Does DNA methylation mediate the effect of maternal smoking on birth weight? Exposure misclassification in mediation analyses for environmental epigenetic studies

Linda Valeri
Harvard Medical School and McLean Hospital
Psychiatric Biostatistics Laboratory

December 14, 2016

Acknowledgments

- Biostatistics Department, HSPH

  Brent Coull

- NIEHS

  Stephanie London

  Sarah Reese

  Shanshan Zhao

- Norwegian Mother and Child Cohort Study (MoBa) participants

The quest for epigenetic links b/w environment and health

- Many studies report that changes in epigenetic marks are associated with broad range of diseases (cancer, asthma, metabolic disorders, reproductive conditions perinatal outcomes).
- In recent years evidence has accumulated that environmental exposures (malnutrition, tobacco smoke, air pollutants, metals, organic chemicals) lead to changes in the epigenome.
- Timing of environmental exposure (for example during key periods of development) might be critical.
- In the popular press 'epigenetics' has become almost synonymous with nutritional and environmental influences on gene expression, intensifying interest in the possible role of epigenetic mechanisms in disease.
- Several studies are currently attempting to provide clearer mechanistic insights and evidence to support this hypothesis.

Cortessis et al., *Human Genetics* , 2012.

The quest for epigenetic links b/w environment and health

Table 1 | **Chemicals and pollutants that affect health and induce epigenetic alterations**

Compound	Species	Ontogenic stage	Epigenetic alteration	Tissues or cell types affected	Phenotypic alterations	Refs
Tobacco smoke	Human	Adult life	Locus-specific DNA methylation and histone modifications; chromatin remodelling machinery	Lung, blood	Lung cancer?	60,61,143
Particulate air pollution	Human, Mouse	Adult life	DNA methylation	Blood, sperm	Unknown	54,69
Asbestos	Human	Adult life	DNA methylation	Pleural tissues	Susceptibility to different diseases	57
Bisphenol A (BPA)	Mouse	Embryonic development	Locus-specific DNA methylation	Systemic	Coat colour distribution of agouti viable yellow (A^{vy}) mice	99
Diethylstilbestrol (DES)	Mouse	Embryonic development	DNA methylation	Gonads	Male sexual function	144,145
Metal ions (such as chromium, cadmiun, nickel, arsenic and methylmercury)	Multiple species	Embryonic development, adult life	DNA methylation; histone modifications (for nickel)	Multiple tissues	Increased susceptibility to diseases such as cancer	Reviewed in REFS 146,147
Vinclozolin	Mouse, rat	Embryonic development	DNA methylation	Male germ cells	Altered gonad development and spermatogenesis in the male offspring	81,82
Methoxychlor	Mouse	Embryonic development, adult life	DNA methylation	Male germ cells	Altered male reproductive system	84
Silica	Human	Adult life	DNA methylation	Blood	Silicosis	148
Benzene	Human	Adult life	DNA methylation	Blood	Increased risk of AML	55
Di- and trichloroacetic acid, trichloroethylene	Mouse	Adult life	Locus-specific DNA methylation	Liver	Increased risk of hepatic cancer	Reviewed in REF. 147

AML, acute myeloid leukaemia

Fail and Fraga, *Nature Reviews Genetics* , 2012.

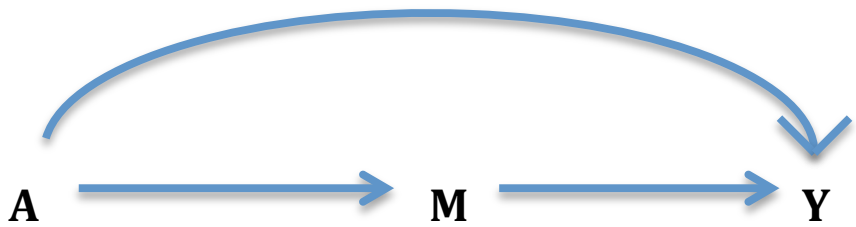
A question about mediating mechanism

Let,

A be an environmental exposure of interest,

M denote an epigenetic intermediate factor,

Y be an health outcome.



Challenges

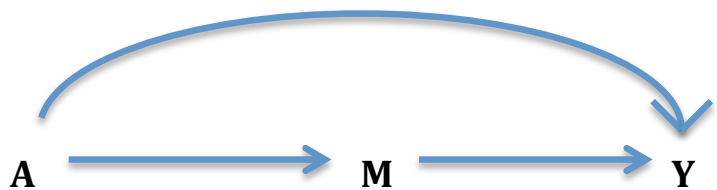
- Epidemiologic research addressing epigenetic mechanisms is constrained by important ethical considerations, that lead to challenges in measurement and in assessment of causal relations.
- Ethical considerations often imply suboptimal choice of tissue collection and the use of *observational* data.
- Depending on the health outcome of interest, long induction periods challenge the study of epigenetic mechanisms.
- In particular, it is challenging to capture and correctly measure exposure history as well as history of epigenetic-state.
- There is considerable potential for confounding of exposure-epigenetic state, exposure-disease, epigenetic state-disease associations.
- Issues of reverse causation are common.

Case Study: A link between smoking and birthweight

- Early stages of embryonic development are periods of particular vulnerability and epigenetic changes might have long-lasting consequences on off-spring health.
- Smoking in adults has been reproducibly associated with alterations in methylation at specific loci (Gao et al., 2015). Similar effects are seen in newborns whose mothers smoked during pregnancy (Joubert et al., 2016).
- These smoking-methylation signals have been used to develop novel biomarkers of exposure.
- Epigenetic changes in several of those sequences gene have been shown to impact child birth weight (Engel et al, 2014).

Case Study: A link between smoking and birthweight

Question: Is the effect on birthweight (Y) of smoking during pregnancy (A) mediated by DNA methylation (M) or is there a direct effect of smoking on birthweight?



A link between smoking and birthweight (Küppers et al.)

- Küppers et al. (2015) conducted in the Dutch GECKO cohort an EWAS in cord blood to examine the association between (self-reported) sustained maternal smoking during pregnancy and DNA methylation.
- Further, they performed a mediation analysis on the 35 top CpGs (epigenome-wide significant at FDR < 0.05).
- Eight CpG sites in the GFI1 gene that showed the most robust indirect effect were then replicated and meta-analyzed in two other cohorts (ALSPAC and GenR).
- The authors were able to replicate the finding of a significant indirect effect of maternal smoking on birthweight mediated by three CpG sites on the GFI1 gene (cg12876356, cg09935388, and cg14179389). DNA methylation was found to explain about $12\% - 19\%$ of maternal smoking on birthweight.
- The authors did not evaluate in the paper potential bias due to mis-report of smoking status.

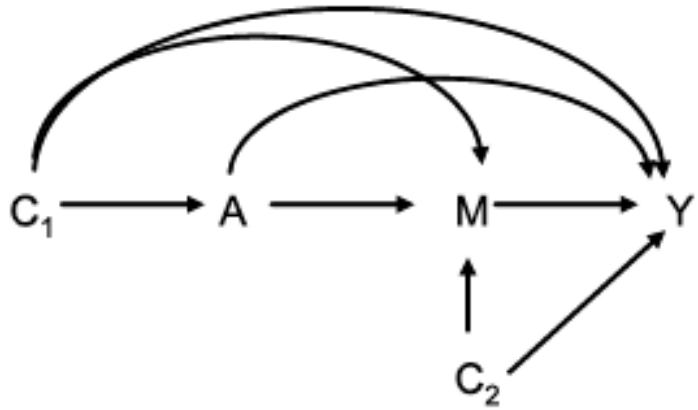
Goals of the presentation

- Quick overview of concepts of direct and indirect effects under the counterfactual framework and a regression approach to mediation analysis
- Exposure misclassification in mediation analysis: bias and type I error
- Case Study: Investigate the role of DNA methylation as mediator of maternal smoking effect on birthweight in the MoBa Cohort

Causal Mediation Analysis: Notation

- Y = birth weight
- A = mother smoking status during pregnancy
- M = cord blood DNA methylation (potentially on the pathway between A and Y)
- C = set of covariates for each individual
- Y_a = birthweight for each individual when intervening to set A to a
- Y_{am} = birthweight when intervening to set A to a and M to m
- M_a = DNA methylation when intervening to set A to a

Causal Effects under Counterfactual Framework



$$NDE = E[Y_{1M_0} - Y_{0M_0}|C]$$

$$NIE = E[Y_{1M_1} - Y_{1M_0}|C]$$

$$TE = NDE + NIE$$

(Robins and Greenland, 1992; Pearl, 2001)

Mediation Analysis under Counterfactual Framework

IDENTIFIABILITY ASSUMPTIONS:

(i) No unmeasured exposure-outcome confounding given C

(ii) **No unmeasured mediator-outcome confounding given C**

(iii) No unmeasured exposure-mediator confounding given C

(iv) No effect of exposure that confounds the mediator-outcome relationship

Note that assumptions (i) and (iii) are satisfied automatically if the exposure is randomized but not (ii) and (iv).

Regression Approach to Mediation Analysis :

$$E(M|A=a, C=c) = \beta_0 + \beta_1 a + \beta_2^{'} c \quad (1)$$

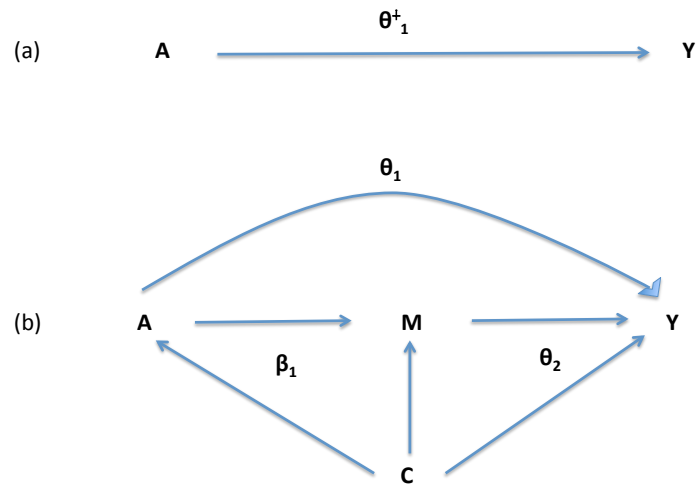
$$E(Y|A=a, M=m, C=c) = \theta_0 + \theta_1 a + \theta_2 m + \theta_4^{'} c \quad (2)$$

$$E(Y|A=a, C=c) = \theta_0^{\dagger} + \theta_1^{\dagger} a + \theta_4^{'\dagger} c \quad (3)$$

$$NDE = \theta_1$$

$$NIE = \theta_2 * \beta_1$$

$$TE = \theta_1^{\dagger}$$



By property of effect decomposition

$$NIE = \theta_1^{\dagger} - \theta_1$$

Note: Estimators for direct and indirect effects can be derived in the presence of exposure-mediator interaction (Valeri and VanderWeele, 2013). We assume throughout no exposure-mediator interaction.

Challenges in Mediation Analysis

- Unmeasured confounding
- Model mis-specification
- Incomplete data and measurement error

Measurement Error in Mediation Analysis: Background

Setting usually considered (Hoyle and Kenny, 1999; VanderWeele, Valeri and Ogburn, 2012):

- Mediator measured with error
- Classical non-differential measurement error
- Linear Models

Common but Overlooked Scenarios

It is often stated that "Measurement error in mediation analysis leads to conservative estimates of the indirect effect."

This intuition may not hold in several settings, such as when:

- The mediator interacts with the exposure (Valeri and VanderWeele, 2014)
- The exposure or confounders are measured with error (Valeri et al., 2016)
- A polytomous mediator is misclassified (Ogburn and VanderWeele, 2013)

Other important issues are currently ignored:

- Validity of Type I error for test of mediation
- Impact of coarsening or "ill defining" variables involved in mediation analysis

Mediation analysis when exposure is misclassified

Let Y denote birthweight, M methylation of CpG site in the GFI1 gene, A maternal smoking status during pregnancy.

$$E(Y|a,m,c) = \theta_0 + \theta_1 a + \theta_2 m + \theta_4^{'} c$$

$$E(M|a,c) = \beta_0 + \beta_1 a + \beta^{'} c$$

Direct and indirect effects are given by:

$$NDE = \theta_1$$

$$NIE = \theta_2 \beta_1$$

Misclassified Exposure: Assumptions

- Let A be a binary true exposure and $A^* = A + U$ the observed exposure, self-reported smoking status

- U is misclassification error with moments dependent on $SN = P(A^* = 1|A = 1)$, $SP = P(A^* = 0|A = 0)$, and prevalence of A^*

- Misclassification error depends on the latent true exposure $Cov(A, U) \neq 0$

- In this setting we can assume perfect specificity ($SP = 1$)

- In this setting the mediator, DNA methylation, is a strong and precisely measured biomarker of an imperfectly measured exposure (Joubert et al, 2012; Reese et al., 2016).

Asymptotic Limit of Regression Parameters

Define the observed outcome and mediator regressions as:

$$E(Y|a^*, m, c) = \theta_0^* + \theta_1^* a^* + \theta_2^* m + \theta_4^{*\prime} c$$

$$E(M|a^*, c) = \beta_0^* + \beta_1^* a^* + \beta_2^{*\prime} c$$

- Estimators of the *mediator regression* parameters are not consistent
- Estimators of the *outcome regression* parameters are not consistent

Asymptotic Limit of Mediator Regression Parameters

Let

β_1 denote the true exposure-mediator association and β_1^* the association in the presence of exposure misclassification.

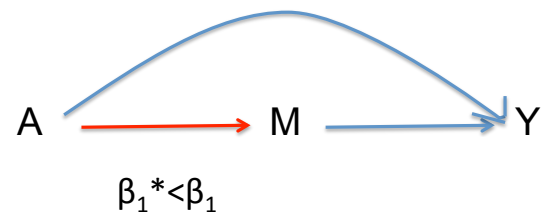
$$Cov(A^*, U) = \{(1 - SP)\frac{P(A=0)}{P(A^*=1)} + (1 - SN)\frac{P(A=1)}{P(A^*=0)}\}P(A^* = 1)P(A^* = 0)$$

Δ^{m*} denote the variance-covariance matrix of the observed centered covariates in the mediator regression and $\delta^m_{i,j}$ denote an element of the inverse of the variance-covariance matrix.

$$plim\beta_1^* \quad = \quad \beta_1 - \beta_1 \delta^m_{A^*,A^*} Cov(A^*, U)$$

Since $Cov(A^*, U) \geq 0$,
$\delta^m_{A^*,A^*} > 0 \rightarrow \beta_1^* < \beta_1$

A-M association is underestimated



Asymptotic Limit of Outcome Regression Parameters

Let
θ_1 denote the true exposure-outcome association and θ_1^* the association in the presence of exposure misclassification.

θ_2 denote the true mediator-outcome association and θ_2^* the association in the presence of exposure misclassification.

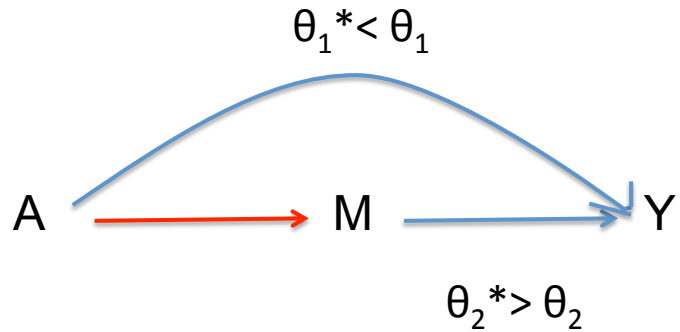
Δ^{y*} denote the variance-covariance matrix of the observed centered covariates in the outcome regression and $\delta^y_{i,j}$ denote an element of the inverse of the variance-covariance matrix.

$$\begin{aligned} plim\theta_1^* &= \theta_1 - \theta_1 \delta^y_{A^*,A^*} Cov(A^*, U) \\ plim\theta_2^* &= \theta_2 - \theta_1 \delta^y_{A^*,M} Cov(A^*, U) \end{aligned}$$

Asymptotic Limit of Outcome Regression Parameters

Since $Cov(A^*, U) \geq 0$,

- $\delta^y_{A^*,A^*} > 0 \rightarrow \theta_1^* < \theta_1$
  A-Y association is underestimated
- $\delta^y_{A^*,M} < 0 \rightarrow \theta_2^* > \theta_2$
  M-Y association is overestimated
- Note that if A and M are uncorrelated $\theta_2^* = \theta_2$.



Asymptotic bias of direct and indirect effects

- Misclassification induced down-ward biased NDE

$$ABIAS(\widehat{NDE}^*) = \theta_1^* - \theta_1 = -\theta_1 \delta^y_{A^*,A^*} Cov(A^*, U)$$

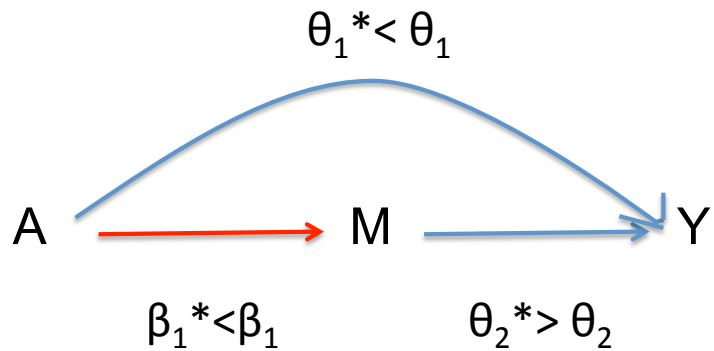
- The bias of NIE can go in either direction

$$ABIAS(\widehat{NIE}^*) = Cov(A^*, U)\beta_1[-\theta_2 \delta^m_{A^*,A^*} + \theta_1 \delta^y_{A^*,M}(\delta^m_{A^*,A^*} Cov(A^*, U) - 1)]$$

Note that if $\beta_1 = 0$ (i.e. A and M are uncorrelated) the (null) indirect effect is unbiased.

Asymptotic bias of direct and indirect effects

- In the presence of exposure misclassification *NDE* is underestimated
- The bias of *NIE* can go in either direction

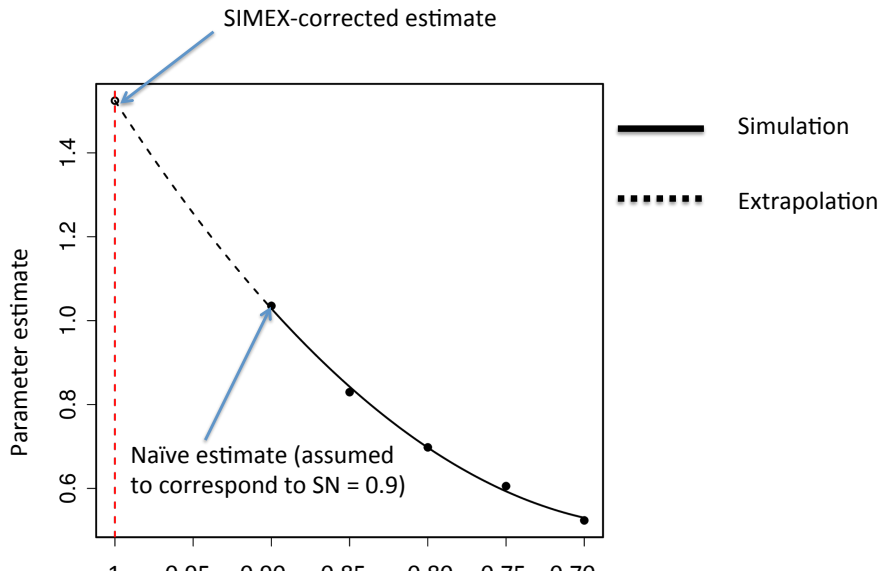


Type I error

- Most widely used test for mediated effect is the Sobel test which constructs a Wald type test from $NIE = \theta_2 * \beta_1$ and delta method standard error.
- The test for indirect effect embeds a composite null hypothesis.
- $H_0 : NIE = 0$ can be achieved if either θ_2 or β_1 or both are equal to zero.
- The type I error is in general not preserved when the exposure is measured with error.
- In particular it is not preserved when DNA methylation is a strong biomarker of the exposure ($\beta_1 \neq 0$).
- This is because the naïve estimator is biased under the null when $\beta_1 \neq 0$.

Misclassification Correction and Sensitivity Analyses: SIMEX

- Several approaches could be adopted such as regression calibration, EM-algorithm, and SIMEX.
- SIMEX is a simulation-based approach for measurement error and misclassification correction (Carroll et al., 1996, Rpackage *simex*.)
- We obtain SIMEX-corrected estimates of the regression parameters and then plug them in the formulas for direct and indirect effects.
- Given a range of SN and SP, the SIMEX approach consists of a simulation and an extrapolation step



Analysis of MoBa Cohort

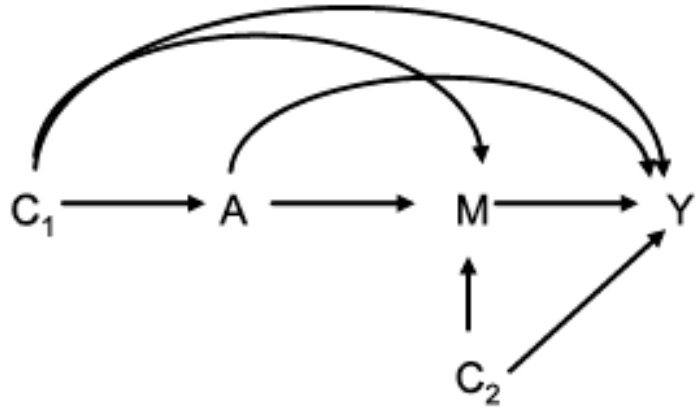
- The Norwegian Mother and Child Cohort Study (MoBa) is a large population based pregnancy study conducted by the Norwegian Institute of Public Health targeting all women in Norway who gave birth between 1999 and 2008.

- Illumina HumanMethylation450K data from cord blood is available on a sub-cohort of MoBa participants born between 2002 and 2004 (N=1068).

- Several measures of self-reported smoking behavior are recorded. Additionally, cotinine measured from maternal plasma at about gestational week 18 of pregnancy is available.

- We proceed studying in the MoBa cohort the three GFI1 CpG sites that were replicated in Küpers et al. (2015) analysis.

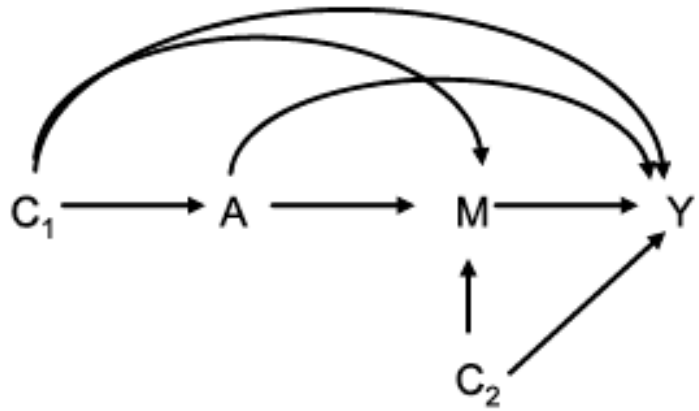
Analysis of MoBa Cohort



QUESTIONS:

1) Is maternal smoking (A) associated with DNA methylation of CpG sites on the GFI1 gene (cg12876356, cg09935388, and cg14179389)?
2) Is DNA methylation of CpG sites on the GFI1 gene (M) associated with birthweight (Y)?
3) Is the effect on birth weight of sustained maternal smoking during pregnancy mediated by DNA methylation of GFI1 CpG sites (NIE) accounting for confounding (C: sex, age, mother education, gestational age, parity, mother bmi)?

Analysis of MoBa Cohort



1) We proceed running regression and mediation analyses ignoring exposure misclassification
2) We then improve exposure measurement by reclassifying smoking status based on auxiliary information on cotinine
3) We finally apply SIMEX misclassification correction approach to assess sensitivity of findings to residual misclassification

Naïve mediator regression analyses: self-reported sustained smoking

- DNA methylation of the three CpG sites is strongly associated with smoking (i.e. DNA-methylation is a strong biomarker for smoking exposure)

Table : Naïve mediator regression given self-reported sustained smoking indicators and confounders in the MoBa Cohort.

	cg09935388	**cg12876356**	**cg14179389**
smoking	-0.12 (-0.13, -0.10)	-0.11 (-0.11,-0.09)	-0.07 (-0.09, -0.05)

Naïve outcome regression analyses: self-reported sustained smoking

- Naïve outcome regression indicates that self-reported maternal sustained smoking during pregnancy is negatively associated with birth weight $\theta_1^{\dagger} = -93.78$, $CI = (-180, -8)$
- Once adjustment is made for CpGs', smoking-birth weight association decreases and ceases to be statistically significant
- DNA methylation of the three CpG sites is marginally associated with increased birth weight

Table : Naïve outcome regression given self-reported sustained smoking indicators, DNA methylation of CpG site, and confounders in the MoBa Cohort.

	cg09935388	**cg12876356**	**cg14179389**
smoking	-63.83 (-154, 26)	-76.92 (-165,11)	-82.94 (-172, 6)
CpG	247.42 (-5,500)	144.54 (-67,356)	144.99 (-192, 482)

Naïve mediation analyses: self-reported sustained smoking

- Naïve outcome regression analyses yield a significant total effect of smoking on birth weight $TE = \theta_1^{\dagger} = -93.78$, $CI = (-180, -8)$
- Naïve mediation analyses yield a not statistically significant direct effect.
- DNA mehylation of CpG site cg09935388 is found to mediate the effect of smoking on birth weight by about 32%
- The indirect effect through CpG sites cg12876356 and cg14179389 is not statistically significant and indicates a proportion mediated of 15% and 12% respectively.

Table : Estimate of nave direct effects (NDE) and nave indirect effects (NIE) and bootstrap 95% confidence interval for each of the three CpGs

EFFECT	**cg09935388**	**cg12876356**	**cg14179389**
nde	-64.5 (-174.6 ,38.3)	-76.6 (-165.3 ,4.0)	-81.5 (-168.6, 37.9)
nie	-29.4 (-65.4, -4.0)	-14.2 (-37.4, 5.7)	-11.1 (-32.6 ,12.1)
pm	32%	15%	12%

Mediation analyses using auxiliary information on cotinine

- We improved classification of smokers using cotinine information at the 18th gestational week
- Outcome regression analyses yield a significant total effect of smoking on birth weight $TE = \theta_1^{\dagger} = -116.04$, $CI = (-196, -36.1)$
- Mediation analyses using a cotinine-reclassified exposure yield a statistically significant and larger direct effect
- DNA mehylation of CpG site cg09935388 is found to mediate the effect of smoking on birth weight by about 24% (but the indirect effect is not statistically significant)

Table : Estimate of nave direct effects (NDE) and nave indirect effects (NIE) and bootstrap 95% confidence interval for each of the three CpGs

EFFECT	**cg09935388**	**cg12876356**	**cg14179389**
nde	-90.1 (-177.6 ,-3.08)	-102 (-186, -18.2)	-108.9 (-193 ,-24.2)
nie	-27.8 (-59.7,5.1)	-13.9 (-37.4,10.7)	-8.3 (-36.4,17.5)
pm	24%	12%	7%

SIMEX analysis

- We assess the effect of residual misclassification in smoking status employing the SIMEX approach (R package "simex" by Lederer and a Küchenhoff, 2013).

- A recent study (Reese et al., 2016) suggests that in the MoBa cohort self-reported smoking behavior could be highly misclassified. and SN could reasonably get to as low as 70%.

- We perform a sensitivity analysis of over a range of $SN = (0.7, 0.8, 0.9)$ and assuming perfect specificity ($SP = 1$).

SIMEX analysis results

Table : Estimate of SIMEX-corrected direct effects (NDE) and indirect effects (NIE) and bootstrap 95% confidence interval for each of the three CpGs for SN=(0.7,0.8,0.9)

Effect	cg09935388	cg12876356	cg14179389
SN=0.7			
NDE	-106.9 (-216.4,8.1)	-118.5 (-209.3,-14.3)	-125.8 (-223.6,-14.5)
NIE	-18.7 (-60.4,21.7)	-8.3 (-36.5,20.6)	-0.7 (-34.0,31.3)
PM	15%	6%	0%
SN=0.8			
NDE	-102.1 (-205.9,16.2)	-111.9 (-211.1,-9.8)	-119.8 (-213.7,-12.5)
NIE	-21.5 (-60.8,17.6)	-9.9 (-35.8,18.1)	-2.6 (-34.9,26.9)
PM	17%	8%	2%
SN=0.9			
NDE	-97.9 (-184.0,-0.9)	-111.4 (-201.1,-14.8)	-115.0 (-212.1,-12.4)
NIE	-23.9 (-59.1,8.3)	-11.2 (-36.3,14.7)	-4.9 (-36.5,23.4)
PM	20%	9%	4%

Summary

- Naïve analyses indicate that DNA methylation of CpG sites on GFI1 might mediate 32% of the effect of sustained maternal smoking on birthweight.
- Using auxiliary information and adjusting for misclassification indicates that *NDE* is under-estimated in the Naïve analyses.
- After improving smokers classification and correcting for misclassification using SIMEX, *NIE* estimate is reduced and ceases to appear statistically significant for each sensitivity analysis parameters
- Proportion mediated (*PM*) could be over-estimated by 50% in case of severe misclassification
- Note that along with measurement error issues, residual unmeasured confounding could bias the estimates as well
- In interpreting this apparently null findings one would also need to take into account the low power of tests for mediation. Statistical development is needed to improve efficiency in mediation analyses in EWAS.

Conclusion

- Mediation analysis can be severely undermined by exposure misclassification

- When the exposure is measured with error the indirect effect and the proportion mediated can be overestimated if the mediator is a strong biomarker of the exposure.

- Type I error may not be preserved in the presence of exposure misclassification

- We discussed a correction strategy for misclassification for which no auxiliary data on the exposure are needed.

- These techniques can be applied to non linear models and when outcome, mediator, or confounders are measured with error.

- It is important that analyses reflect uncertainty around measurement of environmental exposure to avoid spurious findings.

Thank You!

References

- BARON, R.M. KENNY, D.A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. Journal of Personality and Social Psychology 51 1173-1182.
- CARROLL, R.J., RUPPERT, D., STEFANSKI, L.A., CRAINICEANU C.M. (2006). Measurement Error in Nonlinear Models. Chapman Hall/CRC.
- KÜPERS, L. K., et al. (2015) DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *International journal of epidemiology* dyv048.
- PEARL, J. (2001). Direct and Indirect Effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, 411-420.
- ROBINS, J.M. & Greenland S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143-155.
- VALERI, L., COULL, A.B., REESE, S., ZHAO, S., LONDON, S.J. (2016) Does DNA methylation mediate the effect of maternal smoking on birth weight? The impact of exposure measured with error in mediation analyses for environmental epigenetic studies. *Under revision*.
- VALERI, L., and VANDERWEELE, T.J. (2013). Mediation analysis allowing for exposure-mediator interaction and causal interpretation: SAS and SPSS macros.*Psychological Methods.*
- VALERI, L., LIN, X., and VANDERWEELE, T.J. (2014). Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model.*Statistics in Medicine.*
- VANDERWEELE, T.J., VALERI, L., OGBURN, E.L. (2012) The role of measurement error and misclassification in mediation analysis. *Epidemiology* 23: 561-564.

Simulation study

Generated data mimicking the recent study (Küpers et al, 2015) which found that differential methylation of three GFI1 CpGs explained 12-19% of lower birthweight in smoking mothers.
Let $n = 500$, $runs = 10,000$, $C = (age, weight, education)$

- $A \sim Ber(p_A)$, where $p_A = expit(\gamma' C) \sim 50\%$
- $M \sim N(\beta_0 + \beta_1 A + \beta_2 C, 0.1)$, where $\beta_0 = 0.65, \beta_1 = -0.2$
- $Y \sim N(\theta_0 + \theta_1 A + \theta_2 M + \theta_4 C, 300)$, where $\theta_0 = 3685, \theta_1 = 300, \theta_2 = 400$
- A^* is the misclassified exposure generated assuming sensitivity taking values $SN \in (0.90, 1)$.

Simulation study: Estimates

Table : Estimate of natural direct effect (NDE), natural indirect effect (NIE), total effect (TE), proportion mediated (PM) for true binary exposure and for misclassified exposure due to self-reporting ($SP = 1$ and $SN = (0.90, 0.925, 0.95, 0.975)$). We assume that the mediator is a biomarker for the exposure ($\beta_1 \neq 0$) under either the null hypothesis of no mediated effect (due to $\theta_2 = 0$) or under the alternative.

Effects (H_1)	True	$SN = 0.90$	$SN = 0.925$	$SN = 0.95$	$SN = 0.975$
TE	-380	-312	-324	-341	-360
NDE	-302	-212	-227	-247	-272
NIE	-78	-100	-97	-93	-87
PM	20%	32%	30%	28%	24%
Effects (H_0)	True	$SN = 0.90$	$SN = 0.925$	$SN = 0.95$	$SN = 0.975$
TE	-300	-246	-256	-269	-284
NDE	-300	-212	-227	-247	-272
NIE	0	-34	-29	-22	-12
PM	0%	14%	12%	8%	4%

Simulation study: Numerical Bias

Table : Bias and variance of naïve estimates of natural direct effect (NDE), natural indirect effect (NIE), total effect (TE), proportion mediated (PM) for binary misclassified exposure. Misclassification setting assumes perfect specificity and $SN = (0.90, 0.925, 0.95, 0.975)$.

	$SN = 0.90$	$SN = 0.925$	$SN = 0.95$	0.975
H_1 naïve	Rel. Bias (var)	Rel. Bias (var)	Rel. Bias (var)	Rel. Bias (var)
TE	-0.17 (1336)	-0.14 (1305)	-0.10 (1284)	-0.05 (1130)
NDE	-0.30 (1939)	-0.25 (1973)	-0.18 (1942)	-0.10 (1952)
NIE	0.28 (484)	0.26 (526)	0.20 (561)	0.11 (671)
PM	0.55 (0.007)	0.50 (0.007)	0.35 (0.006)	0.20 (0.006)
H_0 naïve	Rel. Bias (var)	Rel. Bias (var)	Rel. Bias (var)	Rel. Bias (var)
TE	-0.18 (1284)	-0.15 (1243)	-0.10(1223)	-0.05(1107)
NDE	-0.30 (1939)	-0.25 (1973)	-0.18 (1941)	-0.10 (1952)
NIE	-36/0 (426)	-31/0 (481)	-23/0 (517)	-13/0 (637)
PM	0.15/0 (0.008)	0.12/0 (0.008)	0.09/0 (0.008)	0.05/0 (0.009)

Simulation study: Type I error

Table : Type I error for Sobel test of the difference method *NIE* when the exposure is misclassified. We consider the case in which the mediator is a biomarker for the exposure ($\beta_1 \neq 0$) and the case in which it is not ($\beta_1 = 0$). Misclassification setting assumes perfect specificity and $SN = (0.90, 0.925, 0.95, 0.975)$.

Binary	True	$SN = 0.90$	$SN = 0.925$	$SN = 0.95$	$SN = 0.975$
$\beta_1 = 0$	0.01%	0.01%	0.01%	0.01%	0.01%
$\beta_1 \neq 0$	4.9%	40%	28%	17%	9%

Simulation study: SIMEX correction

Table : Bias and variance of SIMEX-corrected estimates of natural direct effect (NDE), natural indirect effect (NIE), total effect (TE), proportion mediated (PM). Misclassification setting assumes perfect specificity and $SN = (0.90, 0.925, 0.95, 0.975)$.

	$SN = 0.90$	$SN = 0.925$	$SN = 0.95$	0.975
H_1 SIMEX	Rel. Bias (var)	Rel. Bias (var)	Rel. Bias (var)	Rel. Bias (var)
TE	-0.03 (1879)	-0.02 (1721)	-0.01 (1619)	-0.00 (1273)
NDE	-0.08 (3294)	-0.06 (3122)	-0.03 (2832)	-0.00 (2355)
NIE	0.19 (865)	0.14 (838)	0.06 (836)	0.02 (824)
PM	0.20 (0.008)	0.15 (0.008)	0.10 (0.007)	0.00 (0.007)
H_0 SIMEX	Rel. Bias (var)	Rel. Bias (var)	Rel. Bias (var)	Rel. Bias (var)
TE	-0.02 (1797)	-0.02 (1624)	-0.01 (1530)	-0.00(1241)
NDE	-0.09(3294)	-0.06(3121)	-0.03 (2832)	-0.00 (2355)
NIE	-18/0 (828)	-12/0 (825)	-6/0 (809)	-2/0 (803)
PM	0.07/0 (0.01)	0.04/0 (0.01)	0.02/0 (0.01)	0.00 (0.01)