Harvard University
Harvard University Biostatistics Working Paper Series

Year 2008 *Paper* 75

Estimation of Controlled Direct Effects

Sylvie Goetgeluk* Stijn Vansteelandt†
Els Goetghebeur‡

*Ghent University
†Ghent University
‡Ghent University and Harvard School of Public Health

Estimation of controlled direct effects

Sylvie Goetgeluk, Stijn Vansteelandt

Department of Applied Mathematics and Computer Sciences
Ghent University, 9000 Ghent, Belgium

and Els Goetghebeur

Department of Applied Mathematics and Computer Sciences
Ghent University, 9000 Ghent, Belgium
and Department of Biostatistics
Harvard School of Public Health, Boston, MA 02115, U.S.A.

SUMMARY. When regression models adjust for mediators on the causal path from exposure to outcome, the regression coefficient of exposure is commonly viewed as a measure of the direct exposure effect. This interpretation can be misleading, even with a randomly assigned exposure. This is because adjustment for post-exposure measurements introduces bias whenever their association with outcome is confounded by more than just the exposure. By the same token, adjustment for such confounders stays problematic when these are themselves affected by the exposure. Robins (1999) accommodated this by introducing structural nested direct-effect models with direct effect parameters that can be estimated using inverse probability weighting by a conditional distribution of the mediator. The resulting estimators are consistent, but inefficient and can be extremely unstable when the intermediate variable is absolutely continuous. In this paper, we develop direct effect estimators which are not only more efficient, but also consistent under a less demanding model for a conditional expectation of the outcome. We find the one estimator which avoids inverse probability weighting altoghether to perform best. This estimator is intuitive, computationally straightforward and, as

demonstrated by simulation, competes extremely well with ordinary least squares estimators in settings where standard regression is valid.

KEY WORDS: Direct effect; Indirect effect; Instability; Inverse probability weighting; Pathway; Structural nested model; Surrogate marker.

1 Introduction

Once researchers have established that an exposure affects an outcome, the attention typically turns to understanding the biologic/mechanistic pathways that contribute to this effect. Empirically, this is most naturally approached by disentangling the part of the exposure effect that is explained by intermediate effects of exposure on outcome through given mediators, and by the remaining direct effect. The following examples illustrate this.

Example 1 (Surrogate biomarkers). The pressure of accelerated evaluation of new AIDS therapies has led to the use of CD4 blood count and viral load as endpoints that replace time to clinical events and overall survival. This raises the question whether an effect of treatment on the biomarker provides evidence for a clinical effect (Molenberghs et al., 2004). While a good biomarker need not lie on the causal path from treatment to clinical event, a biomarker which does, is often more trustworthy. For that reason, a number of approaches have been developed to infer whether the effect of treatment on the outcome is entirely mediated by its effect on the biomarker (Frangakis and Rubin, 2002; Taylor et al., 2005). These approaches are particularly of interest in settings where data from a single study are available and prediction-based approaches (Molenberghs et al., 2004) are thus not applicable.

Example 2 (Gender discrimination). Bickel, Hammel and O'Connell (1975) examine data from the University of Berkeley on sex bias in graduate admissions. Noting that study choices are on average different between male and female applicants, the investigation of gender discrimination may be approached by evaluating

whether there is a direct gender effect on admission rates, which is not mediated by study choice.

Example 3 (Zygosity in reproductive epidemiology). Verstraelen et al. (2005) estimate that the odds of preterm birth is 38% (95%CI: 15%-66%) higher in twins conceived after in vitro fertilization versus naturally conceived twins, after controlling for maternal age and parity. Since 95% of all twins conceived after subfertility treatment are dizygotic versus 54% in naturally conceived twins, and since perinatal outcomes tend to be better for dizygous than for monozygous twins, part of this odds ratio is explained by the effect of subfertility treatment on zygosity. Verstraelen et al. (2005) thus infer the effect which subfertility treatment has on preterm birth risk, other than through modifying the dizogytic/monozygotic twinning rate.

Standard regression approaches for direct effects estimate the residual exposure effect that remains on the outcome after adjusting for the given mediator. These approaches tend to be biased by the same token that adjustment for post-randomization measurements may introduce bias in the analysis of randomized experiments (Rosenbaum, 1984). This is so whenever there exist common causes of the mediator and outcome, other than the considered exposure (Cole and Hernan, 2002; Pearl, 2000; Robins, 1986). In some cases, the absence of such common causes is a plausible assumption based on biological grounds. For instance, Verstraelen et al. (2005) used standard adjustment for zygosity to estimate the direct effect of subfertility treatment on preterm birth because it is reasonable to assume that zygosity is not affected by risk factors of preterm birth other than subfertility treatment (and parental fertility) itself. When the presence of common causes of mediator and outcome cannot be precluded, as in most cases of interest, untestable assumptions must be made. In this article, as in Robins (1999) and Petersen, Sinisi and van der Laan (2006), we proceed under the assumption of no unmeasured confounders for the association between mediator and outcome. Intuitively, this assumption is sufficient because the size of the direct effect depends on how

strongly the mediator affects the outcome and inferring the latter requires knowing all common causes of both mediator and outcome. Ten Have et al. (2007) avoid this assumption but assume instead that exposure and mediator do not interact in their effect on the outcome, and that the effect of exposure on the mediator varies by baseline covariates.

Even when all confounders for the association between mediator and outcome have been measured, standard regression techniques are not applicable for estimating the direct of exposure on outcome. They are prone to bias whenever some of these confounders are themselves affected by the treatment. This happens for the same reason that stratifying by the mediator may induce selection bias. van der Laan and Petersen (2005) and Robins (1999) accommodate this via inverse probability of treatment weighting estimators for the parameters indexing marginal structural (MS) models and structural nested direct-effect (SNDE) models, respectively. Both classes of estimators involve inverse probability weighting by a conditional distribution of the mediator. As demonstrated by extensive simulation studies in Section 3, these estimators can be extremely inefficient and unstable when there are strong predictors of the mediators, or when the mediator is absolutely continuous; in the latter case, they are also likely biased by the fact that models for a conditional density are difficult to postulate.

In this paper, we mitigate these problems by developing estimators for direct effect parameters indexing SNDE models, which are asymptotically unbiased as soon as a less demanding model for the conditional expectation of the outcome holds. One of the estimators avoids inverse probability weighting altoghether by using sequential G-estimation. This estimator is intuitive and computationally straightforward. As demonstrated by extensive simulation studies, it competes extremely well with standard ordinary least squares estimators in settings where standard regression is valid, but in contrast, remains valid when some of the considered confounders are themselves affected by the exposure. Our methods also provide insights on how to stabilize estimators based on inverse probability weight-

ing in the presence of extreme weights.

2 Structural nested direct-effect models

2.1 Controlled direct effects

Let Y_{xk} be the potential outcome which a given subject would have experienced under exposure $X = x$ and a fixed value k for the intermediate variable K. Then, as in Robins (1999), we formally define the direct effect on outcome Y of setting exposure $X = x$ (versus $X = 0$), when holding K fixed, as the contrast $Y_{xk} - Y_{0k}$ between the two potential outcomes Y_{xk} and Y_{0k} for the same subject. This is termed a controlled direct effect. In this article, we develop inference for direct effect SNMs (Robins, 1999) which parameterize average controlled direct effects conditionally on pre-exposure covariates S and among subjects with $X = x$:

$$E(Y_{xk} - Y_{0k}|X = x, S) = m(x, k, S; \psi^*) \tag{1}$$

where $m(x, k, S; \psi)$ is a known function, smooth in ψ, satisfying $m(0, k, S; \psi) = 0$ and where ψ^* is an unknown finite-dimensional parameter. For example, assuming that the direct effect of exposure x (versus 0) is linear in x and the same regardless of k and S, we may choose $m(x, k, S; \psi) = \psi x$.

