Harvard University

Harvard University Biostatistics Working Paper Series

Year 2014 *Paper* 176

A Simple Regression-based Approach to Account for Survival Bias in Birth Outcomes Research

Eric J. Tchetgen Tchetgen[*] Kelesitse Phiri[†]

Roger Shapiro[‡]

[*] Harvard University, etchetge@hsph.harvard.edu

[†] Harvard School of Public Health, kphiri@hsph.harvard.edu

[‡] Harvard School of Public Health, rshapiro@hsph.harvard.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/harvardbiostat/paper176

Copyright ©2014 by the authors.

Original Article

A simple regression-based approach to account for survival bias in birth outcomes research

Eric J. Tchetgen Tchetgen[1,2], Kelesitse Phiri[2], Roger Shapiro[3]

[1]Department of Epidemiology,

[2]Department of Biostatistics,

[3]Department of Immunology and Infectious Diseases,

Harvard School of Public Health

Corresponding author: Eric J. Tchetgen Tchetgen, Department of Epidemiology, Harvard School of Public Health 677 Huntington Avenue, Boston, MA 02115.

Running head:Accounting for survival bias in perinatal studies

KEY WORDS: Survival bias, survivor average causal effect, principal stratification, counterfactual, stillbirth

Word count, main text:3161

Word count, abstract: 246

Number of Figures:1

Number of tables:2

Number of Pages:21

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Abstract

In perinatal epidemiology, birth outcomes such as small for gestational age (SGA) may not be observed for a pregnancy ending with a stillbirth. It is then said that SGA is truncated by stillbirth, which may give rise to survival bias when evaluating the effects on SGA of an exposure known also to influence the risk of a stillbirth. In this paper, we consider the causal effects of maternal infection with human immunodeficiency virus (HIV) infection on the risk of SGA, in a sample of pregnant women in Botswana. We hypothesize that previously estimated effects of HIV on SGA may be understated because they fail to appropriately account for the over-representation of live births among HIV negative mothers, relative to HIV positive mothers. A simple yet novel regression-based approach is proposed to adjust effect estimates for survival bias for an outcome that is either continuous or binary. Under certain straightforward assumptions, the approach produces an estimate which may be interpreted as the survivor average causal effect (SACE) of maternal HIV, which is, the average effect of maternal HIV on SGA among births that would be live irrespective of maternal HIV status. The approach is particularly appealing, because it recovers an exposure effect which is robust to survival bias, even if the association between the risk of SGA and that of a stillbirth, cannot be completely explained by adjusting for observed common risk factors. The approach also gives a formal statistical test of the null hypothesis of no survival bias in the regression framework.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Selection bias is a well-known threat to epidemiologic research, and is said to be present if in a data sample, features of primary scientific interest are entangled with features of the selection process that gave rise to the sample, but are not of immediate interest. In such situations, it may not be possible to obtain reliable inferences without explicitly acknowledging the selection process. A particularly virulent form of selection bias, sometimes present in cohort studies concerned with evaluating causal effects of a point exposure, arises when a subset of the cohort dies prior to follow-up, and therefore does not have the outcome of interest ascertained. Then the outcome is said to be truncated by death and inference about causal effects for the outcome in question may be subject to survival bias[1,2]. For instance, in perinatal epidemiology studies, a maternal exposure may have significant teratologic effects which may include causing a stillbirth, thus leading to truncation of birth outcomes defined only for live births, e.g. small for gestational age (SGA), preterm delivery and certain congenital malformations. In the presence of truncation due to stillbirth, maternal exposure effects measured only among live births cannot easily be interpreted causally, if the risk of a stillbirth remains associated with the outcome in view even after adjusting for the exposure and other observed risk factors. This is an example of an effect defined conditional on a post-exposure event (live birth) affected by exposure, a potential source of selection bias due to so-called collider bias.[3,4] In this paper, we consider the causal effect of maternal infection with the human immunodeficiency virus (HIV) infection on the risk of giving birth to an infant who is SGA in a study of 15922 pregnant women in Botswana, Africa. Maternal HIV status elevates the risk of a stillbirth, which may itself have unobserved genetic and environmental common causes with SGA. As a result, inferences about the effects of maternal HIV on SGA may be severely biased even if one has properly accounted for all confounders of the effects of HIV, unless one also appropriately accounts for differential risk of a stillbirth associated with maternal HIV status.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

