American Journal of Epidemiology Advance Access published May 5, 2015

American Journal of Epidemiology
© The Author 2015. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/aje/kwv060

Invited Commentary

Invited Commentary: Boundless Science—Putting Natural Direct and Indirect Effects in a Clearer Empirical Context

Ashley I. Naimi*

* Correspondence to Dr. Ashley I. Naimi, Department of Obstetrics and Gynecology, Faculty of Medicine, McGill University, 687 Pine Avenue West, Room F 432, Montreal, QC H3A 1A1, Canada (e-mail: ashley.naimi@mcgill.ca).

Initially submitted October 17, 2014; accepted for publication January 28, 2015.

Epidemiologists are increasingly using natural effects for applied mediation analyses, yet 1 key identifying assumption is unintuitive and subject to some controversy. In this issue of the *Journal*, Jiang and VanderWeele (*Am J Epidemiol*. 2015;000(00):000–000) formalize the conditions under which the difference method can be used to estimate natural indirect effects. In this commentary, I discuss implications of the controversial "cross-worlds" independence assumption needed to identify natural effects. I argue that with a binary mediator, a simple modification of the authors' approach will provide bounds for natural direct and indirect effect estimates that better reflect the capacity of the available data to support empirical statements on the presence of mediated effects. I discuss complications encountered when odds ratios are used to decompose effects, as well as the implications of incorrectly assuming the absence of exposure-induced mediator-outcome confounders. I note that the former problem can be entirely resolved using collapsible measures of effect, such as risk ratios. In the Appendix, I use previous derivations for natural direct effect bounds on the risk difference scale to provide bounds on the odds ratio scale that accommodate 1) uncertainty due to the cross-world independence assumption and 2) uncertainty due to the cross-world independence assumption and the presence of exposure-induced mediator-outcome confounders.

causal inference; difference method; effect decomposition; epidemiologic methods; logistic regression; mediation analysis; natural direct effects; natural indirect effects

Abbreviations: NDE, natural direct effect; NIE, natural indirect effect; RD, risk difference.

Epidemiologists invariably rely on statistical model parameters to quantify exposure effects. Often, the precise definition of the effect is not given but is assumed to make sense on the implied logic of the model. For example, in a simple unconfounded model of an outcome regressed against an exposure, intuition might suggest that further adjustment for a mediator would block part of the exposure's effect. The exposure effect that remains is interpreted as the "direct" exposure effect, and the magnitude of the change in the exposure effect upon adjustment is interpreted as the "indirect" exposure effect. This regression procedure, often referred to as the difference method, dates back at least to the early 1980s (1), but the central ideas can be found in Sewall Wright's work on path analysis (2). While appealingly simple, it can also be misleading when applied generally.

Robins and Greenland (3) were the first to provide counterfactual-based definitions of direct and indirect effects. They defined pure (now mostly known as natural) direct and indirect effects as a contrast of proportions of potential response types. These authors and several others since have shown how use of standard procedures such as the difference method can lead to bias when estimating direct and indirect effects (4–6). In this issue of the *Journal*, Jiang and VanderWeele (7) bring the topic back full circle and give a formal account of how the difference method sometimes provides unbiased or lower-bound natural indirect effect (NIE) estimates on the risk difference or odds ratio scale. Using the difference method requires several assumptions that the authors note may not always be satisfied, even in randomized experiments. As a work-around, they recommend sensitivity analyses (7; see page 5 of their Web Appendix 1, available at http://aje.oxfordjournals.org/).

In this commentary, I discuss issues surrounding one of the assumptions required for estimating natural effects, sometimes referred to as a "cross-worlds" independence assumption. This issue is distinct from the ambiguous clinical or public health interpretation of natural effects that arise from the use of cross-world counterfactuals to define these effects (8). Here, I note distinctions between the *definition*, *identification*, and *estimation* of NIEs, and I discuss implications for using the difference method as formalized by Jiang and VanderWeele. In particular, I emphasize that even if one accepts them as relevant clinical or public health quantities, and even under ideal conditions of no selection, information, or confounding bias, natural direct and indirect effects will always be compatible with a range of possible values for a given data set, and thus will not be (point-) identifiable. I argue that a simple modification of the approach presented by Jiang and VanderWeele will provide NIE estimate bounds that better reflect the capacity of the available data to support empirical statements on the presence of mediated effects.

