Marginal Structural Models for the Estimation of Direct and Indirect Effects

Tyler J. VanderWeele

Abstract: The estimation of controlled direct effects can be carried out by fitting a marginal structural model and using inverse probability of treatment weighting. To use marginal structural models to estimate natural direct and indirect effects, 2 marginal structural models can be used: 1 for the effects of the treatment and mediator on the outcome and 1 for the effect of the treatment on the mediator. Unlike marginal structural models typically used in epidemiologic research, the marginal structural models used to estimate natural direct and indirect effects are made conditional on the covariates.

(*Epidemiology* 2009;20: 18–26)

Investigators may wish to examine the extent to which the effect of a treatment on some outcome is mediated by an intermediate variable. Methods in the social sciences using structural equation modeling have often been used to perform such analyses.[1–5] However, these methods in general only allow for the definition of, estimation of, and effect decomposition into direct and indirect effects in linear models in which there is no interaction between the mediator and the outcome. The methods have, nevertheless, been routinely used in settings in which interactions or nonlinearities are present. It is not uncommon, for example, to use a regression of the outcome on the treatment including the mediator as a predictor variable to estimate the direct effects of treatment. Often, the direct effect is then subtracted from the total effect of treatment to give an indirect effect. Kaufman et al[6] have conclusively shown that this strategy will fail when there is interaction between the effects of treatment and the mediator on the outcome. In particular, subtracting the direct effect (estimated from the regression) from the total effect does not in general give a quantity that can be interpreted as an indirect or mediated effect. There has recently developed a literature in causal inference that addresses the definition and identification of direct and indirect effects in settings in which interactions and nonlinearities are present.[7–14] This work circumvents many of the criticisms concerning the estimation of direct and indirect effects using structural equations modeling and path analysis. In particular, direct and indirect effects can be defined and estimated, and the total effect can be decomposed into a natural direct and indirect effect, even in settings involving interactions and nonlinear models.

Using ideas developed by Robins and Greenland,[7] Pearl[8] introduced the terminology of "controlled direct effects" and "natural direct and indirect effects." Several authors have considered the question of the identification of these direct and indirect effects.[7–14] Petersen et al[10] explicitly discusses how, if the identification assumptions hold, both controlled direct effects and natural direct effects can be estimated from regression models. In this paper we consider an alternative estimation strategy, one related to the use of marginal structural models.[15–17] Robins[17] and van der Laan and Petersen[12] discuss the use of marginal structural models to estimate direct effects and in this paper we develop more fully an approach of using marginal structural models in the estimation of direct and indirect effects. As is discussed below, the estimation of controlled direct effects using marginal structural models is straightforward. The estimation of natural direct and indirect effects is somewhat more subtle and the approach described below requires the use of 2 marginal structural models, which differ somewhat from marginal structural models as they have typically been used in epidemiology in that the models are made conditional on the covariates. Before we consider the use of marginal structural models for the estimation of direct and indirect effects, we will first review the potential outcomes framework as well as definitions of and identification conditions for direct and indirect effects.

Potential Outcomes

We will index the subjects in the population by i. Let A_i denote the treatment received by subject i and let Y_i denote some posttreatment outcome for subject i. Let Z_i denote the value, for subject i, of some posttreatment intermediate variable that may serve as a mediator for the treatment-outcome relationship. We will assume that the subjects are sampled from a population and thus treat A, Z, and Y as random variables and consequently often suppress the index i. Let Y_a

Submitted 27 November 2007; accepted 2 May 2008.
From the Department of Health Studies, University of Chicago, Chicago, Illinois.
Supplemental material for this article is available with the online version of the journal at www.epidem.com; click on "Article Plus."
Correspondence: Tyler J. VanderWeele, Department of Health Studies, University of Chicago, 5841 S. Maryland Ave., MC 2007, Chicago, IL 60637. E-mail: vanderweele@uchicago.edu.

ISSN: 1044-3983/09/2001-0018
DOI: 10.1097/EDE.0b013e31818f69ce

denote a subject's outcome if treatment A were set, possibly contrary to fact, to a. Note that for each subject there will potentially be a different value Y_a for each level of treatment a. The variables Y_a are referred to as counterfactual outcomes or potential outcomes.[18,19] If A is binary then an individual will have 2 potential outcomes Y_0 and Y_1, what the outcome would have been for the subject if treatment had been set, possibly contrary to fact, to 0 or 1, respectively. In the context of mediation there will also be potential outcomes for the intermediate variable. Let Z_a denote a subject's counterfactual value of the intermediate Z if treatment A were set to the value a. Finally, let Y_{az} denote a subject's counterfactual value for Y if A were set to a and Z were set to z. A brief overview of the counterfactual or potential outcomes framework is given by Hernán.[20] Note that throughout this paper we will assume that some intervention on the mediator is available that can change the value of Z. If the mediator Z cannot be changed through an intervention, an alternative approach based on principal stratification will be needed.[21–23] The principal stratification approach avoids counterfactuals that reference interventions on Z by considering causal effects of the form $Y_a - Y_{a*}$ within strata of individuals for whom for each a, $Z_{a,i} = Z_{a,j}$ for all individuals i and j in that stratum.

To estimate the average causal effect of a treatment we need data on a set of variables X that suffice to control for confounding of the effect of treatment A on outcome Y. Essentially what is required is that within strata of the variables X the different treatment groups have comparable counterfactual outcomes so that any difference between treatment groups must be attributable to the effect of the treatment. Formally, we will use the notation $A \coprod B|C$ to denote that A is independent of B given C. The no unmeasured confounding condition for the effect of A on Y is then $Y_a \coprod A|X$, that is, the counterfactual outcomes are independent of treatment given the covariates X. Under this no unmeasured confounding assumption, the researcher can use the observed outcomes within strata of X, weighted by the probability of X, as a valid estimate of average counterfactual outcomes, that is, $\mathbb{E}[Y_a] = \sum_x \mathbb{E}[Y|A = a, X = x]P(X = x)$. Note that the left hand side of the equation is a counterfactual quantity whereas the right hand side is given entirely in observable quantities. When X contains numerous covariates then the quantity $\mathbb{E}[Y|A = a, X = x]$ is often estimated by means of regression or propensity score analysis.[24]