While truncation by death has in recent years become a prominent topic in causal inference,

Hosted by The Berkeley Electronic Press

and epidemiologists have increasingly become aware of survival bias induced by conditioning on an intermediate event (e.g. live birth) in studies of perinatal epidemiology, still, this practice remains quite common. Truncation by death presents certain challenges that are quite distinct from standard missing outcome problems. Specifically, a missing outcome is in principle nonetheless well defined, albeit being unobserved for a subset of the sample. In contrast, an outcome such as birthweight used in defining SGA is not only unobserved for stillbirths, but cannot be well defined unless the birth is live, which requires care in defining causal contrasts. In this paper, we will focus primarily on the so-called survivor average causal effect (SACE), which is the causal effect of an exposure for the subset of persons that would survive whether exposed or not.[1,2] In the context of maternal HIV, SACE for say birthweight gives the average causal effect (say on the additive scale) of maternal HIV status on birthweight for the subset of infants born alive irrespective of maternal HIV status. For this subset of births, it is arguably the case that an observed association between HIV and birthweight cannot be attributed to survival bias, since the latter cannot operate with a null effect of maternal HIV on stillbirth in the sub-sample. Furthermore, the SACE contrast remains unambiguous despite the presence of truncation by death, since the birth outcome in view remains well defined for infants who would survive under both exposure conditions.

Although one can never know with certainty, whether an observed live birth of an HIV negative mother, would also be a live birth if contrary to fact the mother was HIV positive, it is nonetheless sometimes possible to make population inferences about the SACE under certain assumptions. In this paper, we present a simple approach for estimating the SACE based on a straightforward modification of standard regression analysis routinely used in epidemiologic practice. The proposed modified regression approach can under certain conditions recover a valid estimate of the SACE of HIV on SGA, even if substantial dependence persists between SGA and the risk of a stillbirth after adjusting for common risk factors. To the best of our knowledge, the proposed analytic approach

COBRA A BEPRESS REPOSITORY Collection of Biostatistics Research Archive

to account for survival bias is novel. For simplicity, in the next section, we present the approach in the context of simple linear regression. Next, the approach is extended to handle binary outcomes using logistic regression (for a rare outcome) and log linear regression (for a non-rare outcome). We conclude with a discussion comparing the proposed methodology to some prominent methods in the literature. For illustration, the methods are presented throughout in the context of data from the Botswana Birth Outcomes Surveillance Study, that we use to estimate the effects of maternal HIV infection on SGA .[5]

Estimating SACE for a continuous outcome

In the following, we let A denote a mother's HIV status ($A = 1$ if HIV infected, 0 if HIV-uninfected), S is an indicator of a live birth ($S = 1$ if live birth, 0 if stillbirth), Y is a continuous birth outcome, here birthweight, which is only observed in case of a live birth (i.e. if $S = 1$) and is otherwise undefined. In addition to these variables, one also observes C which includes (pre-exposure) correlates of A, S and Y (among live births with $S = 1$). Note that in a randomized trial in which A is randomly assigned relative to C, the latter may nonetheless remain associated with S, as well as with Y in live births. The causal relationship for the observed data is depicted in the causal direct acyclic graph (DAG) displayed in Figure 1, for an observation with $S = 1$. The causal DAG also includes a variable U, which we will suppose is an unobserved common cause of S and Y. The presence of U ensures that Y and S remain dependent even after conditioning on observed risk factors included in C. We will also consider the counterfactual outcome $S(a)$ which stands for infant stillbirth status under maternal HIV status $a = 0, 1$. For $S(a) = 1$, a live birth under exposure a, we define the corresponding counterfactual birthweight $Y(a)$; however for $S(a) = 0$, a stillbirth under exposure value a, birthweight is undefined. We also make the consistency assumption that $S = S(A)$, and $Y = Y(A)$ if $S = 1$. Throughout, we assume that

