approximation to the finite sample distribution of $(\hat{\gamma}, \hat{\alpha})$ when β is large even if n is moderate. The extra experiment with $\beta = 2$ is meant to show that the asymptotic distribution is still a good approximation when $n = 2000$ at this value of β .

6. Example

In this section, we apply our methods to data from a randomized, double-blind, placebo-controlled phase III trial, which was conducted by VaxGen between 1998 and 2003. This trial sought to evaluate the performance of the vaccine AIDSVAX B/B among a total of 5403 HIV negative at-risk individuals. The ratio of vaccine to placebo assignment was 2:1. Overall, the vaccine was found to be ineffective in preventing HIV infection. Exploratory subgroup analyses, however, suggested the possibility that the vaccine partially prevented infection in non-white and high risk subjects (rgp120 HIV Vaccine Study Group, 2005).

In our analysis we eliminated 21 subjects who did not enrol in the post-infection phase of the study, and three subjects who started antiretroviral therapy before assessment of their HIV viral load. After doing so we were left with 223 infected out of 3823 subjects in the vaccine arm and 121 out of 1920 in the placebo arm. Our analysis is for illustrative purposes only. An adequate analysis would also adjust for potential selection bias that is induced by our censoring rule.

A Wald test of the hypothesis $H_0: \Pr(S = 1|Z = 1) = \Pr(S = 1|Z = 0)$ had a p -value of 0.244, thus raising doubts about the validity of assumptions 1 and 2, and hence the applicability of our methods to the entire study population. However, as previously mentioned, possible partial vaccine efficacy was observed in non-white subjects, suggesting that our assumptions 1–4 are reasonable for this subgroup. Indeed, this subgroup consisted of 28 HIV infected out of 309 subjects in the placebo arm, and 28 out of 602 in the vaccine arm, yielding a p -value of 0.008 for the Wald test of hypothesis H_0 . We therefore estimated the treatment effect on the mean viral load, as measured on the common logarithms scale, at a study visit 2 weeks after infection was diagnosed, among non-white subjects conditional on the base-line covariates age and risk score. Specifically, we estimated the parameters of the model

$$m(Z, X, \gamma) = \gamma_0 + \gamma_1 \text{Age} + \gamma_2 \text{Risk} + \gamma_3 Z + \gamma_4 Z \text{Age} + \gamma_5 Z \text{Risk}.$$

For subjects who did not come in for a visit 2 weeks after HIV diagnosis, we used as outcome the viral load measurement at the visit closest in time. Of the infected subjects, seven out of 56 had a viral load that reached the assay's measurement lower limit. These subjects were assigned viral load levels equal to the lower detection limit $\log_{10}(400)$. Since our methods do not require specification of the viral load distribution, this imputation does not induce model misspecification. Indeed, the mean of the thus-defined outcome can be viewed as the mean of the viral load distribution truncated at the assay measurement limit.

In our analysis, we assumed normal working distributions for common logarithm viral load among always infected $\mathcal{P}_{E,E}$ subjects in both treatment arms. The variance of these distributions was assumed constant over X and estimated as such. The probability $\Pr(S = 1|Z = 0, X)$ was estimated by logistic regression.

Fig 1 and Fig 2 display the locally efficient estimates of (γ_4^*, γ_5^*) , and the vaccine effect

$$m(1, \text{Age}, \text{Risk}) - m(0, \text{Age}, \text{Risk}) = \gamma_3 + \gamma_4^* \text{Age} + \gamma_5^* \text{Risk}$$

for various values of age and risk score respectively, as a function of the sensitivity parameter β in the range $(-2.5, 2.5)$, and their associated pointwise 95% confidence intervals. To interpret the results, note that a difference of 0.5 in the mean of common logarithm viral load is generally considered to be the smallest clinically significant vaccine effect size. The range of $(-2.5, 2.5)$ for β was chosen to reflect possibly severe discrepancies between the distributions of viral load (conditional on the covariates) in the always infected and the protected subjects. Values of β that are close to 2.5 correspond to the case in which the distribution of viral loads in the always infected subjects is severely tilted to the right relative to that of the protected individuals (the reverse being true when β is close to -2.5). We chose the values 0, 1 and 4 for risk score as they reflect low, medium and high risk. The chosen values of 28, 34 and 39 for age correspond to the 25th, 50th and 75th percentiles of its empirical distribution in the non-white sample respectively. Positive and negative values of β respectively correspond to assuming that, the higher and lower the viral load under placebo, the more likely it is that an individual who is

infected under placebo would have also been infected under vaccine or, equivalently, that the distribution of viral load under placebo in the always infected population is to the right and the left of that of the protected population respectively. The confidence intervals in Fig. 1 include the value 0 for nearly all values of β , the only exception being the intervals for γ_4 under $\beta < -1$. This suggests an effect modification by age only under the rather unrealistic assumption that the protected individuals tend to have much higher levels of viral load under placebo than the always infected subjects. This effect modification is reflected in Fig. 2, where for sufficiently negative values of β there is some evidence of a detrimental vaccine effect at age 28 years, but no vaccine effect at the other ages. Fig. 2 also shows that, for other values of β , there is no evidence of a vaccine effect.

