

Association, Causation, and Marginal Structural Models

James M. Robins

Professor of Epidemiology and Biostatistics

Harvard School of Public Health

Boston, MA 02115

robins@epinet.harvard.edu

1. Introduction

The subject-specific data from a longitudinal study consist of a string of numbers. These numbers represent a series of empirical measurements. Calculations are performed on these strings and causal inferences are drawn. For example, an investigator might conclude that the analysis provides strong evidence for “a direct effect of AZT on the survival of AIDS patients controlling for the intermediate variable – therapy with aerosolized pentamidine.” The nature of the relationship between the sentence expressing these causal conclusions and the computer calculations performed on the strings of numbers has been obscure. Since the computer algorithms are well-defined mathematical objects, it is useful to provide formal mathematical definitions for the English sentences expressing the investigator’s causal inferences. In Robins (1986, 1987), I proposed a formal theory of counterfactual (Lewis, 1973) causal inference that extended the Neyman-Rubin-Holland (Holland, 1986) “point treatment” theory to longitudinal studies with time-varying treatments, outcomes, and covariates (concomitants). This theory translates any causal question concerning the overall (net), direct, and/or indirect effects of a possibly time-varying treatment on an outcome into a formal mathematical conjecture about event trees, referred to as causally interpreted structured tree graphs.

Pearl (1995), and Spirtes, Glymour, and Scheines (SGS) (1993) recently developed a formal theory of causal inference based on causal directed acyclic graphs (DAGs). I showed that these causal DAGs are mathematically equivalent to a particular special case of my more general theory (Robins, 1995).

In longitudinal studies, treatment often varies over time. The standard approach to the estimation of the effect of a time-varying treatment on an outcome of interest is to model the outcome at time t as a function of past treatment history. I have shown that this approach may be biased, whether or not one further adjusts for the past history of time-dependent confounding covariates, when these covariates predict subsequent outcome and treatment history and are themselves influenced by past treatment (Robins, 1986). In this setting, I have proposed several methods that can provide, under certain assumptions, valid estimates of the causal effect of a time-varying treatment in the presence of time varying confounding factors. In Section 2, I describe one of these methods of estimation: inverse-probability-of-treatment weighted (IPTW) estimation of the parameters of a marginal structural model (MSM) (Robins, 1998, 1999). These models are particularly useful in clarifying the difference between association and causation, requiring that the reader have only a working knowledge of ordinary linear regression. IPTW estimation can consistently estimate the causal effect of a time-dependent treatment only if all relevant confounding factors have been measured. In an observational study, the data provide no evidence as to either the existence or the magnitude of confounding by additional unmeasured factors. In view of this fact, a data analyst should perform a “sensitivity analysis” to quantify how one’s inference

concerning the causal effect of treatment varies as a function of the magnitude of confounding due to unmeasured factors. In Section 3, I describe how to conduct such a sensitivity analysis.

My goal in this article is to indicate to non-statistical readers the sort of judgements and statistical tools required by practicing epidemiologists-statisticians when attempting to evaluate the evidence for a causal effect of a time-varying treatment, in the context of a single observational study. I do not directly consider the problem of combining epidemiologic evidence across studies or with other types of evidence. In my presentation, I purposely strike an attitude of indifference or obliviousness to the philosophical problems raised by the methods I describe. For example, I freely use counterfactual outcomes without regard to the discomfort that raises for certain philosophers and statisticians. My attitude is much like that of a practicing physicist – these are the tools and concepts I require to get the job that I need to do done. I will describe clearly the job that needs doing and the tools that I and others have developed to do it. This is not to say that I myself do not have my own strong views about the philosophical underpinnings of the concepts I use. Rather, I choose to have my own point of view take a back seat to the task of describing the methods used by myself and collaborators to evaluate causal effects.

2. Marginal Structural Models

2.1. A Regression (Association) Model

I now give a somewhat informal introduction to marginal structural models. I begin with the following setting. Consider an observational study of AIDS patients. Let $A(t)$ be the dose of a treatment of interest, say AZT, at time t with time measured as days since start of follow-up. Let Y be an outcome of interest measured at end-of-follow-up at time $K + 1$. Suppose the study investigator states that his goal is to estimate the causal effect of the time-dependent treatment $A(t)$ on the mean of Y . Let $\bar{A}(t) = \{A(u); 0 \leq u \leq t\}$ be treatment history through t and let $\bar{L}(t) = \{L(u); 0 \leq u \leq t\}$ be the history through t of all measured prognostic factors $L(u)$ for (i.e., predictors of) Y , such as CD4 lymphocyte count, white blood count (WBC), hematocrit, age, gender, etc. Both $A(u)$ and $L(u)$ are recorded daily so the functions $\bar{A}(t)$ and $\bar{L}(t)$ can jump at most once per day. We assume that a decision whether to take treatment on day t is made after knowledge of the covariates recorded in $L(t)$ becomes available, so that $A(t)$ is temporally actually subsequent to $L(t)$. To keep matters simple, we unrealistically assume there is no missing data. In particular, no subject dies prior to day $K + 1$. Suppose Y is a continuous outcome (e.g., Y is the number of milligrams of HIV RNA detectable in a cubic centimeter of blood), and we entertain a model that says the mean (i.e., expectation) of Y given AZT history, $\bar{A} \equiv \bar{A}(K + 1)$, is a linear function of a subject's cumulative AZT dose. We write the model

$$E[Y | \bar{A}] = g(\bar{A}; \gamma) \tag{1}$$

where

$$g(\bar{A}; \gamma) = \gamma_1 + \gamma_2 cum(\bar{A}) . \tag{2}$$

E stands for the expectation operator and $cum(\bar{A}) = \int_0^{K+1} A(t) dt = \sum_{t=0}^K A(t)$ is the subject's cumulative treatment. (As in much of the causality literature, I regard the n subjects as randomly drawn from a near-infinite hypothetical superpopulation of subjects about whom we wish to make inference. Expectations refer to averages in the superpopulation and probability statements to proportions in the superpopulation.) The ordinary least squares (OLS) estimator of γ can then be computed from the ob-

served data $O_i = (\bar{L}_i, \bar{A}_i, Y_i)$, $i = 1, \dots, n$, on the n study subjects using standard software with Y as the outcome variable and $cum(\bar{A})$ as the regressor. That is, the OLS estimator $\hat{\gamma}$ of $\gamma = (\gamma_1, \gamma_2)'$ minimizes the residual sum of squares $\sum_{i=1}^n [Y_i - \gamma_1 - \gamma_2 cum(\bar{A}_i)]^2$. Note that the residual sum of squares does not depend on the patient's prognostic factor history $\bar{L}_i \equiv \bar{L}_i(K+1)$. The OLS estimator $\hat{\gamma}_2$ is an unbiased estimator of the regression (association) parameter γ_2 . We have assumed that the above linear form for the regression function $g(\bar{A}; \gamma)$ is correct, even though, in reality, any model will be incorrect, and will, at best, serve as a good approximation to the unknown regression function $E[Y | \bar{A}]$.

2.2. Causal and Statistical Exogeneity

The question then is when does γ_2 have an interpretation as the causal effect of treatment history on the mean of Y ? To approach this question, imagine that the decision to administer treatment at each time t were made totally at random by the treating physician and all subjects took their prescribed treatment. In that hypothetical case, giving treatment at time t is not expected to be associated with any measured or unmeasured prognostic factors (i.e., in the parlance of epidemiologists, there would be no “confounding”) and therefore γ_2 would intuitively have a causal interpretation. More generally, whenever the conditional probability of receiving treatment on day t given past treatment and prognostic factors history (measured and unmeasured) depends only on past treatment history, we say the treatment process is causally “exogenous” (equivalently, “ancillary”). A formal mathematical definition is provided below after we define counterfactual outcomes. It is well-recognized in the social sciences, econometrics, epidemiologic, and biostatistical literature that γ_2 will have a causal interpretation if $A(t)$ is a causally exogenous (or ancillary) covariate process.

We say that a treatment $A(t)$ is a statistically “exogenous” (“ancillary”) process if the probability of receiving treatment at time t does not depend on the history of measured time-dependent prognostic factors $\bar{L}(t)$ up to t conditional on treatment history prior to t , i.e.,

$$\bar{L}(t) \perp\!\!\!\perp A(t) \mid \bar{A}(t-1), \quad (3)$$

where $A \perp\!\!\!\perp B \mid C$ means that A is independent of B given C . An essentially necessary condition for $A(t)$ to be “causally exogenous” is for it to be “statistically exogenous.” However, that a process is “statistically exogenous” does not imply it is “causally exogenous,” because there may be unmeasured prognostic factors (i.e., confounders) that predict the probability of treatment $A(t)$ at time t given past treatment history. We can test from the data whether $A(t)$ is statistically exogenous but are unable to test whether a statistically exogenous process is causally exogenous. We warn the reader that there is no agreed upon definition of “causally exogenous” or “statistically exogenous” in the literature. I find my definition quite useful and appropriate, but there are other definitions. In particular, the definitions I have given here do not agree with the definition of exogeneity found in the econometric time series literature (Ericsson, Hendry, and Mizon, 1998).

