Supplemental Content (SC)

Multiple mediators approach to study environmental determinants of health disparities

Illustration

To practically illustrate the presented methods, we used a simulated dataset in which we hypothesized 10 000 subjects were recruited to evaluate the contribution of di(2-ethylhexyl) phthalate (ΣDEHP, *E*), an endocrine disruptor chemical found, among the others, in fast food consumption (*B*), in explaining racial/ethnic disparities (*X*) in cardiovascular disease (CVD) risk (*Y*). For simplicity, all X, B, E, and Y, were assumed to be binary (respectively: non-Hispanic blacks vs non-Hispanic white; consumption vs no consumption; high vs low exposure, yes vs no), and all models were logistic regressions.

Summary statistics of the simulated data are presented in Table S1. Exposure, mediators, and outcomes, were all assumed to be binary. Race/ethnicity (*X*) was dichotomized into non-Hispanic U.S. blacks, the subgroup where highest CVD prevalence has been consistently observed, versus non/Hispanic U.S. whites. As potential biomarkers of environmental toxicants (*E*) associated with higher risk of CVD, we selected the molar sum of four urinary metabolites of di(2-ethylhexyl) phthalate (ΣDEHP). This is a common summary measure of di(2-ethylhexyl) phthalate, associated with CVD.(1) We dichotomized this covariate as high vs low exposure (hypothetically set as < or >= 13 ug/l).

Table S1. Summary statistics of simulated data

	Non-Hispanic Blacks	Non-Hispanic white
N (%)	1863 (19)	8137 (81)
CVD cases (%)	486 (26)	1443 (18)
Frequent fast-food consumption (%)	829 (44)	2690 (33)
ΣDEHP urinary concentration, ug/l (sd)	12.9 (2.1)	11.2 (2.1)

We evaluated frequent fast food consumption as the modifiable source of environmental toxicants (B). While this factor does not exhaust all sources of exposure to DEHP, we chose it for purposes of providing a simplified example. Regular fast/food consumption (*B* is also higher among non-Hispanic black,(2) and is a modifiable source of *E*.(3) CVD (*Y*) was simulated as a function of race, fast/food consumption, and ΣDEHP urinary concentration. An exposure-mediator interaction between *X* and *E*, and a mediator-mediator interaction between *E* and *B* were also included in the model to generate *Y*.

We first used mediation analysis to determine the proportion of the X-Y disparity that is mediated by E, without taking B into account. The total effect (i.e. the disparity, *model 1*),

showed a 64% higher odds of CVD among non-Hispanic blacks (OR=1.64; 95% CI: 1.46-1.84). By further adjusting *model 1* for the mediator, ΣDEHP, we estimated the direct effect. The effect acting through all other non-specified pathways that are not including ΣDEHP (i.e. the direct effect), was summarized by a 45% higher odds of CVD (OR=1.45; 95% CI: 1.28-1.64). By using the product method, we then calculated the indirect effect acting through ΣDEHP (OR=1.13; 95% CI: 1.08-1.17). By applying the formulas to calculate the proportion mediated for binary outcomes, we would get to the conclusion that 30% of the racial/ethnic disparity in CVD is due to the higher levels of the ΣDEHP reported in the non-Hispanic black population. This single mediator approach was replicated within the counterfactual framework to allow for exposure-mediator interaction, estimating CDE (OR=1.32; 95% CI: 1.16-1.57) and NIE (OR=1.18; 95% CI: 1.10-1.26). This result shows that by ignoring exposure-mediator interaction when this was truly present, we under-estimated the contribution of the mediator in the association of interest.

We next used the multiple mediator approach to evaluate the joint contribution of fast food consumption and ΣDEHP in the racial/ethnic disparity in CVD. The direct effect and indirect effects are presented in Table S2, calculated with 4 modeling strategies: i) without any interaction; ii) including mediator-mediator interaction; iii) including exposure-mediator interaction; iv) both ii and iii. When both exposure-mediator and mediator-mediator interaction are taken into account, the joint contribution of fast-food consumption and high ΣDEHP concentrations accounts for 37% of the racial/ethnic disparity in CVD. The CDE shows that 52% of the disparity would remain after an intervention setting both the source and the environmental factors to a predefined value (low fast-food consumption and ΣDEHP <13 μg/l, in our case), that is, blocking all pathways including ΣDEHP and fast-food consumption. Results also show that failing to incorporate information on exposure-mediator and mediator-mediator interaction would strongly underestimate the joint mediated effect. As mediators are assumed to be sequential, further disentangling pathway-specific effects cannot be achieved with this general multiple mediators approach.

Table S2: Proportion of disparity due to the joint mediating effect of two sequential mediators under four modeling strategies

	No interaction (i)	Mediator/mediator interaction (ii)	Exposure/mediator interaction (iii)[a]	Both ii and iii[a]
CDE (OR)	1.42	1.42	1.29	1.30
Proportion	84%	79%	52%	52%
Joint NIE (OR)	1.09	1.11	1.16	1.20
Proportion	16%	21%	30%	37%

Results from a simulated population of 10 000 subjects, with binary outcome, two binary mediators, and binary exposure

[a] Proportions in the last two columns do not sum up to 100 because when exposure-mediator interaction is present CDE does not equal NDE.

Confounding assumptions within the counterfactual framework

By placing mediation analysis within the counterfactual framework of causal inference, one can define causal mediation effects (i.e. controlled direct effect – CDE; natural indirect effect - NIE) in a way that is not tied to a specific statistical model and identify them under four assumptions: i) absence of unmeasured confounders of the exposure-outcome association; ii) absence of unmeasured confounders of the mediator-outcome association; iii) absence of unmeasured confounders of the exposure-mediator association; iv) absence of an effect of the exposure on a confounder of the mediator-outcome association. (4) Only assumption i and ii are required to identify the CDE (and therefore the counterfactual disparity measure - CDM). It may often be the case; however, that the assumptions about unmeasured confounding do not hold. In particular, both clinical and observational studies are generally designed to evaluate exposure-outcome associations, and collect (or randomize) a variety of possible confounders of this association. However, to correctly identify direct and indirect effects, control must also be made for confounders of the mediator-outcome association. The presence of residual unmeasured confounding of the mediator-outcome association may strongly limit the interpretation of mediation analysis results.(5–7) To make robust inference in the evaluated settings, performing sensitivity analyses is warranted. Much of the research literature on causal mediation analysis has focused on developing sensitivity analysis techniques that allow evaluating how conclusions might be altered by violating the required assumptions.(8,9)

Relevant extensions of mediation analysis

Non-linearity. An extension of the counterfactual approach to mediation analysis, using the so-called mediation formula, has been proposed to generalize the classical mediation estimators while accounting for arbitrary distributions of outcome and mediator.(10)

Repeated measurements. Environmental factors are often repeatedly assessed over time. This is especially necessary for non-persistent chemicals such as phthalates, which have low to moderate intraclass correlation coefficients,(11) but less of an issue for persistent chemicals and other environmental exposures. Mediation analysis has been extended to incorporate those settings where either one or more exposure, mediator, or outcome vary over time. The method uses the mediation formula to formalize direct and indirect effects in the context of longitudinal data while allowing for exposure-mediator interaction.(12) This approach, however, has not yet been extended to incorporate the presence of multiple mediators.

Multiple independent mediators. The presented multiple mediator approach can also be used to integrate the same conceptual model but with multiple, independent (i.e. non sequential) environmental factors that are supposed to simultaneously contribute to the development of a health disparity. The joint evaluation of multiple environmental factors, however, may require

the use of advanced statistical models that take into account the intra-correlation structure of chemical mixtures.(13) To the best of our knowledge, no method has been proposed to integrate mixtures modeling in a mediation analysis setting.

References

1. Hauser R, Calafat AM. Phthalates and Human Health. Occup Environ Med. 2005 Nov 1;62(11):806–18.
2. Zota AR, Phillips C, Mitro SD. Recent fast food consumption and Bisphenol A and phthalates exposures among the US population in NHANES, 2003-2010. Environmental health perspectives. 2016.
3. James-Todd, T. et al. Urinary phthalate metabolite concentrations and diabetes among women in the National Health and Nutrition Examination Survey (NHANES) 2001-2008. Environ. Health Perspect. 120, 1307–1313 (2012).
4. Loeys T, Moerkerke B, De Smet O, Buysse A, Steen J, Vansteelandt S. Flexible Mediation Analysis in the Presence of Nonlinear Relations: Beyond the Mediation Formula. Multivar Behav Res. 2013 Nov;48(6):871–94.
5. VanderWeele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. Epidemiol Camb Mass. 2014 Mar;25(2):300–6.
6. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. Int J Epidemiol. 2013 Oct 1;42(5):1511–9.
7. Lepage B, Dedieu D, Savy N, Lang T. Estimating controlled direct effects in the presence of intermediate confounding of the mediator–outcome relationship: Comparison of five different methods. Stat Methods Med Res. 2016 Apr 1;25(2):553–70.
8. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychol Methods. 2010;15(4):309–34.
9. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiol Camb Mass. 2010 Jul;21(4):540–51.
10. Albert JM. Distribution-free mediation analysis for nonlinear models with confounding. Epidemiol Camb Mass. 2012 Nov;23(6):879–88.
11. Braun JM, Smith KW, Williams PL, Calafat AM, Berry K, Ehrlich S, et al. Variability of urinary phthalate metabolite and bisphenol A concentrations before and during pregnancy. Environ Health Perspect. 2012 May;120(5):739–45.
12. Bind MA, Vanderweele TJ, Coull BA, Schwartz JD. Causal mediation analysis for longitudinal data with exogenous exposure. Biostat Oxf Engl. 2016 Jan;17(1):122–34.
13. Taylor KW, Joubert BR, Braun JM, Dilworth C, Gennings C, Hauser R, Heindel JJ, Rider CV, Webster TF, Carlin DJ. Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop. Environmental Health Perspectives. 2016 Dec;124(12):A227.

Stata code for data simulation and statistical analyses

```
**********************
* 1. Data Simulation*
**********************

clear all
set more off
set obs 10000
set seed 12
/* Generate race/ethnicity as a binary covariate
with 19% of black-American */
gen x = rbinomial(1,.19)

* Generate a continuous confounder (e.g. Age. Mean: 45 years)
gen c = rnormal(45,5)

/*Generate the binary mediator (yes/no) of fast-food consumption.
Use data from Zota et al, 2016, showing a proportion of 44% fast-food
consumers among black-American, and 33% among other groups. The mediator will also be
 dependent on age*/

gen m1=.
replace m1=rbinomial(1,.5-(c/1000)) if x==1
replace m1=rbinomial(1,.38-(c/1000)) if x==0

/*The following lines generate a second continuous mediator representing the urinary
concentration of a specific chemical. We assume that this covariate is associated
with both race/ethnicity and fast-food consumption (and age).
*/

*Constant
scalar beta0 =6.5
*Main effect of race/ethnicity
scalar beta1 =1.5
*Main effect of fast-food consumption
scalar beta2=0.9
*Main effect of age
scalar beta3=0.1
gen m2 = rnormal(beta0+beta1*x+beta2*m1+beta3*c,2)

*(Note that in real situations environmental chemicals are seldom normally distributed)


/*We will also use a binary version of the mediator, assuming that only urinary
concentrations above 13 are expected to be harmful*/
gen m2cat =0
replace m2cat=1 if m2>13

/*Generates CVD (yes/no) as a function of race/ethnicity (OR= 1.1),
fast-food consumption (OR=1.2), and DEHP urinary concentration
(OR=1.3 for each unit increase of DEHP). Also, we assume an interaction between race and DEHP
(exposure-mediator interaction), and between the two mediators.

scalar beta00 =-3
*CVD around 17%
scalar beta11 =log(1.4)
scalar beta12 =log(1.2)
scalar beta13=log(1.2)
```

```
scalar beta14=log(1.03)
scalar beta3=log(1.23)
scalar beta4=log(1.18)

gen inter=m2cat*x
gen inter2=m2cat*m1

gen y=.
replace y = rbinomial(1,exp(beta00 + beta11 * x + beta12 * m1 + beta13 * m2cat + beta3 * inter +
beta4 * inter2 ///
+beta14*c)/ (1 + exp(beta00 + beta11 * x + beta12 * m1 + beta13 * m2cat + beta3 * inter+ beta4 *
inter2 ///
+beta14*c)))


**********************
* 2. Data description*
**********************


*Summary statistics by CVD status
tab y
tab x y, col
tab m1 y, col
tab m1 x, col
tabstat m2, by(y) stat(mean sd)
tabstat c, by(y) stat(mean sd)

*****************
* 3. Mediation - no interaction*
*****************

* X->Y
logit y x , or
matrix a=e(b)

* X->M2
logit m2cat x, or
matrix b=e(b)

* X, M2 -> Y
logit y x m2cat, or
matrix c=e(b)

*Retrieve coefficients to calculate direct and indirect effects
scalar totaleff=exp(a[1,1])
scalar directeff=exp(c[1,1])
scalar indirecteff_product=((1+exp(b[1,2]))*(1+exp(c[1,2]+b[1,1]+b[1,2])))/ ///
((1+exp(b[1,1]+b[1,2]))*(1+exp(c[1,2]+b[1,2])))
scalar indirecteff_difference=exp(log(totaleff)-log(directeff))
scalar pm=(directeff)*(indirecteff_product-1)/(directeff*indirecteff_product-1)

di totaleff
di directeff
di indirecteff_product
di indirecteff_difference
di pm

*Same results would be obtained with the command paramed
paramed y, avar(x) mvar(m2cat) a0(0) a1(1) m(0) yreg(logistic) mreg(logistic) cvars(c) nointer
```

```
*****************
* 5. Mediation -  interaction*
*****************

* X->Y
logit y x
logit y x, or
matrix a=e(b)

* X->M2
logit m2cat x
logit m2cat x, or
matrix b=e(b)

* X, M2 -> Y
logit y x m2cat inter
logit y x m2cat inter, or
matrix c=e(b)


/* The difference and product method gives slightly different results because
the outcome is binary, and the equivalence requires rare outcomes. */

* The paramed command gives the same results with standard errors
paramed y, avar(x) mvar(m2cat) a0(0) a1(1) m(0) yreg(logistic) mreg(logistic)

*We can also adjust for age
paramed y, avar(x) mvar(m2cat) a0(0) a1(1) m(0) yreg(logistic) mreg(logistic) cvars(c)


* Other components of the 4-way decomposition
(Note that the command med4way is currently under development and not yet online. Please contact
the authors for additional information. Same results can be obtained by available code in SAS)

med4way y x m2cat,  a0(0) a1(1) ///
       m(0) yreg(logistic) mreg(logistic) ///
       reps(100) boot seed(65443) nodeltamethod

