Separating Algorithms From Questions and Causal Inference With Unmeasured Exposures: An Application to Birth Cohort Studies of Early Body Mass Index Rebound

Izzuddin M. Aris, Aaron L. Sarvet, Mats J. Stensrud, Romain Neugebauer, Ling-Jun Li, Marie-France Hivert, Emily Oken, and Jessica G. Young

Correspondence to:
Dr. Izzuddin M. Aris, Division of Chronic Disease Research Across the Lifecourse, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 401 Park Drive, Suite 401E, Boston, MA 02215 (email: Izzuddin_Aris@harvardpilgrim.org).

Author Affiliations
Division of Chronic Disease Research Across the Lifecourse, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts, USA (Izzuddin M Aris, Marie-France Hivert, Emily Oken, and Jessica G Young); Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA (Aaron L. Sarvet and Jessica G Young); Swiss Federal Institute of Technology Lausanne, Lausanne, Switzerland (Mats J. Stensrud); Kaiser Permanente Northern California Division of Research, Oakland, California, USA (Romain Neugebauer); Department of Obstetrics and Gynecology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore, Singapore (Ling-Jun Li); Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts, USA (Marie-France Hivert); Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA (Emily Oken)

Running head: Separating Algorithms from Questions.

Table of Contents

WEB APPENDIX 1

Identification of outcome distributions under proxy representative interventions when baseline covariates, time-varying exposures and the outcome are measured

Here we review assumptions under which the function (1) of factual random variables $(L \equiv L_0, \overline{A}_K, R, Y)$ presented in the main text, for a choice of exposure at each time $k = 0, \ldots, K$ identifies the effect of proxy representative interventions in the study population, that is the difference in the outcome mean had, on each day k, exposure been assigned as a random draw from an intervention density $f^{int}(A_k|\overline{A}_{k-1}, L_0)$ selected as $f(A_k|R = r, \overline{A}_{k-1}, L_0)$ for $r = 1$ versus $r = 0$ with $f(A_k|R = r, \overline{A}_{k-1}, L_0)$ the distribution of A_k in the observational study among those with exposure history $\overline{A}_{k-1}$, baseline covariates L_0 and early BMI rebound status $R = r$. We allow that A_k is a vector of several modifiable exposures at each k (see main text).

To understand this assignment mechanism concretely, we consider a very simplified example. Suppose the population has only one level of L_0 (e.g. $L_0 = 1$), and suppose a short intervention period with time-fixed exposure A_0 taking 3 levels ($A_0 \in \{1, 2, 3\}$). Further, suppose the factual distribution (i.e. without intervention) of A_0 in this study population among those with proxy level $R = 1$ is $f(a_0|L_0 = 1, R = 1) = P(A_0 = a_0|L = 1, R = 1) = 0.3$, 0.5, and 0.2 for $a_0 = 1$, $a_0 = 2$, and $a_0 = 3$ respectively. By contrast, suppose the factual distribution of A_0 in this study population among those with proxy level $R = 0$ is $f(a_0|L_0 = 1, R = 0) = P(A_0 = a_0|L = 1, R = 0) = 0.7$, .2, and 0.1 for $a_0 = 1$, $a_0 = 2$, and $a_0 = 3$, respectively. Now suppose we enroll eligible individuals from this study population in our target trial. To assign the value of A_0 (either 1, 2 or 3) to an eligible individual randomized to the "intervention 1" arm, we would draw a random number from a multinomial distribution with probabilities 0.3, 0.5, and 0.2 for $a_0 = 1$,

$a_0 = 2$, and $a_0 = 3$ respectively (or roll a 3-sided die with these weights). Analogously, to assign the value of A_0 (either 1, 2 or 3) to an eligible individual randomized to the "intervention 2" arm, we would draw a random number from a multinomial distribution with probabilities 0.7, .2, and 0.1 for $a_0 = 1$, $a_0 = 2$, and $a_0 = 3$ respectively (or roll a 3-sided die with these weights).

Let $\mathcal{G}$ be the *set of all possible deterministic interventions* g on A_k that may, at most, depend on $\overline{A}_{k-1}, L_0$. Under a deterministic intervention, every individual in the study population with the same values of $\overline{A}_{k-1}, L_0$ will be assigned the same values of exposure A_k. Denote these values a^g_k which may be a function of past exposure under g and baseline covariates.

For each $g \in \mathcal{G}$, let Y^g be the outcome had, possibly contrary to fact, an individual in this study population been assigned A_k at each $k = 0, \ldots, K$ according to the intervention g and consider the following assumptions for each g:

(1) Consistency: If $\overline{A}_K = \overline{a}^g_K$ then $Y = Y^g$.

(2) Exchangeability: For $k = 0, \ldots, K$

$$Y^g \coprod A_k | \overline{A}_{k-1} = \overline{a}^g_{k-1}, L_0$$

(3) Positivity: For $k = 0, \ldots, K$

$$f_{\overline{A}_{k-1}, L_0}(\overline{a}^g_{k-1}, l_0) \neq 0 \implies$$
$$f_{A_k|, \overline{A}_{k-1}, L_0}(a^g_k | \overline{a}^g_{k-1}, l_0) > 0.$$

Then let $\mathcal{G}' \subseteq \mathcal{G}$ be the *subset of all possible deterministic interventions* g for which the following condition would hold for all k:

$$f_{\overline{A}_{k-1},L_0}(\overline{a}^g_{k-1}, l_0) \neq 0 \implies$$
$$f^{int}(a^g_k|\overline{a}^g_{k-1}, l_0) > 0.$$

That is, the subset under which positivity holds were we to replace the factual treatment distribution at k conditional on past exposure and baseline covariates with the corresponding intervention distribution in the positivity condition above.