Suppose $A(t)$ is discrete and we can correctly model both the probability $f[a(t) \mid \bar{\ell}(t), \bar{a}(t-1)]$ of taking treatment $a(t)$ on day t as a function of past treatment $\bar{a}(t-1)$ and measured prognostic factor history $\bar{\ell}(t)$, and the probability $f[a(t) \mid \bar{a}(t-1)]$ of taking treatment $a(t)$ on day t as a function only of past treatment $\bar{a}(t-1)$ history. Here we use the convention that random variables (i.e., variables whose values can differ from subject to subject) are denoted by upper case letters. Lower case letters denote possible values of the corresponding random variables. Thus, for example, $f[a(t) \mid \bar{a}(t-1)]$ is the proportion of subjects in the superpopulation with treatment $A(t)$ equal to $a(t)$

among subjects with past treatment history $\bar{A}(t-1)$ equal to $\bar{a}(t-1)$. We could then measure the degree to which the treatment process is statistically non-exogenous through day t by the random quantity

$$\mathcal{W}(t) = \prod_{k=0}^t \left\{ f[A(k) | \bar{A}(k-1), \bar{L}(k)] / f[A(k) | \bar{A}(k-1)] \right\} .$$

Informally, the numerator in each term in $\mathcal{W}(t)$ is the probability that a subject received his own observed treatment at time k , $A(k)$, given his own past treatment and prognostic factor history. The denominator is the probability that a subject received his observed treatment conditional on his past treatment history but not further adjusting for his past prognostic factor history. Note that the treatment process is statistically exogenous just in the case that $\mathcal{W}(t) = 1$ for all t . Formally, $f[A(k) | \bar{A}(k-1), \bar{L}(k)]$ and $f[A(k) | \bar{A}(k-1)]$ are random variables obtained by replacing $a(k)$, $\bar{a}(k-1)$ and $\bar{l}(k)$ with the corresponding random variables in the functions $f[a(k) | \bar{a}(k-1), \bar{l}(k)]$ and $f[a(k) | \bar{a}(k-1)]$. Of course, in an observational study, $\mathcal{W}(t)$ is usually unknown and will have to be estimated from the data by specifying and fitting statistical models for the terms in the numerator and denominator of $\mathcal{W}(t)$. However, for pedagogic purposes, assume that $\mathcal{W}(t)$ were known.

2.3. IPTW Estimators

When $A(t)$ is a statistically non-exogenous process, we shall consider estimation by weighted least squares regression in which a subject is given the weight

$$\mathcal{W}^{-1} \equiv [\mathcal{W}(K)]^{-1} .$$

The weighted regression estimator, say $\hat{\beta}$, minimizes the weighted residual sum of squares $\sum_{i=1}^n \mathcal{W}_i^{-1} [Y_i - \gamma_1 - \gamma_2 \text{cum}(A_i)]^2$ and can be computed using standard off-the-shelf software packages. We shall refer to this weighted regression estimator as an inverse-probability-of-treatment weighted (IPTW) estimator. This weighted regression estimator would agree with the usual unweighted estimator $\hat{\gamma}$ just in the case in which $A(t)$ is exogenous. The somewhat surprising result described in detail below is that, if the vector of prognostic factors recorded in $L(t)$ constitutes all relevant time-dependent prognostic factors (i.e., confounders), then, whether or not the treatment process is statistically exogenous, the weighted regression estimator of γ_2 will unbiasedly estimate a quantity β_2 that can be appropriately interpreted as the causal effect of treatment history on the mean of Y . In contrast, when $A(t)$ is statistically non-exogenous, the OLS regression estimator will still estimate the parameter γ_2 , but now γ_2 will have no causal interpretation.

To prove such a claim, we need to give a formal mathematical meaning to the informal concept of the causal effect of treatment history on the mean of Y . To do so, we reinforce some notational conventions we have already been using. We use capital letters to represent random variables and lower case letters to represent possible realizations (values) of random variables. For example, O_i is the random observed data for the i^{th} study subject and o is a possible realization (value) of O_i . Further, we assume that the random vector O_i for each subject is drawn independently from a distribution common to all subjects, i.e. the O_i are independent and identically distributed.. Because the O_i have the same distribution, we often suppress the i subscript. Note the aforementioned superpopulation model implies that the O_i are independent and identically distributed.

2.4. A Marginal Structural Model

Now we introduce counterfactual or potential outcomes. For any fixed non-random treatment history $\bar{a} = \{a(u); 0 \leq u \leq K + 1\}$, let $Y_{\bar{a}}$ be the (possibly counterfactual) random variable representing a subject's outcome had, possibly contrary to fact, the subject been treated with history \bar{a} rather than his observed history \bar{A} . Note the \bar{a} 's are possible realizations of the random variable \bar{A} . For each possible history \bar{a} , we are assuming a subject's response $Y_{\bar{a}}$ is well defined (although generally unobserved). Indeed we only observe $Y_{\bar{a}}$ for that treatment history \bar{a} equal to a subject's actual treatment history \bar{A} , i.e.,

$$Y = Y_{\bar{A}}. \quad (4)$$

The above identity is the fundamental "consistency" assumption that links the counterfactual data $Y_{\bar{a}}$ to the observed data (Y, \bar{A}) . Note that, if on each day t , $A(t)$ can take but one of two values (0 for untreated and 1 for treated) and the study duration K is 300 days, then there are 2^{300} different $Y_{\bar{a}}$ values associated with each subject.

Formally our statement that the effect of treatment history on the mean of Y is a linear function of cumulative treatment is the statement that, for each \bar{a} ,

$$E[Y_{\bar{a}}] = g(\bar{a}; \beta), \text{ where } g(\bar{a}; \beta) = \beta_1 + \beta_2 \text{ cum}(\bar{a}), \quad (5)$$

which we refer to as a marginal structural model (MSM) for the effect of treatment on the mean of Y . This model for $E[Y_{\bar{a}}]$ is a marginal structural model since it is a model for the marginal distribution of counterfactual variables (rather than for the joint distribution, e.g., the correlation matrix, of the $Y_{\bar{a}}$) and, in the econometric and social science literature, causal models (i.e., models for counterfactual variables) are often referred to as structural. Note that a MSM is a model for the overall (i.e., net) effect of the treatment history \bar{a} on the outcome Y , since it is oblivious to particular causal pathways or mechanisms by which the treatment has its effect. The relationship of our MSM (5) to our regression model (1)-(2) can be clearly seen, by expressing our regression model as

$$E[Y_{\bar{a}} | \bar{A} = \bar{a}] = g(\bar{a}; \gamma) \text{ where } g(\bar{a}; \gamma) = \gamma_1 + \gamma_2 \text{ cum}(\bar{a}). \quad (6)$$

Note (6) is equivalent to (1)-(2) since, by (4), we can substitute Y for $Y_{\bar{a}}$ in (6) and obtain $E[Y | \bar{A} = \bar{a}] = g(\bar{a}; \gamma)$ which is equivalent to (1). From (6), we see that a regression model is a model for the conditional mean of $Y_{\bar{a}}$ given $\bar{A} = \bar{a}$.

We now show that the parameter β_2 encodes the magnitude of the average causal effect of the treatment on the outcome. By definition, the causal effect of treatment regime \bar{a} on the outcome Y for a given study subject is the difference $Y_{\bar{a}} - Y_{\bar{0}}$ between her outcome $Y_{\bar{a}}$ when treated with regime \bar{a} and her outcome $Y_{\bar{0}}$ when never treated. Here $\bar{0}$ is the $K + 1$ vector of 0's. Thus $E[Y_{\bar{a}} - Y_{\bar{0}}] = E[Y_{\bar{a}}] - E[Y_{\bar{0}}]$ is the average causal effect of regime \bar{a} in the superpopulation, which under our MSM (5) is $\beta_2 \text{ cum}(\bar{a})$.

β_2 is also of important policy interest. To see why, consider a new subject exchangeable with (i.e., drawn from the same distribution as) the n study subjects. We must decide which treatment history \bar{a} to administer to the new subject. We would like to provide the treatment that minimizes the expected amount of HIV RNA in his blood at end of follow-up. That is, we want to find \bar{a} that minimizes $E[Y_{\bar{a}}]$. [Technically, we would choose to minimize $E(Y_{\bar{a}})$ under a squared error loss function.] Thus, for example, if the parameter β_2 of our causal model is positive, we will withhold AZT treatment from our subject (i.e., we will give him the treatment history $\bar{0}$), since positive β_2 indicates that the expected amount of HIV

RNA in one’s blood at the end of follow-up increases with increasing cumulative AZT dose. In contrast to β_2 , the parameter γ_2 of our association (regression) model (1) may have no causal interpretation. For example, suppose physicians preferentially started AZT on subjects who, as indicated by their prognostic factor history, were doing poorly and that AZT has no causal effect on the mean of Y (i.e., $\beta_2 = 0$). Nonetheless, the mean of Y will increase with cumulative AZT dose (since patients with poor prognostic factor history, say low white blood count, will have higher levels of HIV RNA and will have received more AZT treatment). Thus γ_2 will be positive. In this setting, we say that the parameter γ_2 of the association model lacks a causal interpretation because it is confounded by the association of the prognostic factors $\bar{L}(u)$ with the subsequent treatment $A(u)$. If we made policy decisions as to the optimal AZT dose based on the parameter γ_2 rather than β_2 , we may well be doing many of our patients a potentially fatal disservice. For example γ_2 might be positive even if AZT was beneficial and thus β_2 was negative, if the selection bias (i.e., confounding) due to physicians preferentially treating subjects with low white blood counts is of greater magnitude than the true beneficial effect of AZT on Y as measured by the absolute value of β_2 .