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

conditional on C, there is no unobserved confounding of the effects of A on Y. The survivor average causal effect, SACE, conditional on C is defined as

$$SACE(c) = E(Y(1) - Y(0)|S(1) = S(0) = 1, C = c).$$

Suppose briefly that U were observed. Then, identification of $SACE(c)$ would be possible under the DAG in Figure 1 if it could be interpreted as implying the following counterfactual independence condition:

$$Y(a) \perp\!\!\!\perp S(1-a)|S(a), A = a, C = c, U = u. \tag{1}$$

This independence assumption would hold for instance, if the causal diagram were interpreted as a graphical representation of Pearl's nonparametric structural equations model with independent error,[3] see for instance Tchetgen Tchetgen et al.[4] Then, under condition (1),

$$\begin{aligned}
&E\left(Y(1)|S(1) = S(0) = 1, C = c, U = u\right)\\
&= E\left(Y(1)|S(1) = 1, C = c, U = u\right)\\
&= E\left(Y(1)|A = 1, S(1) = 1, C = c, U = u\right)\\
&= E\left(Y|A = 1, S = 1, C = c, U = u\right),
\end{aligned}$$

where the first equality follows from (1), the second equality follows from no confounding assumption of the effects of A, and the third equality follows from consistency. One can likewise show

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

that

$$E\left(Y(0)|S(1)=S(0)=1,C=c,U=u\right) \tag{2}$$
$$=E\left(Y|A=0,S=1,C=c,U=u\right),$$

and

$$SACE(c)=E\left\{E\left(Y|A=1,S=1,c,U\right)-E\left(Y|A=0,S=1,c,U\right)|c\right\}.$$

However, the causal contrast $SACE(c)$ is generally not identified from the observed data (SY,A,S,C) without an additional assumption, since

$$Y(a) \not\perp\!\!\!\perp S(1-a)|S(a)=1,A=a,C=c, \tag{3}$$

even if assumption (1) holds. To make progress, we show that identification is sometimes possible even if U is not observed, under certain assumptions. In this vein, suppose that conditional on (U,A,C),

$$E\left(Y|S=1,A,C,U\right)=\beta_0+\beta_a A+U+b\left(C\right), \tag{4}$$

where $b\left(C\right)$ is an unrestricted function of the covariates. This model is attractive in its simplicity and flexibility, as it includes a range of model specifications one is likely to use in practice for a continuous outcome, if U were actually observed. For example, standard linear regression corresponds to specifying $b\left(C\right)=\beta_c'C$ as a linear function of C. But semiparametric or nonparametric techniques could also be used to model the confounders, including generalized additive models,

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

splines, polynomial or wavelets. It is straightforward to verify that under model (4) ,

$$SACE(c) = \beta_a$$

does not vary with c. To make further progress, it is helpful to encode the dependence between S and U on the log odds ratio scale using the simple specification:

$$\operatorname{logit} \Pr \{S = 1|A, U, C\} = \alpha' U + v(A, C), \tag{5}$$

which specifies a linear log odds ratio association between U and S conditional on A and C. Under the above model specification, the baseline function $v(A, C)$ =logitPr($S = 1|A, U = 0, C$) is allowed to remain unrestricted. It is straightforward to verify that the null value $\alpha = 0$ implies that $Y(a)$ and $S(1 - a)$ are independent conditional on $S = 1, A = a, C$, i.e.

$$\alpha = 0 \Longrightarrow Y(a) \perp\!\!\!\perp S(1 - a)|S = 1, A = a, C,$$

which would also imply that $SACE(c)$ is nonparametric identified from the observed data even if one does not observe U. Finally, we assume that in the population

$$E\left(U|A, C\right) = E(U|C), \tag{6}$$

which essentially states that A does not directly influence U conditional on C, and is consistent with the causal diagram of Figure 1. However, for stillbirths (with $S = 0$), we expect that U and A will be associated conditional on C, and we denote the corresponding residual $\Delta =$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

$U - E\left(U|A=a, S=0, C\right)$, such that

$$\Delta \perp\!\!\!\perp (A, C) \text{ given } S=0. \tag{7}$$

Specifically, we assume that any association between U and (A, C) among stillbirths must be operating entirely as a location shift. This would be the case, for instance if U were normally distributed with variance σ^2 among stillbirths, however, our model is substantially more flexible.