Following [1], when these three assumptions hold for all $g \in \mathcal{G}'$, then the outcome mean under this intervention rule is equal to

$$\sum_{\overline{a}_K}\sum_{l_0} \mathrm{E}[Y|\overline{A}_K = \overline{a}_K, L_0 = l_0]\prod_{j=0}^{K} f^{int}(a_j|\overline{a}_{j-1}, l_0)f(l_0) =$$
$$\sum_{\overline{a}_K}\sum_{l_0} \mathrm{E}[Y|\overline{A}_K = \overline{a}_K, L_0 = l_0]\prod_{j=0}^{K} f(a_j|R = r, \overline{a}_{j-1}, l_0)f(l_0) =$$
$$\sum_{\overline{a}_K}\sum_{l_0} \mathrm{E}[Y|\overline{A}_K = \overline{a}_K, L_0 = l_0]f(\overline{a}_k|R = r, l_0)f(l_0), \tag{1}$$

with $\overline{A}_{-1} \equiv 0$ by convention. Importantly, these assumptions are completely agnostic as to whether this intervention on A_k at all $k = 0, \ldots, K$ has any effect on early BMI rebound status by the end of the intervention interval R in some or all individuals in the population. As we can see, R plays no role in these assumptions other than the choice of the intervention distribution, here dependent on features of the factual exposure distribution in Project Viva, the observational study. Even had our data come, rather than from an observational study like Project Viva, but instead from a study where treatment was actually assigned according to this intervention, we could not guarantee that this would result in all individuals having $R = r$ ($r = 1$ or $r = 0$) nor is that the goal. Therefore, the causal effects we consider are not generally interpretable as effects of "early BMI rebound" because we only define A_k, $k = 0, \ldots, K$ and not R, as "treatment" ("exposure") and such effects are not the goal of the analysis. Of course these effects are dependent on the particular choice of A_k which we must articulate. In later sections in the appendix, we consider weaker versions of the conditions above that allow exchangeability to depend on other time-varying covariates besides past exposure. As discussed in the main text, in addition to exposure, time-varying covariate changes were not available within the exposure period of interest in our application given the interval nature of measurement.

Identification when only (L, R, Y) are measured

As discussed in the main text, estimation of the function (1) requires measurement of exposure $\overline{A}_K$. We now show that, when exposure is not measured but (L, R, Y) are measured, (1) can be estimated but under the additional assumption of proxy separation given in the main text.

By the main text, a^r_k is any value in the support of A_k under an intervention that assigns A_k according to a proxy representative intervention characterized by $f^{int}(A_j|\overline{A}_{j-1}, L_0) = f(A_j|R = r, \overline{A}_{j-1}, L_0)$. By this, the g-formula [1] in expression (1) is equivalent to

$$\sum_{\overline{a}^r_K}\sum_{l_0} \mathrm{E}[Y|\overline{A}_K = \overline{a}^r_K, L_0 = l_0] \prod_{j=0}^{K} f(a^r_j|R = r, \overline{a}^r_{j-1}, l_0) f(l_0)$$

Under proxy separation and probability laws we can further rewrite (1) as

$$\sum_{l_0}\sum_{\overline{a}^r_K} \mathrm{E}[Y|\overline{A}_K = \overline{a}^r_K, R = r, L_0 = l_0] f(\overline{a}^r_K|R = r, l_0) f(l_0) =$$

$$\sum_{l_0}\sum_{\overline{a}^r_K}\sum_{y} y f(y|\overline{a}^r_K, r, l_0) f(\overline{a}^r_K|r, l_0) f(l_0) =$$

$$\sum_{l_0}\sum_{\overline{a}^r_K}\sum_{y} y f(y|\overline{a}^r_K, r, l_0) f(\overline{a}^r_K|r, l_0) f(l_0) =$$

$$\sum_{l_0}\sum_{\overline{a}^r_K}\sum_{y} y f(y|\overline{a}^r_K, r, l_0) f(\overline{a}^r_K|r, l_0) f(l_0) =$$

$$\sum_{l_0}\sum_{\overline{a}^r_K}\sum_{y} y f(y, \overline{a}^r_K|r, l_0) f(l_0) =$$

$$\sum_{l_0}\sum_{y} y f(y|r, l_0) f(l_0) =$$

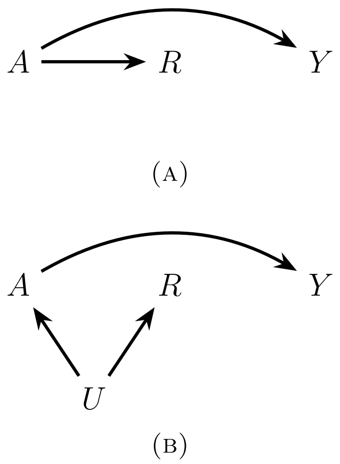
$$\sum_{l_0} \mathrm{E}[Y|R = r, L_0 = l] f(l_0) \tag{2}$$