2.5. The Assumption of No Unmeasured Confounders

Formally, in terms of counterfactuals, we say that the $A(t)$ process is “causally exogenous” if, for all histories \bar{a} , $Y_{\bar{a}}$ is independent of the dose of treatment on day t given past treatment history, i.e.,

$$Y_{\bar{a}} \perp\!\!\!\perp A(t) \mid \bar{A}(t-1) \tag{7}$$

which is mathematically equivalent to the statement that $Y_{\bar{a}}$ is independent of the subject’s entire treatment history, i.e.,

$$Y_{\bar{a}} \perp\!\!\!\perp \bar{A} . \tag{8}$$

Note that even when $A(t)$ is “causally exogenous,” if the treatment has an effect on the outcome, then the observed outcome $Y = Y_{\bar{A}}$ will not be independent of \bar{A} , since $Y_{\bar{A}}$ is a function of a subject’s observed treatment history \bar{A} itself. When $A(t)$ is causally exogenous, we say there is no confounding by either measured or unmeasured factors.

Remark: In the context of the causal DAG theory of Pearl (1995) and Spirtes et al. (1993), we would say a treatment process is causally exogenous if there are no arrows into any of the treatment variables on a causally sufficient DAG representing the data, except for those originating from other treatment variables. This definition implies but is not implied by my definition (7). It is only my weaker definition (7) that is relevant when considering the causal effect of the treatment process \bar{A} on the outcome Y .

Given the covariates recorded in $L(t)$, we say there are no unmeasured confounders if, for each \bar{a} , $Y_{\bar{a}}$ is independent of the treatment $A(t)$ at time t given the past treatment and measured covariate history:

$$Y_{\bar{a}} \perp\!\!\!\perp A(t) \mid \bar{L}(t), \bar{A}(t-1) . \tag{9}$$

With these formalizations, it can then be shown mathematically, that when there are no unmeasured confounders, (i) statistical exogeneity (3) implies causally exogeneity (7), (ii) the IPTW estimator is unbiased for and converges in probability to the parameter β_2 of the marginal structural model (5) for $E[Y_{\bar{a}}]$, and (iii) the probability limit γ_2 of the usual OLS estimator generally differs from the causal parameter β_2 of the MSM unless the treatment process is statistically exogenous.

We shall also refer to the assumption of no unmeasured confounders as the assumption that treatment $A(t)$ is sequentially randomized given the past. The assumption states that, conditional on AZT history and the history of all recorded covariates prior to t , increments in AZT dosage rate at t are independent of the counterfactual random variables $Y_{\bar{a}}$. This assumption will be true if all prognostic factors for, i.e., predictors of, $Y_{\bar{a}}$ that are used by patients and physicians to determine the dosage of AZT at t are recorded in $\bar{L}(t)$ and $\bar{A}(t-1)$. For example, since physicians tend to withhold AZT from subjects with low white blood count (WBC), and in untreated subjects, low white blood count is a predictor of HIV RNA, the assumption of no unmeasured confounders would be false if $\bar{L}(t)$ does not contain WBC history. It is the primary goal of the epidemiologists conducting an observational study to collect data on a sufficient number of covariates to ensure that the assumption of no unmeasured confounders will be at least approximately true.

The assumption of no unmeasured confounders is the fundamental condition that will allow us to draw causal inferences from observational data. It is precisely because it cannot be guaranteed to hold in an observational study and is not empirically testable that it is so very hazardous to draw causal inferences from observational data.

On the other hand, the assumption of no unmeasured confounders is guaranteed to be true in a sequential randomized trial. A sequential randomized trial is a trial in which, at each time t , the dose of treatment is chosen at random by the flip of a coin, with the probability of a heads depending both on past measured covariate $\bar{L}(t)$ and treatment history $\bar{A}(t-1)$. It is because physical randomization guarantees the assumption that most people accept that valid causal inferences can be obtained from a randomized trial. See Rubin (1978), Robins (1986) and Holland (1986) for further discussion. Later we discuss how the consequences of violations of the assumption of no unmeasured confounders can be explored through sensitivity analysis. (To the reader who might try to read my early papers, the assumption of no unmeasured confounders was encoded in Robins (1987, pg. 327) in an event tree referred to as a CISTG randomized with respect to Y for treatment $g = \bar{a}$ given covariates \bar{L} .)

2.6. Why Weighting Controls Confounding

We now explain why weighting by \mathcal{W}^{-1} corrects our regression estimator for the “confounding” due to the measured prognostic factors in $L(t)$. The first point to note is that in the definition of $\mathcal{W}(t)$ we could have replaced the denominator $f[A(t) | \bar{A}(t-1)]$ by any other function of treatment history $\bar{A}(t)$ without influencing the consistency of our weighted estimator of the parameter β_2 of the MSM; only the efficiency (variance) of our estimator would be influenced. (An estimator is consistent for a parameter if, as the sample size increases to infinity, the estimate converges to the parameter in probability.) However, our estimator would be inconsistent if we replaced the numerator by any other function of $\bar{A}(t)$ and $\bar{L}(t)$. Thus one can view weighting by \mathcal{W}^{-1} as weighting by the inverse of a subject’s probability of having his own observed treatment history. This explains why we refer to our weighted regression estimator as an inverse-probability-of-treatment weighted (IPTW) estimator. Now view each person as a member of a pseudo- or ghost population consisting of themselves and $\mathcal{W}^{-1} - 1$ ghosts (copies) of themselves who have been added by weighting. In this new ghost or pseudo population, it is easy to show that $\bar{L}(t)$ does not predict treatment at t given past treatment history, and thus we have created a pseudo-population in which treatment is exogenous. Furthermore, the causal effect of \bar{A} on Y in the ghost population is the same as in the original population. That is, if $E[Y_{\bar{a}}] = g(\bar{a}; \beta)$ in the true population, the same will be true of the ghost population. Hence, we would like to do ordinary least squares regression in the pseudo-population. That is what our weighted regression estimator is doing, since the weights create, as

required, $\mathcal{W}^{-1} - 1$ additional copies of each subject.

Remark: In a sequential randomized trial not only is the assumption (9) of no unmeasured confounders guaranteed to hold, but in addition the treatment probabilities $f[a(k) | \bar{a}(k-1), \bar{\ell}(k)]$ are under the control of the investigator and thus known. However, the probabilities $f[a(k) | \bar{a}(k-1)]$ are not known and must be modelled and then estimated from the data. It follows from the above discussion that even if the model for $f[a(k) | \bar{a}(k-1)]$ is misspecified, the IPTW estimator of β is consistent in a sequential randomized trial. In an observational study, our IPTW estimator will be consistent only if we can specify a correct model for the unknown probabilities $f[a(k) | \bar{a}(k-1), \bar{\ell}(k)]$.

Remark: Fully parametric likelihood based inference: Our inverse-probability-of-treatment weighted estimator for the parameter β of our MSM is so simple to calculate that essentially any computer-literate practicing epidemiologist or social scientist could compute it using available statistical packages. Furthermore, it is guaranteed to be consistent for the parameter β of the MSM (5) when the data are from a sequentially randomized trial. However, from the point of view of a Bayesian or pure likelihoodist, our IPTW estimator seems inappropriate, since it requires that we either know or model the densities $f(A_k | \bar{L}_k, \bar{A}_{k-1})$ and, under the sequential randomization assumption (9), the likelihood function for the parameter β of our MSM does not depend on $f(A_k | \bar{L}_k, \bar{A}_{k-1})$. That is, our IPTW estimator of β_2 is not a likelihood-based estimator. However, as discussed in Appendix 1, a parametric likelihood-based or Bayesian estimator of the parameter β of our MSM can be computationally difficult or intractable, and, more importantly, will be inconsistent for β , even when treatment is statistically and causally exogenous and the $f[a(k) | \bar{a}(k-1), \bar{\ell}(k)] = f[a(k) | \bar{a}(k-1)]$ are known, as would be true in a sequential randomized trial in which the treatment probabilities actually only depended on past treatment history.