med4way y x m2cat c,  a0(0) a1(1) ///
       m(0) yreg(logistic) mreg(logistic) ///
       c(45) reps(100) boot seed(65443) nodeltamethod


*****************
* 6. Multiple mediators - ignoring all interactions
*****************

* X->Y
logit y x, or
matrix a=e(b)

* X->M1
logit m1 x, or
matrix b=e(b)

* X,M1->M2
logit m2cat m1 x, or
matrix c=e(b)

* X, M1, M2 -> Y
logit y m2cat m1 x, or
```

```
matrix d=e(b)

scalar totaleff=exp(a[1,1])
scalar directeff=exp(d[1,3])
scalar indirecteff_product_m1= ///
((1+exp(b[1,2]))*(1+exp(d[1,2]+b[1,1]+b[1,2])))/ ///
((1+exp(b[1,1]+b[1,2]))*(1+exp(d[1,2]+b[1,2])))
scalar indirecteff_product_m2= ///
((1+exp(c[1,3]))*(1+exp(d[1,3]+c[1,1]+c[1,3])))/ ///
((1+exp(c[1,1]+c[1,3]))*(1+exp(d[1,3]+c[1,3])))

scalar pm=log(indirecteff_product_m1)/log(totaleff)+log(indirecteff_product_m2)/log(totaleff)


di totaleff
di directeff
di indirecteff_product_m1
di indirecteff_product_m2
di pm

di exp(log(indirecteff_product_m1)+log(indirecteff_product_m2))



*****************
* 6. Multiple mediators - med-med interactions, but ignoring int-med
*****************

* X->Y
logit y x , or
matrix a=e(b)

* X->M1
logit m1 x, or
matrix b=e(b)

* X,M1->M2
logit m2cat m1 x, or
matrix c=e(b)

* X, M1, M2 -> Y
logit y m2cat m1 x inter2, or
matrix d=e(b)

* X->M1
logit inter2 x, or
matrix f=e(b)

scalar totaleff=exp(a[1,1])
scalar directeff=exp(d[1,3])
scalar indirecteff_product_m1= ///
((1+exp(b[1,2]))*(1+exp(d[1,2]+b[1,1]+b[1,2])))/ ///
((1+exp(b[1,1]+b[1,2]))*(1+exp(d[1,2]+b[1,2])))
scalar indirecteff_product_m2= ///
((1+exp(c[1,3]))*(1+exp(d[1,1]+c[1,1]+c[1,3])))/ ///
((1+exp(c[1,1]+c[1,3]))*(1+exp(d[1,1]+c[1,3])))
scalar indirecteff_product_m1m2= ///
((1+exp(f[1,2]))*(1+exp(d[1,4]+f[1,1]+f[1,2])))/ ///
((1+exp(f[1,1]+f[1,2]))*(1+exp(d[1,4]+f[1,2])))

scalar pm=log(indirecteff_product_m1)/log(totaleff)+log(indirecteff_product_m2)/log(totaleff)+
///
```

```
log(indirecteff_product_m1m2)/log(totaleff)


di totaleff
di directeff
di indirecteff_product_m1
di indirecteff_product_m2
di indirecteff_product_m1m2
di pm
di exp(log(indirecteff_product_m1)+log(indirecteff_product_m2)+log(indirecteff_product_m1m2))


*****************
* 6. Multiple mediators - med-int, ignoring med-med interactions
*****************

* X->Y
logit y x , or
matrix a=e(b)

* X->M1
logit m1 x, or
matrix b=e(b)

* X,M1->M2
logit m2cat m1 x inter, or
matrix c=e(b)

* X, M1, M2 -> Y
logit y m2cat m1 x inter, or
matrix d=e(b)

scalar totaleff=exp(a[1,1])
scalar directeff=exp(d[1,3])
scalar indirecteff_product_m1= ///
((1+exp(b[1,2]))*(1+exp(d[1,2]+d[1,4]+b[1,1]+b[1,2])))/ ///
((1+exp(b[1,1]+b[1,2]))*(1+exp(d[1,2]+d[1,4]+b[1,2])))
scalar indirecteff_product_m2= ///
((1+exp(c[1,3]))*(1+exp(d[1,1]+d[1,4]+c[1,1]+c[1,3])))/ ///
((1+exp(c[1,1]+c[1,3]))*(1+exp(d[1,1]+d[1,4]+c[1,3])))

scalar pm=log(indirecteff_product_m1)/log(totaleff)+log(indirecteff_product_m2)/log(totaleff)
scalar pr=log(directeff)/log(totaleff)

di totaleff
di directeff
di indirecteff_product_m1
di indirecteff_product_m2
di pm
di pr
di exp(log(indirecteff_product_m1)+log(indirecteff_product_m2))

*****************
* 6. Multiple mediators - all interactions
*****************

* X->Y
logit y x , or
matrix a=e(b)

* X->M1
logit m1 x, or
```

```
matrix b=e(b)

* X,M1->M2
logit m2cat m1 x, or
matrix c=e(b)

* X, M1, M2 -> Y
logit y m2cat m1 x inter inter2, or
matrix d=e(b)

* X->M1
logit inter2 x, or
matrix f=e(b)

scalar totaleff=exp(a[1,1])
scalar directeff=exp(d[1,3])

scalar indirecteff_product_m1= ///
((1+exp(b[1,2]))*(1+exp(d[1,2]+d[1,4]+b[1,1]+b[1,2])))/ ///
((1+exp(b[1,1]+b[1,2]))*(1+exp(d[1,2]+d[1,4]+b[1,2])))
scalar indirecteff_product_m2= ///
((1+exp(c[1,3]))*(1+exp(d[1,1]+d[1,4]+c[1,1]+c[1,3])))/ ///
((1+exp(c[1,1]+c[1,3]))*(1+exp(d[1,1]+d[1,4]+c[1,3])))
scalar indirecteff_product_m1m2= ///
((1+exp(f[1,2]))*(1+exp(d[1,5]+d[1,4]+f[1,1]+f[1,2])))/ ///
((1+exp(f[1,1]+f[1,2]))*(1+exp(d[1,5]+d[1,4]+f[1,2])))

scalar pm=log(indirecteff_product_m1)/log(totaleff)+log(indirecteff_product_m2)/log(totaleff)+
///
log(indirecteff_product_m1m2)/log(totaleff)


di totaleff
di directeff
di indirecteff_product_m1
di indirecteff_product_m2
di indirecteff_product_m1m2
di pm
di exp(log(indirecteff_product_m1)+log(indirecteff_product_m2)+log(indirecteff_product_m1m2))
```