Direct and Indirect Effects

We will now introduce definitions concerning direct and indirect effects and conditions to identify them. For illustration we will consider an example given by Robins and Greenland.[7] Let treatment A be smoking, let the outcome Y be cardiovascular disease, and let the intermediate Z denote hyperlipidemia. Pearl gave the following definitions for controlled and natural direct and indirect effects based on interventions on the mediator Z.[8] Robins and Greenland provided related definitions.[7] The controlled direct effect of treatment A on outcome Y comparing $A = a$ with $A = a^*$ and setting Z to z is defined by $Y_{az} - Y_{a^*z}$ and measures the effect of A on Y not mediated through Z, that is, the effect of A on Y after intervening to fix the mediator to some value z. Thus in the smoking example, comparing smoking ($A = 1$) to nonsmoking ($A = 0$) and setting hyperlipidemia to 0 would give $Y_{10} - Y_{00}$, the effect of smoking intervening to eliminate hyperlipidemia, through perhaps either diet or medication. Although we will not in general be able to calculate individual controlled direct effects from data, we will see below that in certain instances we can estimate the average controlled direct effect for a population, $\mathbb{E}[Y_{az} - Y_{a^*z}]$.

The controlled direct effect then represents the effect of treatment on the outcome intervening to fix the intermediate variable to some particular level. In contrast with controlled directed effects, natural direct effects fix the intermediate variable for each individual to the level it would have been under the presence or absence of treatment. The natural direct effect of treatment A on outcome Y comparing $A = a$ with $A = a^*$ intervening to set Z to what it would have been if treatment had been $A = a^*$ is formally defined by $Y_{aZ_{a^*}} - Y_{a^*Z_{a^*}}$. Essentially the natural direct effect assumes that the intermediate Z is set to Z_{a^*}, the level it would have been for each individual had treatment been a^*, and then compares the direct effect of treatment (with the intermediate set to the level Z_{a^*}). Thus in the smoking example, $Y_{1Z_0} - Y_{0Z_0}$ compares the effect of smoking to nonsmoking assuming that hyperlipidemia is set to what it would have been for each subject had he or she not smoked. Corresponding to a natural direct effect is a natural indirect effect. The natural indirect effect comparing $A = a$ with $A = a^*$ and intervening to set treatment A to a is formally defined by $Y_{aZ_a} - Y_{aZ_{a^*}}$. The natural indirect effect assumes that treatment is set to some level $A = a$ and then compares what would have happened if the mediator were set to what it would have been if treatment had been a versus what would have happened if the mediator were set to what it would have been if treatment had been a^*. In the smoking example, $Y_{1Z_1} - Y_{1Z_0}$ measures the effect of smoking on cardiovascular disease mediated by hyperlipidemia; specifically $Y_{1Z_1} - Y_{1Z_0}$ compares the cardiovascular disease status setting hyperlipidemia to what it would have been had the individual smoked versus setting hyperlipidemia to what it would have been had the individual not smoked, assuming that there was an intervention so that the individual did in fact smoke. A total effect can be decomposed into a natural direct and indirect effect. In the smoking example, it is easily verified that the total effect of smoking on cardiovascular disease $Y_1 - Y_0$ can be written as $Y_1 - Y_0 = Y_{1Z_1} - Y_{0Z_0} = (Y_{1Z_1} - Y_{1Z_0}) + (Y_{1Z_0} - Y_{0Z_0})$ where the first expression in the sum is the indirect or mediated effect described previously and the second expression is the natural direct effect described previously. One can thus subtract a

natural direct effect from the total effect to get a natural indirect effect or one can subtract a natural indirect effect from a total direct effect to get a natural direct effect.

The distinction between controlled and natural direct effects is an important one. Pearl argues that controlled direct effects have a prescriptive interpretation and that natural direct effects have a descriptive interpretation.[8] Controlled direct effects are prescriptive, in that, we consider the effect of some treatment A after prescribing or intervening on the intermediate Z; natural direct effects are descriptive, in that, we consider the effect of treatment A if we let the intermediate be whatever it naturally would have been under a particular scenario ($A = a^*$). Controlled direct effects may be of interest in policy settings in which both the treatment and the intermediate will be manipulated. Natural direct effects are of greater interest when attempting to address questions concerning the manner in which treatment A brings about the outcome. A further difference between controlled and natural direct effects is to be noted: the aforementioned effect decomposition works for natural direct and indirect effects but not for controlled direct effect. If one subtracts a controlled direct effect from a total effect the resulting quantity cannot in general be interpreted as an indirect effect unless there is no interaction between the effects of the treatment and the mediator on the outcome.[6] If there is indeed no interaction between the effects of the treatment and the mediator on the outcome, then controlled direct effects and natural direct effects are equivalent because $Y_{az} - Y_{a^*z}$ will be constant for all values of z and thus $Y_{az} - Y_{a^*z} = Y_{aZ_{a^*}} - Y_{a^*Z_{a^*}}$.

As noted in the previous section, it is well understood that to estimate causal effects from observational data, the researcher needs to have data on some set of covariates X that suffice to control for confounding. For controlled direct effects, one needs not just 1 no unmeasured confounding condition but 2. To estimate controlled direct effects one first needs that

$$Y_{az} \amalg A|X \text{ (no unmeasured confounding for } A \sim Y \text{ relationship).} \quad (1)$$

Condition (1) can be interpreted as that conditional on X there is no unmeasured confounding for the treatment-outcome relationship. In addition to condition (1), one also needs a second no unmeasured confounding condition: one needs that there be no unmeasured confounding for the mediator-outcome relationship. Formally, we require

$$Y_{az} \amalg Z|A, X, W \text{ (no unmeasured confounding for } Z \sim Y \text{ relationship).} \quad (2)$$