2.7. Why Standard Regression Adjustment Fails To Control Confounding

One might suppose that an alternative approach to controlling confounding by measured covariates is to fit by OLS a regression model that adjusts for confounder history $\bar{L} \equiv \bar{L}_K$ such as

$$E[Y | \bar{A}, \bar{L}] = \theta_0 + \theta_1 \text{cum}(\bar{A}) + \theta_2 \text{cum}(\bar{L}) + \theta_3 L(K) + \theta_4 L(K-1)$$

where, for notational convenience, we assume that $L(k)$ is univariate. However, even under the assumption of no unmeasured confounders, the parameter θ_1 can differ from the causal parameter β_2 of our MSM. This may not seem particularly disturbing, since one may hope that the parameter θ_1 represents the direct effect of treatment on Y not mediated through pathways involving the covariates \bar{L} while the parameter β_2 of our MSM represents the overall (direct and indirect) effects of treatment \bar{A} on Y . Unfortunately, this is not necessarily so. Indeed, it is easy to show that the parameter θ_1 can differ from zero even under the causal null hypothesis that treatment history \bar{a} has no causal effect on the outcome, either directly or through pathways including the covariate history \bar{L} . This is true even though the association (regression) model for $E[Y | \bar{A}, \bar{L}]$ is correctly specified and the assumption (9) of no unmeasured confounders is true. The problem is that covariate $\text{cum}(\bar{A})$ depends on a subject's entire treatment history including, for example, $A(0)$. However, $A(0)$ may affect the time-dependent covariates $L(k)$ and $L(k-1)$, for example. A regression model that adjusts for a covariate $L(k)$ (say, through the term $\text{cum}(\bar{L})$) that is both affected by earlier treatment $A(0)$ and itself predicts the outcome can result in a biased estimate of the treatment effect (even under the null hypothesis of no direct, indirect, or overall treatment effect). See Robins (1986) and Rosenbaum (1984) for further discussion. To summarize,

standard regression methods adjust for variables by including them in the model as regressors. These standard methods fail to appropriately adjust for confounding due to measured confounders $L(k)$ when treatment is time-varying since (i) $L(k)$ is a confounder for later treatment and thus must be adjusted for, but (ii) may also be affected by earlier treatment and thus cannot be adjusted for. The solution to this conundrum is to adjust for the time-dependent covariates $L(k)$ by using them to calculate the weights \mathcal{W}^{-1} rather than by adding the covariates to the regression model as regressors.

2.8. Estimation of Direct Effects

Suppose $A(t) = (A_P(t), A_Z(t))$ is comprised of two dichotomous treatments: AZT treatment $A_Z(t)$ and aerosolized pentamidine (AP) treatment $A_P(t)$. Let $cum(\overline{A_P})$ and $cum(\overline{A_Z})$ denote cumulative treatment with AP and AZT respectively. Then in the MSM

$$E[Y_{\bar{a}}] = g(\bar{a}; \beta), \text{ where } g(\bar{a}; \beta) = \beta_1 + \beta_2 cum(\overline{a_Z}) + \beta_3 cum(\overline{a_P}) + \beta_4 cum(\overline{a_Z}) cum(\overline{a_P}),$$

β_3 represents the direct effect of AP on the mean of Y when AZT is withheld, and $\beta_3 + \beta_4$ represents the direct effect of AP when all subjects receive continuous treatment with AZT. IPTW estimation of β is performed as above.

2.9. Formal Mathematical Justifications

The following lemma is the key to providing a formal, mathematical explanation of why weighting by \mathcal{W}^{-1} corrects our regression estimator for “confounding.”

Lemma 2.1. Under the sequential randomization assumption (9), $E(Y_{\bar{a}})$ is unique function $c(\bar{a})$ of \bar{a} such that $E[q(\overline{A})(Y - c(\overline{A})) / \mathcal{W}] = 0$ for all functions $q(\overline{A})$ where the expectation exists.

Consistency of our weighted estimator for the parameter β of our MSM (5) then follows from the fact that (i) the IPTW estimator $\hat{\beta}$ minimizing the weighted residual sum of squares can also be characterized as the solution to the weighted least squares “normal” equation

$$0 = n^{-1} \sum_{i=1}^n U_i(\beta_1, \beta_2) \tag{10}$$

where

$$U(\beta_1, \beta_2) = \mathcal{W}^{-1} (1, cum(\overline{A}))' [Y - \beta_1 - \beta_2 cum(\overline{A})], \tag{11}$$

and (ii) the probability limit of our weighted least squares “normal” equation is $E[q(\overline{A})(Y - c(\overline{A})) / \mathcal{W}] = 0$ with $q(\overline{A}) = (1, cum(\overline{A}))'$ and $c(\overline{A}) = g(\overline{A}; \beta) = Y - \beta_1 - \beta_2 cum(\overline{A})$.

2.9.1. A Proof based on the G-computation algorithm formula

We now sketch a proof of Lemma (2.1). It is somewhat more technical than the remainder of the paper and can be skipped on first reading. Given the assumption of no unmeasured confounders and a particular positivity assumption on the joint distribution of the observables, Robins (1987, Corollary to Theorem 1; 1997, Theorem 3.2) shows the mean of the dichotomous variable $Y_{\bar{a}}$ is non-parametrically identified from the joint distribution F_O of the observed data O by the non-parametric g-computation algorithm

functional $b(\bar{a})$ of Robins (1986). Specifically, $E(Y_{\bar{a}}) = b(\bar{a})$ where

$$b(\bar{a}) \equiv \int \cdots \int E(Y \mid \bar{\ell}_K, \bar{a}_K) \prod_{k=0}^K f(\ell_k \mid \bar{\ell}_{k-1}, \bar{a}_{k-1}) d\mu(\ell_k) \quad (12)$$

μ is a dominating measure, and for notational convenience we have written $\bar{z}(k)$ as \bar{z}_k and $z(k)$ as z_k . If covariates L_k are discrete random variables, the integrals in (12) are simply sums and (12) can be written

$$\sum_{\ell_0, \dots, \ell_K} E[Y \mid \bar{\ell}_K, \bar{a}_K] \prod_{k=0}^K f(\ell_k \mid \bar{\ell}_{k-1}, \bar{a}_{k-1}) .$$

The required positivity assumption (which we shall assume is true) is that, given $\bar{a} = (a_0, \dots, a_K)$, for each possible $\bar{\ell}_k$

$$\text{if } f(\bar{\ell}_k, \bar{a}_{k-1}) > 0, \text{ then } f(a_k \mid \bar{\ell}_k, \bar{a}_{k-1}) > 0, \quad (13)$$

which essentially says that if any set of subjects at time k have the opportunity of continuing on the treatment regime \bar{a} under consideration, at least some will take that opportunity.

The g-computation algorithm functional $b(\bar{a})$ can be characterized using the language of directed acyclic graphs. Specifically, $b(\bar{a})$ is the marginal mean of Y in the manipulated subgraph of the directed acyclic graph (DAG) G representing the observed data O in which all arrows into the treatment variables $\bar{A} = (A_1, \dots, A_K)$ have been removed and \bar{A} is set to \bar{a} with probability 1 (Spirtes et al., 1993). More specifically, let DAG G be the complete DAG with temporally ordered vertex set $O = \{L_0, A_0, L_1, A_1, \dots, A_K, Y\}$ and let DAG $G_{\bar{a}}$ be the subgraph of G in which all arrows into the $A_k, k = 0, \dots, K$ have been cut. Then $b(\bar{a})$ is the marginal mean of Y based on a distribution for O represented by DAG $G_{\bar{a}}$ in which $f(A_k \mid \bar{A}_{k-1}, \bar{L}_k)$ is replaced by a degenerate density that takes the value a_k with probability 1 while the conditional density of each other variable in the set O given its parents remains as in F_O .

We say that the distribution of $O = \{L_0, A_0, L_1, A_1, \dots, A_K, Y\}$ is standardly parameterized if, for each variable in O , we have specified a parametric (or semiparametric) model for the conditional distribution of that variable given its temporal predecessors (the past) and the parameters of each conditional model are variation-independent of those of any other conditional model. When our goal is to estimate the effect of a sequential (time-dependent) treatment \bar{A} on an outcome Y , Lemma 1 and Theorem 2 of Robins and Wasserman (1997) imply that inference procedures based on the standard parameterization will fail. Specifically, they prove that common choices for the parametric families in a standard parameterization often lead to joint densities such that the g-computation formula $b(\bar{a})$ for $E(Y_{\bar{a}})$ can never satisfy the causal null hypothesis that $E(Y_{\bar{a}})$ is the same for all \bar{a} . In particular, as discussed above, the causal null hypothesis does not imply that $Y \perp\!\!\!\perp \bar{A}_K \mid \bar{L}_K$. As a consequence, in large samples, the causal null hypothesis, even when true, will be falsely rejected regardless of the data. Robins and Wasserman propose reparameterizing the distribution of O using structural nested models. MSMs represent an alternative reparameterization that also overcomes the fatal deficiencies of the standard parameterization.

The following characterization due to Robins et al. (1999, Theorem 7.1) of the g-computation algorithm functional $b(\bar{a})$ defined in (12) above indicates why weighting by \mathcal{W}^{-1} controls confounding by measured covariates.