Acknowledgments

This work was supported by National Institutes of Health grants R29 GM48704 (Jemiai and Rotnitzky), R01 AI32475 (Rotnitzky) and 1 RO1 AI054165-01 (Gilbert).

Appendix A

A.1. Proof of theorem 1

Consider any joint distribution of (Y, S, Z, X) satisfying condition (7). To prove theorem 1 we must exhibit a joint distribution for $(S(0), S(1), Y(0), Y(1), Z, X, Y, S)$ satisfying condition (1), and the conditions defining model $\mathcal{A}(g)$. We do this in the absence of covariates X , since the construction can then be repeated within each level of X . We define the candidate joint distribution in the following steps:

- a. given $(Z = s, Y, S)$, $(Y(z), S(z)) \equiv (Y, S)$, i.e. the distribution of $(Y(z), S(z))$ given $(Z = z, Y, S)$ is a point mass at (Y, S) ;
- b. $\Pr\{S(0) = 1 | S(1) = 1, S, Y, Y(1), Z = 1\} \equiv 1$,
- c. $\Pr\{S(0) = 1 | S(1) = 0, S, Y, Y(1), Z = 1\} \equiv \{\Pr(S = 1 | Z = 0) - \Pr(S = 1 | Z = 1)\} / \Pr(S = 0 | Z = 1)$ (note that this is well defined because, by condition (7), the right-hand side is a positive number);
- d. $\Pr\{S(1) = 1 | S(0) = 0, S, Y(0), Y, Z = 0\} \equiv 0$;
- e. $\Pr\{S(1) = 1 | S(0) = 1, S, Y(0), Y, Z = 0\} \equiv \omega\{r_0 + g(Y)\}$ where r_0 is the unique solution to the equation

$$E[\omega\{r_0 + g(Y)\} | S = 1, Z = 0] = \frac{\Pr(S = 1 | Z = 1)}{\Pr(S = 1 | Z = 0)}. \quad (16)$$

So far, we have constructed a joint distribution for $(S(1), S(0), Y(z), S, Y)$, given $Z = z$, $z = 0, 1$. We finalize our construction by defining the candidate distribution of $Y(1-z)$ given $(S(1), S(0), Y(z), S, Y, Z = z)$, $z = 0, 1$, as

$$f\{Y(1-z) | S(1), S(0), Y(z), Y, S, Z = z\} \equiv f\{Y(1-z) | S(1), S(0), S, Z = 1-z\}. \quad (17)$$

We now show that the candidate distribution satisfies condition (1) and the restrictions defining model $\mathcal{A}(g)$. Condition (1) is satisfied by construction in step (a). To show that condition (2) holds, first note that step (b) implies that $\Pr\{S(0) = 1|S(1) = 1, Z = 1\} = 1$ and therefore that $\Pr\{S(0) = S(1) = 1|Z = 1\} = \Pr\{S(1) = 1|Z = 1\}$. In addition,

$$\begin{aligned}\Pr\{S(0)=S(1)=1|Z=0\} &= E[\omega\{r_0+g(Y)\}|S=1, Z=0] \Pr\{S=1|Z=0\} \\ &= \Pr\{S=1|Z=1\} = \Pr\{S(1)=1|Z=1\},\end{aligned}$$

where the first equality is by steps (e) and (a), the second is by equation (16) and the third is by step (a). Therefore, $\Pr\{S(0) = 1, S(1) = 1|Z = 0\} = \Pr\{S(0) = 1, S(1) = 1|Z = 1\}$. Also, $\Pr\{S(1) = 1, S(0) = 0|Z = 0\} = 0$ by step (d) and $\Pr\{S(1) = 1, S(0) = 0|Z = 1\} = 0$ by step (b). Consequently, $\Pr\{S(1) = 1, S(0) = 0|Z = 0\} = \Pr\{S(1) = 1, S(0) = 0|Z = 1\} = 0$, and we have

$$\begin{aligned}\Pr\{S(0)=1|Z=1\} &= \sum_{j=0}^1 \Pr\{S(0)=1|S(1)=j, Z=1\} \Pr\{S(1)=j|Z=1\} = \Pr\{S(1)=1|Z=1\} \\ &\quad + \frac{\Pr\{S=1|Z=0\} - \Pr\{S=1|Z=1\}}{\Pr\{S=0|Z=1\}} \Pr\{S(1)=0|Z=1\} = \Pr\{S(0)=1|Z=0\}\end{aligned}$$

where the second equality is by steps (b) and (c) and the third is by step (a).