A number of alternative definitions for direct effects have been proposed in the literature. Principal stratification direct effects (Rubin, 2004) measure the average causal effect of setting exposure $X = x$ (versus $X = 0$) among subjects for whom the mediator was not affected by X; that is, $E(Y_x - Y_0|K_x = K_0)$, where K_x is the counterfactual value of the mediator K corresponding to setting $X = x$. These are more suitable when the potential outcomes Y_{xk} are ill defined (e.g. due to the mediator not being manipulatable), but have the drawback that inference typically relies on a small subsample of the population (Robins, Rotnitzky and Vansteelandt, 2007). Standardized direct effects (Didelez, Dawid and Geneletti, 2006) are obtained by averaging controlled direct effects over a chosen

distribution function for the mediator, which does not depend on the exposure; that is, $\int E(Y_{xk} - Y_{0k})f^*(K = k)dk$ for a chosen density function $f^*(K)$. These are more suitable in settings where it is not realistic to fix the mediator at the same value for all subjects. Natural or pure direct effects (Pearl, 2001, Robins and Greenland, 1992, Petersen, Sinisi and van der Laan, 2006; Didelez, Dawid and Geneletti, 2006) form a special case obtained by setting $f^*(K) = f\{K(0)\}$; that is, $E(Y_{xK(0)} - Y_{0K(0)})$. Because the exposure-free level forms a natural reference level for each subject, there exist studies in which natural direct effects may be regarded as well-defined even when the controlled direct effect is not for some k (Petersen, Sinisi and van der Laan, 2006). In most cases, however, it is difficult to conceive of interventions that capture the notion of natural direct effects (Didelez, Dawid and Geneletti, 2006), so that controlled and standardized direct effects may be the more useful for practical use. See Didelez, Dawid and Geneletti (2006) and Robins, Rotnitzky and Vansteelandt (2007) for further discussions.

2.2 Inverse Probability of Intermediate Weighted estimators

Inference for ψ^* in model (1) is developed by Robins (1999) and briefly reviewed here from a different perspective. Suppose first that the potential outcome $Y_k \equiv Y_{Xk}$ following setting $K = k$ is observed for every subject and every value k on the support of K. Further, assume that, as expressed by the causal diagram of Figure 1, S contains all confounders for the association between X and Y_k so that

$$Y_{xk} \amalg X|S \qquad \forall(x,k) \tag{2}$$

Then, for each k, model (1) is a structural nested mean model (Robins, 1994) which can be fitted by G-estimation (Robins, Mark and Newey, 1992). That is, ψ^* can be estimated as the value ψ such that, after subtracting the direct effect $m(X,k,S;\psi)$ from Y_k, no dependence on X remains, conditionally on S. Specifically, for given

k, all unbiased estimating functions for ψ^* in the model given by restrictions (1) and (2) for the given k, with Y_k observed, are of the form

$$\Delta\{d_k(X,S)|S\}\,\{Y_k - m(X,k,S;\psi) - q_k(S)\} \tag{3}$$

where $d_k(X,S)$ is an arbitrary vector function of the dimension of ψ, $q_k(S)$ is an arbitrary scalar function and where for any 2 random variables A and B, we define $\Delta\{A|B\} \equiv A - E(A|B)$. For example, we may choose $d_k(X,S) = X$, which, as we will show later, corresponds to the optimal choice for $d_k(X,S)$ when model (1) is linear in x and independent of k and S (i.e. $m(X,k,S;\psi) = \psi X$). That (3) is unbiased at $\psi = \psi^*$ follows because $E(Y_k - m(X,k,S;\psi)|X,S) = E(Y_{0k}|X,S) = E(Y_{0k}|S)$ under the model given by restrictions (1) and (2). It then follows that all unbiased estimating functions for ψ^* in model (1)-(2) (for all k) with Y_k observed are of the form

$$\int \Delta\{d_k(X,S)|S\}\,\{Y_k - m(X,k,S;\psi) - q_k(S)\}\,dk \tag{4}$$

Estimating equations based on (4) yield no feasible estimators for ψ^* because Y_k is unknown for each k except the observed realization of K. Multiplying each term in (4) with $I(K=k)$ yields an observed data estimating function, which in general no longer has mean zero because subjects with $K = k$ may form a selective subgroup. To correct for this, we make the additional assumption, which is expressed by the diagram of Figure 1, that (X,L,S) contains all confounders for the association between K and Y so that

$$Y_{xk} \amalg K|X = x, L, S \qquad \forall x,k \tag{5}$$

This assumption allows for inversely weighting each term in the estimating function (4) by the conditional distribution $f(K|L,S,X)$ of K given X, L and S, as in

$$\begin{aligned}&\int \frac{I(K=k)}{f(K=k|L,S,X)}\Delta\{d_k(X,S)|S\}\,\{Y_k - m(X,k,S;\psi) - q_k(S)\}\,dk \\ &= \frac{\Delta\{d_K(X,S)|S\}}{f(K|L,S,X)}\,\{Y - m(X,K,S;\psi) - q_K(S)\}\end{aligned} \tag{6}$$

Estimating function (6) has mean zero at $\psi = \psi^*$ under model (1) because the conditional mean of $\frac{I(K=k)}{f(K=k|L,S,X)}$, given (L, S, X, Y_k) equals 1 under assumption (5). The solution to the estimating equation

$$0 = \sum_{i=1}^{n} \frac{\Delta\{d_{K_i}(X_i, S_i)|S_i\}}{f(K_i|L_i, S_i, X_i)} \{Y_i - m(X_i, K_i, S_i; \psi) - q_{K_i}(S_i)\} \tag{7}$$

is therefore a consistent and asymptotically normal (CAN) estimator of ψ^*, provided that $f(K = k|L, S, X) > 0$ with probability 1 for all k in the support of K (and that similar, weak regularity conditions hold as in Robins, Mark and Newey (1992, Theorem 1A)).

Solving (7) requires that we specify parametric models

$$f(K|L, S, X) = f(K|L, S, X; \alpha^*) \tag{8}$$

$$E(d_K(X, S)|S) = E(d_K(X, S)|S; \beta^*) \tag{9}$$

where $f(K|L, S, X; \alpha)$ is a conditional density function, smooth in α, $E(d_K(X, S)|S; \beta)$ is a function of S, smooth in β, and (α^*, β^*) is an unknown finite-dimensional parameter. For example, we may assume that the conditional distribution of K given (L, S, X) is normal with mean $\alpha_0 + \alpha_1 L + \alpha_2 S + \alpha_3 X$ and constant standard deviation σ_K and, with $d_K(X, S) = X$, that $E(X|S; \beta) = \beta_0 + \beta_1 S$. Consistent estimators $\hat{\alpha}$ for α^* and $\hat{\beta}$ for β^* can be obtained via standard regression.

Throughout we let $\mathcal{A}$ be the model for the observed data defined by the model restrictions (1), (8) and (9), and the no unmeasured confounders assumptions (2) and (5). It follows from the previous discussion that a CAN estimator $\hat{\psi}_{IPIW}$ for the direct effect parameter ψ^* under model $\mathcal{A}$ can be obtained by solving

$$0 = \sum_{i=1}^{n} U_{i,IPIW}(d, q; \psi, \hat{\alpha}, \hat{\beta}) \tag{10}$$

where

$$U_{i,IPIW}(d, q; \psi, \alpha, \beta) = \frac{\Delta\{d_{K_i}(X_i, S_i)|S_i; \beta\}}{f(K_i|L_i, S_i, X_i; \alpha)} \{Y_i - m(X_i, K_i, S_i; \psi) - q_{K_i}(S_i)\} \tag{11}$$

For given k, optimal choices for $d_k(X, S)$ and $q_k(S)$ which lead to a semi-parametric efficient estimator of ψ^* in the model given by restrictions (1) and (2) (for the given

k) and with Y_k observed, have been derived by Robins (1994). When the potential outcome variance $\text{Var}(Y_k|X,S)$ is constant in (X,S), these choices equal

$$\begin{aligned} d_k(X,S) &= \frac{\partial m(X,k,S;\psi)}{\partial \psi} \\ q_k(S) &= E(Y_k - m(X,k,S;\psi)|S) \end{aligned}$$

where the latter can be calculated using the law of iterated expectations

$$E(Y_k - m(X,k,S;\psi)|S) = E\left[E(Y|K=k,X,L,S) - m(X,k,S;\psi)|S)\right]$$

The same choices may not lead to a semi-parametric efficient estimator of ψ^* under model $\mathcal{A}$, in which Y_k is not observed for each subject. However, we recommend using the above choices for practical use because we conjecture that they will generally yield reasonable efficiency, while calculating the semi-parametric efficient estimator under model $\mathcal{A}$ is much more tedious as it requires solving integral equations.