Lemma 2.2. $b(\bar{a})$ defined in (12) is the unique function $c(\bar{a})$ of \bar{a} such that $E[q(\bar{A})(Y - c(\bar{A})) / \mathcal{W}] = 0$ for all functions $q(\bar{A})$ for which the expectation exists.

Lemma 2.1 is an immediate corollary of Lemma 2.2.

2.9.2. A purely causal proof

Under a mild strengthening of our assumption of sequential randomization (no unmeasured confounders), a simple, self-contained, quite revealing, purely “causal” proof of Lemma 2.1 can be obtained that does not use the fact that $E(Y_{\bar{a}})$ is given by the g-computation algorithm formula $b(\bar{a})$ of Eq. (12). Let $Y_{\bar{\mathcal{A}}} = \{Y_{\bar{a}}; \bar{a} \in \bar{\mathcal{A}}\}$ be the set of all counterfactuals. Here $\bar{\mathcal{A}}$ is the support of the random variable \bar{A} . Suppose we strengthen our assumption of no unmeasured confounders to

$$Y_{\bar{\mathcal{A}}} \amalg A_k \mid \bar{L}_k, \bar{A}_{k-1} .$$

That is, the treatment A_k is jointly independent of the set of all counterfactuals given the measured past. Then we have the following proof of Lemma 2.1. Denote the total factual and counterfactual data by $Z = (Y_{\bar{\mathcal{A}}}, \bar{A}, \bar{L})$ and the observed data by $O = (Y \equiv Y_{\bar{\mathcal{A}}}, \bar{A}, \bar{L})$. With the strengthened assumption of no unmeasured confounders, we can factor the true joint density of Z that generated the data as

$$f(Z) = f(Y_{\bar{\mathcal{A}}}) \prod_{k=0}^K f(L_k \mid \bar{L}_{k-1}, \bar{A}_{k-1}, Y_{\bar{\mathcal{A}}}) \prod_{k=0}^K f(A_k \mid \bar{L}_k, \bar{A}_{k-1}) .$$

Now let $f^*(A_k \mid \bar{A}_{k-1})$ be a density for A_k given \bar{A}_{k-1} . It need not equal the true density $f(A_k \mid \bar{A}_{k-1})$. Let $f^*(Z)$ be a joint density for Z that differs from the true joint density $f(Z)$ only in that $f^*(A_k \mid \bar{L}_k, \bar{A}_{k-1}) = f^*(A_k \mid \bar{A}_{k-1})$ so that A_k is statistically and thus causally exogenous were the data generated under $f^*(Z)$. Thus,

$$f^*(Z) = f(Y_{\bar{\mathcal{A}}}) \prod_{k=0}^K f(L_k \mid \bar{L}_{k-1}, \bar{A}_{k-1}, Y_{\bar{\mathcal{A}}}) \prod_{k=0}^K f^*(A_k \mid \bar{A}_{k-1}) .$$

Now, we see at once that $E(Y_{\bar{a}}) = E^*(Y_{\bar{a}})$ since $f(Z)$ and $f^*(Z)$ have the same marginal distribution for $Y_{\bar{a}}$. Second, a simple calculation shows that \bar{A} is causally exogenous under $f^*(z)$ [i.e., $Y_{\bar{a}} \amalg^* \bar{A}$]. Thus, we have that $E^*[Y_{\bar{a}}] = E^*[Y_{\bar{a}} \mid \bar{A} = \bar{a}] = E^*[Y_{\bar{\mathcal{A}}} \mid \bar{A} = \bar{a}] = E^*[Y \mid \bar{A} = \bar{a}]$ where the first equality is by independence, the second by the properties of conditional expectations, and the third by the consistency assumption $Y_{\bar{\mathcal{A}}} \equiv Y$. Hence, we have formally proved the result informally argued above that when \bar{A} is causally exogenous, the marginal mean of $Y_{\bar{a}}$ is given by the regression function $E^*[Y \mid \bar{A} = \bar{a}]$ of Y on $\bar{A} = \bar{a}$. Now it is a standard statistical result that the regression function $E^*(Y \mid \bar{A} = \bar{a})$ is characterized as the unique function $c(\bar{a})$ solving $E^*\{q(\bar{A}) [Y - c(\bar{A})]\} \equiv \int q(\bar{A}) (Y - b(\bar{A})) f^*(Z) d\mu(Z) = 0$ for all $q(\bar{A})$ where μ is a dominating measure. But, $\int q(A) (Y - c(\bar{A})) f^*(Z) d\mu(Z) = \int q(A) (Y - c(\bar{A})) \frac{f^*(Z)}{f(Z)} f(Z) d\mu(Z) = E\left[q(A) (Y - c(\bar{A})) \frac{f^*(Z)}{f(Z)}\right]$ where the first equality is by algebra and the second is by the definition of an expectation with respect to $f(Z)$. But, by definition, $\frac{f^*(Z)}{f(Z)} = \mathcal{W}^{-1}$ if we choose $f^*(A_k \mid \bar{A}_{k-1})$ equal to $f(A_k \mid \bar{A}_{k-1})$. Lemma 2.1 then follows, by $E^*(Y \mid \bar{A} = \bar{a}) = E(Y_{\bar{a}})$.

The proof also makes clear that consistency of our weighted estimator does not require that we choose $f^*(A_k \mid \bar{A}_{k-1}) = f(A_k \mid \bar{A}_{k-1})$.

3. Sensitivity Analysis for Unmeasured Confounders

3.1. A Sensitivity Analysis Methodology

We have seen that under the assumption of no unmeasured confounders, our inverse-probability-of-treatment-weighted estimator $\widehat{\beta}$ consistently estimates the parameter β of an MSM that encodes the strengths of the causal effect of treatment history on the outcome Y . However, the assumption of no unmeasured confounders itself is not testable based on the observed data O . Thus, in an observational study, we never know whether our assumption of no unmeasured confounders is true. In a randomized or sequential randomized study, we will know it is true only because we physically randomized using a coin or some other physical randomization device. The data themselves give no indication. Thus, in observational studies, it is of interest to conduct a sensitivity analysis to quantify how our inferences concerning the causal effect of a treatment on the outcome vary as a function of the magnitude of (non-identifiable) confounding by unmeasured factors. In this section, we describe how to conduct such a sensitivity analysis.

Our first task is to develop a measure that quantifies in a useful way the degree of confounding due to unmeasured factors. Consider the function

$$q_m(\bar{\ell}_m, \bar{a}, a_m^*) = E[Y_{\bar{a}} | \bar{\ell}_m, \bar{a}_{m-1}, a_m] - E[Y_{\bar{a}^*} | \bar{\ell}_m, \bar{a}_{m-1}, a_m^*]. \quad (14)$$

Again, for notational convenience, we have written $\bar{z}(k)$ as \bar{z}_k and $z(k)$ as z_k . To help understand why this function is a useful measure of the magnitude of confounding due to unmeasured factors, fix a treatment history $\bar{a} = \bar{a}_K$ through the end of the study. Fix a time m . Now consider the subgroup of the study population with a particular covariate history $\bar{\ell}_m$ through day m who has followed the particular treatment history \bar{a}_{m-1} through day $m-1$ consistent with the treatment regime \bar{a} . Consider next both the subset of this subgroup who continued at day m with the treatment a_m consistent with the regime \bar{a} and the subset who left the regime \bar{a} on day m and took instead treatment a_m^* . Then $q_m(\bar{\ell}_m, \bar{a}, a_m^*)$ is precisely the difference of the mean of the counterfactual $Y_{\bar{a}}$ between these two subsets. In particular, if $a_m = a_m^*$, $q_m(\bar{\ell}_m, \bar{a}, a_m) = 0$, since then the second of the two subsets did not actually leave the regime on day m and the two subsets are identical.

Now if the assumption (9) of no unmeasured confounders holds, we see that $q_m(\bar{\ell}_m, \bar{a}, a_m^*)$ is identically zero, since the assumption of no unmeasured confounders implies $Y_{\bar{a}}$ is mean-independent of A_m given the past $\bar{L}_m \bar{A}_{m-1}$. Indeed, $q_m(\bar{\ell}_m, \bar{a}, a_m^*)$ is a natural measure of the magnitude of non-comparability with respect to the mean of $Y_{\bar{a}}$ of the two subsets mentioned above due to unmeasured confounding.