This concludes the proof that $(S(0), S(1)) \perp\!\!\!\perp Z$. That $(Y(0), Y(1)) \perp\!\!\!\perp Z|S(0), S(1)$ follows because, by construction in expression (17), $Y(0)$ and $Y(1)$ are conditionally independent given $(S(0), S(1), Z)$ and, in addition, $f\{Y(1-z)|S(0) = 1, S(1), Z = z\}$ is equal to $f\{Y(1-z)|S(0) = 1, S(1), Z = 1-z\}$, $z = 0, 1$. This concludes the proof that condition (2) holds.

To show that assumptions 2 and 3 hold it suffices to show that $\Pr\{S(0) = 1\} > 0$ and $0 < \Pr\{S(1) = 1|S(0) = 1\} < 1$. Now, $\Pr\{S(0) = 1\} = \Pr\{S(0) = 1|Z = 0\}$ by condition (2) and the right-hand side is equal to $\Pr\{S = 1|Z = 0\}$ by step (a) which in turn is greater than 0 by condition (7). Furthermore,

$$\begin{aligned}\Pr\{S(1)=1|S(0)=1\} &= \Pr\{S(1)=1|S(0)=1, Z=0\} \\ &= E[\omega\{r_0+g(Y)\}|S(0)=1, Z=0]\end{aligned}$$

where the first equality is by condition (2) and the second is by step (e). Thus, $0 < \Pr\{S(1) = 1|S(0) = 1\} < 1$ follows because $0 > \omega\{r_0 + g(y)\} > 1$ for any y . This shows that assumptions 2 and 3 hold. Condition 1 follows by step (d) and condition (2). Finally, assumption 4 follows directly from step (e) after application of Bayes rule.

A.2. Proof of theorem 2

That restrictions 1 and 2 are restrictions on the observed data distribution that are implied by model $\mathcal{C}(g)$ follows easily from expressions (5) and (6) respectively. To show that these are the only restrictions that are imposed by model $\mathcal{C}(g)$ we must exhibit, for any distribution of (Y, S, Z, X) satisfying restrictions 1 and 2, a joint distribution of $(X, Z, S, Y, S(0), S(1), Y(0), Y(1))$ that satisfies condition (1) and the restrictions defining model $\mathcal{C}(g)$. Such a distribution can be constructed within each level of X , exactly as in the proof of theorem 1, but replacing $\omega\{r_0 + g(Y)\}$ with $\omega\{r(X; \alpha^*) + g(Y, X)\}$, which satisfies equation (16) conditional on X by restriction 2.

A.3. Proof of theorem 3

Since, under model $\mathcal{D}(g)$, condition (13) holds, the model can be parameterized with the known function $g(\cdot, \cdot)$, the known conditional probability function $\Pr(Z = z|X = \cdot)$, the unknown $q \times 1$ and $p \times 1$ parameters γ and α , and an infinite dimensional parameter indexing $f(X)$, $\Pr(S = 1|Z = 1, X)$, $f(\varepsilon_0|S = 1, Z = 0, X)$ and $f(\varepsilon_1|S = 1, Z = 1, X)$ where $\varepsilon_j \equiv Y - E(Y|S = 1, Z = z, X)$, $z \in \{0, 1\}$. The set of influence functions of RAL estimators of (γ, α) under the model is the set

$\{E(AS_{\text{eff}}^T)^{-1}A: A \text{ is a } (q + p) \times 1 \text{ random vector with } j\text{th entry } A_j \in \Lambda^\perp\}$ where S_{eff} is the efficient score for (γ, α) and $\Lambda^\perp = \Lambda_x^\perp \cap \Lambda_s^\perp \cap \Lambda_0^\perp \cap \Lambda_1^\perp$. The sets $\Lambda_x, \Lambda_s, \Lambda_0$ and Λ_1 are the infinite dimensional parameter-specific tangent sets comprised of products of constant conformable vectors times the scores for the parameters in any regular parametric submodel for $f(X)$, $\Pr(S = 1|Z = 1, X)$, $f(\varepsilon_0|S = 1, Z = 0, X)$ and $f(\varepsilon_1|S = 1, Z = 1, X)$ respectively (for a precise definition of the nuisance tangent space, see for example Newey (1990)). Throughout, a set superscripted with ' $\perp$ ' denotes the orthogonal complement of that set and $\Pi(\cdot)$ denotes the projection operator in the Hilbert space $\mathcal{L}_2^0(F_O)$ of zero-mean, finite variance, random variables with covariance inner product under the observed data law F_O .