In general, there may be some variables W that do not confound the treatment-outcome relationship but that do confound the mediator-outcome relationship. Condition (2) states that conditional on treatment A and the confounding variables X and W, there is no unmeasured confounding of the mediator-outcome relationship. If conditions (1) and (2) hold controlled direct effects can be estimated from the data.[8,10,25,26] When attempts are made to estimate direct effects by including the mediator in a regression of the outcome on the treatment, it is often forgotten that one must control not only those variables that confound the treatment-outcome relationship but also those that confound the mediator-outcome relationship. When the confounders of the mediator-outcome relationship are not controlled for, this leads to biased estimates for the controlled direct effect.[27] See the work of Hernández-Díaz et al[28] for an interesting application of how this observation was used to solve the "birth-weight paradox" in perinatal epidemiology. In general, it will be assumed that conditions (1) and (2) hold for all a and z; however, the conditions only need to hold for those values of a and z concerning which comparison is being made. Similar remarks hold for the other identification conditions in this paper. We will discuss in the following section how marginal structural models can be used to estimate controlled direct effects if conditions (1) and (2) hold. In the eAppendix (available in the online version of this manuscript) we also give empirical expressions for average controlled direct effects that follow from Robins' g-formula.[25,26]

To estimate natural direct and indirect effects we need conditions (1) and (2) to hold but we also need 2 additional no unmeasured confounding assumptions. In addition to conditions (1) and (2) we require

$$Z_a \amalg A|X \text{ (no unmeasured confounding for the } A \sim Z \text{ relationship).} \quad (3)$$

Condition (3) can be interpreted as that conditional on X there is no unmeasured confounding for the treatment-mediator relationship. Condition (1) required no unmeasured confounding of the treatment-outcome relationship; condition (3) requires no unmeasured confounding of the treatment-mediator relationship. We also require

$$Y_{az} \amalg Z_{a^*}|X \text{ (no } Z \text{ confounders which are effects of } A \sim Y \text{).} \quad (4)$$

The interpretation of condition (4) is somewhat difficult. Condition (4) requires that there be no variable that is a consequence of treatment that confounds the mediator-outcome relationship.[8] If there is a consequence of treatment that confounds the mediator-outcome relationship then we cannot in general estimate natural direct effects; we could however still potentially estimate controlled direct effects because condition (2) allows for a set of variables W that do not confound the treatment-outcome relationship but which do

confound the mediator-outcome relationship. We will discuss below how to adjust appropriately for such variables by using marginal structural models. However, if conditions (1)–(4) hold then we can estimate natural direct and indirect effects.[8,10] In the eAppendix, we give empirical expressions for average natural direct and indirect effects; we furthermore discuss below how marginal structural models can be used to estimate natural direct effects and indirect effects if conditions (1)–(4) hold. In summary then, if condition (1) holds then we can estimate total causal effects, if conditions (1) and (2) hold we can estimate controlled direct effects, and if conditions (1)–(4) hold then we can estimate natural direct and indirect effects.

One final comment merits attention. As noted previously, if there is no interaction between the effects of the treatment and the mediator on the outcome, then controlled direct effects and natural direct effects can be shown to be equivalent.[7,9] In this case, conditions (1) and (2) suffice to estimate controlled direct effects and natural direct effects; conditions (3) and (4) are not necessary.

Using Marginal Structural Models to Estimate Controlled Direct Effects

A marginal structural model is a model for the expected value of a certain counterfactual outcome.[15–17] For example, a marginal structural model for $\mathbb{E}[Y_{az}]$ might take the form

$$\mathbb{E}[Y_{az}] = \alpha_0 + \alpha_1 a + \alpha_2 z + \alpha_3 az. \quad (5)$$

A marginal structural model differs from a regression model, in that, the model is for counterfactual outcomes not observed outcomes and in that the model given in (5) for example is marginal not conditional on any covariates. In contrast, a regression model for Y on A and Z would take the form

$$\mathbb{E}[Y|A = a, Z = z] = \beta_0 + \beta_1 a + \beta_2 z + \beta_3 az. \quad (6)$$

Because a marginal structural model is a model for the counterfactual outcomes and because most of the counterfactual outcomes are not observed, certain no-unmeasured-confounding assumptions are required, and, in the case of time-varying confounders, special estimation techniques such as inverse probability of treatment weighting will often have to be used to estimate the coefficients of a marginal structural model.[15–17] The coefficients from regression model (6), (β_0, β_1, β_2, β_3) will only correspond to those of the marginal structural model in (5), (α_0, α_1, α_2, α_3) if the effects of treatment A and mediator Z on outcome Y are unconfounded. If treatment-outcome and mediator-outcome relationships are confounded but if data are available for the confounding variables (if for example conditions (1) and (2) above hold) then the coefficients of the marginal structural model can still be consistently estimated using inverse probability of treatment weighting techniques.

Once the coefficients of the marginal structural model in (5) have been estimated it is straightforward to calculate controlled direct effects.[12,17] For example, if the marginal structural model (5) holds then the average controlled direct effect is given by

$$\begin{aligned}\mathbb{E}[Y_{az} - Y_{a^*z}] &= (\alpha_0 + \alpha_1 a + \alpha_2 z + \alpha_3 az) \\ &\quad - (\alpha_0 + \alpha_1 a^* + \alpha_2 z + \alpha_3 a^* z) \\ &= \alpha_1(a - a^*) + \alpha_3(a - a^*)z.\end{aligned}$$

More generally, a marginal structural model for $\mathbb{E}[Y_{az}]$ might involve any function of a and z, that is, we may have $\mathbb{E}[Y_{az}] = g(a, z)$ for any function g. Once the marginal structural model is fit, the estimation of controlled direct effects is again straightforward because $\mathbb{E}[Y_{az} - Y_{a^*z}] = g(a, z) - g(a^*, z)$.