When the intermediate variable is absolutely continuous, the above method requires inverse weighting by a density. The inverse weighting estimator $\hat{\psi}_{IPIW}$ is then likely to have serious finite sample bias because statistical models for a density are difficult to postulate and small misspecifications in the tails of the density can have a large effect on the direct-effects estimates through their influence on the inverse weights. Furthermore, the large variability of the inverse weights may then seriously distort the precision of the estimate. An ad hoc approach to stabilize the inverse weights is to multiply the estimating function (6) by $f(K|S)$, because observations with extreme values for $f(K|L,S,X)$ are likely also extreme in terms of $f(K|S)$ and may therefore have a more stable ratio of both. The resulting estimating function remains unbiased because $d_K(X,S)$ is an arbitrary function of K, X and S, and the conditional expectation in $E(d_K(X,S)|S)$ is only w.r.t. X. Therefore, from now on, we will replace the weights $1/f(K|L,S,X)$ by the stabilized weights $f(K|S)/f(K|L,S,X)$. However, as we will show in several simulation studies in Section 3, this ad hoc stabilization will often not suffice

to obtain well-behaved estimators in moderate sample sizes. Alternatively, one could truncate the weights. However, one may argue that truncated weights are deliberately misspecified weights and, as such, may impact the consistency of the direct-effects estimator. In the next sections, we will therefore develop estimators which allow misspecification (and thus truncation) of the weights.

2.3 Doubly-robust estimators

To obtain estimators with better performance in the presence of unstable weights, note (using similar arguments as in van der Laan and Robins, 2003) that, up to asymptotic equivalence, all CAN estimators for ψ^* under model $\mathcal{A}$ can be obtained by solving estimating equations of the form

$$0 = \sum_{i=1}^{n} U_{i,IPIW}(d, q; \psi, \alpha, \beta) - \Delta\left\{\phi(K_i, L_i, X_i, S_i)|L_i, X_i, S_i\right\} \tag{12}$$

where $\phi(K_i, L_i, X_i, S_i)$ is an arbitrary vector function of the dimension of ψ. Part 1 of Theorem 1 below shows that for given $d_K(X, S)$ and $q_K(S)$, the optimal choice of $\phi(K_i, L_i, X_i, S_i)$ that leads to estimators of ψ^* with minimum asymptotic variance, equals

$$\phi_{opt}(K_i, L_i, X_i, S_i) \equiv E(U_{i,IPIW}(d, q; \psi, \alpha, \beta)|K_i, L_i, X_i, S_i) \tag{13}$$

In the proof of Theorem 1, we further show that this yields the following estimating function for ψ

$$\begin{aligned} &\Delta\{d_K(X, S)|S; \beta\}W(\alpha)\Delta\left\{Y|K, L, X, S\right\} \qquad (14)\\ &+ \int \Delta\{d_K(X, S)|S; \beta\}\left\{E(Y|K, L, X, S) - m(X, K, S; \psi) - q_K(S)\right\} f(K|S)dK \end{aligned}$$

where $W(\alpha) = f(K|S)/f(K|L, S, X; \alpha)$ and where the conditional density $f(K|S)$ may be replaced by an estimate. Using this estimating function requires that we specify a parametric model

$$E(Y|K, L, X, S) \;=\; E(Y|K, L, X, S; \gamma^*) \tag{15}$$

where $E(Y|K,L,X,S;\gamma)$ is a function of (K,L,X,S), smooth in γ, and γ^* is an unknown finite-dimensional parameter. A consistent estimator $\hat{\gamma}$ for γ^* can be obtained using standard regression techniques. For example, for the linear model

$$E(Y|K,L,X,S;\gamma) = \gamma_0 + \gamma_1 K + \gamma_2 L + \gamma_3 X + \gamma_3 S \tag{16}$$

and with $m(X,S,K;\psi) = \psi X$, $d_K(X,S) = d(X,S)$ and $q_K(S) = 0$, the estimating function (14) has the relatively simple form

$$\Delta\{d(X,S)|S;\beta\}\left[W(\alpha)\Delta\{Y|K,L,X,S;\gamma\} + E(Y|K = E(K|S),L,X,S;\gamma) - \psi X\right]$$

Part 2 of Theorem 1 shows that the solution $\hat{\psi}_{DR}$ to an estimating equation based on (14) has the interesting feature of being a consistent estimator of ψ^* when either model (8) holds or model (15), but not necessarily both. We therefore call $\hat{\psi}_{DR}$ a doubly-robust estimator of ψ^*.

Theorem 1. 1. The solution $\hat{\psi}_{DR}$ to equation

$$0 = \sum_{i=1}^{n} U_{i,DR}(d,q;\psi,\hat{\alpha},\hat{\beta},\hat{\gamma}) \tag{17}$$

where

$$\begin{aligned} U_{i,DR}(d,q;\psi,\alpha,\beta,\gamma) &= \Delta\{d_{K_i}(X_i,S_i)|S_i;\beta\}W_i(\alpha)\Delta\left\{Y_i|K_i,L_i,X_i,S_i;\gamma\right\} \\ &+ \int \Delta\{d_{K_i}(X_i,S_i)|S_i;\beta\}\left\{E(Y_i|K_i,L_i,X_i,S_i;\gamma) - m(X_i,K_i,S_i;\psi)\right. \\ &\left. -q_{K_i}(S_i)\right\} f(K_i|S_i)dK_i \end{aligned} \tag{18}$$

is locally efficient among all CAN estimators under model $\mathcal{A}$ that use the given choice of $d_K(X,S)$ and $q_K(S)$, in the sense that it is efficient in this class when model (15) is correctly specified.

2. $\hat{\psi}_{DR}$ is a consistent estimator of ψ^* under model $\mathcal{A} \cup \mathcal{B}$, where $\mathcal{B}$ is the model for the observed data defined by the model restrictions (1), (15) and (9), and the no unmeasured confounders assumptions (2) and (5).

2.4 Unweighted estimators

The attractiveness of the doubly-robust estimator $\hat{\psi}_{DR}$ lies not only in it (typically) being more efficient than the simpler inverse weighting estimator $\hat{\psi}_{IPIW}$. Its main attraction lies in the fact that it avoids reliance on a difficult-to-postulate model for the density of the mediator when a simpler model for a conditional expectation of the outcome holds. In this section, we completely avoid reliance on the model for the density of the mediator by setting $f(K|L,S,X)$ equal to $f(K|S)$ in the estimating function (17) of the doubly-robust estimator. The implication of this is to set all weights equal to 1, which leads to an unweighted estimating equation. The corresponding estimators $\hat{\psi}_{UW}$ solve

$$0 = \sum_{i=1}^{n} U_{i,UW}(d, q; \psi, \hat{\beta}, \hat{\gamma}) \tag{19}$$

where $U_{i,UW}(d,q;\psi,\beta,\gamma)$ is defined as $U_{i,DR}(d,q;\psi,\alpha,\beta,\gamma)$, but with $W_i(\alpha)$ replaced by 1. For example, when choosing a linear conditional mean model for Y as in (16), $m(X,K,S;\psi) = \psi X$, $d_K(X,S) = X$, $q_K(S) = \gamma_1 E(K|S)$, we obtain the simple form

$$0 = \sum_{i=1}^{n} \Delta\{X_i|S_i;\beta\}\,(Y_i - \gamma_1 K_i - \psi X_i) \tag{20}$$

Note that this estimating equation is very intuitive as it expresses that, after subtracting the effect $\gamma_1 K_i$ of the mediator and the direct effect ψX_i of the exposure from the outcome, no association with X_i should remain after adjustment for the confounder S_i. As such, the solution $\hat{\beta}$ for β to equation (20) with γ_1 replaced by a consistent estimate $\hat{\gamma}_1$, can be viewed as a sequential G-estimator (i.e., it is obtained by G-estimation applied to the residual outcome $Y_i - \hat{\gamma}_1 K_i$ that remains after removing the effect of the mediator from the outcome). The solution to (19) generalizes such sequential G-estimators by allowing for nonlinear models (15) for the outcome.

By the fact that the solutions to (17) are consistent estimators for ψ^* under model $\mathcal{A} \cup \mathcal{B}$, solving (19) gives a consistent estimator for ψ^* under model $\mathcal{B}$. In

the simulation study of Section 3, we will show that the resulting estimator has the desirable property of being very stable and efficient as a result of avoiding the inverse weighting, but is no longer doubly-robust. In the following sections, we briefly introduce alternative estimators which are designed to perform well in the presence of extreme weights and protect the double robustness property.