Consider arbitrary correctly specified parametric submodels $f(X; \phi)$, $\Pr(S = 1|Z = 1, X; \eta_2)$, $f(\varepsilon_0|S = 1, Z = 0, X; \eta_0)$ and $f(\varepsilon_1|S = 1, Z = 1, X; \eta_1)$ where $(\phi^*, \eta_0^*, \eta_1^*, \eta_2^*)$ correspond to the true distribution. The likelihood that is based on a single observation O is proportional to $\mathcal{L}(\eta, \alpha, \gamma, \phi) = f\{\varepsilon_0(\gamma)|S = 1, Z = 0, X; \eta_0\}^{S(1-Z)} f\{\varepsilon_1(\gamma)|S = 1, Z = 1, X; \eta_1\}^{SZ} Q(\alpha, \gamma, \eta_0, \eta_2) f(X; \phi)$ where $\eta \equiv (\eta_0, \eta_1, \eta_2)$,

$$Q(\alpha, \lambda, \eta_0, \eta_2) \equiv \{\Gamma(\eta_2) R(\alpha, \lambda, \eta_0)^{Z-1}\}^S \{1 - \Gamma(\eta_2) R(\alpha, \lambda, \eta_0)^{Z-1}\}^{1-S},$$

$$\Gamma(\eta_2) \equiv P(S = 1|Z = 1, X; \eta_2),$$

$$\begin{aligned} R(\alpha, \lambda, \eta_0) &\equiv E[\omega\{r(X; \alpha) \\ &\quad + g(X \\ &\quad , Y)\}|S \\ &\quad = 1, Z = 0, X; \eta_0, \gamma] \end{aligned}$$

$$\text{and } \varepsilon_z(\gamma) \equiv Y - E(Y|S = 1, Z = z, X = x; \gamma), z \in \{0, 1\}$$

Thus, with S_{η_j} denoting the score for η_j , $j = 0, 1, 2$, evaluated at the truth and ω^* denoting $\omega\{r(X; \alpha^*) + g(X, Y)\}$ we have

$$\begin{aligned} S_{\eta_2} &= \{S - E(S|Z \\ &\quad , X)\} \frac{\partial}{\partial \eta_2} \text{logit}\{\Gamma(\eta_2)|_{\eta_2 = \eta_2^*}\} u_0(X)^{1-Z} \end{aligned}$$

where $\text{logit}(v) = \log\{v/(1-v)\}$ and $u_0(X) = \{1 - E(S|Z = 1, X)\}/\{1 - E(S|Z = 0, X)\}$. Also, after some algebra and using the fact that

$$\frac{\partial}{\partial \eta_2} R(\alpha^*, \gamma^*, \eta_0) \Big|_{\eta_0 = \eta_0^*} = E(\omega^* A_0 | S=1, Z=0, X)$$

where

$$A_0 \equiv a_0(\varepsilon_0, X) \equiv \frac{\partial}{\partial \eta_0} \log\{f(\varepsilon_0 | S=1, Z=0, X; \eta_0)\} \Big|_{\eta_0 = \eta_0^*}$$

it can be shown that

$$S_{\eta_0} = (1 - Z) [S A_0 - \{S - E(S|Z=0, X)\} E(\omega^* A_0 | S=1, Z=0, X) / u_1(X)]$$

where $u_1(X) = R(\alpha^*, \gamma^*, \eta_0^*) \{1 - E(S|Z=0, X)\}$. From these expressions we can deduce that $\Lambda_s = \{ \{S - E(S|Z, X)\} u_0(X)^{1-Z} b(X) : b(X) \text{ is arbitrary} \} \cap \mathcal{L}_2^0(F_O)$ and that

$$\Lambda_0 = \mathcal{L}_2^0(F_O) \cap \{(1 - Z) [S A_0 - \{S - E(S|Z=0, X)\} E(\omega^* A_0 | S=1, Z=0, X) / u_1(X)] : A_0 = a_0(\varepsilon_0, X) \text{ satisfies } E(\omega^* \varepsilon_0 A_0 | S=1, Z=0, X) = E(A_0 | S=1, Z=0, X) = 0\}.$$

We arrive at the last set after noting that restriction 1 implies that $E(\omega^* \varepsilon_0 A_0 | S=1, Z=0, X) = 0$ and that A_0 , being a score in a model for a conditional distribution given $X, Z=0$ and $S=1$, has conditional mean 0. Also, a similar argument leads to

$$\Lambda_1 = \mathcal{L}_2^0(F_O) \cap \{S Z A_1 : A_1 \equiv a_1(\varepsilon_1, X) \text{ satisfies } E(\varepsilon_1 A_1 | S=1, Z=1, X) = E(A_1 | S=1, Z=1, X) = 0\}.$$

In addition, $\Lambda_x = \mathcal{L}_2^0(F_O) \cap \{c(X) : E\{c(X)\} = 0\}$. We next show that

$$\Lambda_s^\perp = \mathcal{L}_2^0(F_O) \cap \{ \{D_1 - E(D_1|X, Z, S)\} + h(X) - E\{h(X)\} + \{S \omega^{*1-Z} - E(S \omega^{*1-Z} | X)\} \times \{Z - E(Z|X)\} d_3(X) + Z - E(Z|X)\} d_2(X) ; D_1 \equiv d_1(Y, X, Z, S), d_2, h \text{ and } d_3 \text{ arbitrary} \}.$$