A paper by Robins et al contains extensive detail about fitting a marginal structural model and intuition as to why the inverse probability of treatment weighting procedure works.[16] Here we will provide a brief outline. The inverse probability of treatment weighting technique to fit a model such as (5) makes use of a regression of Y on A and Z but handles confounding not by including covariates in the regression model but by weighting. To fit a marginal structural model with 2 intervention variables, A and Z, 2 sets of weights are needed. For each individual i consider the following weights

$$w_i^A = \frac{P(A = a_i)}{P(A = a_i|X = x_i)}$$

and

$$w_i^Z = \frac{P(Z = z_i|A = a_i)}{P(Z = z_i|A = a_i, X = x_i, W = w_i)}.$$

The denominator of the w_i^A is the probability of receiving the treatment the individual in fact received conditional on the covariate X taking the value x_i. The denominator of the w_i^Z is the probability of having the value of the mediator that the individual in fact had conditional on $A = a_i$, $X = x_i$ and $W = w_i$. The inclusion of the probabilities in numerator is in fact optional but tends to lead to more efficient estimation.[16] If A and Z are binary, these numerator and denominator probabilities can be estimated by a logistic regression; if they are categorical or ordinal by a multinomial or ordinal logistic regression; and if continuous, the probabilities can be replaced by values from a probability density function.[16] Robins has shown that to a fit a marginal structural model such as that in (5), a weighted regression of Y on A and Z (including an $A * Z$ interaction term because one is present in model [5])

in which each individual is weighted by $w_i^A * w_i^Z$ will give valid estimates for the coefficients in the marginal structural model given in (5).[15] See Robins et al[16] for further details. Brumback et al[29] describe sensitivity analysis techniques for marginal structural models and these could also be used for the marginal structural model in (5) to conduct sensitivity analyses of controlled direct effects estimates to the no unmeasured confounding assumptions given in (1) and (2).

An additional comment merits attention. In certain cases, traditional regression methods will suffice to estimate controlled direct effects. If the set of variables X that suffice to control for confounding of the treatment-outcome relationship also suffice to control for confounding of the mediator-outcome relationship (ie, if W can be chosen to be empty) then a regression of Y on A, Z and X will allow for the estimation of controlled direct effects.[10] If, however, there is a consequence of treatment that confounds the mediator-outcome relationship so that W cannot be chosen to be empty, then this regression approach will not work. The approach described above using marginal structural model will, however, still give valid estimates. The use of marginal structural model and inverse probability of treatment weighting does require a positivity assumption that the probabilities in the denominator of the weights are nonzero.[15–17] In cases in which this assumption is violated, Robins' structural nested models can still be used.[30,31] Ten Have et al have described a method to estimate direct effects by using structural nested models in settings in which there is no interaction between the effects of the treatment and the mediator on the outcome.[32] Robins provides extensive discussion of the estimation of controlled direct effects by using structural nested direct effect models.[11]

Using Marginal Structural Models to Estimate Natural Direct and Indirect Effects

Unlike the estimation of controlled direct effects such as $\mathbb{E}[Y_{az} - Y_{a^*z}]$ that required a single marginal structural model for $\mathbb{E}[Y_{az}]$, the estimation of natural direct effects $\mathbb{E}[Y_{aZ_a} - Y_{aZ_{a^*}}]$ described here will require 2 marginal structural models, 1 related to counterfactuals of the form Y_{az} and 1 related to counterfactuals of the form Z_a; see the work of van der Laan and Petersen for related discussion.[12] Also unlike the marginal structural model for the estimation of controlled direct effects, the marginal structural models for natural direct effects will be made conditional on the baseline covariates X. For example instead of using a marginal structural model such as (5) for $\mathbb{E}[Y_{az}]$ we will use a marginal structural model of the form

$$\mathbb{E}[Y_{az}|X = x] = \theta_0 + \theta_1 a + \theta_2 z + \theta_3 az + \theta_4' x \qquad (7)$$

or more generally, $\mathbb{E}[Y_{az}|X = x] = g(a, z, x)$. Note that x may be multivariate and if so θ_4 will denote a vector of coefficients. We will also use a second marginal structural model for the counterfactual Z_a:

$$\mathbb{E}[Z_a|X = x] = \gamma_0 + \gamma_1 a + \gamma_2' x \qquad (8)$$

or more generally, $\mathbb{E}[Z_a|X = x] = h(a, x)$. We will be able to use these 2 marginal structural models together to estimate natural direct effects because of the following result. The proof is given in Appendix 1.

Result 1. Suppose that $\mathbb{E}[Y_{az}|X = x] = g(a, z, x)$, that $\mathbb{E}[Z_a|X = x] = h(a, x)$ and that g is linear in z then if condition (4) holds then $\mathbb{E}[Y_{aZ_{a^*}}|x = x] = g(a, h(a^*, x), x)$.

Note that the requirement that g is linear in z is satisfied by the conditional structural model given in (7). More generally, the requirement that g is linear in z means that there will be no quadratic or higher-order terms in z. There can be interactions between z and a as in (7) or between z and x or between z and a and x but simply not terms such as z^2 or z^3 or $\log(z)$ and so forth. van der Laan and Petersen discuss estimation of natural direct effects when this linearity assumption does not hold; the linearity assumption allows for simplification in the estimation procedure.[12] As noted above we can estimate natural direct and indirect effects if assumptions (1)–(4) hold. In using structural models to estimate natural direct and indirect effects, assumption (4) will be needed to be able to apply Result 1; assumptions (1) and (2) are needed to fit the marginal structural model for $\mathbb{E}[Y_{az}|X = x]$; assumption (3) is needed to fit the marginal structural model for $\mathbb{E}[Z_a|X = x]$. We will discuss below how one goes about estimating the marginal structural models given in (7) and (8). We will first illustrate how one can use Result 1 to estimate natural direct effects from the marginal structural models once they have been fit. If, for example, the marginal structural models (7) and (8) hold then the natural direct effect is given by

$$\begin{aligned}
&\mathbb{E}[Y_{aZ_{a^*}} - Y_{a^*Z_{a^*}}|X = x] \\
&\quad = g(a, h(a^*, x), x) - g(a^*, h(a^*, x), x) \\
&\quad = \{\theta_0 + \theta_1 a + \theta_2 h(a^*, x) + \theta_3 a h(a^*, x) + \theta_4' x\} \\
&\quad - \{\theta_0 + \theta_1 a^* + \theta_2 h(a^*, x) + \theta_3 a^* h(a^*, x) + \theta_4' x\} \\
&\quad = \{\theta_1 a + \theta_3 a h(a^*, x)\} - \{\theta_1 a^* + \theta_3 a^* h(a^*, x)\} \\
&\quad = \{\theta_1 a + \theta_3 a(\gamma_0 + \gamma_1 a^* + \gamma_2' x)\} \\
&\quad - \{\theta_1 a^* + \theta_3 a^*(\gamma_0 + \gamma_1 a^* + \gamma_2' x)\} \\
&\quad = (\theta_1 + \theta_3\gamma_0 + \theta_3\gamma_1 a^* + \theta_3\gamma_2' x)(a - a^*).
\end{aligned}$$