2.5 Stabilized doubly-robust estimators

Using arguments similar to Robins et al. (2007), we will stabilize the doubly-robust direct effects estimator by substituting ψ in expression (13) by an estimator $\tilde{\psi}$ which is consistent under model $\mathcal{B}$. We denote the resulting estimator with $\hat{\psi}_{SDR}$. When considering closed-form estimators for ψ^* obtained from expression (17), it can be seen that the impact of this is that the weights $W_i(\alpha)$ appear both in the numerator and denominator. For example, with $m(X, K, S; \psi) = \psi X$, $d_K(X, S) = X$, $q_K(S) = 0$ and a linear conditional mean model for Y as in (16), we then obtain

$$\begin{aligned}\hat{\psi}_{SDR} &= \frac{\sum_{i=1}^{n} \Delta\{X_i|S_i;\beta\}\left[W_i(\alpha)\left\{\Delta\{Y_i|K_i,L_i,X_i,S_i;\gamma\}+\tilde{\psi}X_i\right\}\right]}{\sum_{i=1}^{n} W_i(\alpha)X_i\Delta\{X_i|S_i;\beta\}} \\ &\quad + \frac{\sum_{i=1}^{n}\left\{\tilde{\psi}X_i - E(Y_i|K_i=E(K_i|S_i),L_i,S_i,X_i;\gamma)\right\}}{\sum_{i=1}^{n} W_i(\alpha)X_i\Delta\{X_i|S_i;\beta\}} \qquad (21)\end{aligned}$$

The resulting estimator is generally more stable than the doubly-robust estimator

$$\frac{\sum_{i=1}^{n} \Delta\{X_i|S_i;\beta\}\left[W_i(\alpha)\Delta\{Y_i|K_i,L_i,X_i,S_i;\gamma\}+E(Y_i|K_i=E(K_i|S_i),L_i,S_i,X_i;\gamma)\right]}{\sum_{i=1}^{n} X_i\Delta\{X_i|S_i;\beta\}}$$

of Section 2.3 by the fact that subjects with extreme weights $W_i(\alpha)$ in the numerator of (21) will also make the denominator of (21) extreme. The stabilized doubly-robust estimator $\hat{\psi}_{SDR}$ is a consistent estimator of ψ^* under model $\mathcal{A}$, even when model (15) is misspecified and thus even when $\tilde{\psi}$ is an inconsistent estimator, because estimating equation (12) is unbiased under model $\mathcal{A}$ regardless of $\phi(K_i, L_i, X_i, S_i)$ (and thus in particular when the unknown parameters indexing $\phi(K_i, L_i, X_i, S_i)$ are replaced by inconsistent estimators). Likewise, $\hat{\psi}_{SDR}$ is a

consistent estimator of ψ^* under model $\mathcal{B}$, even when model (8) is misspecified, because estimating equation (17) is unbiased under model $\mathcal{B}$ and because $\tilde{\psi}$ is a consistent estimator of ψ^* under model $\mathcal{B}$. It follows that $\hat{\psi}_{SDR}$ is a doubly-robust estimator of ψ^*.

Alternatively, we may improve the finite-sample behavior of $\hat{\psi}_{DR}$ by adapting ideas in Tan (2006) for inverse weighting estimators to inverse weighting estimating functions. Specifically, we modify the doubly-robust estimating equation for ψ^* as

$$0 = \sum_{i=1}^{n} U_{i,IPIW}(d, q; \psi, \alpha, \beta) - \kappa \Delta \left\{ \phi_{opt}(K_i, L_i, X_i, S_i) | L_i, X_i, S_i \right\} \tag{22}$$

and determine an 'optimal' choice of κ that leads to improved efficiency. Note that the choice $\kappa = 1$ yields the estimator $\hat{\psi}_{DR}$, which may be an inefficient doubly-robust estimator whenever model (15) is incorrectly specified.

Let for notational convenience $\xi \equiv \Delta \left\{ \phi_{opt}(K_i, L_i, X_i, S_i) | L_i, X_i, S_i \right\}$ and $\eta \equiv U_{i,IPIW}(d, q; \psi, \alpha, \beta)$. For arbitrary random variable A, define $\hat{E}(A)$ as the sample average $\sum_{i=1}^{n} A_i/n$. Then choosing κ equal to $\kappa_{opt} = \hat{E}^{-1}(\xi\xi')\hat{E}(\xi\eta')$ yields an estimator $\hat{\psi}(\kappa_{opt})$ with minimal variance among all estimators $\hat{\psi}(\kappa)$ that solve (22) for given κ. This can be seen from the following 2 arguments. First, the variance $E(\eta^2 - 2\kappa\eta\xi + \kappa^2\xi^2)$ of the estimating function $\eta - \kappa\xi$ is minimized at κ_{opt}. Second, the estimator obtained by solving the corresponding estimating equation itself has minimal variance among all estimators $\hat{\psi}(\kappa)$ because

$$Var(\hat{\psi}(\kappa)) \approx \frac{1}{n} E\left(\frac{\partial \eta}{\partial \psi}\right)^{-1} Var\left(\eta - \kappa\xi\right) E\left(\frac{\partial \eta}{\partial \psi}\right)^{-1'}$$

and thus the variance of these estimators is proportional to the variance of their estimating function. The estimator $\hat{\psi}(\kappa_{opt})$ is however not doubly-robust because κ_{opt} may not converge to 1 under a correctly specified model for (15).

Choosing κ to equal

$$\kappa_{\mathrm{dr}} \equiv \hat{E}^{-1}(\xi\chi')\hat{E}(\xi\eta')$$

with

$$\chi \equiv \Delta\{d_{K_i}(X_i, S_i) | S_i; \beta\} W_i(\alpha) \left[E\left\{ Y_i - m(X_i, K_i, S_i; \psi) | X_i, K_i, L_i, S_i \right\} - q_{K_i}(S_i) \right]$$

accommodates this. Indeed, $\kappa_{\rm dr}$ converges to 1 when the model for (15) is correctly specified, because $\chi = E(\eta|X, K, L, S)$ and thus $E^{-1}(\xi\chi')E(\xi\eta') = 1$ under such correctly specified model. It follows that the estimator $\hat{\psi}(\kappa_{\rm dr})$ is a doubly-robust estimator. Further,

$$E(\xi\chi') \;\; = \;\; E\left[\xi\left\{\xi + E(\eta|X, L, S)\right\}'\right] = E(\xi\xi')$$

if model (8) is correctly specified, because ξ has conditional mean zero, given (X, L, S), under that model. It follows that $\kappa_{\rm dr} - \kappa_{opt}$ converges to zero under model (8), suggesting that $\hat{\psi}(\kappa_{\rm dr})$ has minimal variance among all estimators $\hat{\psi}(\kappa)$ under that model. Throughout this paper, we will refer to $\hat{\psi}(\kappa_{\rm dr}) \equiv \hat{\psi}_{IDR}$ as an improved doubly-robust estimator.

Finally, combining the ideas leading to the estimators $\hat{\psi}_{SDR}$ and $\hat{\psi}_{IDR}$ leads to yet a final estimator that we will refer to as the stabilized, improved doubly-robust estimator. The resulting estimator is obtained by substituting κ with $\kappa_{\rm dr}$ in (22) and ψ in expression (13) by an estimator $\tilde{\psi}$ which is consistent under model $\mathcal{B}$. We will denote it as $\hat{\psi}_{SIDR}$.

3 Simulation study

We generate 1000 datasets of size 1500 according to the data generating mechanism of Figure 1, but without confounder S. In a first simulation experiment, we postulate linear models for all variables in the diagram: $X = 1 + \epsilon_X$, $L = 1+\lambda X+0.8U+\epsilon_L$, $K = 0.5L-0.5X+\epsilon_K$ and $Y = \delta(-1+2X+0.5K+U+\epsilon_Y)$ for mutually independent, normally distributed variates U, ϵ_X, ϵ_L, ϵ_K and ϵ_Y with mean zero and standard deviations 1, 0.5, 1, 0.3 and 0.5, respectively and with $\delta = 1$. A characteristic feature of this simulation experiment is that there is a strong association between X and Y along the path $X - L - U - Y$. We considered both the cases $\lambda = 1.5$ and $\lambda = 0$ to represent settings where L is/is not affected by X. As such, we represent both settings where standard regression

methods are/are not applicable for estimating the direct effect (i.e. 2δ) of X on Y (which is not mediated by K). Assuming a correctly specified structural nested direct-effects model with $m(X, K; \psi) = \psi X$, the following estimators were calculated in each simulation, corresponding to the choices $d_K(X) = X$ and $q_K = 0$: the Inverse Probability of Intermediate Weighting (IPIW) estimator of Section 2.2, the doubly-robust (DR) estimator of Section 2.3, the unweighted (UW) estimator of Section 2.4, the stabilized doubly-robust (SDR) estimator, the improved doubly-robust (IDR) estimator and the stabilized, improved doubly-robust (SIDR) estimator of Section 2.5. Finally, we also reported the estimated coefficient for X in a linear regression model for Y, given X, K and L. We chose the following correctly specified working models: a normal conditional distribution for K given L and X with mean linear in L and X and constant residual standard deviation, and a linear regression model for Y with mean linear in X, K and L.

Because of outlying values for a number of estimators, Table 1 reports both the average and median bias, the average and median bootstrap standard error, the empirical standard deviation of the estimates and corresponding (robust) MCD-estimator, the p-value of the Wilcoxon rank test whether the median direct effect estimate differs from zero, and the coverage of standard 95% bootstrap confidence intervals. Here, bootstrap estimates are based on 1000 bootstrap samples.

We find that the standard linear regression analysis (LM) yields severely biased estimates when the confounder L is affected by X, while all other estimators are approximately unbiased. The IPIW estimator is unstable in the sense that it suffers from many outlying values. The DR estimator is considerably more stable and more efficient. Slightly higher efficiency is observed for the improved doubly-robust estimator, but the best results are obtained using the unweighted estimator. The simulation experiment where L is not affected by X reveal that the latter estimator competes very well with the standard regression analysis. Indeed, it is only slightly less efficient, but has the advantage of remaining unbiased when L is affected by X.

In a second simulation experiment (see Table 2), we investigate the impact of model misspecification by generating K as $\exp(0.5L - 0.5X + \epsilon_K)$, with all remaining variables generated as before. Estimators were obtained using the same working models that were previously used. We now obtain extremely unstable IPIW and DR estimates as a result of the estimated density of the intermediate taking extremely small values for some subjects (due to the skewness of the data). The improved doubly-robust estimators perform considerably better, with the stabilized improved DR estimator having the best performance. However, as a result of remaining instability, bootstrap standard errors could not be obtained in all simulated datasets. Overall, the most efficient estimates are again obtained via the unweighted estimator.

In a third simulation experiment (see Table 3), we misspecified working model (15) by generating Y as $Y = \delta(-1 + 2X + 0.5(K - E(K)) - 3(K - E(K))^2 + U + \epsilon_Y)$, with $\delta = 0.7$ to obtain the same variability in Y as in the first simulation expiriment. Results are now similar to those of the first simulation experiment. Curiously, also the unweighted estimator, while fully relying on the misspecified working model (15), remains unbiased. When L is not affected by X, this can be understood from the following arguments. Fitting the outcome model (16) (with S empty) then yields valid estimates for the direct effect ψ^* of X on Y, even when the association between K and Y is misspecified, because the conditional mean of X is linear in K and L under the considered data-generating mechanism (see e.g. Robins, Mark and Newey, 1992). From the form of the normal equations for the parameters indexing model (16), it thus follows that $\Delta(X)(Y - \gamma_0^* - \gamma_1^* K - \psi^* X - \gamma_2^* L)$, with γ_0^*, γ_1^* and γ_3^* the limiting values of the ordinary least squares estimators for γ_0, γ_1 and γ_3 under model (16), has mean zero. In particular, because L is not affected by X and thus independent of X under our model, we have that the estimating function of the unweighted estimator,

$$\Delta(X)(Y - \gamma_1^* K - \psi^* X),$$

is unbiased, even when the association between K and Y is misspecified. It can be seen with some algebra that this result continues to hold when L is affected by X and (X, K, L) is multivariate normal.

In a fourth simulation experiment (see Table 4), we misspecified both working models by generating K and Y as in the previous 2 simulation experiments, respectively, but with $\delta = 0.04$. As expected, all estimators are now biased. Note that, while the additional misspecification of the working model (8) has no immediate impact on the unweighted estimator (because this estimator avoids inverse probability weighting), it also becomes biased because the robustness property of this estimator (see previous paragraph) only holds for linear models. Note however, that the unweighted estimator is still outperforming the other estimators both in terms of precision and bias. With a correctly specified working model (8) for the intermediate, but a misspecified outcome model (see Table 5, simulation experiment 5), as expected, the unweighted estimator continues to behave poorly as it does not make use of the working model for the intermediate. However, the (stabilized) improved doubly-robust estimators now outperform the others, both in terms of bias and precision (but the bootstrap confidence intervals are poor in terms of coverage). The usefulness of the latter estimators is most apparent when the intermediate is non-normal and, additionally, this is acknowledged via the working model (8). We conjecture that these stabilized doubly-robust estimators will be more competitive with the unweighted estimator in settings where the weights are more stable, such as may happen when the mediator is binary.

4 Data analysis

De Sutter et al. (2006) estimate the effect of single versus double embryo transfer (SET versus DET) on birth weight using a survey of 557 SET and 396 DET patients who entered the subfertility program at the Ghent University hospital and who delivered a singleton child of at least 500 grams after fresh embryo transfer

in a first, second or third cycle between January 2003 and May 2007. The mean gestational age (GA) of singleton babies is 273.9 days (SD 12.4). The mean birth weight (BW) is 3231.8 grams (SD 565.4). De Sutter et al. (2006) observed birth weights to be 120 grams (95% confidence interval 44 - 197) lower on average in babies born after double than single enmbryo transfer. In response to criticism that the analysis was not adjusted for gestational age, Delbaere et al.(2007) argue that such adjustment would remove a possible indirect effect of SET/DET on birth weight through gestational age, and would introduce bias because gestational age may be affected by SET/DET and is associated with birth weight. At the same time, the debate raises the question whether the effect of SET/DET on birth weight is entirely mediated through gestational age.

To address this question, we assume that the causal diagram of Figure 1 represents the data generating mechanism, with S representing measured baseline confounders (embryo quality, duration of infertility, maternal age, female and male pathology, gravida and type of conception (IVF/ICSI)) for the association between SET/DET and pregnancy outcomes, and L representing measured confounders (complications during pregnancy, vaginal blood loss, preterm contractions, preterm rupture of the membranes and growth retardation) for the association between gestational age and birth weight. The diagram allows for the presence of unmeasured confounders U for the association between these confounders and outcome Y. Note that the analysis is restricted to women who deliver a singleton baby and that an implicit assumption in the analysis is thus that the loss of an embryo (in early pregnancy) in women with DET treatment is not associated with gestational age and birth weight.

Of all variables listed above as potential baseline confounders for the assocation between embryo transfer (SET/DET) and the outcome BW, only maternal age, embryo quality, duration of infertility and IVF/ICSI treatment showed a significant association with SET/DET and/or BW. Thus, only these variables are included as confounders S. For similar reasons, only preterm contractions, preterm rupture

of the membranes and growth retardation are included as confounders L. Due to the many missing values for duration of infertility (33.6%), we first performed the analyses assuming that infertility duration does not confound the association between GA and BW. This leaves us with 895 complete observations.

To estimate the direct effect of SET/DET on BW, which is not mediated by GA, we use the approaches proposed in Section 2. Based on the results of the simulation experiments in the previous section, we use the sequential G-estimator (i.e. the unweighted estimator with a linear conditional model for Y) as the primary estimator in the analysis. Since GA is skewly distributed to the left, we transformed it via a Box-Cox transformation so that we could assume a normal distribution for model (8), with mean $\alpha_0 + \alpha_1 L + \alpha_2 S + \alpha_3 X$ and constant residual standard deviation σ_K. Further, we postulated $m(X, K, S; \psi) = \psi X$, chose $d_K(X, S) = X$ and $q_K(S) = 0$ and we specified linear models $E(X|S; \beta) = \beta_0 + \beta_S$ and $E(Y|K, L, X, S) = \gamma_0 + \gamma_1 K + \gamma_2 L + \gamma_3 X + \gamma_4 S$ for the conditional expectations of the exposure SET/DET (X) and the outcome BW (Y). Table 6 summarizes the estimates obtained from the different estimation methods, along with bootstrap standard errors and confidence intervals based on 1000 bootstrap samples. As expected, after removing the indirect effect through GA, we now estimate the average birth weight to be merely 60 grams (95% confidence interval 14 - 136) lower on average in babies born after double than single enmbryo transfer. While the difference in birth weight is no longer significant after controlling for GA, the confidence interval does not exclude the possibility of important differences exceeding 100 grams.

5 Discussion

Estimating the direct effect of an exposure on an outcome, which is not mediated by some given variable, requires adjustment not only for prognostic factors of the outcome that are associated with the exposure, but additionally for those

associated with the mediator. In practice, several of these prognostic factors may only arise after the exposure was administered and thus possibly be affected by it. In such settings, standard regression methods may yield biased estimates of the direct exposure effect.

While methods based on inverse probability weighting have been proposed to accommodate this problem, they require inverse weighting by a density when the mediator is discrete with many levels or absolutely continuous. Inefficient effect estimators with large bias are then typically obtained. In this article, we have proposed an estimation approach which avoids the inverse weighting altogether. This estimator competes remarkably well with ordinary least squares estimators in settings where these are valid (i.e. in settings where prognostic factors of the outcome which are predictive of the mediator, are not themselves affected by the exposure), but remains valid in settings where the ordinary least squares estimator fails. The proposed estimator requires postulating a working model for the expected outcome in function of exposure, mediator and prognostic factors. It is robust against misspecification of this working model when the exposure, mediator and its prognostic factors have a multivariate normal distribution, but not otherwise. In view of this, we have derived doubly-robust estimators which allow for misspecification, provided that a working model for a conditional density of the mediator is correctly specified. On the basis of the simulation studies, we recommend the unweighted estimator and the (stabilized) improved doubly-robust estimator when the mediator is absolutely continuous. For a dichotomous mediator, less variable inverse weights are expected, and thus a relatively much better performance of the (stabilized) improved doubly-robust estimator.

A number of restrictions are implicit in our approach. First, we have implicitly assumed that the mediator may affect the outcome, but is not itself affected by it. In many practical studies, mediator and outcome may mutually affect each other over time. We plan to accommodate this in future work by allowing for repeated measurements on mediator and outcome. Second, we have implicitly

assumed that controlled direct effects are well defined. Standardized direct effects (Didelez, Dawid and Geneletti, 2006) are more broadly useful and can be obtained by averaging the controlled direct-effect estimates in this article over a chosen mediator density.

ACKNOWLEDGEMENT

The authors acknowledge support from IAP research network grant nr. P06/03 from the Belgian government (Belgian Science Policy).

REFERENCES

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975), "Sex Bias in Graduate Admissions: Data from Berkeley," *Science*, 187, 398-404.

Cole, S.R., and Hernan, M.A. (2002), "Fallibility in estimating direct effects," *International Journal of Epidemiology*, 31, 163-165.

De Sutter, P., Delbaere, I., Gerris, J., Verstraelen, H., Goetgeluk, S., Van der Elst, J., Temmerman, M., and Dhont, M. (2006), "Birthweight of singletons after assisted reproduction is higher after single- than after double-embryo transfer," *Human Reproduction*, 21, 2633-2637.

Delbaere, I., Vansteelandt, S., De Bacquer, D., Verstraelen, H., Gerris, J., De Sutter, P., and Temmerman, M. (2007), "Should we adjust for gestational age when analysing birth weights? The use of z-scores revisited," *Human Reproduction*, 22, 2080-2083.

Didelez, V., Dawid, A.P., and Geneletti, S. (2006), "Direct and Indirect Effects of Sequential Treatments," *Proceedings of the 22nd Annual Conference on Uncertainty in Artifical Intelligence*, 138-146

Frangakis, C.E., and Rubin, D.B. (2002), "Principal stratification in causal inference," *Biometrics*, 58, 21-29.

Molenberghs, G., Burzykowski, T., Alonso, A., and Buyse, M. (2004), "A perspective on surrogate endpoints in controlled clinical trials," *Statistical Methods in Medical Research*, 13, 177-206.

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference,* Cambridge, U.K.: Cambridge University Press.

Pearl, J. (2001), "Direct and Indirect Effects," In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, ed. M. Kaufmann, San Francisco, CA, pp. 411-420.

Petersen, M.L., Sinisi, S.E., and van der Laan, M.J. (2006), "Estimation of direct causal effects," *Epidemiology*, 17, 276-284.

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. (1994), "Correcting for non-compliance in radomized trials using structural nested mean models," *Communications in Statistics*, 23, 2379-2412.

Robins, J.M. (1999), "Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models," In *Computation, Causation, and Discovery*, eds. C. Glymour, and G.F. Cooper, AAAI Press/The MIT Press, pp. 349-405.

Robins, J.M., Mark, S.D., and Newey, W.K. (1992), "Estimating exposure effects by modelling the expectation of exposure conditional on confounders," *Biometrics*, 48, 479-495.

Robins, J.M., and Greenland, S. (1992), "Identifiability and exchangeability for direct and indirect effects," *Epidemiology*, 3, 143-155.

Robins, J.M., Rotnitzky, A., and Vansteelandt, S. (2007), "Discussion of 'Principal stratification designs to estimate input data missing due to death' by C.E. Frangakis, D.B. Rubin, M.-W. An, and E. MacKenzie', *Biometrics*, 63, 650-653.

Robins, J.M., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007), "Performance of double-robust estimators when 'inverse probability weights are highly variable," *Statistical Science*, in press.

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* 147, 656-666.

Rubin, D.B. (2004), "Direct and indirect causal effects via potential outcomes," *Scandinavian Journal of Statistics*, 31, 161-170.

Tan, Z.Q. (2006), "A distributional approach for causal inference using propensity scores," *Journal of the American Statistical Association*, 101, 1619-1637.

Taylor, J.M.G., Wang, Y., and Thiebaut, R. (2005), "Counterfactual links to the proportion of treatment effect explained by a surrogate marker," *Biometrics*, 61, 1102-1111.

Ten Have, T.R., Joffe, M.M., Lynch, K.G., Brown, G.K., Maisto, S.A. and Beck, A.T. (2007), "Causal Mediation Analyses with Rank Preserving Models," *Biometrics*, 63, 926-934.

van der Laan, M. J., and Robins, J. M. (2003), *Unified Methods for Censored Longitudinal Data and Causality*, New-York: Springer-Verlag.

van der Laan, M.J., and Petersen, M.L. (2005), "Direct Effect Models," *U.C. Berkeley Division of Biostatistics Working Paper Series*, Paper 187.

Verstraelen, H., Goetgeluk, S., Derom, C., Vansteelandt, S., Derom, R., Goetghebeur, E., and Temmerman, M. (2005), "Preterm birth in twins following subfertility treatment: a population-based cohort study," *British Medical Journal*, 331, 1173-1176.