To do so, first note that any element of $\mathcal{L}_2^0(F_O)$ can be written as

$$\begin{aligned}
D_1 - E(D_1|X, Z, S) + \{Z - E(Z|X)\} d_2(X) \\
+ h(X) \\
- E\{h(X)\} \\
+ \{S - E(S|X, Z)\} \{Z t_1(X) \\
+ t_0(X)(1 - Z)\}
\end{aligned}$$

for some $D_1 \equiv d_1(Y, X, Z, S)$, $h(X)$, $d_2(X)$, $r_0(X)$ and $r_1(X)$. Next, note that any element of Λ_s , being a function of (X, Z, S) , is uncorrelated with $D_1 - E(D_1|X, Z, S)$. In addition, since any element of Λ_s is a linear combination of (or a limit of linear combinations of) scores under submodels for $f(S|X, Z)$, it must have mean 0 given (X, Z) and therefore it is uncorrelated with $\{Z - E(Z|X)\} d_2(X) + h(X) - E\{h(X)\}$. So it remains to determine the subset of random variables of the form $\{S - E(S|X, Z)\} \{Z t_1(X) + t_0(X)(1 - Z)\}$ that are uncorrelated with the elements of Λ_s . These satisfy, for all $b(X)$,

$$\begin{aligned}
0 = E[\{S - E(S|X, Z)\} \{Z t_1(X) \\
+ t_0(X)(1 \\
- Z)\} \{S \\
- E(S|Z, X)\} u_0(X)^{1-Z} b(X)]
\end{aligned}$$

which after some calculations can be seen to be equivalent to

$$\begin{aligned}
t_0(X) &= -t_1(X) E(\omega^*|S \\
&= 1, Z=0, X) E(Z|X) \{1 - E(Z|X)\}^{-1}
\end{aligned}$$

from where we can write

$$\begin{aligned}
\{S - E(S|X, Z)\} \{Z t_1(X) + t_0(X)(1 - Z)\} &= \{E(\omega^*|S=1, Z=0, X)^{1-Z} - \omega^{*1-Z}\} t_1^*(X) S \{Z - E(Z|X)\} \\
&+ t_1^*(X) \{Z - E(Z|X)\} S \omega^{*1-Z} - t_1^*(X) E(S|X, Z) \\
&\times E(\omega^*|S=1, Z=0, X)^{1-Z} \{Z - E(Z|X)\}
\end{aligned}$$

with $t_1^*(X) = t_1(X) \{1 - E(Z|X)\}^{-1}$. But the first term of the right-hand side of the last identity is of the form $D_1 - E(D_1|X, Z, S)$. Also the last term is of the form $d_2(X) \{Z - E(Z|X)\}$ because $E(S|X, Z) E(\omega^*|S=1, Z=0, X)^{1-Z}$ does not depend on Z . Finally, the postulated form for $\Lambda_s^\perp$ is obtained after noting that $t_1^*(X)$ is unrestricted because $t_1(X)$ is unrestricted.

We now characterize the elements of $\Lambda_s^\perp$ that are also in $\Lambda_0^\perp \cap \Lambda_1^\perp$. First note that the elements of Λ_0 and Λ_1 are linear combinations of (or limits of linear combinations of) scores for submodels for $f(Y|Z, S, X)$ and/or $f(S|Z, X)$. Therefore, they are uncorrelated with $\{Z - E(Z|X)\} d_2(X) + h(X) - E\{h(X)\}$ since this is a function of (X, Z) only. In addition, it is straightforward to check that the elements of Λ_0 and Λ_1 are also uncorrelated with random variables of the form $S \omega^{*1-Z} \{Z - E(Z|X)\} d_3(X)$. Thus, it remains to find the subset of the random variables of the form $D_1^* \equiv D_1 - E(D_1|X, Z, S)$ that are orthogonal to both Λ_0 and Λ_1 . Now, write $D_1^* = SZ \{M_1 - E(M_1|X, Z=1, S=1)\} + S(1 - Z) \{M_0 - E(M_0|X, Z=0, S=1)\}$ for some $M_1 \equiv m_1(Y,$

X) and $M_0 \equiv m_0(Y, X)$. It can be easily checked that D_1^* is uncorrelated with the elements of Λ_0 if and only if, for all $A_0 \equiv a_0(\varepsilon_0, X)$ defined as in the set Λ_0 , $0 = E[S(1-Z)\{M_0 - E(M_0|X, Z=0, S=1)\}A_0]$. Equivalently, M_0 satisfies $0 = \text{cov}\{m_0(Y, X), a_0(\varepsilon_0, X)|Z=0, S=1, X\}$ for all A_0 . Reasoning again as in Chamberlain (1987) or Robins and Rotnitzky (1995) we conclude that there is a $d_0(X)$ such that $m_0(Y, X) = d_0(X)\omega^*\varepsilon_0$. Similarly, we can arrive at the expression $m_1(Y, X) = d_1(X)\varepsilon_1$ for some $d_1(X)$. We conclude that the elements of $\Lambda_s^\perp \cap \Lambda_0^\perp \cap \Lambda_1^\perp$ are of the form

$$S(1-Z)d_0(X)\omega^*\varepsilon_0 + SZd_1(X)\varepsilon_1 + \{Z - E(Z|X)\}d_2(X) + h(X) - E\{h(X)\} + [S\omega^{*1-Z} - E\{S\omega^{*1-Z}|X\}] \times \{Z - E(Z|X)\}d_3(X).$$