The average natural direct effect is then given by $\mathbb{E}[Y_{aZ_a} - Y_{aZ_{a^*}}] = \sum_x \mathbb{E}[Y_{aZ_a} - Y_{aZ_{a^*}}|X = x]P(X = x) = (\theta_1 + \theta_3\gamma_0 + \theta_3\gamma_1 a^* + \theta_3\gamma_2'\mathbb{E}[X])(a - a^*)$. Thus natural direct effects can be computed directly from the coefficients of the mar-

ginal structural models given in (7) and (8) and the distribution of the confounding variables X. More generally, for other marginal structural models of the form $\mathbb{E}[Y_{az}|X = x] = g(a, z, x)$ and $\mathbb{E}[Z_a|X = x] = h(a, x)$ average natural direct effect can be estimated by $\mathbb{E}[Y_{aZ_{a^*}} - Y_{a^*Z_{a^*}}] = \sum_x \{g(a, h(a^*, x), x) - g(a^*, h(a^*, x), x)\}P(X = x)$.

We note that Result 1 can also be used to estimate natural indirect effects. This can be done either by estimating natural direct effect effects and subtracting these from total effects or it can be calculated by using Result 1 directly. For example, if the marginal structural models (7) and (8) hold then the natural indirect effect is given by

$$\begin{aligned}
&\mathbb{E}[Y_{aZ_a} - Y_{aZ_{a^*}}|X = x] \\
&\quad = g(a, h(a, x), x) - g(a, h(a^*, x), x) \\
&\quad = \{\theta_0 + \theta_1 a + \theta_2 h(a, x) + \theta_3 a h(a, x) + \theta_4' x\} \\
&\quad - \{\theta_0 + \theta_1 a + \theta_2 h(a^*, x) + \theta_3 a h(a^*, x) + \theta_4' x\} \\
&\quad = \{\theta_2 h(a, x) + \theta_3 a h(a, x)\} - \{\theta_2 h(a^*, x) + \theta_3 a h(a^*, x)\} \\
&\quad = \{\theta_2(\gamma_0 + \gamma_1 a + \gamma_2 x) + \theta_3 a(\gamma_0 + \gamma_1 a + \gamma_2' x)\} \\
&\quad - \{\theta_2(\gamma_0 + \gamma_1 a^* + \gamma_2 x) + \theta_3 a(\gamma_0 + \gamma_1 a^* + \gamma_2' x)\} \\
&\quad = \theta_2\gamma_1(a - a^*) + \theta_3\gamma_1 a(a - a^*).
\end{aligned}$$

Because $\mathbb{E}[Y_{aZ_a} - Y_{aZ_{a^*}}|X = x] = \theta_2\gamma_1(a - a^*) + \theta_3\gamma_1 a(a - a^*)$ does not depend on x it follows that the average natural indirect effect is also given by $\mathbb{E}[Y_{aZ_a} - Y_{aZ_{a^*}}] = \theta_2\gamma_1(a - a^*) + \theta_3\gamma_1 a(a - a^*)$. More generally, for other marginal structural models of the form $\mathbb{E}[Y_{az}|X = x] = g(a, z, x)$ and $\mathbb{E}[Z_a|X = x] = h(a, x)$ the average natural indirect effect can be estimated by $\mathbb{E}[Y_{aZ_a} - Y_{aZ_{a^*}}] = \sum_x \{g(a, h(a, x), x) - g(a, h(a^*, x), x)\}P(X = x)$.

It is important to remember that the application of Result 1, which allows the use of marginal structural models (7) and (8) to estimate natural direct and indirect effects, requires that assumption (4) holds. If assumption (4) is violated because there is a consequence of treatment that confounds the mediator-outcome relationship, then the procedure described here and the expressions for natural direct and indirect effects based on the structural model coefficients will not be valid.

Robins et al (section 9) discuss fitting marginal structural models of the form (7) and (8).[16] The procedure for fitting these models is very similar to that of fitting a regular marginal structural model. For the model given in (7), weights can be estimated just as described in the previous section, based on the denominator probability that A takes the value it in fact does (conditional on X) and on the denominator probability that Z takes the value it in fact does (conditional on X, W, and A). Numerator probabilities in the weights can also be made conditional on X[16] and the weight for w_i^A will then reduce to 1. The only difference in the procedure for fitting the conditional structural model (7) compared with fitting the marginal structural model (5) is that when running the weighted regression for Y, Y is regressed not simply on A, Z, and AZ but on A, Z, AZ, and X. The inverse probability of treatment weighting technique will give valid estimates for the coefficients in model (7) provided the no unmeasured confounding conditions (1) and (2) hold. For the model given in (8), weights are estimated based on the probability that A takes the value it in fact does (conditional on X) and a weighted regression of Z on A and X is used. The covariates X are included in the final weighted regressions not in attempt to control for confounding (this is taken care of by the weighting)[16]; rather, the covariates X are included in the final weighted regressions so that the structural model is made conditional on $X = x$ to allow for the application of Result 1 in the estimation of natural direct and indirect effects. We note that if W can be chosen to be empty then the marginal structural models can be fit by using traditional regression methods; inverse probability of treatment weighting methods are not needed. See Petersen et al[10] for further discussion of the estimation of natural direct effects using regression methods. We note further that Pearl[8] provides somewhat more general identification conditions for natural direct effects than those given in (1)–(4). The approach described here of using marginal structural models to estimate natural direct and indirect effects could also be applied to Pearl's more general identification conditions but the details of fitting the conditional structural models would be somewhat different. In particular, if assumption (4) holds conditional on some subset of the baseline covariates X, then the marginal structural models (7) and (8) can be modified so that they are conditional on the subset, rather than the entirety of X; confounding beyond this subset can be addressed by the inverse probably of treatment weighting.

Petersen et al. replace condition (4) with the slightly more general condition $(Y_{az} - Y_{a^*z}) \coprod Z_{a^*}|X$. The approach described here of using marginal structural models will also apply to this weaker condition for natural direct effects. van der Laan and Petersen give an identification results for natural direct effects in the setting of time-varying treatments.[12,33] In Appendix 2 we show how the approach described of estimating direct and indirect effects by using marginal structural models can be extended to the estimation of direct and indirect effects of time-varying treatments. Repeated-measures marginal structural models with time-varying weights can be employed to estimate natural direct and indirect effects in the setting of time-varying treatments.[34]

DISCUSSION

In this paper we have discussed a set of conditions needed to estimate direct and indirect effects. We have noted that to estimate controlled direct effects there must be no unmeasured confounding of the treatment-outcome relation-

ship and also of the mediator-outcome relationship.[7,8,10] To estimate natural direct and indirect effects these 2 no unmeasured confounding conditions must also hold and, in addition, there must in general be no unmeasured confounding of the treatment-mediator relationship and there must be no consequence of treatment that confounds the mediator-outcome relationship.[8] We have also discussed how marginal structural models can be used to estimate controlled direct effects and we have shown that natural direct and indirect effects can be estimated by using 2 marginal structural models that are made conditional on the baseline covariates.

The approach of using marginal structural models to estimate direct and indirect effects is in principle applicable to binary, categorical, and continuous outcomes. For controlled direct effects, it suffices to choose an appropriate marginal structural model for the outcome under consideration. For example, if the outcome is binary a logistic regression marginal structural model could be used; if the outcome is continuous a linear marginal structural model can be used. To estimate natural direct and indirect effects from structural models we imposed the condition that the model for the expected counterfactual outcome be linear in the intermediate. Although this will probably not be a substantial constraint for continuous outcomes, modeling binary outcomes using models with linear links can be problematic. For example, Wacholder[35] notes that the convergence properties of maximum likelihood estimators for Bernoulli regressions with linear links are often poor. Thus, in practice, the approach described here will be useful for controlled direct effects regardless of the type of outcome but, when natural direct and indirect effects are of interest, will only be able to be easily implemented for continuous outcomes. van der Laan and Petersen discuss estimation of natural direct effects when the linearity assumption does not hold.[12]

This work could be extended in a number of directions. First, our discussion has focused only on the estimation of controlled direct effects and natural direct and indirect effects, not on their standard errors. In all cases standard errors for direct and indirect effects could be computed using bootstrap methods. Other methods for estimating standard errors could also be explored. Second, estimates of direct and indirect effects from marginal structural models could be compared with those from regression models to examine issues of efficiency and of robustness to violations in assumptions. Finally, our analysis has been restricted to nonclustered cohorts. In related work, definitions and identification conditions have been developed for the estimation of direct and indirect effects in settings in which individuals are clustered in groups or neighborhoods and in which the treatment is administered at the neighborhood level.[36] Furthermore, Hong and Raudenbush[37] have recently developed a class of multilevel marginal structural models that generalize the marginal structural models of Robins[15–17] to data that are clustered. Future work could extend the ideas in this paper to consider how these multilevel marginal structural models could be used to estimate direct and indirect effects in neighborhood-based research.

REFERENCES

1. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*. 1986;51:1173–1182.
2. Bollen KA. Total, direct and indirect effects in structural equation models. In: Clogg CC, ed. *Sociological Methodology*. Washington, DC: American Sociological Association; 1987:37–69.
3. Holland PW. Causal inference, path analysis, and recursive structural equations models. In: Clogg CC, ed. *Sociological Methodology*. Washington, DC: American Sociological Association; 1988:449–484.
4. Sobel ME. Effect analysis and causation in linear structural equation models. *Psychometrika*. 1990;55:495–515.
5. Raudenbush SW, Sampson R. Assessing direct and indirect effects in multilevel designs with latent variables. *Sociol Methods Res*. 1999;28:123–153.
6. Kaufman JS, MacLehose RF, Kaufman S. A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiol Perspect Innov*. 2004;1:4.
7. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3:143–155.
8. Pearl J. Direct and indirect effects. In: *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence.* San Francisco, CA: Morgan Kaufmann; 2001:411–420.
9. Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. In: Green P, Hjort NL, Richardson S, eds. *Highly Structured Stochastic Systems*. New York: Oxford University Press; 2003:70–81.
10. Peterson ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. *Epidemiology.* 2006;17:276–284.
11. Robins JM. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In: Glymour C, Cooper GF, eds. *Computation, Causation, and Discovery*. Menlo Park, CA, Cambridge, MA: AAAI Press/The MIT Press; 1999:349–405.
12. van der Laan MJ, Petersen ML. Estimation of direct and indirect causal effects in longitudinal studies. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 155. 2004. Available at: http://www.bepress.com/ucbbiostat/paper155.
13. Didelez V, Dawid AP, Geneletti S. Direct and Indirect Effects of Sequential Treatments. In: *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence.* Arlington, VA: AUAI Press; 2006:138:–146.
14. Geneletti S. Identifying direct and indirect effects in a non-counterfactual framework. *J R Stat Soc. Series B*. 2007;69:199–216.
15. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME, Berry D, eds. *Statistical Models in Epidemiology: The Environment and Clinical Trials*. New York: Springer-Verlag; 1999:95–134.
16. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550–560.
17. Robins JM. Association, causation, and marginal structural models. *Synthese*. 1999;121:151–179.
18. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66:688–701.
19. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat*. 1978;6:34–58.
20. Hernán MA. A definition of causal effect for epidemiological studies. *J Epidemiol Comm Health*. 2004;58:265–271.
21. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002;58:21–29.
22. Rubin DB. Direct and indirect effects via potential outcomes. *Scand J Stat*. 2004;31:161–170.
23. VanderWeele TJ. Simple relations between principal stratification and direct and indirect effects. *Stat Prob Lett.* 2008;78:2957–2962.
24. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41–55.

25. Robins JM. A new approach to causal inference in mortality studies with sustained exposure period - application to control of the healthy worker survivor effect. *Math Model*. 1986;7:1393–1512.
26. Robins JM. Addendum to a new approach to causal inference in mortality studies with sustained exposure period - application to control of the healthy worker survivor effect. *Comput Math Appl*. 1987;14:923–945.
27. Cole SR, Hernán MA. Fallibility in estimating direct effects. *International J Epidemiol*. 2002;31:163–165.
28. Hernández-Díaz S, Schisterman EF, Hernán MA. The birth weight "paradox" uncovered? *Am J Epidemiol*. 2006;164:1115–1120.
29. Brumback BA, Hernán MA, Haneuse SJPA, Robins JM. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Stat Med*. 2004;23:749–767.
30. Robins JM, Blevins D, Ritter G, et al. G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology*. 1992;3:319–336.
31. Hernán MA, Cole SR, Margolick JB, Cohen MH, Robins JM. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiol Drug Saf.* 2005;14:477–491.
32. Ten Have TR, Joffe MM, Lynch KG, et al. Causal mediation analyses with rank preserving models. *Biometrics*. 2007;63:926–934.
33. van der Laan MJ, Petersen ML. Direct effect models. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 187. 2005. Available at: http://www.bepress.com/ucbbiostat/paper187.
34. Hernán MA, Brumback B, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Stat Med*. 2002;21:1689–1709.
35. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol*. 1986;123:174–184.
36. VanderWeele TJ. Direct and indirect effects of clustered and longitudinal treatments. *Sociol Res Methods.* [under review].
37. Hong G, Raudenbush SW. Causal inference for time-varying instructional treatments. *J Educ Behav Stat*. 2008;33:333–362.

APPENDIX 1

Proof of Result 1

Proof. By iterated expectations we have that $\mathbb{E}[Y_{aZ_{a^*}}|X = x] = \sum_z \mathbb{E}[Y_{aZ_{a^*}}|X = x, Z_{a^*} = z]P(Z_{a^*} = z|X = x) = \sum_z \mathbb{E}[Y_{az}|X = x, Z_{a^*} = z]P(Z_{a^*} = z|X = x) = \sum_z \mathbb{E}[Y_{az}|X = x]P(Z_{a^*} = z|X = x)$, the final equality holding by (4). Substituting $g(a, z, x)$ for $\mathbb{E}[Y_{az}|X = x]$ gives $\sum_z g(a, z, x)\, P(Z_{a^*} = z|X = x)$. Because g is linear in z we can rewrite this as $g(a, \sum_z zP(Z_{a^*} = z|X = x), x) = g(a, \mathbb{E}[Z_{a^*}|X = x], x) = g(a, h(a^*, vx), x)$.

APPENDIX 2

Using Marginal Structural Models to Estimate Direct and Indirect Effects for Time-Varying Treatments

In the longitudinal setting we will use the following definitions and notation. We index time by $s = 0, 1, \ldots, S$. Let Y denote the outcome of interest after time period S. Let A_s denote the treatment or exposure at time s; note that this treatment variable may change over time. Let X_s denote covariates immediately prior to time s. Let Z_s denote the intermediate variable between time s and time $s + 1$. Let $\overline{A}_s$ denote the vector $(A_0, A_1, \ldots, A_s)$ and similarly let $\overline{X}_s$ and $\overline{Z}_s$ denote X and Z, respectively from time periods 0 through s. We will use $\overline{A}$, $\overline{X}$, and $\overline{Z}$ to denote $\overline{A}_S$, $\overline{X}_S$, and $\overline{Z}_S$. We will consider static treatment regimes in which treatment history $\overline{A}$ is set to some fixed regime $\bar{a}$ which does not vary with the covariate or intermediate histories $\overline{X}$ and $\overline{Z}$. Comparisons of the effects of dynamic treatment regimes which do depend on covariate or intermediate histories can be addressed with structural nested models.[11,15,30,31]

Let $Y_{\bar{a}\bar{z}}$ denote a subject's counterfactual outcome if treatment $\overline{A}$ were set to $\bar{a}$ and if $\overline{Z}$ were set to $\bar{z}$. Let $Z_{\bar{a}_s}$ denote the counterfactual value of the intermediate Z_s if $\overline{A}_s$ were set to $\bar{a}_s$. Let $\overline{Z}_{\bar{a}_s}$ denote $Z_{\bar{a}_s}$ from time periods 0 through s and let $\bar{Z}_{\bar{a}}$ denote $\overline{Z}_{\bar{a}_s}$. For this longitudinal setting we can define the controlled direct effect comparing $\overline{A} = \bar{a}$ with $\overline{A} = \overline{a^*}$ and setting $\overline{Z}$ to $\bar{z}$ by $Y_{\bar{a}\bar{z}} - Y_{\overline{a^*}\bar{z}}$. We can define the natural direct effect comparing $\overline{A} = \bar{a}$ with $\overline{A} = \overline{a^*}$ intervening to set Z to it would have been if treatment had been $\overline{a^*}$ by $Y_{\overline{a\bar{Z}_{a^*}}} - Y_{\overline{a^*}\overline{\bar{Z}_{a^*}}}$. We can define the natural indirect effect comparing $\overline{A} = \bar{a}$ with $\overline{A} = \overline{a^*}$ and intervening to set treatment $\overline{A}$ to $\bar{a}$ by $Y_{\bar{a}\bar{Z}_{\bar{a}}} - Y_{\bar{a}\bar{Z}_{a^*}}$. It follows from the results of Robins[25,26] that if the following 2 conditions hold:

$$Y_{\bar{a}\bar{z}} \amalg A_s|\overline{X}_s,\ \overline{A}_{s-1},\ \overline{Z}_{s-1} \text{ for all } s \tag{9}$$

and

$$Y_{\bar{a}\bar{z}} \amalg Z_s|\overline{X}_s,\ \overline{A}_s,\ \overline{Z}_{s-1} \text{ for all } s \tag{10}$$

then controlled direct effects are identified. These conditions are longitudinal generalizations of conditions (1) and (2) in the text. Condition (9) requires that for each time-period s the effect of treatment A_s on outcome Y is unconfounded given the covariate history $\overline{X}_s$ up until time s, the treatment history $\bar{A}_{s-1}$ up until time $s - 1$ and the mediator history $\overline{Z}_{s-1}$ up until time s. Condition (10) requires that for every time-period s the effect of the mediator Z_s on outcome Y is unconfounded given the covariate history $\overline{X}_s$ up until time s, the treatment history $\overline{A}_s$ up until time s and the mediator history $\overline{Z}_{s-1}$ up until time s. Consider a marginal structural model for $Y_{\bar{a}\bar{z}}$:

$$\mathbb{E}[Y_{\bar{a}\bar{z}}] = g(\bar{a}, \bar{z}).$$

If conditions (9) and (10) hold then this marginal structural model can be fit using inverse probability of treatment weighting estimation.[15–17] Once the marginal structural model is fit then average controlled direct effects can be estimated from the marginal structural model since $\mathbb{E}[Y_{\bar{a}\bar{z}} - Y_{\overline{a^*}\bar{z}}] = g(\bar{a}, \bar{z}) - g(\bar{a}^*\bar{z})$. In order to estimate natural direct and

indirect effects for a time-varying treatment, in addition to conditions (9) and (10) we also need

$$Z_{\bar{a}_s} \amalg A_s | \overline{X}_s, \overline{A}_{s-1}, \overline{Z}_{s-1} \text{ for all } s \tag{11}$$

and

$$Y_{\bar{a}\bar{z}} \amalg Z_{\bar{a}^*} | X_0 \tag{12}$$

van der Laan and Petersen note that if conditions (9)–(12) hold then natural direct effects are identified.[33] Condition (11) requires that for each time-period s the effect of treatment A_s on the mediator Z_s is unconfounded given the covariate history $\overline{X}_s$ up until time s, the treatment history $\overline{A}_{s-1}$ until time $s-1$ and the mediator history $\overline{Z}_{s-1}$ up until time s. Condition (12) is a longitudinal generalization of condition (4) in the text. It will be violated if for any s there is an effect of treatment A_s which itself confounds the relationship between the mediator Z_s and the outcome Y. Suppose that we had 2 marginal structural models conditional on the covariates X_0, 1 for $Y_{\bar{a}\bar{z}}$ and 1 for $\overline{Z}_{\bar{a}_s} = (Z_{a_0}, Z_{\bar{a}_1}, \ldots, Z_{\bar{a}_S})$ so that $\mathbb{E}[Y_{\bar{a}\bar{z}}|X_0 = x_0] = g(\bar{a}, \bar{z}, x_0)$ and $\mathbb{E}[\overline{Z}_{\bar{a}}|X_0 = x_0] = h(\bar{a}, x_0)$. If conditions (9) and (10) hold then the marginal structural model $\mathbb{E}[Y_{\bar{a}\bar{z}}|X_0 = x_0] = g(\bar{a}, \bar{z}, x_0)$ can be fit using inverse probability of treatment weighting.[15–17] If condition (11) holds, then the model $\mathbb{E}[\overline{Z}_{\bar{a}}|X_0 = x_0] = h(\bar{a}, x_0)$ can also be fit using inverse probability of treatment weighting. Note, however, that since $\overline{Z}_{\bar{a}} = (Z_{a_0}, Z_{\bar{a}_1}, \ldots, Z_{\bar{a}_S})$ is multivariate, the model $\mathbb{E}[\overline{Z}_{\bar{a}}|X_0 = x_0] = h(\bar{a}, x_0)$ is a repeated measures structural model and appropriate time-varying weights must be used in the estimation.[34] Finally, suppose that condition (12) holds and suppose also that g is linear in $\bar{z}$ then $\mathbb{E}[Y_{\bar{a}\overline{Z}_{\bar{a}^*}}|X_0 = x_0] = \sum_{\bar{z}}\mathbb{E}[Y_{\bar{a}\bar{z}}|X_0 = x_0, \overline{Z}_{\bar{a}^*} = \bar{z}]P(\overline{Z}_{\bar{a}^*} = \bar{z}|X_0 = x_0) = \sum_{\bar{z}}\mathbb{E}[Y_{\bar{a}\bar{z}}|X_0 = x_0, \overline{Z}_{\bar{a}^*} = \bar{z}]P(\bar{Z}_{\bar{a}^*} = \bar{z}|X_0 = x_0) = \sum_{\bar{z}}\mathbb{E}[Y_{\bar{a}\bar{z}}|X_0 = x_0]P(\overline{Z}_{\bar{a}^*} = \bar{z}|X_0 = x_0) = \sum_{\bar{z}}g(\bar{a}, \bar{z}, x_0)P(\overline{Z}_{\bar{a}^*} = \bar{z}|X_0 = x_0) = g(\bar{a}, \sum_{\bar{z}}\bar{z}P(\overline{Z}_{\bar{a}^*} = \bar{z}|X_0 = x_0), x_0) = g(\bar{a}, \mathbb{E}[\bar{Z}_{\bar{a}^*}|X_0 = x_0]\ x_0) = g(\bar{a}, h(\overline{a^*}, x_0), x_0)$. Thus if conditions (9)–(12) hold and g is linear in $\bar{z}$, the marginal structural models can be fit and average natural direct effects can be estimated by $\mathbb{E}[Y_{\bar{a}\overline{Z}_{\bar{a}^*}} - Y_{\bar{a}^*\overline{Z}_{\bar{a}^*}}] = \sum_{x_0}\mathbb{E}[Y_{\bar{a}\overline{Z}_{\bar{a}^*}} - Y_{\bar{a}^*\overline{Z}_{\bar{a}^*}}|X_0 = x_0]P(X_0 = x_0) = \sum_{x_0}\{g(\bar{a}, h(\overline{a^*}, x_0), x_0) - g(\overline{a^*}, h(\overline{a^*}, x_0), x_0)\}P(X_0 = x_0)$. Average natural indirect effects can be estimated by $\mathbb{E}[Y_{\bar{a}\overline{Z}_{\bar{a}}} - Y_{\bar{a}\overline{Z}_{\bar{a}^*}}] = \sum_{x_0}\mathbb{E}[Y_{\bar{a}\overline{Z}_{\bar{a}}} - Y_{\bar{a}\overline{Z}_{\bar{a}^*}}|X_0 = x_0]P(X_0 = x_0) = \sum_{x_0}\{g(\bar{a}, h(\bar{a}, x_0), x_0) - g(\bar{a}, h(\overline{a^*}, x_0), x_0)\}P(X_0 = x_0)$.