Suppose now that instead of imposing the assumption $q_m(\bar{\ell}_m, \bar{a}, a_m^*) = 0$ of no unmeasured confounders, we impose the assumption that

$$q_m(\bar{\ell}_m, \bar{a}, a_m^*) \text{ is some specified non-zero function.} \quad (15)$$

Robins et al. (1999) prove that assumption (15) places no restrictions on the joint distribution F_O of the observed data, and thus the assumption cannot be rejected by any statistical test. We say that a parameter or function is identified if it can be determined from the joint distribution of the observed data. The result just stated implies that the selection bias function $q_m(\bar{\ell}_m, \bar{a}, a_m^*)$ is not identified. Indeed, the observed data distribution places no restrictions on the possible values of the function except for the structural requirement that $q_m(\bar{\ell}_m, \bar{a}, a_m^*)$ equals zero if a_m^* equals the treatment a_m specified by \bar{a} . However, Robins et al. (1999) prove that for each choice of $q_m(\bar{\ell}_m, \bar{a}, a_m^*)$, the mean of $Y_{\bar{a}}$ is identified

(determined) by the joint distribution F_O of the data. Specifically, they prove

Lemma 3.1. Under assumption (15), $E(Y_{\bar{a}})$ is identified for each \bar{a} from the joint distribution of the observed data $O = (Y, \bar{L}_K, \bar{A}_K)$. Specifically, $E(Y_{\bar{a}})$ is the unique function $c(\bar{a})$ such that

$$0 = E \left[\mathcal{W}^{-1} q(\bar{A}) \left\{ Y - \sum_{m=0}^K \int q_m(\bar{L}_m, \bar{A}, a_m^*) dF(a_m^* | \bar{L}_m, \bar{A}_{m-1}) - c(\bar{A}) \right\} \right] \quad (16)$$

for all functions $q(\bar{A})$ for which the expectation is finite.

Note our previous Lemma 2.1 is the special case of Lemma 3.1 with q_m chosen to be identically zero. Indeed, Lemma 3.1 can be interpreted as saying that Lemma 2.1 remains true if we replace the observed data Y on each subject by the selection-bias-corrected version $Y - \sum_{m=0}^K \int q_m(\bar{L}_m, \bar{A}, a_m^*) dF(a_m^* | \bar{L}_m, \bar{A}_{m-1})$. Note that the integral in (16) becomes the sum

$$\sum_{a_m^*} q_m(\bar{L}_m, \bar{A}, a_m^*) f(a_m^* | \bar{L}_m, \bar{A}_{m-1})$$

when A_m is a discrete random variable. Indeed, in the special case in which A_m is a dichotomous $(0, 1)$ treatment indicator, the sum evaluates to $q_m(\bar{L}_m, \bar{A}, 1) pr(A_m = 1 | \bar{L}_m, \bar{A}_{m-1})$ for a subject with observed treatment $A_m = 0$ and to $q_m(\bar{L}_m, \bar{A}, 0) pr(A_m = 0 | \bar{L}_m, \bar{A}_{m-1})$ for subject with observed treatment $A_m = 1$.

Thus, given the distribution F_O of the observed data and a particular choice of $q_m(\bar{\ell}_m, \bar{a}, a_m^*)$, the causal dose-response relationship $E(Y_{\bar{a}}) = c(\bar{a})$ is determined. It follows that for any given distribution F_O of the observed data O , we can tabulate in a sensitivity analysis the dose-response relationship $c(\bar{a})$ as a function of the non-identified selection bias function q_m . In practice, the distribution F_O of O is unknown, but the distribution can be estimated by the empirical distribution F_n of the data that puts probability mass $1/n$ on each of the observations O_i for subjects $i = 1, \dots, n$. The reader should not be discouraged that we only provide a sensitivity analysis. Since the function q_m represents the magnitude of confounding (i.e., selection bias) due to unmeasured factors, it would not be desirable or scientifically reasonable for q_m to be identified in the absence of further knowledge of these factors. Our sensitivity analysis formalizes this desideratum; we cannot identify the selection bias function q_m , but we can identify the dose-response function $c(\bar{a})$ as a function of q_m . Since the data contain no independent evidence about q_m , final substantive conclusions would depend upon the functions q_m that are considered plausible by relevant subject-matter experts. A convenient approach to carrying out a practical sensitivity analysis is as follows.

(i) First specify a relatively parsimonious model for the dose-response function in which the magnitude and shape of the dose-response function can be summarized by reporting the value of just a few parameters. A particularly simple example would be our model

$$E(Y_{\bar{a}}) = \beta_1 + \beta_2 cum(\bar{a})$$

in which β_2 summarizes the effect of treatment on the outcome;

(ii) Second, choose a simple parameterized form for the selection bias function q_m with parameter α , such as

$$q_m(\bar{\ell}_m, \bar{a}, a_m^*; \alpha) = \alpha [a_m - a_m^*] \quad (17)$$

so that the choice $\alpha = 0$ corresponds to the assumption of no unmeasured confounders. Analyzing the data under the assumption that α is positive implies that less healthy subjects (subjects with high counterfactual HIV RNA $Y_{\bar{a}}$) are given higher doses of treatment (even after we control for past confounder history \bar{L}_m and past treatment history \bar{A}_{m-1}). This would be the case if physicians give higher doses of AZT to patients who have unfavorable prognostic factors for the outcome and these prognostic factors were not recorded in \bar{L}_m for data analysis. In contrast, a negative value for α implies that healthy subjects were preferentially given higher doses of treatment even after adjusting for past treatment and measured covariate history. To implement our sensitivity analysis, we choose a large number of values for α , and for each choice of α separately, we obtain an estimate $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ of β by solving the empirical analog of Eq. (16) with $q(\bar{A}) = (1, \text{cum}(\bar{A}))'$ and $c(\bar{A}) = \beta_1 + \beta_2 \text{cum}(\bar{A})$. That is, we solve the weighted least squares normal equations with each subject's observed outcome Y replaced by their selection-bias-corrected outcome $Y(\alpha) \equiv Y - \sum_{m=0}^K \int q_m(\bar{L}_m, \bar{A}, a_m^*; \alpha) dF(a_m^* | \bar{L}_m, \bar{A}_{m-1})$, i.e., we solve

$$0 = \sum_{i=1}^n U_i(\beta_1, \beta_2)$$

where

$$U(\beta_1, \beta_2) = \mathcal{W}^{-1} (1, \text{cum}(\bar{A}))' [Y(\alpha) - \beta_1 - \beta_2 \text{cum}(\bar{A})] .$$

Solving this equation is equivalent to minimizing the sum of weighted squared selection-bias-corrected residuals

$$\sum_{i=1}^n \mathcal{W}_i^{-1} [Y_i(\alpha) - \beta_1 - \beta_2 \text{cum}(A_i)]^2 .$$

Remark: The subject-specific selection bias correction term $\sum_{m=0}^K \int q_m(\bar{L}_m, \bar{A}, a_m^*; \alpha) dF(a_m^* | \bar{L}_m, \bar{A}_{m-1})$ can be difficult to compute, especially for continuous treatment (e.g., the dose of a drug recorded in milligrams). In that case, one can obtain an estimated version by sampling. Specifically, one replaces the above selection bias correction term by $J^{-1} \sum_{j=1}^J \sum_{m=0}^K q_m(\bar{L}_m, \bar{A}, a_{mj}^*; \alpha)$ where the $a_{mj}^*, j = 1, \dots, J$ are, for each m , J -independent draws from the conditional distribution $f(a_m | \bar{L}_m, \bar{A}_{m-1})$. Again, in an observational study, \mathcal{W} and $f(a_m | \bar{L}_m, \bar{A}_{m-1})$ will not be known and will have to be modelled and estimated from the data in a preliminary step.

The final result of such a sensitivity analysis will be graph such as that shown in Figure 1. The confidence intervals surrounding the point estimate $\hat{\beta}_2$ on the graph represent uncertainty due to sampling variability and can be computed easily using standard statistical software. Since the functional form of q_m is not identified, it follows that one may wish to report the results of several different sensitivity analyses with different functional forms for $q_m(\bar{L}_m, \bar{a}, a_m^*; \alpha)$ and also consider choosing the parameter α to be multidimensional.

The biggest challenge in conducting a sensitivity analysis is not the technical computational details described above. Rather, it is the choice of one or more sensible parameterized sensitivity analysis functions q_m whose meaning can be explained to relevant subject-matter experts with sufficient clarity that the experts are able to provide a plausible range for the parameter α encoding the magnitude and direction of selection bias (i.e., unmeasured confounding). Hard as this challenge may sound, I believe

it would be a worthwhile exercise if it leads to results of studies being summarized by plots such as that given in Figure 1 rather than being summarized by the single confidence interval at $\alpha = 0$, as would be done in the absence of a formal sensitivity analysis. This belief of mine is in line with the well-known adage that “it is not what you don’t know that hurts you; it’s the things you think you know but don’t.” Paraphrasing Freedman Rothenberg and Sutch (1984), I believe reporting sensitivity analysis graphs like Fig. 1 in scientific papers rather than simply reporting the single confidence interval at $\alpha = 0$ (corresponding to the assumption of no unmeasured confounders) will decrease the stock of things we think we know, but don’t.