Result 1: Under assumptions $(1), (4), (5), (6)$ *and* (7) *we have that*

$$E(Y|A, C, S=1) = \beta_0^* + \beta_a A + \beta_{ac} Q + b^*\left(C\right), \tag{8}$$

where $b^*\left(C\right)$ is an unrestricted function of C, the unknown coefficient β_{ac} is defined in the appendix, and

$$Q = \left(1 - \pi\left(A, C\right)\right),$$

$$\pi\left(A, C\right) = \Pr(S=1|A, C),$$

Our first Result 1 establishes that under the (semi-linear) model (4), with U satisfying (1) and (7), $SACE(c) = \beta_a$ is identified from the observed data (A, C, SY) despite the fact that U is unobserved and (3) holds. Suppose for a moment that $\pi\left(a, c\right) = \Pr(S=1|a, c)$ is known for all a and c, so that Q given in Result 1 is observed. Suppose also for simplicity that one has correctly specified a linear model for $b^*\left(C\right) = \beta_c^{*\prime} C$. Then according to the Result, β_a is identified by the regression coefficient of A in a standard regression model of Y on (A, C, Q), and unbiased estimates of the regression coefficients $\beta = (\beta_0, \beta_a, \beta_c^{*\prime}, \beta_{ac})'$ can be obtained via ordinary least squares. The regression model (1) is similar to the underlying data generating regression (4), with

Hosted by The Berkeley Electronic Press

COBRA A BEPRESS REPOSITORY Collection of Biostatistics Research Archive

the important distinction that U is substituted with the factor Q. The result shows that under the stated assumptions, this latter factor is essentially needed in the regression model, to account for survival bias. Under a normal model for Δ with mean zero and variance σ^2 among stillbirths, the regression coefficient for Q, β_{ac} is equal to $\sigma^2\alpha$, the product of the log-odds ratio association between U and S, and the variance of U. Thus, as intuition would dictate, survival bias vanishes, i.e. $\beta_{ac} = 0$ if either, as previously discussed, $\alpha = 0$, or alternatively, if $\sigma^2 = 0$ and therefore there is no unmeasured predictor of Y conditional on $A = 0, S = 1$ and C. In addition to reporting β_a, the parameter β_{ac} is also of interest as it quantifies the extent to which survival bias might be operating on the mean difference scale, so that a test of the null hypothesis $\beta_{ac} = 0$ amounts to a test of no selection bias.

In practice, $\pi(a, c)$ will seldom be known a priori, in which case, one may proceed with estimation in two-stages, whereby a first stage estimation of $\pi(a, c)$ can be obtained via standard logistic regression of S on (A, C) by maximum likelihood which produces an estimator $\widehat{\pi}(a, c) = \widehat{\Pr}(S = 1|a, c)$ which may in turn be used to construct the estimate $\widehat{Q} = (1 - \widehat{\pi}(A, C))$. An asymptotically unbiased estimator of β is then obtained by a second stage regression of Y on $(A, C, \widehat{Q})$. For valid inference, it is necessary to acknowledge the additional uncertainty associated with the first stage estimation of $\pi(A, C)$. Thus, we recommend the nonparametric bootstrap to construct valid confidence interval for β_a or other parameters of interest.

Estimating SACE for a binary outcome

In the following, we now suppose that Y is a binary birth outcome, such as SGA. An appropriate target of inference for such a binary outcome in the context of truncation by death is given by the risk ratio SACE measure :

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

$$SACE(c) = \frac{\Pr(Y(1) = 1|S(1) = S(0) = 1, C = c)}{\Pr(Y(0) = 1|S(1) = S(0) = 1, C = c)}.$$

Similar to the model used for a continuous outcome, we will suppose that the following log-linear model generated the data

$$\log \Pr\left(Y = 1|S = 1, A, C, U\right) = \beta_0 + \beta_a A + U + b_c\left(C\right) \tag{9}$$

where $b_c\left(C\right)$ is an unrestricted function, U is an unobserved correlate of Y and S as depicted in the causal DAG in Figure 1, so that conditions (1) and (7) hold, and the following Result 2 is the log-linear analog of Result 1.