Relation to previous work

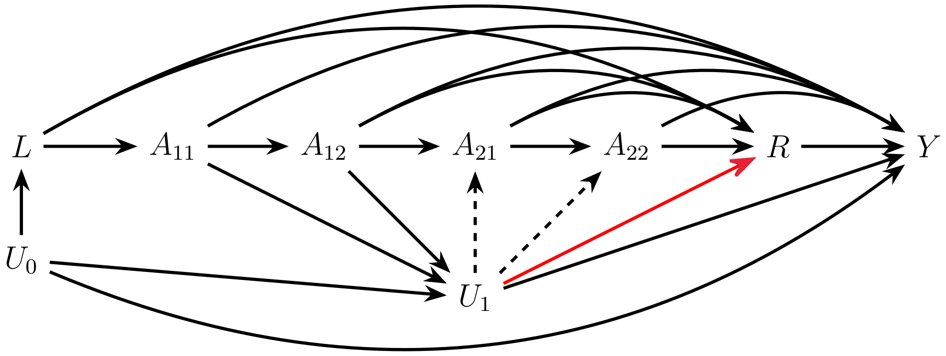
Our definition of a proxy representative intervention as an intervention that assigns $\overline{A}_K$ according to $f(\overline{A}_K|R = r, L_0)$ is related to previous definitions of a so-called representative intervention, where the proxy is restricted to a particular coarsening of exposure. For example for $K = 0$, and exposure A_0 number of minutes of exercise per day and R an indicator of exercising at least 30 minutes on that day [2, 3, 4, 5, 6]. In this case, the proxy separation assumption holds by definition. Therefore, only exchangeability, positivity and consistency with respect to A_k are required to identify effects of corresponding representative interventions when exposure is unmeasured but the proxy is measured (this is importantly *not* the case when exposure is time-varying and either the proxy is time-varying or under weaker versions of exchangeability that require measures of baseline and time-varying covariates as shown in [6] and as we discuss in later sections). Further, in a trial that actually assigned exposure according to a representative intervention, it is guaranteed that this intervention will control the value of R. By contrast, this is not the case for a proxy representative intervention where any relation between exposure and the proxy is an assumption made by the investigator whether in the context of a trial or observational study. [5] also noted this case but limited explicit consideration to problems where the proxy is still viewed as a "treatment". Here we do not view the proxy as a treatment but rather a measured variable that might be leveraged for identification of a treatment effect when that treatment is difficult to measure.

Importantly, our definition of proxy separation requires no particular assumption on the underlying causal structure between exposure and the proxy, including their temporal order. For example, the simplified causal directed acyclic graphs (DAGs) in Web Figure 1 depicts a trivialized data generating assumption where, for added simplicity, L_0 is taken as constant (and therefore not depicted), a short intervention interval ($K = 0$) is considered with $A_0 \equiv A$ and univariate and U is unmeasured. In Web Figure 1A, A causes R while in Web Figure 1B A and R are only associated through U. However, both graphs in Web Figure 1 are consistent with the assumptions of proxy separation and exchangeability as defined above, under the assumption that these graphs represent underlying nonparametric structural equation models [7, 1, 8]. Notably, both graphs also make the assumption that R in fact has no effect on the outcome but is only associated with the outcome through A, even conditional on L_0. This may be a reasonable assumption in many settings where the proxy is a derived variable like BMI rebound, obesity status and other variables that have been controversial with respect to the consistency assumption. In Web Figure 1A exchangeability is unaffected by adding an arrow from R to Y, however, in Web Figure 1B, this arrow would violate exchangeability.

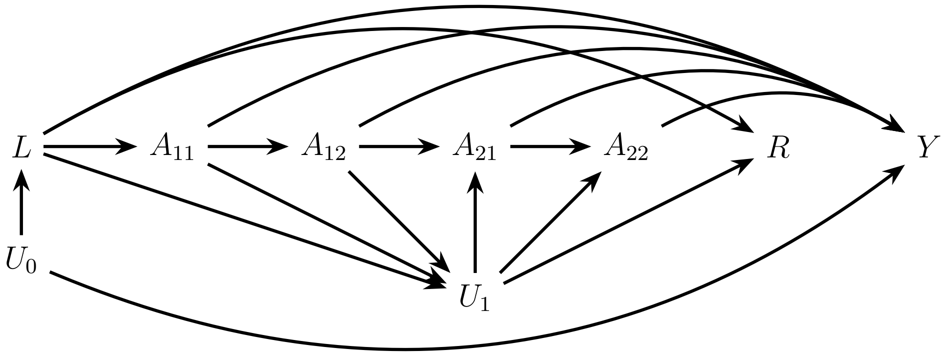
Web Figures 2 and 3 depict more realistic settings allowing a higher dimensional $A_k = (A_{1k}, A_{2k})$ and $K = 1$. Web Figure 2 is the same as Web Figure 1 in the main text. Web Figure 3 depicts an alternative scenario, similar to Web Figure 1B, where R is not a cause of Y and A_{21} and A_{22} are only associated with R through a common cause (here U_1).



Web Figure 1. Simplified assumptions on the data generating mechanism for constant L_0. Both figures are consistent with the proxy separation assumption and exchangeability as defined in this appendix when exposure is time-fixed ($K = 0$).



Web Figure 2. A data generating scenario allowing multivariate time-varying exposure. With all solid arrows present, exchangeability requires dashed arrows absent. With all solid black arrows present, proxy separation may fail by the presence of the red arrow.



Web Figure 3. An alternative data generating scenario that is not reasonable for our application but is consistent with both the exchangeability and proxy separation assumptions. In this case R, the proxy for treatment, is not caused by treatments at time 2 (they are only associated through the common cause U_1) and also R is not a cause of the outcome. An investigator drawing this graph would clearly have no interest in R as a treatment but could leverage this treatment proxy to identify an effect on Y of a proxy representative intervention on exposures unmeasured in the data set.

Measured time-varying exposure proxies and confounders

Consider a modified proxy representative intervention such that exposure is assigned at each $k = 0, \ldots, K$ as a random draw from an intervention density $f^{int}(A_k|\overline{L}_k, \overline{A}_{k-1})$ selected as $f(A_k|R_k = r_k, \overline{L}_k, \overline{A}_{k-1})$ where R_k is an exposure proxy in interval k and L_k other covariates in that interval. This differs from the proxy representative interventions considered above and in the main text in that the proxy changes by interval k (e.g. as opposed to BMI rebound by the end of the intervention interval, define the proxy R_k as an indicator of obesity in interval k) and also assignment depends on time-varying covariate history (as opposed to only baseline covariates as above and in the main text).