APPENDIX: PROOF OF THEOREM 1

Part 1 of Theorem 1 is immediate upon applying Theorem 1.2 in van der Laan and Robins (2003). With $F \equiv f(K|L,S,X;\alpha)$, $M \equiv m(X,S,K;\psi)$, $Q \equiv q_K(S)$ and $D \equiv d_K(X,S)$, this yields estimating function

$$\begin{aligned}
&\Delta\{D|S;\beta\}W(\alpha)(Y-M-Q)\\
&-[E(\Delta\{D|S;\beta\}W(\alpha)(Y-M-Q)|K,L,X,S)\\
&-E\{E(\Delta\{D|S;\beta\}W(\alpha)(Y-M-Q)|K,L,X,S)|L,X,S\}]
\end{aligned}$$

The first two terms can be rewriten as

$$\sum_{i=1}^{n}\Delta\{D|S;\beta\}W(\alpha)(Y-E(Y|K,L,X,S))$$

The third term equals

$$E[\Delta\{D|S;\beta\}W(\alpha)\{E(Y|K,L,X,S)-M-Q\}|X,L,S]$$

which can be calculated as

$$\int\Delta\{D|S;\beta\}\{E(Y|K,L,X,S)-M-Q\}f(K|S)dK$$

To prove Part 2 of Theorem 1, we assume that the regularity conditions of Theorem 1A in Robins, Mark and Newey (1992) hold for $U_{i,DR}(d,q;\psi,\alpha,\beta,\gamma)$, the estimating function $G_i(\gamma)$ for γ and $A_i(\alpha)$ for α. By standard Taylor expansion arguments,

we have that

$$\begin{aligned} 0 &= n^{-1/2}\sum_{i=1}^{n} U_{i,DR}(d,q;\psi^*,\tilde{\alpha},\beta,\tilde{\gamma}) + E\left\{\frac{\partial}{\partial\psi}U_{i,DR}(d,q;\psi=\psi^*,\tilde{\alpha},\beta,\tilde{\gamma})\right\}\sqrt{n}(\widehat{\psi}-\psi^*) \\ &\quad -E\left\{\frac{\partial}{\partial\gamma}U_{i,DR}(d,q;\psi^*,\tilde{\alpha},\beta,\gamma=\tilde{\gamma})\right\}E^{-1}\left\{\frac{\partial}{\partial\gamma}G_i(\gamma=\tilde{\gamma})\right\}G_i(\tilde{\gamma}) \\ &\quad -E\left\{\frac{\partial}{\partial\gamma}U_{i,DR}(d,q;\psi^*,\alpha=\tilde{\alpha},\beta,\tilde{\gamma})\right\}E^{-1}\left\{\frac{\partial}{\partial\alpha}A_i(\alpha=\tilde{\alpha})\right\}A_i(\tilde{\alpha}) + o_p(1) \qquad (23) \end{aligned}$$

where $o_p(1)$ denotes a random variable converging to 0 in probability, and where $\tilde{\gamma}$ and $\tilde{\alpha}$ are the probability limits of the estimators for γ^* and α^*.

First note that $U_{i,DR}(d,q;\psi,\tilde{\alpha},\beta,\tilde{\gamma})$ has mean zero at $\psi=\psi^*$ under model $\mathcal{A}$, even when model (15) for the conditional expectation of the outcome is misspecified. This is because the first term in (12) has mean zero at ψ^* under model $\mathcal{A}$ by construction and the second term is a mean zero function under model $\mathcal{A}$ for each choice of $\phi(K_i,L_i,X_i,S_i)$ and thus in particular for $\phi_{opt}(K_i,L_i,X_i,S_i)$. We now show that $U_{i,DR}(d,q;\psi,\tilde{\alpha},\beta,\tilde{\gamma})$ has mean zero at $\psi=\psi^*$ under model $\mathcal{B}$, even when model (8) for the conditional density of the mediator is misspecified. Using the potential outcomes framework, we may rewrite the estimating function in (17) as

$$\begin{aligned} U_{DR} &= \int \frac{I(K=k)}{F}\Delta\{D|S\}(Y_k-M-Q)dk \\ &\quad -E\left(\int \frac{I(K=k)}{F}\Delta\{D|S\}(Y_k-M-Q)dk|X,K,L,S\right) \\ &\quad +E\left[E\left(\int \frac{I(K=k)}{F}\Delta\{D|S\}(Y_k-M-Q)dk|X,K,L,S\right)|X,L,S\right] \end{aligned}$$

We rewrite the first term as

$$\int\left[(Y_k-M-Q)\Delta\{D|S\}+\left\{\frac{I(K=k)}{F}-1\right\}(Y_k-M-Q)\Delta\{D|S\}\right]$$

The second term equals

$$\int\frac{I(K=k)}{F}E(Y_k-M-Q|K=k,X,L,S)\Delta\{D|S\}dk$$

and the third term can be further simplified to

$$
\begin{aligned}
& E\left[\int \frac{I(K=k)}{F} E(Y_k - M - Q|K=k, X, L, S)\Delta\{D|S\}dk|X, L, S\right] \\
&= \int E(Y_k - M - Q|K=k, X, L, S)\Delta\{D|S\}E\left(\frac{I(K=k)}{F}|X, L, S\right)dk \\
&= \int E(Y_k - M - Q|K=k, X, L, S)\Delta\{D|S\}dk
\end{aligned}
$$

Adding these 3 terms yields

$$
\begin{aligned}
& \int \left[(Y_k - M - Q)\Delta\{D|S\} + \left\{\frac{I(K=k)}{F} - 1\right\}(Y_k - M - Q)\Delta\{D|S\}\right. \\
& \left. - \left\{\frac{I(K=k)}{F} - 1\right\} E(Y_k - M - Q|K=k, X, L, S)\Delta\{D|S\}\right] dk \\
&= \int (Y_k - M - Q)\Delta\{D|S\}dk \\
& + \int \left\{\frac{I(K=k)}{F} - 1\right\}(Y_k - E(Y_k|K=k, X, L, S))\Delta\{D|S\}dk
\end{aligned}
$$

The first term was shown to have mean zero at ψ^* in Section 2.2. The second term has mean zero when, as in model $\mathcal{B}$, the conditional expectation of Y is correctly specified, since Y_k is independent of K conditionally on X, L and S provided that, as model $\mathcal{B}$ postulates, L represents all common causes of K and Y. We conclude that $U_{i,DR}(d, q; \psi, \tilde{\alpha}, \beta, \tilde{\gamma})$ has mean zero at $\psi = \psi^*$ under model $\mathcal{A} \cup \mathcal{B}$. Further, note that $\tilde{\gamma} = \gamma^*$, and thus that $E\left\{\partial U_{i,DR}(d, q; \psi^*, \alpha = \tilde{\alpha}, \beta, \tilde{\gamma})/\partial\alpha\right\} = 0$ and $E\{G_i(\tilde{\gamma})\} = 0$ when model (15) is correctly specified. Likewise, $\tilde{\alpha} = \alpha^*$, and thus $E\left\{\partial U_{i,DR}(d, q; \psi^*, \tilde{\alpha}, \beta, \gamma = \tilde{\gamma})/\partial\gamma\right\} = 0$ and $E\{A_i(\tilde{\alpha})\} = 0$ when model (8) is correctly specified. Under the regularity conditions of Theorem 1A in Robins, Mark and Newey (1992), it now follows from the asymptotic unbiasedness of $\sqrt{n}(\widehat{\psi} - \psi^*)$ under model $\mathcal{A} \cup \mathcal{B}$ that $\widehat{\psi}$ is a consistent estimator of ψ^* under model $\mathcal{A} \cup \mathcal{B}$.

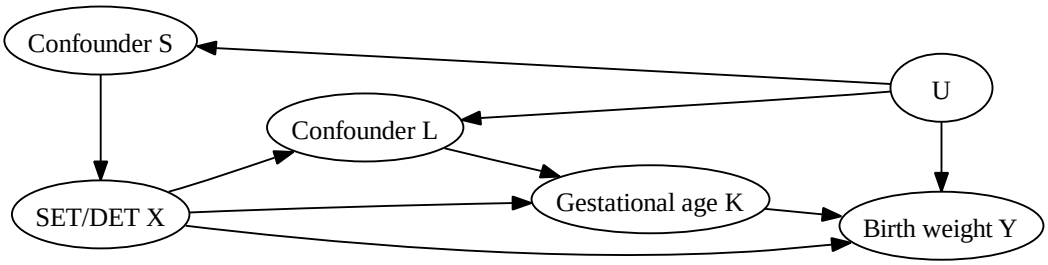

Figure 1: *Causal Diagram*

Table 1: *Simulation Experiment 1*

with effect of X on L					
	bias (mean/med)	boot SE (mean/med)	SE (emp/MCD)	p-value bias= 0	boot cov
IPIW	-0.49/-0.024	44.98/0.20	13.59/0.28	0.01	92.3
DR	0.029/-0.001	1.36/0.12	1.37/0.17	0.73	96.1
UW	-0.001/-0.001	0.061/0.061	0.061/0.061	0.51	95.0
SDR	-0.055/0.00	26.53/0.15	1.91/0.21	0.51	94.1
IDR	-0.011/-0.010	0.16/0.10	0.21/0.13	0.11	93.9
SIDR	-0.016/-0.006	1.013/0.12	0.40/0.15	0.43	95.3
LM	-0.73/-0.73	0.068/0.068	0.071/0.072	0.00	0
without effect of X on L					
	bias (mean/med)	boot SE (mean/med)	SE (emp/MCD)	p-value bias= 0	boot cov
IPIW	0.18/0.032	49.33/0.26	1.79/0.37	0.00	94.4
DR	0.009/0.005	1.23/0.12	1.25/0.18	0.60	95.5
UW	0.004/0.007	0.07/0.07	0.071/0.071	0.059	95.3
SDR	0.034/0.002	18.92/0.16	2.59/0.23	0.71	94.0
IDR	0.005/0.002	0.15/0.10	0.17/0.13	0.19	94.7
SIDR	-0.003/0.006	1.88/0.13	0.24/0.16	0.28	95.3
LM	0.003/0.003	0.062/0.062	0.063/0.064	0.078	95.3

Table 2: *Simulation Experiment 2*

with effect of X on L					
	bias (mean/med)	boot SE (mean/med)	SE (emp/MCD)	p-value bias= 0	boot cov
IPIW	10.82/9.1	437/10.0	23.8/5.3	0.00	60.0
DR	8.3 10^{50}/-0.13	1.4 10^{64}/1.8 10^{9}	2.7 10^{51}/1.9	0.59	100.0
UW	-0.001/-0.001	0.059/0.059	0.059/0.059	0.78	95.1
SDR	0.00/0.018	41.3/0.97	2.1/0.77	0.80	94.2
IDR	17.1/0.037	-/-	520/1.8	0.00	-
SIDR	-0.022/-0.002	-/-	0.58/0.088	0.86	-
LM	-0.73/-0.73	0.061/0.061	0.063/0.063	0.00	0
without effect of X on L					
	bias (mean/med)	boot SE (mean/med)	SE (emp/MCD)	p-value bias= 0	boot cov
IPIW	-15.0/10.2	4153/39.0	936.4/16.2	0	84.8
DR	$-3.3/10^{50}$/-4.4	2.5 10^{65}/8.9 10^{11}	1.1 10^{53}/-	0.012	100.0
UW	0.002/0.000	0.062/0.062	0.062/0.062	0.23	96
SDR	-6.7/0.022	594/4.9	151.8/2.4	0.88	96.6
IDR	3.6/0.033	-/-	353.62/2.60	0.001	-
SIDR	-0.074/0.005	-/-	3.13/0.12	0.17	-
LM	0.002/0.001	0.053/0.053	0.053/0.053	0.41	95.3

Table 3: *Simulation Experiment 3*

with effect of X on L					
	bias (mean/med)	boot SE (mean/med)	SE (emp/MCD)	p-value bias= 0	boot cov
IPIW	-0.63/-0.027	84.1/0.33	20.8/0.43	0.9	91.7
DR	-0.017/-0.010	1.9/0.17	2.0/0.25	0.94	96.1
UW	-0.002/-0.004	0.096/0.095	0.099/0.098	0.35	93.0
SDR	0.11/-0.008	44.6/0.21	3.7/0.31	0.55	94.7
IDR	-0.004/-0.007	0.23/0.13	0.26/0.16	0.45	95.1
SIDR	0.059/-0.008	0.53/0.15	1.8/0.20	0.95	95.2
LM	-0.51/-0.52	0.12/0.12	0.12/0.12	0.00	1.2
without effect of X on L					
	bias (mean/med)	boot SE (mean/med)	SE (emp/MCD)	p-value bias= 0	boot cov
IPIW	0.068/-0.004	106/0.35	7.5/0.51	0.01	92.9
DR	-0.035/-0.005	2.4/0.18	2.4/0.26	0.66	95.1
UW	0.001/0.003	0.12/0.12	0.12/0.12	0.63	95.6
SDR	-0.17/-0.014	56.9/0.23	3.02/0.35	0.22	94.5
IDR	-0.008/0.001	0.23/0.14	0.29/0.18	0.83	94.6
SIDR	-0.008/-0.007	0.65/0.17	0.50/0.22	0.97	95.3
LM	0.000/0.002	0.12/0.12	0.12/0.12	0.80	95.5

Table 4: *Simulation Experiment 4*

with effect of X on L					
	bias (mean/med)	boot SE (mean/med)	SE (emp/MCD)	p-value bias= 0	boot cov
IPIW	-92.87/-37.12	8590.71/51.34	405.46/63.32	0.00	53.2
DR	-8.210^{52} / -42581	$1.3\ 10^{66}$ / $3.5\ 10^{43}$	$2.0\ 10^{54}$/-	0.0	100.0
UW	1.85/0.58	0.067/0.066	0.067/0.066	0.00	0.0
SDR	-1.65/-20.01	493.06/2.70	40.03/1.96	0.00	95.7
IDR	-16.75/0.52	224.65/113.73	319.29/0.11	0.57	96.7
SIDR	1.83/0.55	3.49/0.14	0.44/0.083	0.00	93.1
LM	1.78/-1.16	0.054/0.055	0.065/0.064	0.00	0
without effect of X on L					
	bias (mean/med)	boot SE (mean/med)	SE (emp/MCD)	p-value bias= 0	boot cov
IPIW	16.6/-19.9	7999/65.6	1507/39.5	0	79.6
DR	$4.2\ 10^{52}$/2211408	$3.0\ 10^{67}$/ $1.2\ 10^{13}$	$1.2\ 10^{51}$ / 2.37	0.00	100.0
UW	-0.34/-0.33	0.11/0.097	0.15/0.095	0.00	0
SDR	9.4/-11.2	3514/0.097	826/21.9	0.00	76.2
IDR	6.6/0.52	-/-	107/1.78	0.00	-
SIDR	-0.24/0.55	-/-	1.3/0.12	0.00	-
LM	-0.34/-0.30	0.16/0.14	0.15/0.089	0.00	27.3

Table 5: *Simulation Experiment 5*

with effect of X on L					
	bias (mean/med)	boot SE (mean/med)	SE (emp/MCD)	p-value bias= 0	boot cov
IPIW	-0.30/-0.035	13.68/0.091	5.79/0.092	0.00	92.7
DR	0.43/0.46	1.13/0.35	5.04/0.36	0.00	61.5
UW	0.64/0.57	0.18/0.14	0.29/0.15	0.00	0
SDR	201/0.41	137.52/0.43	49.38/0.47	0.00	70.3
IDR	0.042/0.0014	0.13/0.097	0.22/0.11	0.57	74.7
SIDR	0.039/-0.00097	0.14/0.084	0.22/0.088	0.00	77.8
LM	-1.30/-1.16	0.097/0.085	0.55/0.056	0.00	44.4
without effect of X on L					
	bias (mean/med)	boot SE (mean/med)	SE (emp/MCD)	p-value bias= 0	boot cov
IPIW	-0.008/0.022	2.78/0.033	1.25/0.035	0.00	39.6
DR	-0.17/-0.24	0.39/0.11	2.34/0.12	0.00	58.8
UW	-0.34/-0.30	0.084/0.060	0.15/0.072	0.00	0.30
SDR	-0.17/-0.22	7.2/0.14	3.10/0.15	0.00	73.3
IDR	0.0003/-0.008	0.069/0.056	0.10/0.049	0.037	64.9
SIDR	0.009/0.009	0.081/0.053	0.096/0.055	0.00	54.5
LM	-0.34/-0.30	0.024/0.021	0.14/0.063	0.00	21.3

Table 6: *Data Analysis Results*

	Without infertility duration			With infertility duration		
	$\hat{\psi}$	boot SE	95% CI	$\hat{\psi}$	boot SE	95% CI
IPIW	78.20	100.16	[-143.41;304.21]	91.90	144.41	[-224,78;338.44]
DR	-67.77	40.44	[-141.19;14.42]	-84.11	53.75	[-190.64;14.79]
UW	-59.64	36.70	[-136.49;13.97]	-70.76	47.92	[-156.37;14.98]
SDR	-67.69	40.38	[-141.22;14.38]	-83.82	53.54	[-189.42;15.37]
IDR	-69.52	42.46	[-154.40;18.41]	-86.07	54.87	[-181.93;15.03]
SIDR	-69.45	42.33	[-153.08;18.18]	-85.77	54.59	[-181.63;14.70]
LM	-44.59	33.49	[-115.06;28.88]	-71.14	45.18	[-148.02;6.27]