Result 2: Under assumptions $(1), (5), (6), (7)$ *,and* (9), *we have that,*

$$\log \Pr(Y = 1|A, C, S = 1) = \beta_0^* + \beta_a A + b^*\left(C\right) + \beta_{ac} Q, \tag{10}$$

with Q *and* β_{ac} *as defined in Result 1.*

Result 2 gives a simple parametrization for the log-linear regression of Y on (A, C) for live births, which allows one to recover under the assumptions stated in the result, the risk ratio SACE of A. Similar to the linear case, the adjustment is achieved by extending the standard log-linear model with the extra term Q with regression coefficient β_{ac} encoding on the log risk ratio scale the extent to which survival bias may be operating. For estimation, it is convenient when Y is rare within levels of A and C, to use a logit link in lieu of the log link, with $b^*\left(C\right)$ modeled as the linear function $\beta_c^{*\prime} C$ with unknown parameter β_c^*, and fit the following regression,

$$\operatorname{logit} \Pr(Y = 1|A, C, S = 1) = \theta_0^* + \theta_a A + \theta_c^{*\prime} C + \theta_{ac} \widehat{Q}, \tag{11}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

via standard maximum likelihood, where $\widehat{Q}$ is obtained in a first stage regression as described in the previous section. We then have that under the rare disease assumption $\theta \approx \beta$, where $\theta = (\theta_0^*, \theta_a, \theta_c^{*\prime}, \theta_{ac})$ and $\beta = (\beta_0^*, \beta_a, \beta_c^{*\prime}, \beta_{ac})$. As in the linear case, we recommend the nonparametric bootstrap for inference.

If Y is not rare, one may adopt one of several existing methods to estimate the risk ratio regression (11) with the log link replacing the logit link, including the log-binomial model of Wacholder,[6] the Poisson regression approach of Zou,[7] and the semiparametric locally efficient approach of Tchetgen Tchetgen.[8]

Data application

In this section, we illustrate the methods described above to account for survival bias induced by stillbirths using data from a birth outcomes surveillance study of 9,504 (30%) HIV-infected and 22,609 HIV-uninfected women in Botswana. The primary aim of the study was to assess whether maternal HIV status and use of highly active antiretroviral therapy during pregnancy, respectively, is associated with adverse birth outcomes, including stillbirths, preterm delivery, SGA and congenital anomalies.[5] Details regarding the study sample, data extraction and analysis are available in the original paper.[5] Briefly, the study included all women who delivered live births or stillbirths at a gestational age $\geq$ 20 weeks at 6 government facilities in Botswana between May 2009 and April 2011. Study information, obtained from maternal obstetric records, included maternal demographics, medical history, antiretroviral use, and birth outcomes.

For our purpose, we consider the effect of maternal HIV status on SGA in a complete-case reanalysis (i.e. all pregnancies with complete information on HIV status and potential confounders); 6210 (39%) HIV-infected and 9712 HIV-uninfected pregnancies were identified. To estimate the risk ratio SACE of HIV on SGA, first we ran a prediction model for the probability of a live birth,

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

$\pi(A,C) = Pr(S=1|A,C)$, where C includes the following correlates of A, S and Y: presence of past adverse pregnancy outcome(s), maternal age (15-20, 21-34, 35-50 years), educational status (none/primary vs. secondary/tertiary), marital status (single/widowed/divorced vs. married) and occupation (employed vs. unemployed). Table 1 shows results for the first stage regression fit. We then used the simple modification of the logistic regression given in Result 2 (equation (9)), adjusting for the same covariates used in the first stage model. Our results indicate that survival bias may be operating in this study and suggest that the association between maternal HIV and SGA using standard logistic regression are likely conservative when compared to the SACE estimate. In fact, the SACE point estimate is somewhat larger than the standard risk ratio estimate (SACE risk ratio standard risk ratio=2.1, 95%CI=[1.7,2.6] compared to standard risk ratio=1.7, 95%CI=[1.6,1.9]), indicating that, if not appropriately accounted for, survival bias may attenuate the estimated effects of HIV on SGA. Although the standard estimate is clearly smaller than the SACE estimate, their confidence intervals overlap, mainly due to the fact that the SACE estimator was considerably more variable. This was further reflected in the formal test of the null hypothesis of no survival bias, i.e. $\beta_{ac} = 0$ in equation (9) which failed to reject at the 0.05 level (p-value=0.07), but is nonetheless somewhat suggestive of the presence of selection bias.