Analogously consider the following weaker exchangeability condition, and corresponding positivity and consistency conditions, for each $g \in \mathcal{G}$ where $\mathcal{G}$ is now the *set of all possible deterministic interventions* g on A_k that may, at most, depend on $\overline{L}_k, \overline{A}_{k-1}$: For all $k = 0, \ldots, K$:

(1) Consistency: If $\overline{A}_k = \overline{a}^g_k$ then $\overline{L}_{k+1} = \overline{L}^g_{k+1}$ with $L_{K+1} \equiv Y$ and $L^g_{K+1} \equiv Y^g$

(2) Exchangeability: For all $k = 0, \ldots, K$

$$Y^g \coprod A_k|\overline{L}_k, \overline{A}_{k-1} = \overline{a}^g_{k-1}$$

(3) Positivity:

$$f_{\overline{L}_k, \overline{A}_{k-1}}(\bar{l}_k, \overline{a}^g_{k-1}) \neq 0 \implies$$

$$f_{A_k|\overline{L}_k, \overline{A}_{k-1}}(a^g_k|\bar{l}_k, \overline{a}^g_{k-1})$$

Again, following [1], when these three assumptions hold for the subset of all $g \in \mathcal{G}$ such that positivity holds were we to replace the factual exposure distribution at k conditional on $\overline{A}_{k-1}, \overline{L}_k$, with that under the chosen intervention rule $f^{int}(A_k|\overline{A}_{k-1}, \overline{L}_k) = f(A_k|R_k =$

$r_k, \overline{A}_{k-1}, \overline{L}_k)$, then the outcome mean under this intervention rule is equal to

$$(3) \qquad \sum_{\overline{a}_K} \sum_{\overline{l}_k} \mathrm{E}[Y|\overline{A}_K = \overline{a}_K, \overline{L}_k = \overline{l}_k] \prod_{j=0}^{K} f^{int}(a_j|\overline{a}_{j-1}, \overline{l}_k) f(l_k|\overline{a}_{k-1}, \overline{l}_{k-1}),$$

selecting $f^{int}(A_j|\overline{A}_{j-1}, \overline{L}_j) = f(A_j|R = r, \overline{A}_{j-1}, \overline{L}_j)$ and $\overline{A}_{-1} \equiv \overline{L}_{-1} \equiv 0$ by convention.

[6] considered these modified proxy representative interventions (which coincide with those of the main text for time-fixed exposure, $K = 0$), when each R_k is selected as an actual coarsening of A_k (e.g. for A_k minutes of exercise on day k, choose R_k an indicator of exercising 30 minutes in that interval). In this case, an analogously modified proxy separation assumption is guaranteed to hold, where $a_k^{\overline{r}}$ is any value in the support of A_k under this modified proxy representative intervention:

Modified proxy separation: Within each possibly observed covariate history $\overline{L}_k$ and each exposure history consistent with intervention $\overline{A}_{k-1} = \overline{a}_{k-1}^{\overline{r}}$, if $A_k = a_k^{\overline{r}}$ then $R_k = r_k$, $k = 0, \ldots, K$.

[6] showed that, given this modified proxy separation assumption, the g-formula (3) indexed by this modified proxy representative intervention, which is a stochastic intervention on a time-varying exposure, can be estimated with an inverse probability weighted algorithm that is identical to what we would implement had we considered the proxy at each time k as the actual exposure (even though conceptually it is not, for proxy representative interventions). However, in this more general time-varying scenario, the exposure history $\overline{A}_{k-1}$ will "act" as past time-varying confounders in the algorithm (e.g. they are generally required in the weight denominator). Therefore, modified proxy separation is not sufficient to avoid the need to measure exposure in this general time-varying setting. We refer the reader to [6] for details and proofs with related results discussed in [4].

To see this in a simple setting, suppose constant L_0 and two time points ($K = 1$): By $a_k^{\overline{r}}$ any value in the support of A_k under this modified proxy representative intervention,

(3) can be written as

$$\sum_{a_1^{\overline{r}}, a_0^{\overline{r}}, l_1} \mathrm{E}[Y|\overline{A}_1 = \overline{a}_1^{\overline{r}}, L_1 = l_1] f(a_1^{\overline{r}}|R_1 = r_1, l_1, a_0^{\overline{r}}) f(l_1|a_0^{\overline{r}}) f(a_0^{\overline{r}}|R_0 = r_0)$$

Under modified proxy separation and probability laws we can further rewrite (3) as

$$\sum_{a_1^{\overline{r}}, a_0^{\overline{r}}, l_1} \mathrm{E}[Y|a_1^{\overline{r}}, R_1 = r_1, l_1, a_0^{\overline{r}}] f(a_1^{\overline{r}}|R_1 = r_1, l_1, a_0^{\overline{r}}) f(l_1|a_0^{\overline{r}}) f(a_0^{\overline{r}}|R_0 = r_0) =$$

$$\sum_{a_1^{\overline{r}}, a_0^{\overline{r}}, l_1} \sum_{y} f(y|a_1^{\overline{r}}, R_1 = r_1, l_1, a_0^{\overline{r}}) f(a_1^{\overline{r}}|R_1 = r_1, l_1, a_0^{\overline{r}}) f(l_1|a_0^{\overline{r}}) f(a_0^{\overline{r}}|R_0 = r_0) =$$

$$\sum_{a_0^{\overline{r}}, l_1} \sum_{y} f(y|R_1 = r_1, l_1, a_0^{\overline{r}}) f(l_1|a_0^{\overline{r}}) f(a_0^{\overline{r}}|R_0 = r_0) \tag{4}$$

We can see that the expression still depends on A_0 by the last step. Additional restrictive assumptions would need to be made to remove A_0 from the expression, for example, the strong assumption that $f(y|R_1 = r_1, l_1, a_0^{\overline{r}}) = f(y|R_1 = r_1, l_1, r_0)$; that is, the outcome distribution does not depend on past exposure A_0 conditional on proxy and covariate history.

When distinctions between proxies and exposures are less clear: more transparent analysis of interval cohorts with causal inference goals

In our running example, by stating our question in terms of interventions on exercise and diet, it is clear that the variable we have put in our estimation algorithm in the role of exposure (early BMI rebound) is *not* the exposure and we must be relying on additional assumptions - beyond the “usual” exchangeability, consistency and positivity assumptions – to claim we have answered our question even without concerns of sampling variability. However, this may need to be made more transparent in other settings.

For example, in many cohort studies, even when time-updated measurements are available all changes in exposure are not possible to measure. A cohort study that contains a measure of minutes of exercise per day may have a single self-reported measure at the end of a long interval (possibly years). While we might understand this as a form of measurement error, we might alternatively view the single exposure measure in interval k as itself an exposure proxy, “acting” as exposure in the algorithm. Counterfactual outcomes and interventions however, may be specified more precisely in terms of intervention on exposure at a finer interval than the measurement interval. Reasoning about identification might then proceed as in this appendix and in the main text.

Web Appendix 2: Description of steps in TMLE, parametric g-computation and inverse probability weighted estimators.

For our primary analysis, we applied TMLE, an approach which has been described in detail elsewhere [9,10]. The function $\sum_l E[Y|R=r, L=l, \delta=0]\, f(l)$ can be estimated using the following implementation of this approach implemented in the R package *tmle*.[3] For r=1 and r=0:

1. Obtain an estimate $\bar{Q}_n^0\ (R=r, L)$ *of* $E(Y|R=r, L, \delta=0)$; that is, the outcome regression for each level of baseline covariates in the sample $L=L_i,\ i=1,\ldots,n$.

2. Obtain estimates $P_n(R=r|L)$ *and* $P_n(\delta=0|R=r,L)$ of the probability of having rebound status r for covariate level L and of being uncensored among those with status r covariate level L respectively, for each $L=L_i,\ i=1,\ldots,n$.

3. Fit a logistic regression model with dependent variable Y, independent variable $H_r = \frac{I(R=r)\ I(\delta=0)}{P_n(R=r|L)\ P_n(\delta=0|R=r,\ L)}$ based on the estimates in step 2 and offset term logit $\bar{Q}_n^0\ (R=r, L)$ from step 1 to obtain $\hat{\varepsilon_r}$, the estimated coefficient on H_r in this model.

4. Compute for each $L=L_i$, $logit\ \bar{Q}_n^*\ (R=r, L) = logit\ \bar{Q}_n^0\ (R=r, L) + \hat{\varepsilon}_r H_r$, with $\bar{Q}_n^0\ (R=r, L)$ from Step 1 and H_r and $\hat{\varepsilon_r}$, from step 3.

The final estimate of $\sum_l E[Y|R=r, L=l, \delta=0]\, f(l)$ is computed as a sum over i=1 to n of: $\bar{Q}_n^*\ (R=1, L_i)$ - $\bar{Q}_n^*\ (R=0, L_i)$ from step 4 plugging in r=1 and r=0, respectively. The use of a logistic link in Step 4 ensures estimates remain bounded within the parameter space [11].

A parametric g-computation estimator of $\sum_l E[Y|R=r, L=l, \delta=0]\, f(l)$ requires only step 1 of the TMLE algorithm above. This approach requires that $E(Y|R=r, L, \delta=0)$ is consistently estimated. An IPW estimator requires only step 2 of the TMLE algorithm above and can be computed as the estimated coefficient of a weighted linear regression with dependent variable Y, independent variable R and weights $w = \frac{R \times (1-\delta)}{P(R=1|L)\ P(\delta|R=1,L)} + \frac{(1-R)\times(1-\delta)}{P(R=0|L)\ P(\delta|R=0,L)}$. This approach requires that $P(R=r|L)$ and $P(\delta=0|R=r, L)$ are consistently estimated. Alternatively, stabilized weights with generally lower variability can be used that multiply w by estimates of $P(R=r)*P(\delta=0|R=r)$ for an individual with $R=r$. We implemented these methods using both standard parametric models and SuperLearner [12] for nuisance parameter estimation.

Web Appendix 3: Stata code for G-computation

```
***Y = observed outcome
***A = R, the proxy (change of notation from main text)
***L1,L2,L3 etc = list of covariates

**Perform regression with main effects
glm Y A L1 L2 L3 L4 L5, fam(gaussian)

gen aa=A
replace A = 1
predict double Y1, mu
replace A = aa
drop aa

gen aa=A
replace A = 0
predict double Y0, mu
drop aa

gen ATE = Y1 - Y0
sum ATE

**Generate 95% confidence intervals for 1000 bootstrapped samples of ATE
set matsize 1000
matrix t1 = J(1000, 1, .)
forvalues i=1/1000 {
bsample n
glm Y A L1 L2 L3 L4 L5, fam(gaussian)
gen aa=A
replace A = 1
predict double Y1, mu
replace A = 0
predict double Y0, mu
replace A = aa
drop aa
gen ATE = Y1 - Y0
sum ATE
matrix t1[`i', 1] = r(mean)
}
svmat t11
centile t11, centile(2.5 97.5)
```

Note: ***n*** in “bsample” represents number of complete observations in the sample

Web Appendix 4: Stata code for stabilized inversed probability weighting

```
***Y = observed outcome
***A = R, the proxy (change of notation from main text)
***L1,L2,L3 etc = list of covariates

**Perform regression to estimate P(A = 1)
logit A
predict pA, pr
gen pn_A = (pA*A) + ((1-pA)*(1-A))

**Perform regression to estimate P(A = 1| L=L1,L2,L3)

logit A L1 L2 L3 L4 L5
predict pA_L, pr
gen pd_A = (pA_L*A) + ((1-pA_L)*(1-A))

**Calculate stabilized weights
gen w_A = pn_A / pd_A

***Calculating censoring weights for missing outcomes (C)
**Generate missing outcome indicator
gen C=.
replace C=0 if Y==.
replace C=1 if Y!=.

**Perform regression to estimate probability of having an observed outcome P(C=1 | A)
logit C A
predict p_C_A, pr
gen pn_C = (p_C_A*C) + ((1-p_C_A)*(1-C))

**Perform regression to estimate P(C | A,L)
logit C A L1 L2 L3 L4 L5
predict p_C_A_L, pr
gen pd_C = (p_C_A_L*C) + ((1-p_C_A_L)*(1-C))

**Calculate stabilized weights for missing outcome
gen w_C = pn_C / pd_C

**Calculate stabilized weights for A and missing outcome
gen w = w_A * w_C

**Calculate weighted regression using only those with C=1
regress Y A [pw=w]

**Generate 95% confidence intervals for 1000 bootstrapped samples
set matsize 1000
matrix t2 = J(1000, 1, .)
forvalues i=1/1000 {
bsample n
logit A
predict pA, pr
gen pn_A = (pA*A) + ((1-pA)*(1-A))
logit A L1 L2 L3 L4 L5
predict pA_L, pr
gen pd_A = (pA_L*A) + ((1-pA_L)*(1-A))
gen w_A = pn_A / pd_A
gen C=.
replace C=0 if Y==.
replace C=1 if Y!=.
logit C A
predict p_C_A, pr
gen pn_C = (p_C_A*C) + ((1-p_C_A)*(1-C))
logit C A L1 L2 L3 L4 L5
predict p_C_A_L, pr
gen pd_C = (p_C_A_L*C) + ((1-p_C_A_L)*(1-C))
gen w_C = pn_C / pd_C
gen w = w_A * w_C
regress Y A [pw=w]
matrix t2[`i', 1] = _b[A]
}
svmat t21
centile t21, centile(2.5 97.5)
```

Web Appendix 5: R code for TMLE

```
##Y = observed outcome
##A = R, the proxy (change of notation from main text)
##L1,L2,L3 etc = list of covariates

##Load libraries
library(tmle)

## TMLE
data <- read.csv(file="insert working directory for data file here",header=TRUE)
tmleSL <- tmle(Y=data$Y, A=data$A, W=data[,2:13], Delta = data$C, Qform =
Y~A+L1+L2+L3+L4+L5, gform = A~L1+L2+L3+L4+L5, g.Deltaform = Delta ~
A+L1+L2+L3+L4+L5)
tmle.SL <- tmleSL[["psi"]][["A"]]
```

Note: bold and italicized code is dependent on the covariates that were used for adjustment in the model.

Web Appendix 6: R code for G-computation with SuperLearner

```
##Load libraries
library(SuperLearner)
library(gam)
library(biglasso)
library(bartMachine)
require(rJava)

##Specify algorithms for SuperLearner libraries
SL.library1 <- c("SL.glm", "SL.step","SL.glm.interaction")
SL.library2 <- c("SL.glm", "SL.step","SL.glm.interaction", "SL.bartMachine", "SL.biglasso",
"SL.gam")

### G-computation + SuperLearner (parametric algorithms only)
data <- subset(data,(!is.na(data$Y)))
newData <- rbind(cbind(data[,2:13], A=1), cbind(data[,2:13], A=0))
SL.fit <- SuperLearner(Y=data$Y, X=data[,2:14], SL.library=SL.library1, family="gaussian",
method="method.NNLS", newX=newData, verbose=TRUE)
data$Y1.pred <- SL.fit$SL.predict[1:564]
data$Y0.pred <- SL.fit$SL.predict[565:1128]
gcomp.ATE.SL <- data$Y1.pred - data$Y0.pred
gcomp.SL <- mean(gcomp.ATE.SL)

### G-computation + SuperLearner (parametric + non-parametric algorithms)
data <- subset(data,(!is.na(data$Y)))
newData <- rbind(cbind(data[,2:13], A=1), cbind(data[,2:13], A=0))
SL.fit <- SuperLearner(Y=data$Y, X=data[,2:14], SL.library=SL.library2, family="gaussian",
method="method.NNLS", newX=newData, verbose=TRUE)
data$Y1.pred <- SL.fit$SL.predict[1:564]
data$Y0.pred <- SL.fit$SL.predict[565:1128]
gcomp.ATE.SL <- data$Y1.pred - data$Y0.pred
gcomp.SL <- mean(gcomp.ATE.SL)
```

Note: bold and italicized code is dependent on total number of observations in the sample, and covariates used for adjustment in the model.

Web Appendix 7: R code for IPW with SuperLearner

```
##Load libraries
library(SuperLearner)
library(gam)
library(biglasso)
library(bartMachine)
require(rJava)

##Specify algorithms for SuperLearner libraries
SL.library1 <- c("SL.glm", "SL.step","SL.glm.interaction")
SL.library2 <- c("SL.glm", "SL.step","SL.glm.interaction", "SL.bartMachine", "SL.biglasso",
"SL.gam")

## IPW + SuperLearner (parametric algorithms only)
data <- read.csv(file="insert working directory for data file here",header=TRUE)
gA <- glm(A ~ 1 , family = binomial, data = data)
pA <- predict(gA, type = "response")
pA_num <- (pA*data$A) + ((1-pA)*(1-data$A))
SL.gA <- SuperLearner(Y=data[,14], X=data[,2:13], SL.library=SL.library1, family="binomial",
method="method.NNLS", verbose=TRUE)
p_SL.gA <- SL.gA$SL.predict
pA_denom <- (p_SL.gA*data$A) + ((1-p_SL.gA)*(1-data$A))
w_A <- pA_num / pA_denom
g_C <- glm(C ~ A , family = binomial, data = data)
p_C <- predict(g_C, type = "response")
p.C_num <- (p_C*data$C) + ((1-p_C)*(1-data$C))
SL.g_C <- SuperLearner(Y=data[,16], X=data[,2:14], SL.library=SL.library1, family="binomial",
method="method.NNLS", verbose=TRUE)
p_SL.g_C <- SL.g_C$SL.predict
p.C_denom <- (p_SL.g_C*data$C) + ((1-p_SL.g_C)*(1-data$C))
w_C <- p.C_num / p.C_denom
w <- w_A*w_C
ipw.ATE.SL <- glm(Y ~ A, family = gaussian, data=data, weights = w)

## IPW + SuperLearner (parametric + non-parametric algorithms)
data <- read.csv(file="insert working directory for data file here",header=TRUE)
gA <- glm(A ~ 1 , family = binomial, data = data)
pA <- predict(gA, type = "response")
pA_num <- (pA*data$A) + ((1-pA)*(1-data$A))
SL.gA <- SuperLearner(Y=data[,14], X=data[,2:13], SL.library=SL.library2, family="binomial",
method="method.NNLS", verbose=TRUE)
p_SL.gA <- SL.gA$SL.predict
pA_denom <- (p_SL.gA*data$A) + ((1-p_SL.gA)*(1-data$A))
w_A <- pA_num / pA_denom
g_C <- glm(C ~ A , family = binomial, data = data)
p_C <- predict(g_C, type = "response")
p.C_num <- (p_C*data$C) + ((1-p_C)*(1-data$C))
SL.g_delta <- SuperLearner(Y=data[,16], X=data[,2:14], SL.library=SL.library2,
family="binomial", method="method.NNLS", verbose=TRUE)
p_SL.g_C <- SL.g_C$SL.predict
p.C_denom <- (p_SL.g_C*data$C) + ((1-p_SL.g_C)*(1-data$C))
w_C <- p.C_num / p.C_denom
w <- w_A*w_C
ipw.ATE.SL <- glm(Y ~ A, family = gaussian, data=data, weights = w)
```

Web Appendix 8: R code for TMLE with SuperLearner

```
##Load libraries
library(tmle)
library(SuperLearner)
library(gam)
library(biglasso)
library(bartMachine)
require(rJava)

##Specify algorithms for SuperLearner libraries
SL.library1 <- c("SL.glm", "SL.step","SL.glm.interaction")
SL.library2 <- c("SL.glm", "SL.step","SL.glm.interaction", "SL.bartMachine", "SL.biglasso",
"SL.gam")

## TMLE + SuperLearner (parametric algorithms only)
data <- read.csv(file="insert working directory for data file here",header=TRUE)
tmleSL <- tmle(Y=data$Y, A=data$A, W=data[,2:13], Delta = data$C, Q.SL.library = SL.library1,
g.SL.library = SL.library1)
tmle.SL <- tmleSL[["psi"]][["A"]]

## TMLE + SuperLearner (parametric + non-parametric algorithms)
data <- read.csv(file="insert working directory for data file here",header=TRUE)
tmleSL <- tmle(Y=data$Y, A=data$A, W=data[,2:13], Delta = data$C, Q.SL.library = SL.library2,
g.SL.library = SL.library2)
tmle.SL <- tmleSL[["psi"]][["A"]]
```

Note: bold and italicized code is dependent on the covariates that were used for adjustment in the model.

Web Table 1: Characteristics of children included and excluded from the analytic cohort.

Maternal factors	Excluded n=353	Included n=649
Pre-pregnancy BMI (kg/m^2)	24.1 (5.1)[a]	24.8 (5.1)
College education	73[a]	71
Smoking history		
Never	69	72
Smoked before pregnancy	20	21
Smoked during pregnancy	12	7
Gestational weight gain (kg)	15.8 (5.2)	15.4 (5.5)
Abnormal glucose tolerance (IH, GIGT, GDM)	16	17
Pregnancy hypertension (GH, PE, CH)	12	10
Paternal BMI (kg/m^2)	26.0 (3.2)	26.5 (4.0)
Paternal college education	69	66
Child factors		
Birth weight-for-gestational-age z-score (SD units)	0.20 (0.97)	0.23 (0.94)
Gestational age at delivery (weeks)	39.6 (1.7)	39.5 (1.6)
No breastfeeding initiation	9	10
Male sex	57	52
White race/ethnicity	73	68

[a] Mean (SD) or %.

BMI: body mass index; IH: isolated hyperglycemia; GIGT: gestational impaired glucose tolerance; GDM: gestational diabetes mellitus; GH: gestational hypertension; PE: pre-eclampsia; CH: chronic hypertension; SD: standard deviation.

Web Table 2: Distribution of stabilized weights according to different methods for nuisance parameters.

	Mean	Median	Min, Max
Fat-mass index[a]			
Parametric model[b]	1.00	0.91	0.27, 2.97
SuperLearner v1[c]	0.99	0.92	0.27, 2.94
SuperLearner v2[d]	0.99	0.92	0.31, 3.01
Blood pressure z-score[e]			
Parametric model	1.00	0.94	0.34, 2.70
SuperLearner v1	1.00	0.93	0.34, 2.56
SuperLearner v2	0.99	0.95	0.43, 2.26
HOMA-IR[f]			
Parametric model	1.00	0.92	0.35, 3.59
SuperLearner v1	1.00	0.93	0.34, 3.62
SuperLearner v2	0.98	0.92	0.41, 3.16
Metabolic risk score[g]			
Parametric model	1.00	0.93	0.39, 3.07
SuperLearner v1	0.99	0.93	0.36, 2.71
SuperLearner v2	0.98	0.93	0.44, 2.46

[a] Covariates *L* include maternal smoking status, education level, BMI, paternal education level, paternal BMI, gestational weight gain (GWG), birth weight-for-gestational-age, breastfeeding initiation, child sex, race/ethnicity and BMI in early childhood.

[b] Parametric model with no interaction terms for nuisance parameters.

[c] SuperLearner with prediction algorithms for generalized linear models, interaction terms and stepwise modeling for nuisance parameters.

[d] SuperLearner with prediction algorithms for generalized linear models, interaction terms, stepwise modeling, penalized regression models, Bayesian additive regression trees and generalized additive models for nuisance parameters.

[e] Covariates *L* include maternal smoking status, education level, BMI, pregnancy hypertension, total GWG, birth weight-for-gestational-age, gestational age, child race/ethnicity and BMI in early childhood.

[f] Covariates *L* include maternal smoking status, education level, BMI, glucose tolerance, paternal education level, paternal BMI, child sex, race/ethnicity and BMI in early childhood.

[g] Covariates *L* include maternal smoking status, education level, BMI, glucose tolerance, paternal education level, paternal BMI, child race/ethnicity and BMI in early childhood.

Web Table 3: Sensitivity analysis of effect estimates of proxy representative interventions indexed by *R*=1 versus *R*=0.

	Fat-mass index[a,b]		SBP z-score[c]		DBP z-score[c]		HOMA-IR[d]		Metabolic risk score[e]	
	(kg/m^2)		(SD units)		(SD units)		(units)		(SD units)	
	N = 500		N = 498		N = 498		N = 306		N = 302	
Estimation method	Effect estimate	95% CI	Effect estimate	95% CI	Effect estimate	95% CI	Effect estimate	95% CI	Effect estimate	95% CI
Parametric model[f]										
Main IPW finding	-1.34	-1.88, -0.84	0.06	-0.11, 0.22	0.09	-0.04, 0.22	-0.09	-0.23, 0.05	-0.06	-0.17, 0.05
Sensitivity analysis[g]	-1.31	-1.88, -0.77	0.07	-0.10, 0.23	0.09	-0.03, 0.22	-0.09	-0.24, 0.05	-0.06	-0.16, 0.04
SuperLearner v1[h]										
Main IPW finding	-1.43	-1.75, -0.81	0.07	-0.08, 0.23	0.08	-0.05, 0.20	-0.09	-0.20, 0.06	-0.06	-0.14, 0.06
Sensitivity analysis[g]	-1.43	-1.98, -0.88	0.07	-0.09, 0.22	0.08	-0.04, 0.21	-0.10	-0.23, 0.04	-0.06	-0.16, 0.04
SuperLearner v2[i]										
Main IPW finding	-1.45	-1.79, -0.86	0.07	-0.08, 0.24	0.09	-0.04, 0.20	-0.09	-0.21, 0.05	-0.06	-0.15, 0.04
Sensitivity analysis[g]	-1.43	-1.99, -0.89	0.07	-0.08, 0.22	0.08	-0.04, 0.20	-0.09	-0.23, 0.05	-0.06	-0.16, 0.03

[a] Fat-mass (kg) / height (m)^2

[b] Adjusted for maternal smoking status, education level, BMI, paternal education level, paternal BMI, gestational weight gain (GWG), birth weight-for-gestational-age, breastfeeding initiation, child sex, race/ethnicity and BMI in early childhood.

[c] Adjusted for maternal smoking status, education level, BMI, pregnancy hypertension, total GWG, birth weight-for-gestational-age, gestational age, child race/ethnicity and BMI in early childhood.

[d] Adjusted for maternal smoking status, education level, BMI, glucose tolerance, paternal education level, paternal BMI, child sex, race/ethnicity and BMI in early childhood.

[e] Adjusted for maternal smoking status, education level, BMI, glucose tolerance, paternal education level, paternal BMI, child race/ethnicity and BMI in early childhood.

[f] Parametric model with no interaction terms for nuisance parameters.

[g] Sensitivity analysis where individuals were censored upon missing *R*, with weights incorporating these censoring weights as a function of *L*.

[h] SuperLearner with prediction algorithms for generalized linear models, interaction terms and stepwise modeling for nuisance parameters.

[i] SuperLearner with prediction algorithms for generalized linear models, interaction terms, stepwise modeling, penalized regression models, Bayesian additive regression trees and generalized additive models for nuisance parameters.

Web References

1. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period: application to the healthy worker survivor effect." Mathematical Modelling, vol. 7, pp. 1393-1512, 1986. [Errata (1987) in Computers and Mathematics with Applications 14, 917-921. Addendum (1987) in Computers and Mathematics with Applications 14, 923-945. Errata (1987) to addendum in Computers and Mathematics with Applications 18, 477.]
2. Taubman S, Mittleman M, Robins J, and Hernan MA. Alternative approaches to estimating the effects of hypothetical interventions," in JSM Proceedings, Health Policy Statistics Section, Alexandria, VA: American Statistical Association, 2008.
3. Stitelman O, Hubbard A, and Jewell N. The impact of coarsening the explanatory variable of interest in making causal inferences: Implicit assumptions behind dichotomizing variables," U.C. Berkeley Division of Biostatistics Working Paper Series, 2010. Working Paper 264.
4. Picciotto S, Hernan MA, Page JH, Young JG, and Robins JM. Structural nested cumulative failure time models to estimate the effects of interventions," Journal of the American Statistical Association, vol. 107, no. 499, pp. 886-900, 2012.
5. Vanderweele T and Hernan MA. Causal inference under multiple versions of treatment," Journal of Causal Inference, vol. 1, no. 1, pp. 1-20, 2013.
6. Young JG, Logan R, Robins J, and Hernan MA. Inverse probability weighted estimation of risk under representative interventions in observational studies," Journal of the American Statistical Association, vol. 114, no. 526, pp. 938-947, 2019.
7. Pearl J. Causality. Cambridge, UK: Cambridge University Press, 2000.
8. Robins J and Richardson T. Alternative graphical causal models and the identification of direct effects. Causality and psychopathology: Finding the determinants of disorders and their cures, pp. 103-158, 2010.
9. Van Der Laan MJ, Rubin D. Targeted maximum likelihood learning. The International Journal of Biostatistics. 2006;2(1).
10. Schuler MS, Rose S. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. Am J Epidemiol. 2017;185(1):65-73.
11. Gruber S, van der Laan M. tmle: An R Package for Targeted Maximum Likelihood Estimation. Journal of Statistical Software. 2012;51(13):35.
12. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007;6:Article 25.