DEFINING NATURAL EFFECTS

Natural direct effects (NDEs) and NIEs are defined using potential outcomes (9). For an exposure (A), mediator (M), and outcome (Y), the potential outcome Y_{am} represents the outcome that would be observed if the exposure A and mediator M were set to some values a and m, respectively. Similarly, M_a represents the mediator that would be observed if the exposure were set to a. For a binary outcome, the NIE on the risk difference (RD) scale can then be defined as

$$\mathrm{RD}^{\mathrm{NIE}} = E(Y_{aM_a} - Y_{aM_{a'}}),$$

where, for example, $a = 1$ and $a' = 0$. This contrast measures the effect of switching the mediator from the value it would have taken under exposure to $a = 1$ to the value it would have taken under no exposure, all while holding everyone's exposure fixed to $a = 1$. While one can question their relevance for estimating clinical or public health intervention effects (8), natural effects are well-defined mathematical objects, permitting a rigorous derivation of the assumptions needed to identify them.

IDENTIFYING NATURAL EFFECTS

As Jiang and VanderWeele note (7), using the difference method requires no exposure-mediator interactions, in addition to the standard assumptions necessary to identify natural direct and indirect effects (assumptions 2–5 in their Web Appendix 1). One of these is the assumption that the outcome that would have been observed under a and m is conditionally independent of the mediator that would have been observed under a':

$$Y_{am} \perp M_{a'} \mid A = a. \tag{1}$$

This assumption—referred to as a cross-worlds independence assumption—has been the source of some controversy (8, 10, 11). Though it is typically described as requiring no uncontrolled mediator-outcome confounders affected by the exposure, it is much stronger than this description suggests, because it requires independence between 2 variables that can never be observed to occur together (i.e., Y_a and $M_{a'}$) (12). This assumption is not guaranteed to hold, even in an ideal setting with no confounding, measurement error, or selection bias. For this reason, Robins and Greenland stated that natural direct and indirect effects cannot be (point-) identified, even in an ideal randomized experiment (3).

ESTIMATING NATURAL EFFECTS

Nevertheless, if the following 3 conditions are met: 1) there is no exposure-mediator interaction on the relevant scale of interest, 2) there is no uncontrolled confounding of the exposure-outcome, exposure-mediator, and mediator-outcome relationships, and 3) the cross-worlds independence assumption holds, Jiang and VanderWeele (7) show that the very simple difference method will consistently estimate the NIE on the risk difference scale, or will yield a lower bound on the odds ratio scale. As they detail in their Web Appendix 2, the approach relies on a decomposition of the total effect (TE) into natural direct and indirect components:

$$\begin{aligned}\mathrm{RD}^{\mathrm{NIE}} &= E(Y_a - Y_{a'}) - E(Y_{aM_{a'}} - Y_{a'M_{a'}})\\ &= \mathrm{RD}^{\mathrm{TE}} - \mathrm{RD}^{\mathrm{NDE}}.\end{aligned}$$

Thus, if the following 2 linear regression models are correct:

$$\begin{aligned}E(Y|A) &= \alpha_0 + \alpha_1 a,\\ E(Y|A,M) &= \beta_0 + \beta_1 a + \beta_2 m,\end{aligned}$$

it follows that the NIE can be computed as $\mathrm{RD}^{\mathrm{NIE}} = \hat{\alpha}_1 - \hat{\beta}_1$, where $\hat{\alpha}_1$ and $\hat{\beta}_1$ are, for example, maximum likelihood or ordinary least squares estimates.

Yet, in light of the uncertainty regarding condition 3 above (8), it will always be the case that a given data set (from an observational study or randomized trial) is compatible with a range of values for $\mathrm{RD}^{\mathrm{NDE}}$ and thus $\mathrm{RD}^{\mathrm{NIE}}$ (13). Consequently, a researcher may use the difference method and erroneously conclude that an indirect effect is present only because the NDE was underestimated due to a violation of condition 3. Moreover, because this uncertainty derives from how natural direct and indirect effects are defined, use of more complex methods (e.g., marginal structural models (14)) would not resolve the problem. Further still, this error would not be resolved using a more focused study design (such as a randomized trial in which the exposure and mediator assignment mechanisms were specifically chosen to minimize violation of condition 3) because of the cross-world nature of this key assumption. Indeed, point-identification of natural effects can only be done in a randomized crossover trial with no period or carryover effects (3). Otherwise, it would be necessary to conduct a trial in which one could literally turn back time and observe what the mediator value would have been for the same individual under exposed and nonexposed states.

BOUNDING NATURAL EFFECTS

When using the difference method with a binary mediator, one solution to this problem is to estimate upper and lower bounds for the NDE that are compatible with a given data set, and use them as the second term of the equation for $\mathrm{RD}^{\mathrm{NIE}}$.

Several authors have derived bounds for the NDE under a range of circumstances (12, 13, 15–18). For example, Robins and Richardson (13) present sharp bounds for the NDE on the additive scale (see their Appendix C). Using these bounds, under conditions 1 and 2 defined above, one can point-identify an upper- and lower-bound $\mathrm{RD}^{\mathrm{NIE}}$ with the difference method as

$$\mathrm{RD}^{\mathrm{NIE}}_{\mathrm{lower}} = \mathrm{RD}^{\mathrm{TE}} - \mathrm{RD}^{\mathrm{NDE}}_{\mathrm{upper}}, \quad \mathrm{RD}^{\mathrm{NIE}}_{\mathrm{upper}} = \mathrm{RD}^{\mathrm{TE}} - \mathrm{RD}^{\mathrm{NDE}}_{\mathrm{lower}}, \tag{2}$$

where $\mathrm{RD}^{\mathrm{TE}}$ is as previously defined and where

$$\mathrm{RD}^{\mathrm{NDE}}_{\mathrm{upper}} = \min\{1 - E(M|A=0), E(Y|A=1, M=0)\} + \min\{E(M|A=0), E(Y|A=1, M=1)\} - E(Y|A=0)$$

and

$$\mathrm{RD}^{\mathrm{NDE}}_{\mathrm{lower}} = \max\{0, [1 - E(M|A=0)] + [E(Y|A=1, M=0) - 1]\} + \max\{0, E(M|A=0) + E(Y|A=1, M=1) - 1\} - E(Y|A=0).$$

Subtracting the upper bound of the NDE from the total would yield a conservative estimate of the NIE and would protect against the ambiguities that derive from assuming independencies about potential outcomes under incompatible exposure states. If conditions 1 and 2 hold and a positive NIE remains after using the difference method with the upper bound of the NDE, researchers can make clearer judgements about the role of condition 3 as an alternative explanation for their positive results.

NONCOLLAPSIBILITY QUANDARIES

Making sense of these issues becomes more challenging when odds ratios are used and thus noncollapsibility may be in play. Two potential complications merit consideration. First, while Jiang and VanderWeele's notation is clear on this point, less technical readers should be aware that mediator-outcome confounders must be included in Jiang and VanderWeele's first model (i.e., without the mediator) in order for their result 2 to hold. Second, the authors note that their models 3 and 4 "may not be compatible with each other" (7, p. 000); but they also assert that "logistic regression models with and without the mediator may both hold at least approximately" (7, p. 000), suggesting that the difference method may often yield NIE estimates that are close to the truth. However, previous work has shown that this approximation depends on a rather strict parametric condition: that the mediator follows a homoscedastic normal distribution (19) or (the less well known) Bridge distribution (20).

In a given research context, it is difficult to gauge how close the approximation may be, especially if the mediator is highly skewed or binary, further warranting the use of bounds for natural effects. In the Appendix, I derive bounds for the NDE on the odds ratio scale that can be used to better accommodate the uncertainty inherent in NIE estimates. Importantly, however, all issues related to noncollapsibility disappear when one uses collapsible measures of effect, such as the risk difference or risk ratio. Several methods for estimating risk ratios in a range of settings are available (21–24).

THE RECANTING WITNESS PROBLEM

A final concern merits deliberate and careful consideration in applied settings: the situation where confounders of the mediator-outcome relationship may be affected by the exposure, as depicted in Figure 1. Such mediator-outcome confounders (C_{MY} in Figure 1) have been referred to as "recanting witnesses" (25). This term, coined by Avin et al. (26), is meant to indicate that along the $X \rightarrow C_{MY} \rightarrow Y$ path in Figure 1, the variable C_{MY} operates as though it is in, for example, an unexposed state. However, along the path $X \rightarrow C_{MY} \rightarrow M \rightarrow Y$, this same variable operates as though it is in an exposed state. Thus, C_{MY} "recants its testimony," so to speak, changing the story about whether it is operating in an exposed state or an unexposed state, depending on the path from the exposure to the outcome. Though not generally recognized, the presence of such variables limits the utility of natural effects to scenarios in which the exposure affects the mediator within a short time span (27).

While this limitation markedly curtails the realm of applicability for natural effects, their use in epidemiologic research is nevertheless increasing (8). One potential solution would be to apply bounds that account for both the cross-world independence assumption and the presence of recanting witnesses. Such bounds have been derived for additive measures of effect (12). I extend these bounds to the odds ratio scale in the Appendix.

Though it is shrouded in some controversy, use of natural direct and indirect effects is becoming more common in epidemiology. Reporting bounds for these effects would do

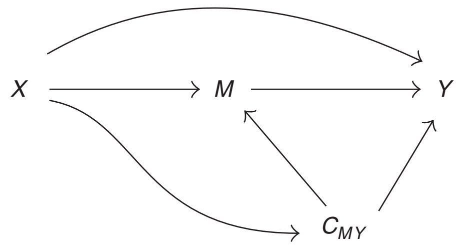

Figure 1. Mediation setting with exposure X, mediator M, and outcome Y in which confounders of the mediator-outcome relationship C_{MY} are affected by the exposure. In the context of natural effects, such confounders are sometimes referred to as "recanting witnesses" in that they operate as though they are exposed in the path $X \rightarrow C_{MY} \rightarrow Y$ but operate as though they are unexposed in the path $X \rightarrow C_{MY} \rightarrow M \rightarrow Y$. Natural effects are not identifiable in the presence of such variables. This lack of identifiability is independent of the identifiability that stems from the cross-world independence assumption.

much to quell the controversy and place them in a clearer empirical context.

ACKNOWLEDGMENTS

Author affiliation: Department of Obstetrics and Gynecology, Faculty of Medicine, McGill University, Montreal, Quebec, Canada (Ashley I. Naimi).

I thank Drs. Jay S. Kaufman, Bianca DeStavola, and Robert W. Platt for helpful comments on a previous version of this article.

Conflict of interest: none declared.

REFERENCES

1. Judd CM, Kenny DA. Process analysis: estimating mediation in treatment evaluations. *Eval Rev*. 1981;5(5):602–619.
2. Wright S. The method of path coefficients. *Ann Math Stat*. 1934;5(3):161–215.
3. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3(2): 143–155.
4. Cole SR, Hernán MA. Fallibility in estimating direct effects. *Int J Epidemiol*. 2002;31(1):163–165.
5. 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(1):4.
6. Ten Have TR, Joffe MM. A review of causal estimation of effects in mediation analyses. *Stat Methods Med Res*. 2012; 21(1):77–107.
7. Jiang Z, VanderWeele TJ. When is the difference method conservative for assessing mediation? *Am J Epidemiol*. 2015; 000(00):000–000.
8. Naimi AI, Kaufman JS, MacLehose RF. Mediation misgivings: ambiguous clinical and public health interpretations of natural direct and indirect effects. *Int J Epidemiol*. 2014;43(5): 1656–1661.
9. Rubin DB. Causal inference using potential outcomes. *J Am Stat Assoc*. 2005;100(469):322–331.
10. Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. In: Green PJ, Hjort NL, Richardson S, eds. *Highly Structured Stochastic Systems*. New York, NY: Oxford University Press; 2003: 70–80.
11. Richardson TS, Robins JM. *Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality*. (Technical report no. 128). Seattle, WA: Center for Statistics and the Social Sciences, University of Washington; 2013. http://www.csss.washington.edu/Papers/wp128.pdf. Accessed August 26, 2013.
12. Tchetgen Tchetgen EJ, Phiri K. Bounds for pure direct effect. *Epidemiology*. 2014;25(5):775–776.
13. Robins J, Richardson T. Alternative graphical causal models and the identification of direct effects. In: Keyes KM, Ornstein K, Shrout PE, eds. *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*. New York, NY: Oxford University Press; 2011:103–158.
14. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects [published correction appears in *Epidemiology*. 2009;20(4):629]. *Epidemiology*. 2009;20(1): 18–26.
15. Cai Z, Kuroki M, Pearl J, et al. Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*. 2008;64(3):695–701.
16. Kaufman S, Kaufman JS, Maclehose RF. Analytic bounds on causal risk differences in directed acyclic graphs involving three observed binary variables. *J Stat Plan Inference*. 2009;139(10): 3473–3487.
17. Sjölander A. Bounds on natural direct effects in the presence of confounded intermediate variables. *Stat Med*. 2009;28(4): 558–571.
18. Chiba Y, Taguri M. Alternative monotonicity assumptions for improving bounds on natural direct effects. *Int J Biostat*. 2013; 9(2):235–249.
19. VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. *Am J Epidemiol*. 2010; 172(12):1339–1348.
20. Tchetgen Tchetgen EJ. A note on formulae for causal mediation analysis in an odds ratio context. *Epidemiol Method*. 2014;2(1): 21–31.
21. Zou G. A modified Poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004;159(7): 702–706.
22. Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and differences. *Am J Epidemiol*. 2005; 162(3):199–200.
23. Chu H, Cole SR. Estimation of risk ratios in cohort studies with common outcomes: a Bayesian approach. *Epidemiology*. 2010; 21(6):855–862.
24. Tchetgen Tchetgen E. Estimation of risk ratios in cohort studies with a common outcome: a simple and efficient two-stage approach. *Int J Biostat*. 2013;9(2):251–264.
25. Tchetgen Tchetgen EJ, VanderWeele TJ. Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*. 2014;25(2):282–291.
26. Avin C, Shpitser I, Pearl J. Identifiability of path-specific effects. In: Bacchus F, Jaakkola T, eds. *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*. (Technical report R-321). Arlington, VA: AUAI Press; 2005:357–363.
27. VanderWeele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*. 2014;25(2): 300–306.
28. Richardson TS, Robins JM. Analysis of the binary instrumental variable model. In: Dechter R, Geffner H, Halpern J, eds. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*. London, United Kingdom: College Publications; 2010:415–444.
29. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003;14(6):680–686.
30. Localio AR, Margolis DJ, Berlin JA. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *J Clin Epidemiol*. 2007;60(9): 874–882.

(Appendix follows)

APPENDIX

Based on Robins and Richardson's Appendix C (13), one can derive the x-adjusted bound for the natural direct effect (NDE) on the odds ratio (OR) scale when the mediator is binary:

$$\mathrm{OR}^{\mathrm{NDE}}_{a,a'|x} = \frac{P(Y_{aM_{a'}} = 1|x)/[1 - P(Y_{aM_{a'}} = 1|x)]}{P(Y_{a'} = 1|x)/[1 - P(Y_{a'} = 1|x)]}. \tag{A1}$$

For $a = 1$ and $a' = 0$, the numerator of equation A1 can be rewritten as $B/(1 - B)$, where

$$B = P(Y_{a,m=0} = 1|M_{a'} = 0, x)P(M = 0|A = 0, x) + P(Y_{a,m=1} = 1|M_{a'} = 1, x)P(M = 1|A = 0, x).$$

One can link the counterfactual quantities $P(Y_{a,m=0} = 1|M_{a'} = 0, x)$ and $P(Y_{a,m=1} = 1|M_{a'} = 1, x)$ to the observed data by noting that the probability of the observed outcome, conditional on the observed exposure and mediator, can be written as a weighted combination of the potential outcomes:

$$\begin{aligned} P(Y = 1|A = 1, M = 0, x) &= P(Y_{a,m=0} = 1|M_{a'} = 0, x)P(M = 0|A = 0, x) + P(Y_{a,m=0} = 1|M_{a'} = 1, x)P(M = 1|A = 0, x), \\ P(Y = 1|A = 1, M = 1, x) &= P(Y_{a,m=1} = 1|M_{a'} = 0, x)P(M = 0|A = 0, x) + P(Y_{a,m=1} = 1|M_{a'} = 1, x)P(M = 1|A = 0, x). \end{aligned}$$

These combinations of potential outcomes can be used to obtain upper and lower limits for natural effects based on the observed data. Following section 2.2 in the article by Richardson and Robins (28) and the derivations shown in Appendix C of the article by Robins and Richardson (13), the x-adjusted upper and lower bounds for the numerator of equation A1 is $B'/(1 - B')$, where

$$\begin{aligned} {B'}_{\mathrm{upper}} = {} & \min\{P(M = 0|A = 0, x), P(Y = 1|A = 1, M = 0, x)\} \\ & + \min\{P(M = 1|A = 0, x), P(Y = 1|A = 1, M = 1, x)\} \end{aligned} \tag{A2}$$

and

$$\begin{aligned} {B'}_{\mathrm{lower}} = {} & \max\{0, P(M = 0|A = 0, x) + P(Y = 1|A = 1, M = 0, x) - 1\} \\ & + \max\{0, P(M = 1|A = 0, x) + P(Y = 1|A = 1, M = 1, x) - 1\}. \end{aligned} \tag{A3}$$

The x-adjusted upper bound for the NDE on the odds ratio scale can thus be computed from the observed data as

$$\mathrm{OR}^{\mathrm{NDE_{upper}}}_{a,a'|x} = \frac{{B'}_{\mathrm{upper}}/(1 - {B'}_{\mathrm{upper}})}{P(Y = 1|A = 0, x)/[1 - P(Y = 1|A = 0, x)]},$$

the log of which can be used instead of θ_1, as documented in the article by Jiang and VanderWeele (7), to provide a lower bound for the natural indirect effect (NIE) on the log odds scale. It follows from Jiang and VanderWeele's result 2 (7) that the lower $\mathrm{OR}^{\mathrm{NDE}}$ bound estimated in this way reflects the uncertainty due to violations of our condition 3 (see main text), as well as noncollapsibility of the odds ratio. To obtain $\mathrm{OR}^{\mathrm{NIE_{upper}}}_{a,a'|x}$, one simply replaces B'_{upper} with B'_{lower} in the above equation and uses the resulting $\mathrm{OR}^{\mathrm{NDE_{lower}}}_{a,a'|x}$ estimate instead of θ_1, as documented by Jiang and VanderWeele (7). When marginal quantities are preferred, one can marginalize the probabilities in the min( ) and max( ) arguments of equations A2 and A3 over x via inverse probability weighting (29) or marginal standardization (30).

In their paper, Tchetgen Tchetgen and Phiri (12) derive similar bounds on the additive scale that additionally account for uncertainty due to the presence of exposure-induced mediator-outcome confounders, as displayed in Figure 1 (see main text). These bounds can be extended to the odds ratio scale by noting that B' can be rewritten as

$$\begin{aligned} B'^{*}_{\mathrm{upper}} = {} & \min\{P(M_{a=0} = 0|x), P(Y_{a=1,m=0} = 1|x)\} \\ & + \min\{P(M_{a=0} = 1|x), P(Y_{a=1,m=1} = 1|x)\} \end{aligned} \tag{A4}$$

and

$$\begin{aligned} B'^{*}_{\mathrm{lower}} = {} & \max\{0, P(M_{a=0} = 0|x) + [P(Y_{a=1,m=0} = 1|x) - 1]\} \\ & + \max\{0, P(M_{a=0} = 1|x) + [P(Y_{a=1,m=1} = 1|x) - 1]\}, \end{aligned} \tag{A5}$$

where

$$P(Y_{a,m} = 1|x) = \sum_{c_{MY}} P(Y|a, m, c_{MY}, x) f(c_{MY}|a, x),$$

$$P(Y_a|x) = P(Y|a, x), \text{ and}$$

$$P(M_a|x) = P(M|a, x).$$

Similarly, one could marginalize the probabilities in the min( ) and max( ) arguments of equations A4 and A5 over *x* to obtain marginal *x*-adjusted estimates.