Discussion

In this paper, we have considered an approach for inference about the effect of an exposure for individuals who would be alive at follow-up irrespective of their exposure value.[1,2] SACE is an instance of what has become known in the literature as a principal strata causal effect,[2] and such effects are generally not identified without certain assumptions. A principal stratum is formally defined by conditioning on a collection of counterfactual outcomes under possibly conflicting exposure values, for example, SACE corresponds to the effect for persons in the stratum $\{S(a=0) = S(a=1) = 1\}$;

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

however there also are other strata to consider, i.e. $\{S(a=0)S(a=1)\neq 1\}$. For identification, a strategy which figures prominently in existing literature on SACE is monotonicity, which essentially states that there is no person in the population for whom the exposure is protective, i.e. no person exist with $\{S(a=0)=0, S(a=1)=1\}$. This is a strong assumption, and although it can sometimes be falsified empirically, it can never be established with certainty. Even when appropriate, monotonicity alone does not suffice for identification and a variety of additional assumptions have appeared in the literature, that permit identification of SACE under monotonicity. A common strategy essentially amounts to a version of ignorability, either conditional on pre-exposure risk factors,[9] or conditional on both pre- and post-exposure risk factors.[4,10] An alternative strategy that is sometimes adopted entails performing a sensitivity analysis,[11–16] or bounds can sometimes be obtained.[17,18] Zhang et al replace monotonicity with strong distributional assumptions, that the outcome is normally distributed within principal strata, a strategy which is of little use for binary outcome.[18] Another approach is given by Ding et al who assume a form of exclusion restriction for an observed pre-exposure covariate to obtain nonparametric identification of SACE.[19] Specifically, Ding et al assume that one has observed a pre-exposure correlate of survival, which is independent of the observed outcome conditional on principal strata.[19] Interestingly, this assumption can be stated as follows, using the current formulation with C as the pre-exposure correlate of survival:

$$E(Y(a)|a, S(1-a)=s, S(a)=1, C=c) = E(Y(a)|a, S(1-a)=s, S(a)=1);\ a=0,1. \quad (12)$$

We note that this assumption essentially requires that C is itself a causal effect of the principal strata, which is also known a priori not to directly influence the outcome. However, in practice, one will generally expect pre-exposure correlates of survival to either be direct causes of survival or the effect of an unobserved cause of survival. Consequently, (12) is unlikely to hold in practice

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

for most correlates of survival and therefore the methods described in their paper based on this assumption are of little substantive interest. We should also note that Ding et al offer an alternative approach in which the exclusion restriction (12) can be relaxed, provided that C takes on three or more values, and additional parametric assumptions can be made; however implementation of the approach is not straightforward.[19] The approach proposed in the current paper complements that of Ding et al[19] by virtue of relaxing assumption (12) while allowing for binary or more general C without imposing neither monotonicity, nor their exclusion restriction. The proposed regression based approach is particularly advantageous in its ease of implementation in standard software, and provides a simple analytic framework for investigators to assess the extent to which survival bias may be operating in a given analysis. We used the approach to demonstrate the impact survival bias due to stillbirth may have in a perinatal epidemiology application using an HIV setting. We found that the effect of maternal HIV status on SGA was somewhat larger for the live births whose survival status was not affected by maternal HIV infection (SACE), when compared to the standard effect estimate. Finally, we note that in principle, the formulation used herein technically accommodates discrete or continuous exposures, although the SACE interpretation may be ambiguous for such exposures.[1]

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Appendix

Proof of Result 1: We observe that

$$
\begin{aligned}
E\left(Y|a,C,S=1\right) &= E\left(Y|a,C,S\left(a\right)=1\right)\ \text{(consistency)}\\
&= E\left\{E(Y|a,C,U,S\left(a\right)=1)|a,C,S\left(a\right)=1\right\}\\
&= \beta_1 a + E\left(U|a,C,S\left(a\right)=1\right) + b\left(C\right)\text{(assumption (4))}\\
&= \beta_1 a + \left[E\left(U|a,C,S\left(a\right)=1\right) - E\left(U|a,C,S\left(a\right)=0\right)\right]\Pr\left\{S\left(a\right)=0|a,C\right\}\\
&+ E\left(U|a,C\right) + b\left(C\right)\\
&= \beta_1 a + \left[E\left(U|a,C,S\left(a\right)=1\right) - E\left(U|a,C,S\left(a\right)=0\right)\right]\Pr\left\{S\left(a\right)=0|a,C\right\}\\
&+ E\left(U|C\right) + b\left(C\right)
\end{aligned}
$$