The subset of these which are also orthogonal to the elements of Λ_x is of the form

$$\begin{aligned} & S(1-Z)d_0(X)\omega^*\varepsilon_0 \\ & + SZd_1(X)\varepsilon_1 \\ & + \{Z - E(Z|X)\}d_2(X) \\ & + \{S\omega^{*1-Z} \\ & - E(S\omega^{*1-Z}|X)\}\{Z - E(Z|X)\}d_3(X) \end{aligned}$$

for arbitrary $d_0(X)$, $d_1(X)$, $d_2(X)$ and $d_3(X)$, or equivalently the form $h(O; d, \alpha^*, \gamma^*) = d(X)q(O)$. That the set of influence functions is as postulated follows immediately from this expression. This concludes the proof of part (a) of theorem 3.

We now prove part (c) before turning to the proof of part (b). In model $\mathcal{C}(g)$, the orthogonal complement to the nuisance tangent set is the subset of the elements of the form as in the last display that are also uncorrelated with scores for $\Pr(Z|X)$. We now show that this subset is comprised of all random variables of the form

$$\begin{aligned} & S(1-Z)d_0(X)\omega^*\varepsilon_0 \\ & + SZd_1(X)\varepsilon_1 \\ & + \{S\omega^{*1-Z} \\ & - E(S\omega^{*1-Z}|X)\}\{Z - E(Z|X)\}d_3(X) \end{aligned} \quad (18)$$

for arbitrary $d_1(X)$, $d_0(X)$ and $d_3(X)$. This follows by noting that

- a. a score for $\Pr(Z|X)$ has the form $\{Z - E(Z|X)\}t(X)$ for some $t(X)$,
- b. $S(1-Z)d_0(X)\omega^*\varepsilon_0 + SZd_1(X)\varepsilon_1$ has mean 0 given (X, Z) and is therefore uncorrelated with $\{Z - E(Z|X)\}t(X)$ and
- c. $\{Z - E(Z|X)\}d_2(X) + S\omega^{*1-Z}\{Z - E(Z|X)\}d_3(X)$ is uncorrelated with $\{Z - E(Z|X)\}t(X)$ if and only if

$$\begin{aligned} 0 &= d_2(X)E[\{Z - E(Z|X)\}^2|X] + d_3(X)E[E(S\omega^{*1-Z}|X)\{Z - E(Z|X)\}^2|X] \\ &= E[\{Z - E(Z|X)\}^2|X]\{d_2(X) + d_3(X)E(S\omega^{*1-Z}|X)\} \end{aligned}$$

from where it follows that $d_2(X) = -d_3(X)E(S\omega^{*1-Z}|X)$.

The proof of part (c) is concluded by noting that random variables of the form (18) can be written as $h(O; d, \alpha^*, \gamma^*)$ with the last column of $d(X)$ equal to zero. Turning now to part (b), the efficient score for (α, γ) is the element $d_{\text{eff}}(X) q(O; \alpha^*, \gamma^*)$ of the nuisance tangent space which satisfies $\text{var} \{d_{\text{eff}}(X) q(O; \alpha^*, \gamma^*)\}^{-1} \geq \text{var} \{h(O; d, \alpha^*, \gamma^*)\}$ for all $(p+q) \times 4$ functions $d(X)$. It is easily checked that this inequality is satisfied when $d_{\text{eff}}(X)$ is as in equation (10). This concludes the proof of part (b).

References

- Angrist J, Imbens GW, Rubin DB. Identification of causal effect using instrumental variables. *J. Am. Statist. Ass* 1996;91:444–455.
- Bickel, JP.; Klaassen, CAJ.; Ritov, Y.; Wellner, JA. Efficient and Adaptive Estimation for Semiparametric Models. Baltimore: Johns Hopkins University Press; 1993.
- Chamberlain G. Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometr* 1987;34:305–334.
- Cox, DR. Planning of Experiments. New York: Wiley; 1958.
- Dawid AP. Conditional independence in statistical theory (with discussion). *J. R. Statist. Soc. B* 1979;41:1–31.
- Dawid AP. Causal inference without counterfactuals. *J. Am. Statist. Ass* 2000;95:407–437.
- Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002;58:21–29. [PubMed: 11890317]
- Gilbert PB, Bosch RJ, Hudgens MG. Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* 2003;59:531–541. [PubMed: 14601754]
- Hayden D, Pauler DK, Schoenfeld D. An estimator for treatment comparisons among survivors in randomized trials. *Biometrics* 2005;61:305–310. [PubMed: 15737107]
- Hudgens MG, Hoering A, Self SG. On the analysis of viral load endpoints in HIV vaccine trials. *Statist. Med* 2003;22:2281–2298.
- Jemiai, Y. Doctoral Dissertation. Boston: Department of Biostatistics, Harvard School of Public Health; 2005. Asymptotic properties of an estimator of treatment effects on an outcome only existing if a post-randomization event has occurred.
- Jemiai, Y.; Rotnitzky, A. Sharp bounds and sensitivity analysis for treatment effects in the presence of censoring by death. Schering-Plough Wrkshp Development and Approval of Oncology Drug Products. 2003. Available from <http://www.hsph.harvard.edu/live/sp-2003-rotnitzky.html>
- Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. New York: Wiley; 1980.
- Klaassen CAJ. Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist* 1987;15:1548–1562.
- Newey WK. Semiparametric efficiency bounds. *J. Appl. Econometr* 1990;5:99–135.
- Neyman J. On the application of probability theory to agricultural experiments: essay on principles, section 9 (Engl. transl.). *Statist. Sci* 1990;5:465–480.
- rgp120 HIV Vaccine Study Group. Placebo-controlled trial of a recombinant glycoprotein 120 vaccine to prevent HIV infection. *J. Infect. Dis* 2005;191:654–665. [PubMed: 15688278]
- Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Math. Modllng* 1986;7:1393–1512.
- Robins JM. An analytic method for randomized trials with informative censoring: part I. *Lifetime Data Anal* 1995;1:241–254.
- Robins JM, Greenland S. Comment on “Causal inference without counterfactuals,” by A. P. Dawid. *J. Am. Statist. Ass* 2000;95:431–435.
- Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 1992;48:479–495. [PubMed: 1637973]
- Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Statist. Ass* 1995;90:122–129.
- Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann. Statist* 1978;6:34–58.

- Rubin DB. Statistics and causal inference: which ifs have causal answers. *J. Am. Statist. Ass* 1986;81:961–962.
- Rubin DB. More powerful randomization-based p -values in double-blind trials with noncompliance. *Statist. Med* 1998;17:371–389.
- Rubin DB. Comment on “Causal inference without counterfactuals,” by A. P. Dawid. *J. Am. Statist. Ass* 2000;95:435–437.
- Rubin DB. Causal inference through potential outcomes and principal stratification: application to studies with ‘censoring’ due to death. *Statist. Sci* 2006;21:299–309.
- Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semipara-metric nonresponse models. *J. Am. Statist. Ass* 1999;94:1096–1120.
- Shepherd BE, Gilbert PB, Jemiai Y, Rotnitzky A. Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics* 2006;62:332–342. [PubMed: 16918897]
- Zhang JL, Rubin DB. Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. *J. Educ. Behav. Statist* 2003;28:353–368.

Fig. 1. Locally efficient estimates of (a) γ_4^* and (b) γ_5^* as a function of the sensitivity parameter β and their associated pointwise 95% confidence intervals

Fig. 2.

$$m(1, \text{Age}, \text{Risk}) - m(0$$

Locally efficient estimate of the vaccine effect $m(1, \text{Age}, \text{Risk}) = \gamma_3 + \gamma_4^* \text{Age} + \gamma_5^* \text{Risk}$ for various values of age and risk score as a function of the sensitivity parameter β and their associated pointwise 95% confidence intervals

Table 1

Mean and median of $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)$, coverage probability CP and median length of the associated 95% confidence intervals in two 1000-run simulation studies with randomization probability $P(Z=1) = 0.5$, probability of the event $\mathbf{E}$ under treatment 0 of $P(S=1|Z=0) = 0.25$, covariate distribution $X \sim N(38, 36)$ and outcome variance $\text{var}(Y|X) = 1$

Method	β	Results for $\gamma_0 = 2.3$				Results for $\gamma_1 = 0.05$				Results for $\gamma_2 = 0$				Results for $\gamma_3 = 0$			
		Mean	Median	CP	Length	Mean	Median	CP	Length	Mean	Median	CP	Length	Mean	Median	CP	Length
$n = 20000$																	
OLS	0.1	2.21	2.21	0.91	0.50	0.05	0.05	0.93	0.01	0.09	0.08	0.93	0.81	0	0	0.95	0.02
INE	0.1	2.30	2.30	0.95	0.52	0.05	0.05	0.96	0.01	0	-0.01	0.96	0.81	0	0	0.96	0.02
EFF	0.1	2.30	2.30	0.96	0.50	0.05	0.05	0.95	0.01	0	-0.01	0.96	0.81	0	0	0.95	0.02
OLS	1	1.09	1.08	0	0.55	0.07	0.07	0	0.01	1.21	1.22	0	0.86	-0.02	-0.02	0.01	0.02
INE	1	2.30	2.30	0.95	0.72	0.05	0.05	0.94	0.02	0	0.02	0.95	0.98	0	0	0.95	0.02
EFF	1	2.30	2.30	0.95	0.69	0.05	0.05	0.95	0.02	0.01	0.01	0.96	0.96	0	0	0.96	0.02
OLS	3	-3.05	-3.04	0	0.76	0.16	0.16	0	0.02	5.34	5.34	0	1.05	-0.11	-0.11	0	0.03
INE	3	2.29	2.29	0.87	1.33	0.05	0.05	0.85	0.03	0.01	0	0.89	1.52	0	0	0.88	0.04
EFF	3	2.28	2.28	0.95	1.12	0.05	0.05	0.95	0.03	0.01	0.01	0.95	1.34	0	0	0.95	0.03
$n = 2000$																	
OLS	0.1	2.21	2.21	0.93	1.57	0.05	0.05	0.94	0.04	0.10	0.12	0.94	2.53	0	0	0.94	0.06
INE	0.1	2.30	2.29	0.94	1.64	0.05	0.05	0.94	0.04	0.01	0.03	0.94	2.58	0	0	0.94	0.07
EFF	0.1	2.29	2.29	0.92	1.60	0.05	0.05	0.94	0.04	0.02	0.04	0.95	2.55	0	0	0.94	0.07
OLS	1	1.09	1.09	0.23	1.70	0.07	0.07	0.46	0.04	1.22	1.26	0.55	2.68	-0.02	-0.02	0.72	0.07
INE	1	2.27	2.26	0.94	2.31	0.05	0.05	0.94	0.06	0.04	0.07	0.94	3.11	0	0	0.94	0.08
EFF	1	2.23	2.24	0.93	2.18	0.05	0.05	0.94	0.06	0.08	0.12	0.94	3.03	0	0	0.94	0.08
OLS	2	-0.78	-0.79	0	2.02	0.11	0.11	0	0.05	3.10	3.12	0.03	2.99	-0.06	-0.06	0.10	0.08
INE	2	2.17	2.18	0.95	3.19	0.05	0.05	0.94	0.08	0.14	0.14	0.94	3.90	0	0	0.94	0.10
EFF	2	2.14	2.15	0.94	2.96	0.05	0.05	0.94	0.07	0.17	0.20	0.93	3.69	0	-0.01	0.93	0.09
OLS	3	-3.04	-3.03	0	2.37	0.16	0.16	0	0.06	5.36	5.37	0	3.31	-0.11	-0.11	0	0.08
INE	3	2.02	2.06	0.89	4.21	0.06	0.06	0.89	0.10	0.30	0.31	0.92	4.76	-0.01	-0.01	0.92	0.11
EFF	3	1.85	1.99	0.90	3.87	0.06	0.06	0.90	0.09	0.47	0.35	0.91	4.55	-0.01	-0.01	0.92	0.11

Table 2

Mean and median of $(\hat{\alpha}_0, \hat{\alpha}_1)$, coverage probability CP and median length of the associated 95% confidence interval in two 1000-run simulation studies with randomization probability $P(Z = 1) = 0.5$, probability of the event **E** under treatment 0 of $P(S = 1|Z = 0) = 0.25$, covariate distribution $X \sim N(38, 36)$ and outcome variance $\text{var}(Y|X) = 1$

Method	β	Results for a_0				Results for $a_1 = 0.069$			
		Mean	Median	CP	Length	Mean	Median	CP	Length
$n = 20000$									
$a_0 = -2.2$									
INE	0.1	-2.24	-2.22	0.96	2.98	0.07	0.07	0.97	0.08
EFF	0.1	-2.19	-2.21	0.96	2.22	0.07	0.07	0.95	0.06
$a_0 = -5.5$									
INE	1	-5.56	-5.50	0.92	3.38	0.07	0.07	0.92	0.10
EFF	1	-5.46	-5.44	0.95	2.63	0.07	0.07	0.94	0.07
$a_0 = -9.9$									
INE	3	-10.84	-9.90	0.80	13.68	0.10	0.07	0.81	0.40
EFF	3	-9.56	-9.63	0.92	7.29	0.06	0.06	0.93	0.22
$n = 2000$									
$a_0 = -2.2$									
INE	0.1	-2.74	-2.47	0.98	9.93	0.09	0.08	0.99	0.28
EFF	0.1	-2.14	-2.05	0.96	7.38	0.07	0.06	0.96	0.21
$a_0 = -5.5$									
INE	1	-6.39	-5.78	0.97	11.34	0.10	0.07	0.98	0.32
EFF	1	-5.29	-5.24	0.95	8.56	0.06	0.06	0.95	0.24
$a_0 = -8.2$									

Method	β	Results for a_0				Results for $a_1 = 0.069$			
		Mean	Median	CP	Length	Mean	Median	CP	Length
INE	2	-9.67	-8.16	0.93	18.88	0.11	0.07	0.93	0.54
EFF	2	-7.69	-7.62	0.92	12.64	0.05	0.05	0.92	0.36
$a_0 = -9.9$									
INE	3	-11.51	-8.13	0.84	37.24	0.12	0.01	0.86	1.09
EFF	3	-7.75	-7.99	0.87	24.31	0.01	0.01	0.85	0.71