3.2. Comparison with other approaches to sensitivity analysis

We quantified the net confounding on the mean of the outcome $Y_{\bar{\alpha}}$ through the selection bias function q_m without any reference to the unmeasured common causes U of treatment and the outcome that are the source of this confounding. A large body of previous work, originating with Cornfield (1959) on sensitivity analysis in causal inference models with time-independent treatments has tried to directly model the effect of these unmeasured causes U . In such a sensitivity analysis, one varies the association of U with the outcome Y (within levels of treatment and measured confounders) and the association of U with the treatment (within levels of measured confounders) (Schlesselman, 1978; Rosenbaum and Rubin, 1983; Lin et al., 1998). In contrast, in our approach, we simply model the association of the mean of the counterfactual outcome variable with the treatment (within levels of measured confounders). The advantage of our approach is that (i) there are many fewer sensitivity parameters to vary, and (ii) the (essentially impossible) decision as to whether to view U as univariate or multivariate, continuous or discrete is done away with. A link between the two approaches is that the counterfactual variables can be considered the ultimate unmeasured confounder U . This reflects the fact that, given the counterfactuals and treatment, other unmeasured covariates fail to predict the observed outcome (and thus are superfluous and can be dispensed with), since the observed outcome variable is a deterministic function of the treatment and the counterfactual outcome.

In our opinion, the unmeasured confounder U approach should be preferred to our counterfactual approach only in circumstances, where (i) U represents a known confounder (e.g., cigarette smoking) that for logistical reasons was not measured in a particular study, and furthermore, (ii) there exists reasonable historical and/or biological knowledge about the magnitude of association of U with both the outcome (conditional on treatment and measured confounders) and the treatment (conditional on measured confounders). In contrast, when U is to represent all possible unmeasured factors, we believe that it is substantively easier for subject-matter experts to give their opinions about the plausible magnitude of the association of the mean of the counterfactual outcome with treatment than about the question of whether any unmeasured confounders U are continuous or discrete, single or multidimensional, and the associations of such confounders with treatment and the outcome.

We have seen that our counterfactual approach leads to extremely simple computations that can be carried out with standard software. In contrast, as discussed by Lin et al. (1998), there can be formidable computational difficulties associated with the approach based on positing an unmeasured covariate U .

3.3. Alternatives to sensitivity analysis

Competitors to sensitivity analysis as means for summarizing uncertainty due to confounding by unmeasured factors about causal effects in observational studies include formal Bayesian inference and computing bounds for the causal effect. Although I am a proponent of computing bounds in placebo-controlled ran-

domized trials with all or none non-compliance (see Robins, 1989), nonetheless, in observational studies the bounds are, in general, too wide to be very useful since they always include the null hypothesis of no treatment effect.

I view a Bayesian analysis as complementary to a sensitivity analysis. A sensitivity analysis such as that given in Fig. 1 accurately reports what we can learn from the data in a statement of the form “if this is the degree of confounding due to unmeasured factors, then this is what we can conclude from the data.” If a decision has to be made, we need to go further, and a natural direction would be to place a prior distribution on the non-identified parameter α (and indeed on the functional form of q_m as well). Thus, I regard a sensitivity analysis as essential pre-processing for a full Bayesian analysis. Mathematical details are described in Sec. 11 of Robins et al. (1999), although the discussion there is restricted to rather simple settings because of technical problems with implementing non-parametric Bayesian procedures.

In conclusion, IPTW estimation of MSMs when combined with a sensitivity analysis is a useful approach to causal inference in longitudinal data. However, there are a number of difficulties which I have not mentioned, some of which are briefly considered in Appendix 2. They are discussed in detail in Robins (1999) and Robins et al. (1999). I hope I have succeeded in the task that I set: to describe clearly the job that needs doing and to begin to describe some of the tools that I and others have developed to do it.

Appendix 1: Parametric Likelihood-Based Inference

In this appendix we assume that the MSM (5) is correct and the assumption (9) of no unmeasured confounders is true. Then the likelihood for the observed data is $\prod_{i=1}^n f(O_i; \beta, \rho, \alpha)$, where

$$f(O; \beta, \rho, \alpha) = \mathcal{L}_1(\beta, \rho) \mathcal{L}_2(\alpha), \tag{A1.1}$$

$$\mathcal{L}_1(\beta, \rho) = \int \cdots \int f(Y_{\bar{A}}; \beta, \rho_1) \prod_{k=0}^K f(L_k | \bar{L}_{k-1}, \bar{A}_{k-1}, Y_{\bar{A}}; \rho_2) \prod_{\{Y_{\bar{a}}; \bar{a} \neq \bar{A}\}} dY_{\bar{a}},$$

$$\mathcal{L}_2(\alpha) = \prod_{k=0}^K f(A_k | \bar{L}_k, \bar{A}_{k-1}; \alpha),$$

ρ_2 is a unknown parameter indexing the unknown density $f(L_k | \bar{L}_{k-1}, \bar{A}_{k-1}, Y_{\bar{A}})$, ρ_1 is an unknown parameter that together with the parameter β of (5) indexes the unknown joint density of the counterfactuals $\{Y_{\bar{a}}; \bar{a} \in \bar{A}\}$, $\rho = (\rho_1, \rho_2)$, and the parameter α indexes the densities $f(A_k | \bar{L}_k, \bar{A}_{k-1})$. In an observational study α will be unknown, but in a sequential randomized trial α will be known since the randomization probabilities $f(A_k | \bar{L}_k, \bar{A}_{k-1})$ are known by design.

To perform fully parametric likelihood-based likelihood inference on β , we specify parametric models $f(Y_{\bar{A}}; \beta, \rho_1)$ and $f(L_k | \bar{L}_{k-1}, \bar{A}_{k-1}, Y_{\bar{A}}; \rho_2)$ with ρ a finite dimensional parameter. The factorization (A1.1) implies that the maximum likelihood estimator of (β, ρ) is the same whether $f(A_k | \bar{L}_k, \bar{A}_{k-1})$ is completely unknown, is completely known (as in a sequential randomized trial), or follows a parametric model $f(A_k | \bar{L}_k, \bar{A}_{k-1}; \alpha)$ depending on an unknown parameter α . From a Bayesian perspective, the factorization (A1.1) implies that, if α and (β, ρ) are *a priori* independent, then they will be *a posteriori* independent and the posterior distribution for β will be the same whether $f(A_k | \bar{L}_k, \bar{A}_{k-1})$ is completely unknown, completely known, or follows a parametric model. Thus, for the purposes of likelihood-based

frequentist inference or Bayesian inference (with independent priors) concerning β , $\mathcal{L}_1(\beta, \rho)$ is often referred to as the likelihood for (β, ρ) and $\mathcal{L}_2(\alpha)$ is referred to as the ancillary part of the likelihood. If the models $f(Y_{\bar{\mathcal{A}}}; \beta, \rho_1)$ and $f(L_k | \bar{L}_{k-1}, \bar{A}_{k-1}, Y_{\bar{\mathcal{A}}}; \rho_2)$ are correctly specified, the MLE of β will be more efficient (i.e., have smaller variance) than any IPTW estimator. Unfortunately, if either model is misspecified, both the MLE and the posterior mean of β will generally be inconsistent (even if the treatment process is statistically exogenous and completely known). Since, when $K = 300$ and A_k is dichotomous, the dimension of $Y_{\bar{\mathcal{A}}}$ is 2^{300} , specification of nearly correct parametric models $f(Y_{\bar{\mathcal{A}}}; \beta, \rho_1)$ and $f(L_k | \bar{L}_{k-1}, \bar{A}_{k-1}, Y_{\bar{\mathcal{A}}}; \rho_2)$ appears to be an essentially hopeless task.

Remark: In addition to the problem of model misspecification, in order to evaluate the likelihood function (A1.1), it is necessary to compute $2^{300} - 1$ dimensional integrals, since one must integrate out all of the unobserved random variables $Y_{\bar{\mathcal{A}}}$. However, in contrast to the intractable problem of model misspecification, the need to calculate high dimensional integrals can be overcome, if one assumes a model under which each $Y_{\bar{\mathcal{A}}}$ is a deterministic function of \bar{a} and an one error random variable (Robins and Wasserman, 1999), which Robins (1997) refers to as a rank preserving structural distribution model (RPSDM) with no local treatment interaction. The joint distribution of the $Y_{\bar{\mathcal{A}}}$ is then massively degenerate and the integrals in the likelihood function (A1.1) are eliminated. However the assumptions underlying a RPSDM are usually biologically implausible.

In contrast, our IPTW estimator of β is easy to compute and is guaranteed to be consistent, if $f(A_k | \bar{L}_k, \bar{A}_{k-1})$ is known (as in a sequential randomized trial), or if the model $f(A_k | \bar{L}_k, \bar{A}_{k-1}; \alpha)$ is correctly specified. It follows that the estimation of a MSM model in a sequential randomized trial is a leading example of an estimation problem that is essentially intractable from a Bayesian or likelihoodist point of view, but for which there exist simple consistent weighted estimators. Robins and Ritov (1997) provide an in depth discussion of this issue and its implications for the likelihood principle. In the technical parlance of statistical literature, our IPTW estimator is a semiparametric estimator rather than a likelihood-based estimator.

I would hope this example would lead the philosophy community to question the emphasis on Bayesian and likelihood-based approaches to inference that has been passed on from the statistical community. When a likelihood function becomes too complicated and high dimensional, there is no way to directly extract accurate information from it, even though there are other methods, such as IPTW estimation, which allow one to accurately extract the information on a parameter β of interest when the ancillary part of the likelihood $\mathcal{L}_2(\alpha)$ is either known or can be accurately modelled.

Remark: It should be pointed out that, in an observational study, in which there are (i) a large number of treatment periods K (ii) subjects go on and off treatment frequently, and (iii) treatment affects the covariates L_k , then, even if in truth (but unknown to the data analyst) treatment is statistically exogenous, the analyst can obtain estimated weights \widehat{W}^{-1} that differ significantly from one due to misspecification of the model $f(A_k | \bar{L}_k, \bar{A}_{k-1}; \alpha)$. In that case the analyst will falsely conclude that treatment is non-exogenous; furthermore, unbeknownst to the analyst, the OLS estimator will be consistent for the parameter β_2 of the MSM, while the IPTW estimator will be inconsistent due to the aforementioned model misspecification.

Appendix 2: Difficulties with IPTW Estimation of MSMs

Although IPTW estimation of MSMs is a useful approach to causal inference from longitudinal data, there are a number of major difficulties which I have not mentioned. They are discussed in detail in Robins (Robins, 1999; Robins et al., 1999). Here, I briefly touch on a few, mainly to provide a sense of what is at issue.

(i) A dynamic treatment regime g is a treatment regime wherein a subject's dose of treatment on day k depends on the evolution of their measured time-dependent covariates \bar{L}_k through day k . For example, treatment regime “take AZT only when your white blood count has exceeded 600 for the past two weeks” is a dynamic regime. Heretofore, we have only considered the counterfactuals $Y_{\bar{a}}$ associated with non-dynamic regime \bar{a} . Formally, a treatment regime g is a collection of $K + 1$ functions $g = (g_0, \dots, g_K)$ where $g_k : \bar{\mathcal{L}}_k \rightarrow \mathcal{A}_k$ maps the support $\bar{\mathcal{L}}_k$ of \bar{L}_k to the support \mathcal{A}_k of A_k . Let Y_g be the counterfactual outcome of a subject when following regime g . If, for each k , $g_k(\bar{\ell}_k)$ is a constant a_k not depending on $\bar{\ell}_k$, we say that regime g is non-dynamic and write Y_g as $Y_{\bar{a}}$. Under consistency and positivity assumptions, $E(Y_g)$ is given by the RHS of (12), with $\bar{a}_{k-1} \equiv g(\bar{\ell}_{k-1}) \equiv (g_0(\bar{\ell}_0), \dots, g_{k-1}(\bar{\ell}_{k-1}))$ and $\bar{a}_k \equiv g(\bar{\ell}_k)$ being the treatment histories prescribed by regime g .

The optimal treatment regime may well be dynamic. For example, suppose our goal is to find the regime that minimizes $E(Y_g)$. If, for example, taking AZT is beneficial when one's white count is greater than 600 but is harmful when one's white count is less than 600, the optimal regime will be dynamic. In practice, optimal regimes for medical treatments are almost always dynamic, since it is important to stop treatment when one becomes toxic. For example, low white blood counts are a sign of toxicity to AZT, and continuing AZT will depress a subject's white count further, leading to infection and death.

Unfortunately, one cannot, in general, use IPTW estimators of a MSM to estimate $E(Y_g)$ for dynamic regimes. Robins (1997, Robins, 1999; Robins et al., 1999) described many different approaches based on MSMs, structural nested models (SNMs), and the g-computation algorithm to estimate $E(Y_g)$. These methods are beyond the scope of this paper.

(ii) There are substantive settings, such as studies of the effect of an occupational exposure on health in a cohort of factory workers, in which methods based on MSMs cannot be used. This reflects the fact that if our positivity assumption that $f(\bar{a}_{k-1}, \bar{\ell}_k) > 0$ implies $f(a_k | \bar{\ell}_k, \bar{a}_{k-1}) > 0$ is only true for a single value of a_k , then inference based on IPTW estimation of MSMs is not possible. In an occupational study, suppose L_k is employment status (i.e., whether a subject is on or off work at time k), and A_k is the dose of a potentially toxic exposure such as formaldehyde. Since subjects off work, i.e., $L_k = 0$, only receive exposure level $a_k = 0$, our positivity assumption is only true for $a_k = 0$ and for no other exposure level, thus precluding the use of our IPTW estimation methods. In such a setting, exposure effects can still be estimated using g-estimation of structural nested models (Robins, 1997).

(iii) If Y is a non-negative random variable, the sensitivity analysis method based on (14) may fail since it can lead to negative values for $E[Y_{\bar{a}}]$. Hence, for non-negative Y , one should base a sensitivity analysis on

$$q_m(\bar{\ell}_m, \bar{a}, a_m^*) = \ln E[Y_{\bar{a}} | \bar{\ell}_m, \bar{a}_{m-1}, a_m] - \ln E[Y_{\bar{a}} | \bar{\ell}_m, \bar{a}_{m-1}, a_m^*] .$$

where \ln is the natural logarithm. Inference proceeds as above except that now $Y(\alpha) = Y \prod_{m=0}^K Q_m(\alpha)$ where $Q_m(\alpha) = \exp[-q(\bar{L}_m, \bar{A}, 1 - A_m; \alpha)] \{1 - f(A_m | \bar{L}_m, \bar{A}_{m-1})\}$ when A_m is a dichotomous (0, 1) variable and $Q_m(\alpha) \equiv \int \exp\{-q(L_m, \bar{A}, a_m; \alpha)\} dF(a_m | \bar{L}_m, \bar{A}_{m-1})$ more generally. Even this approach can fail when Y is a dichotomous (0, 1) random variable, since it can lead to values for the $E[Y_{\bar{a}}]$ that exceed one. The point is that although, under the assumption of no unmeasured confounders, we can estimate the parameters of a logistic MSM, such as $E[Y_{\bar{a}}] = \{1 + \exp[-\beta_1 - \beta_2 \text{cum}(\bar{a})]\}^{-1}$ by IPTW weighted logistic regression (Robins, 1999; Robins et al., 1999), there is no convenient method of sensitivity analysis based on replacing a subject's observed Y by a selection-bias corrected-version $Y(\alpha)$ that is guaranteed to respect the fact that $E[Y_{\bar{a}}]$ cannot exceed one.

BIBLIOGRAPHY:

- Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B., and Wynder, E.L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22: 173-203.
- Ericsson, N.R., Hendry, D.F., and Mizon, G.E. (1998). Exogeneity, cointegration, and economic policy analysis. *Journal of Business and Economic Statistics*, 16: 370-388.
- Freedman, D., Rothenberg, T., and Sutch, P. (1984). On energy policy models. *Journal of Business and Economic Statistics*, 1, 24-36.
- Holland, P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81: 945-961.
- Lewis, D. (1973). Causation. *Journal of Philosophy*, 70: 556-567.
- Lin, D.Y., Psaty, B.M., and Kronmal, R.A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, 54: 948-963.
- Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, 82: 669-688.
- Robins, J.M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods – Application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7: 1393-1512.
- Robins, J.M. (1987). Addendum to “A new approach to causal inference in mortality studies with sustained exposure periods – Application to control of the healthy worker survivor effect.” *Computers and Mathematics with Applications*, 14: 923-945.
- Robins, J.M. (1995). Comments on Judea Pearl’s paper, “Causal diagrams for empirical research.” *Biometrika*, 82: 695-698.
- Robins, J.M. (1997). Causal inference from complex longitudinal data. In: **Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics (120)**, M. Berkane, Editor. NY: Springer Verlag, 69-117.
- Robins, J.M. (1998). Marginal structural models. *1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, pp. 1-10.
- Robins J.M. (1999). Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference. **Statistical Models in Epidemiology**, Ed. E. Halloran, NY: Springer-Verlag (to appear).
- Robins J.M., Rotnitzky A. and Scharfstein D. (1999). Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. **Statistical Models in Epidemiology**, Ed. E. Halloran, NY: Springer-Verlag (to appear).
- Robins, J.M. 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, Providence Rhode Island, August 1-3, 1997*. Dan Geiger and Prakash Shenoy (Eds.), Morgan Kaufmann, San Francisco, pp. 409-420.
- Rosenbaum, P.R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A, General*, 147: 656-666.
- Rosenbaum, P.R., and Rubin, D.B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B*, 11: 212-218.
- Rubin, D.B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6: 34-58.

Schlesselman J.J. (1978). Assessing effects of confounding variables. *American Journal of Epidemiology*, 108: 3-8.

Spirtes, P., Glymour, C., Scheines, R. (1993). **Causation, Prediction, and Search. Lecture Notes in Statistics 81.** New York: Springer-Verlag.

Acknowledgements: David Freedman provided helpful comments. Support was provided through NIH Grants RO1-A132475 and R01-CA74112.

Figure 1 goes here (in file)-Remember to include