where we use the fact that

$$E\left(U|a,C\right) = E\left(U|a,C,S\left(a\right)=1\right)\Pr\left\{S\left(a\right)=1|a,C\right\} + E\left(U|a,C,S\left(a\right)=0\right)\Pr\left\{S\left(a\right)=0|a,C\right\}$$

implies that

$$E\left(U|a,C,S\left(a\right)=1\right) = \left[E\left(U|a,C,S\left(a\right)=1\right) - E\left(U|a,C,S\left(a\right)=0\right)\right]\Pr\left\{S\left(a\right)=0|a,C\right\} + E\left(U|C\right)$$

and

$$E\left(U|a,C\right) = E\left(U|C\right)$$

by assumption.

Next, we proceed as in Theorem 1 of Tchetgen Tchetgen and Wirth (2013), which gives under

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

assumption (5)

$$
\begin{aligned}
E\left(U|a,C,S=1\right) &= \frac{E\left(U\exp\left(\alpha U\right)|a,C,S=0\right)}{E\left(\exp\left(\alpha U\right)|a,C,S=0\right)} \\
&= \frac{\partial \log E\left(\exp\left(\alpha U\right)|a,C,S=0\right)}{\partial\alpha} \\
&= \frac{\partial \log\left[\exp(E\left(U|a,C,S=0\right))E\left(\exp\left(\alpha\Delta\right)|a,C,S=0\right)\right]}{\partial\alpha} \text{ no unobserved confounding of } A \\
&= E\left(U|a,C,S=0\right) + \frac{\partial \log\left[E\left(\exp\left(\alpha\Delta\right)|a,C,S=0\right)\right]}{\partial\alpha} \\
&= E\left(U|a,C,S=0\right) + \frac{\partial \log\left[E\left(\exp\left(\alpha\Delta\right)\right)\right]}{\partial\alpha}
\end{aligned}
$$

therefore

$$
\begin{aligned}
&E\left(U|a,C,S=1\right) - E\left(U|a,C,S=0\right) \\
&= \frac{\partial \log\left[E\left(\exp\left(\alpha\Delta\right)\right)\right]}{\partial\alpha} \\
&\equiv \beta_{ac}
\end{aligned}
$$

We may therefore conclude that

$$
\begin{aligned}
E\left(Y|a,C,S=1\right) &= \beta_0 + \beta_1 a + \beta_{ac} \\
&= \beta_0 + \beta_1 a + \beta_{ac}\Pr\left\{S=0|a,C\right\} \\
&+ b^*\left(C\right)
\end{aligned}
$$

where $b^*\left(C\right) = E\left(U|C\right) + b\left(C\right)$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Proof of Result 2: We observe that

$$
\begin{aligned}
E\left(Y|a,C,S=1\right) &= E\left(Y|a,C,S\left(a\right)=1\right) \ \ \text{(consistency)} \\
&= E\left\{E(Y|a,C,U,S\left(a\right)=1)|a,C,S\left(a\right)=1\right\} \\
&= E\left(\exp\left(U\right)|a,C,S\left(a\right)=1\right)\exp(\beta_1 a+b_c\left(C\right)) \ \text{(assumption \ (4))} \\
&= \exp(\beta_1 a+b_c\left(C\right)+E\left(U|C,S\left(a\right)=1\right))E\left(\exp\left(\Delta\right)|a,C,S=1\right) \\
&= \exp(\beta_1 a+b_c\left(C\right)+\beta_{ac}\Pr\left\{S=0|a,C\right\} \\
&+ E\left(U|C\right))E\left(\exp\left(\Delta\right)|S=1\right) \\
&= \exp(\beta_1 a+b_c^*\left(C\right)+\beta_{ac}\Pr\left\{S=0|a,C\right\})
\end{aligned}
$$

where

$$
b_c^*\left(C\right)=b_c\left(C\right)+\log E\left(\exp\left(\Delta\right)|S=1\right)+E\left(U|C\right)
$$

References

[1] Robins JM.(1986) A new approach to causal inference in mortality studies with sustained exposure period—application to control of the healthy worker survivor effect. Math Model.7:1393–1512.

[2] Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics. 2002;58(1):21–29.

[3] Pearl, J. Causality: Models, Reasoning, and Inference, 2nd ed. New York: Cambridge University Press; 2009.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

[4] Tchetgen Tchetgen EJ, Glymour M, Weuve J, Shpitser I.(2012a) To weight or not to weight? On the relation between inverse-probability weighting and principal stratification for truncation by death. Epidemiology;23(4):644-6.

[5] Chen JL, Ribaudo HJ, Souda S, Parekh N, Ogwu A, Lockman S, Powis K, Dryden-Peterson S, Creek T, Jimbo W. et al. Highly Active Antiretroviral Therapy and Adverse Birth Outcomes Among HIV-Infected Women in Botswana. The Journal of Infectious Diseases 2012;206:1695–705.

[6] Wacholder S.(1986) Binomial regression in GLIM: estimating risk ratios and risk differences. Am J Epidemiol;123:174 –184.

[7] Zou GY.(2004) A modified Poisson regression approach to prospective studies with binary data. Am J Epidemiol.;159:702–706.

[8] Tchetgen Tchetgen E.J. (2013) Estimation of risk ratios in cohort studies with a common outcome: a simple and efficient two-stage approach.Int J Biostat. 2013 May 7;9(2):251-64. doi: 10.1515/ijb-2013-0007.

[9] Hayden D, Pauler DK, Schoenfeld D.(2005) An estimator for treatment comparisons among survivors in randomized trials. Biometrics. 61(1):305–310.

[10] Tchetgen Tchetgen EJ. Identification and estimation of survivor average causal effects. 2014. Featured Article. Statistics in Medicine. In Press.

[11] Hudgens MG, Halloran ME. Causal vaccine effects on binary postinfection outcomes. Journal of the American Statistical Association. 2006;101(473):51–64.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

[12] Shepherd BE, Gilbert PB, Jemiai Y, Rotnitzky A (2006) Sensitivity analyses comparing outcomes only existing in a subset selected postrandomization, conditional on covariates, with application to HIV vaccine trials. Biometrics. 62(2):332–342.

[13] Jemiai Y, Rotnitzky A, Shepherd BE, Gilbert PB. (2007) Semiparametric estimation of treatment effects given base-line covariates on an outcome measured after a post-randomization event occurs.J R Stat Soc Series B Stat Methodol. 1;69(5):879-901.

[14] Gilbert PB, Bosch RJ, Hudgens MG.(2003) Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. Biometrics. 59(3):531–540.Shepherd BE, Gilbert PB, Lumley T. (2007) Sensitivity analyses comparing time-to-event outcomes existing only in a subset selected postrandomization. J Am Stat Assoc.102(478):573–582.

[15] Chiba, Y. and VanderWeele, T. J. (2011). A simple methd for principal strata effects when the outcome has been truncated due to death. American Journal of Epidemiology, 173:745-751.

[16] Zhang JL, Rubin DB.(2003) Estimation of causal effects via principal stratification when some outcomes are truncated by "death." J Educ Behav Stat.28(4):353–368.

[17] Imai, K. (2008). Sharp bounds on the causal effects in randomized experiments with 'truncation-by-death'. Statistics and Probability Letters, 78:144-149.

[18] Ding P, Geng Z, Yan W and Zhou X (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. J Am Stat Assoc. 106 (496):1578-1591.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

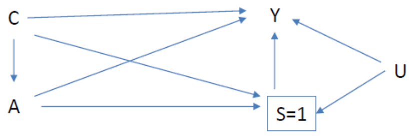

Figure. 1 Causal diagram depicting pre-exposure covariates C, exposure A (maternal HIV), birth outcome Y (SGA) for a survivor with S=1 (livebirth) and with U encoding unobserved common causes of Y and S

Table 1: Prediction model for live births

Maternal characteristics	Adjusted RR	95% CI
HIV-infected	0.6	0.5, 0.7
Age		
15 - 20	ref	ref
21 - 34	0.7	0.4, 1.2
35 - 50	0.5	0.3, 0.8
Education		
none/primary	ref	ref
secondary/tertiary	1.1	0.8, 1.3
Marital status	1.0	0.8, 1.2
Occupation	0.9	0.7, 1.0

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive