

Supplemental Material for:**Statistical Methods to Study Timing of Vulnerability with Sparsely Sampled Data on Environmental Toxicants**

Brisa Ney Sánchez (1), Howard Hu (2), Heather J. Litman (3), Martha Maria Téllez-Rojo (4).

(1) Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA

(2) Department of Environmental Health Sciences, University of Michigan School of Public Health, Ann Arbor, MI, USA

(3) New England Research Institutes, Watertown, MA, USA

(4) Division of Statistics, Center for Evaluation Research and Surveys, National Institute of Public Health, Cuernavaca, Morelos, MX

Corresponding author:

Brisa N Sánchez, PhD
University of Michigan School of Public Health
Department of Biostatistics
1415 Washington Heights
Ann Arbor, MI 48109
Telephone: (734) 763-2451
Fax: (734) 763-2215
E-mail: brisa@umich.edu

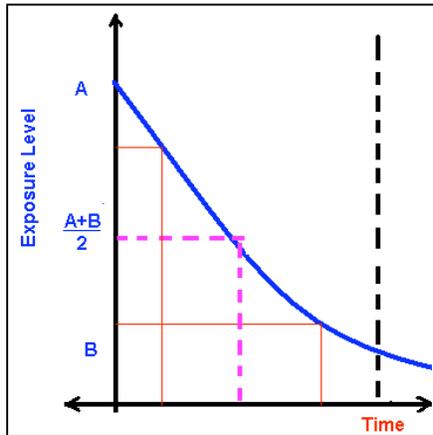
CONTENTS

pg.

EXAMPLES OF BIAS IN MULTIPLE REGRESSION COEFFICIENTS DUE TO NON CONSTANT EXPOSURE EFFECTS OR VARIATION IN TIMING OF EXPOSURE.....	3
MODEL FITTING IN SAS.....	4
Data Layout	4
Method 1: Separate and Simultaneously Adjusted Linear Regression Models	5
Method 2: Multiple Source Predictors with GEE.....	5
Method 2: Multiple Source Predictors with ML.....	6
Method 3: Individual's patterns of exposure in relation to outcome.....	7
Method 4: Population patterns of exposure given the outcome	9
SENSITIVITY ANALYSES.....	10
Methods 1 and 2.....	10
Method 3.....	11
Method 4.....	12
REFERENCES	13
SAS MACROS	14
FIGURES	
Supplemental Material, Figure 1: Consequences of non constant exposure levels	3
Supplemental Material, Figure 2: Consequences of non constant exposure effect or variation in <i>timing</i> of exposure measurements	3
Supplemental Material, Figure 3: Data Layouts.....	4
Supplemental Material, Figure 4: Estimated weight function $w(t)$	8
Supplemental Material, Figure 5: Exposure patters obtained from Method 3	11
Supplemental Material, Figure 6: Relative exposure estimated from Method 4 after adjusting for child blood lead in outcome model	12
TABLES	
Supplemental Material, Table 1: Results of Methods 1 and 2 including child blood lead at 24 months as a confounder	10
Supplemental Material, Table 2: Including child blood lead at 24 months as another window ...	10
Supplemental Material, Table 3. Including fixed quadratic effect of time in exposure model and child blood lead in outcome model.....	11
Supplemental Material, Table 4. Model parameter estimates after including time-varying ceramic use as a predictor of prenatal exposure and child's blood lead at 24 months as a predictor of outcome.....	12

EXAMPLES OF BIAS IN MULTIPLE REGRESSION COEFFICIENTS DUE TO NON CONSTANT EXPOSURE EFFECTS OR VARIATION IN TIMING OF EXPOSURE

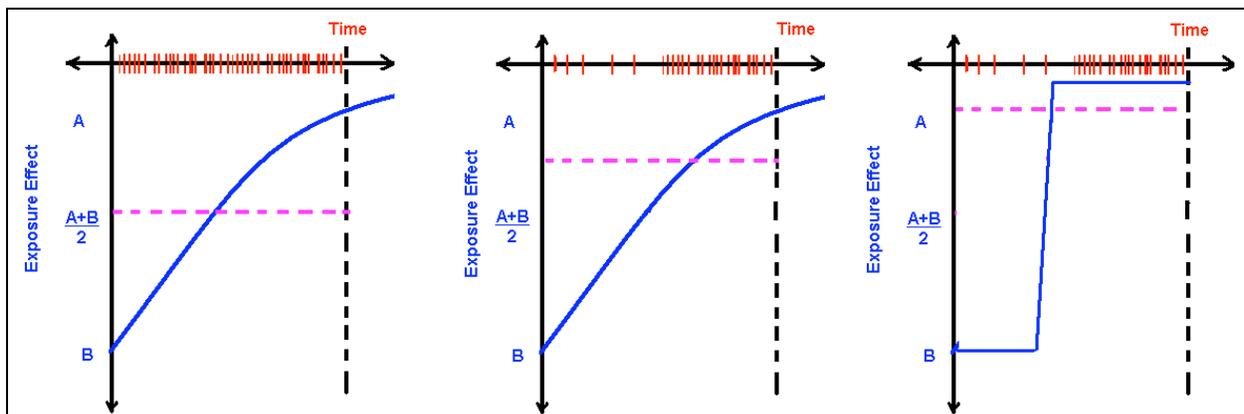
Finding clinically meaningful windows of vulnerability is a goal of this type of research. However, as the figures below show, predefining windows from the outset coupled with variations in the sampling times, may hinder our ability to detect such windows.



Supplemental Material, Figure 1

Figure shows non constant exposure levels (solid blue curve) within a time window (delineated by black vertical dashed line). For measures taken near the midpoint of the time window (pink dashed line) the observed measure will approximately represent average exposure within the window $(A+B)/2$. For all other times, the measurement over- or underestimates average exposure. Depending on the nature of the exposure pattern (e.g., increasing instead of decreasing), measurements taken after the midpoint of the time window may underestimate the average exposure. If the

pattern was known, then the error in exposure measurement could be corrected. This error will introduce measurement error bias in regression coefficients. Since the error could potentially be differential, for example, when more highly exposed participants are also measured later in the time window, the bias in regression coefficients may not always be toward zero.



Supplemental Material, Figure 2. Figure shows a non constant exposure effect (solid blue curve) within a time window (delineated by black vertical dashed line), and two different distributions of sampling times (red vertical dashes on time axis). Left panel shows that under uniform distribution of sampling times, the estimated average effect (horizontal, pink dashed line) would be somewhere between the maximum and minimum effect within the window. Middle panel shows that if the distribution of sampling times is shifted towards an end of the window, then, in this case, the estimated effect would underestimate the average effect within the window. Right panel shows an extreme example where true effect is non-zero only for approximately 1/3 of the predefined time window, but, under skewed sampling times within the predefined window, the estimated effect is close to zero.

MODEL FITTING IN SAS

We outline the fitting procedures for each of the methods discussed in the parent paper. We also discuss extensions to the models presented; namely, including covariates in the exposure models for Methods 3 and 4 (e.g. lead-glazed ceramics use), and incorporating categorical health outcomes for Method 4.

Data Layout

The data available for each individual are: the health outcome Y_i , exposure measurements $X_i=(X_{i1}, \dots, X_{iK})$ taken within time windows $1, \dots, K$, at actual (continuous) $Time = t_1, \dots, t_{K_i}$. Covariates for the outcome model are denoted Z_i in the parent paper. Here in the supplemental material, we separate these covariates into two sets: Z_{Y_i} , and Z_{C_i} . Covariates Z_{Y_i} would be those that are independent predictors of the outcome, but not confounders. Covariates that are predictors (or common causes) for both Y and X (and therefore confounders), are called Z_C . In the lead example, maternal education could predict both MDI24 and maternal exposure, hence it would be a covariate in Z_C ; other variables, for example, measures of the home environment, which would predict MDI24, but would not be a cause of exposure, would belong to Z_Y . Other covariates may also be available, for example, variables that may predict the exposure. For example, use of lead glazed ceramics in the lead example would be a predictor of exposure, but would not (directly) influence MDI24. We call these Z_X , since they predict exposure, X . The extra sets of covariates Z_X will be used in Model 3 and 4. The data needs to be organized in various ways depicted in Supplemental Material, Figure 2, depending on the estimation method used.

Wide Layout													
<i>ID</i>	<i>Y</i>	<i>X1</i>	<i>X2</i>	...	<i>XK</i>	<i>Z_C</i>	<i>Z_Y</i>	<i>Z_X</i>	<i>Time1</i>	<i>Time2</i>	...	<i>TimeK</i>	<i>Complete</i>
<i>i</i>	Y_i	X_{i1}	X_{i2}		X_{iK}	Z_{C_i}	Z_{Y_i}	Z_{X_i}	t_{i1}	t_{i2}		t_{iK}	1
<i>j</i>	Y_j	.	.		X_{jK}	Z_{C_j}	Z_{Y_j}	Z_{X_j}	.	.		t_{jK}	0

Layout 1									Layout 2							
<i>ID</i>	<i>Y</i>	<i>X</i>	<i>T</i>	<i>Z_C</i>	<i>Z_Y</i>	<i>Z_X</i>	<i>Time</i>	<i>Window</i>	<i>ID</i>	<i>Resp</i>	<i>Outcome</i>	<i>Exposure</i>	<i>Time</i>	<i>Z_C</i>	<i>Z_Y</i>	<i>Z_X</i>
<i>i</i>	Y_i	X_{i1}	1	Z_i	Z_{Y_i}	Z_{X_i}	t_{i1}	1	<i>i</i>	Y_i	1	0	99	Z_i	Z_{Y_i}	99
<i>i</i>	Y_i	X_{i2}	2	Z_i	Z_{Y_i}	Z_{X_i}	t_{i2}	2	<i>i</i>	X_{i1}	0	1	t_{i1}	Z_i	99	Z_{X_i}
:	:	:	:	Z_i	:	:	:		<i>i</i>	X_{i2}	0	1	t_{i2}	Z_i	99	Z_{X_i}
<i>i</i>	Y_i	X_{iK}	<i>K</i>	Z_i	Z_{Y_i}	Z_{X_i}	t_{iK}	<i>K</i>	:	:	:	:	:	Z_i	99	:
									<i>i</i>	X_{iK}	0	1	t_{iK}	Z_i	99	Z_{X_i}

Supplemental Material, Figure 3. Data Layout for individual i (and j) used in various estimation approaches. In the Wide Layout, individual j has some exposure measures missing, hence the variable "Complete" is coded as 0; since individual i has complete data, then the indicator variable is coded as 1. In Layout 2, "Outcome" and "Exposure" are dummy variables indicating if the variable *Resp* is the health outcome Y or an exposure X ; the values 99 can be substituted for any other number except ".", since they are not actually used in the computations.

Method 1: Separate and Simultaneously Adjusted Linear Regression Models

Fitting separate and simultaneously adjusted regressions only requires standard use of linear regression. For example, for the first exposure window, the SAS code would be:

```
proc reg data=WideLayout;
title "Regression for window 1";
model Y = X1 ZC ZY ;
run;

proc reg data=WideLayout;
title "Simultaneously adjusted regression";
model Y = X1 X2 ... XK ZC ZY ;
run;
```

Method 2: Multiple Source Predictors with GEE

The objective is to jointly estimate the exposure associations, β_{1k} , from the regressions $Y_i = \beta_{0k} + \beta_{1k}X_{ki} + \beta_{2k}Z_i + \varepsilon_{ki}$, at each time window $k=1, 2, \dots, K$ (where Z_i are covariates Z_{Yi} , and Z_{Ci} , and β_{2k} is a $1 \times p$ row vector). These regressions can be estimated by implementing a non-standard version of generalized estimating equations (GEE) where an artificial multivariate outcome is created for each individual by repeating the outcome, Y_i , K times (Horton et al., 1999; Litman et al., 2007a; Pepe et al., 1999). In matrix notation, the model is $E(\tilde{Y}_i | \tilde{X}_i = \tilde{X}_i^T)$, where

$$\tilde{Y}_i = \begin{pmatrix} Y_i \\ Y_i \\ \vdots \\ Y_i \end{pmatrix}_{K \times 1} \quad \tilde{X}_i = \begin{pmatrix} 1 & X_{1i} & Z_i^T & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & X_{2i} & Z_i^T & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & X_{Ki} & Z_i^T \end{pmatrix}_{K \times (2+p)K}, \text{ and}$$

$$\tilde{\beta} = (\beta_{01} \quad \beta_{11} \quad \beta_{21} \quad \beta_{02} \quad \beta_{12} \quad \beta_{22} \quad \dots \quad \beta_{0K} \quad \beta_{1K} \quad \beta_{2K})^T.$$

To fit the model in SAS, the data for the for i^{th} individual should be formatted in "long format" with K records for each individual, one for each exposure window, and copies of the outcome and covariates repeated for each exposure window as shown in Layout 1 in Supplemental Material, Figure 3. The exposure coefficients for each exposure window can then be estimated with SAS PROC GENMOD, with a working independence assumption and model based standard errors (Litman et al., 2007b):

```
proc genmod data=DataLayout1;
title "Multiple informant with GEE";
class id Window;
model Y=Window X*Window ZC*Window ZY*Window / noint;
repeated subject=id / type=ind modelse;
run;
```

Alternatively, empirical standard errors (Pepe et al., 1999) may be used by deleting the option `modelse` from the `repeated` statement. Tests of the differences in exposure (and covariate) associations across time windows can be obtained by adding the "main effects" of X (and Z) to the model, and adding the `type3` option. The tests for the interactions will have $K-1$ degrees of freedom, and the p-values can be obtained from the type III tests output. In the case of missing data, the option `withinsubject=Window` is needed in the `repeated` statement.

Method 2: Multiple Source Predictors with ML

A key difference between GEE and ML estimating approaches for multiple source predictors is the ability of the ML approach to incorporate missing data. The general idea is to first estimate the joint distribution of the outcome and exposure measures ($Y_i, X_{i1}, \dots, X_{iK}$) given covariates Z_i (both Z_{Yi} , and Z_{Ci}), and subsequently find the conditional mean of Y_i given each X_{ij} . For example, for $E(Y_i | X_{ij}, Z_i) = \beta_{0j} + \beta_{1j} X_{ij} + \beta_{2j} Z_i$, the estimate of β_{1j} is given by $Cov(Y_i, X_{ij}|Z_i)/Var(X_{ij}|Z_i)$, and the estimate of β_{0j} is given by $E(Y|Z_i) - \beta_{1j} * E(X_{ij}|Z_i)$.

Litman et al. (2007a) describe an EM algorithm to find ML estimates in the case when two predictors are available (e.g., two exposure windows). They implemented the algorithm in the R software, but note that their approach to developing the algorithm becomes increasingly complex when there are more multiple source predictors (i.e., more exposure windows), particularly in the presence of missing data and constraints in the variance-covariance matrix. Using the macros that appear at the end of this Supplemental Material, the approach can also be implemented in SAS, where it is easier to include more exposure windows. We give the case when the regression coefficients may differ across three exposure windows, but the set up can be easily extended to include more predictors. When the model is unconstrained, i.e., assuming outcome-exposure associations differ across windows, the macros would be used as follows to obtain estimates and bootstrap standard errors:

```

title "Multiple informant with MLE, assuming different regression
coefficients across windows";

%MultipleInformantMLE(data=WideLayout, out=FullModel,
                      X1=X1, X2=X2, X3=X3, Y=Y, id=ID, covariates= ZY*T);

%BootstrapForStandardErrors(BootstrapReplicates=1000, data= WideLayout,
                             results=BootstrapResults,
                             N_complete=120, N_missing=49, NfullSample=169,
                             variables= Y X1 X2 X3 ZC ZY ZX ,
                             strata=Complete, X1=X1, X2=X2, X3=X3, Y=Y, id=ID,
                             covariates= ZY*T);

%SummaryBootstrap( actualresults=FullModel, results=BootstrapResults ,
                  finalresults=FinalFullModel );

```

The output from the macro prints the estimated parameters, bootstrap standard errors and confidence intervals.

The following code is needed to estimate the model when constraints are imposed (i.e., correlation between outcome and exposure is the same across all windows). The output of the **MultipleInformantMLEConstraints** macro includes a likelihood ratio test to compare the constrained to unconstrained model.

```

title "Multiple informant with MLE, assuming the same correlation
between exposure and outcome across windows";

%MultipleInformantMLEConstraints(data=WideLayout, out=FullModel,
                                X1=X1, X2=X2, X3=X3, Y=Y, id=ID, covariates= ZY*T,
                                type=lin(8), LinCovParms=LinCovParms8);

```

Method 3: Individual's patterns of exposure in relation to outcome

The third method can be estimated in SAS PROC NLMIXED. The code below shows a sample that follows the example in the paper, where the only subject specific parameters are the intercept and linear rate of change for each individual.

```
proc nlmixed data=DataLayout2;
title "Joint estimation for Method 3";
title2 "subject-specific random slope and intercept for exposure";
parms -starting values-;
theta0i= mutheta0 + errtheta0i;          *features are population mean;
thetal1= muthetal + errthetal1;         *plus individual variation;
meanY = B0 + B1*theta0i + B2*thetal1 + B3*ZY + B4*ZC;
loglikOutcome = (-1/2)*((Resp - meanY)/sigma)**2 - log(sigma);
meanXik = theta0i + thetal1*time ;
loglikExposure = (-1/2)*((Resp - meanXik)/tau)**2 - log(tau);
model Response ~ general(Outcome*loglikOutcome +
                          Exposure*loglikExposure);
random errtheta0i errthetal1 ~ normal([0 , 0],
                                       [omega1**2,
                                       omega12 , omega2**2]) subject=ID;
run;
```

One will usually need to obtain starting values for the parms statement. For example (parms B0=0 B1=1 B2=0 B3=0 B4=0 muthetal=0 mutheta2=0 omega1=1 omega2=1 omega12=0). Improved starting values for muthetal, mutheta2, omega1, omega2, and omega12 can be obtained by first fitting a random effects model to the exposure data only (i.e., a repeated measures model for the X's using, for example, PROC MIXED).

It may be of interest to include covariates that predict exposure features. For example, we included use of lead glazed ceramics at baseline as a predictor of the intercept θ_{0i} (See Supplementary Table 3). Modeling the features as dependents on covariates, e.g. $\theta_{0i} = \theta_0 + \gamma Z_X + \delta_{0i}$, may be useful in helping better predict the exposure features for participants with missing data; information across participants with similar characteristics can be borrowed by using predictors for the features. This can be easily accomplished by modifying the statements for theta0i and thetal1

```
theta0i= theta0 + gamma0*ZX +errtheta0i; * predicting features ;
thetal1= thetal + gamma1*ZX +errthetal1; * with time fixed covariates;
```

Note that the total variation in the random effects θ_{0i} is still used in the outcome for the model, i.e. the statement meanY = B0 + B1*theta0i + B2*thetal1 + B3*ZY + B4*ZC; remains unchanged.

Furthermore, including non linear exposure features may be of interest. For example, exposure over time as a quadratic function could be possible, i.e. $X_{ik} = \theta_{0i} + \theta_{1i} t_{ik} + \theta_{2i} t_{ik}^2 + \varepsilon_{ik}$. This is only a possibility if most participants have 4 or more exposure measures. Other semi- or nonparametric shapes would also be possible. However, the number of exposure measures per participant would have to be greater. The model for the outcome would be $Y_i = \beta_0 + \beta_{11} \theta_{0i} + \beta_{12} \theta_{1i} + \beta_{13} \theta_{2i} + \beta_2 Z_i + \varepsilon_i$. Interpreting this type of model could be done by graphing the

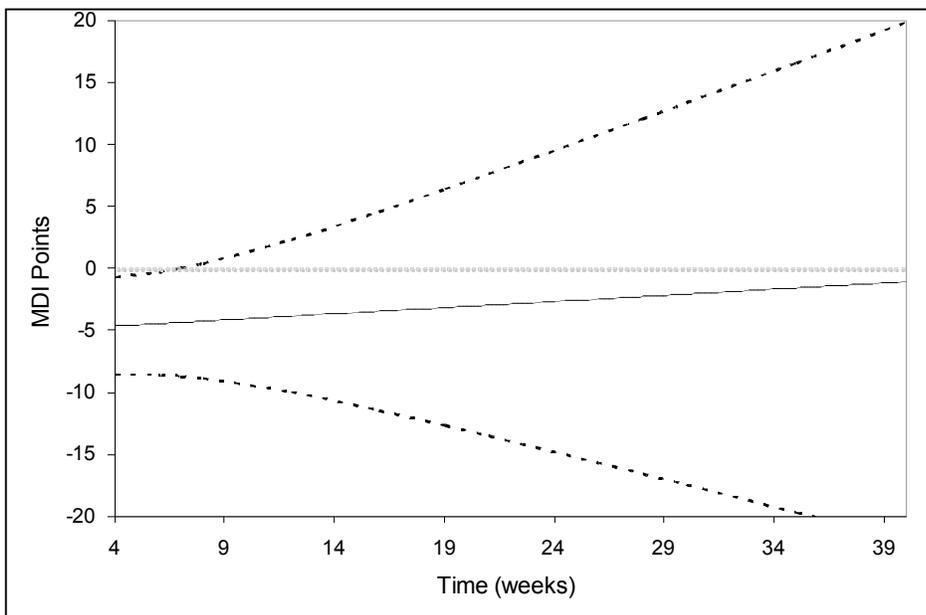
estimated exposure patterns annotated with outcome (See Supplemental Material, Figure 5). Coding-wise, this is straightforward to implement, the necessary changes would be:

```

theta2i= theta2 + errtheta2i;                                *add this line;
*modify the following lines;
meanXik = theta0i + thetali*time + theta2i*time*time;
meanY = B0 + B1*theta0i + B2*thetali + B3*theta2i + B4*ZY + B5*ZC;
random errtheta0i errthetali errtheta2i ~ normal([0 , 0, 0],
                                                [omega1**2,
                                                 omega12 , omega2**2
                                                 omega13 , omega23, omega3**2]) subject=ID;

```

Alternatively, for interpretation, it may be more advantageous to transform the estimated population coefficients into a coefficient function $w(t)$ that measures the relative effect of an exposure increment at time t compared to other times. The function $w(t)$ is constructed as $w(t) = \beta_{11} + \beta_{12}t + \beta_{13}t^2$ (see James 2002 for a full discussion on why $w(t)$ is estimated in this manner). For the analyses in the parent paper, subject-specific exposure patterns were linear, hence $w(t)$ was linear, $w(t) = \beta_{11} + \beta_{12}t$ (see Supplemental Material, Figure 4).



Supplemental Material, Figure 4

Estimated weight function $w(t)$, with point wise confidence intervals, for Method 3 analysis in parent paper. Wide confidence intervals due to high standard error of slope parameter β_{12} .

Even if person-specific non linear patterns cannot be modeled (e.g., due to limited sample size per participant), including a population average non-linear pattern may be advantageous to improve model fit. In the lead example we included a population average U-shaped curve by modeling i.e. $X_{ik} = \theta_{0i} + \theta_{1i}t_{ik} + \theta_{2i}t_{ik}^2 + \varepsilon_{ik}$ (note θ_2 is not random). See Supplemental Material, Table 3.

Method 4: Population patterns of exposure given the outcome

The code below shows how to estimate the fourth approach in SAS PROC MIXED. The approach uses the data formatted as in Layout 1, with an additional variable Y^c (Y_c). The variable Y_c are the residuals obtained by fitting a model to the outcome, where only the covariates (and not exposure X) are used as predictors.

```
proc mixed data=DataLayout1;
title "Quadratic model with continuous residual interaction";
class ID;
model X = time time*time Y_c time*Y_c time*time*Y_c /solution;
random int / subject=ID s;
estimate "Exposure when outcome is 10 points above the mean (Y_c = 10)
and time is 7 weeks "
intercept 1 time 7 time*time 49
Y_c 10 time*Y_c 70 time*time*Y_c 490 / cl ;
estimate "Exposure when outcome is 10 points below the mean (Y_c =-10)
and time is 7 weeks "
intercept 1 time 7 time*time 49
Y_c -10 time*Y_c -70 time*time*Y_c -490 / cl ;
run;
```

The structure of the estimate statement is

```
Estimate "Exposure when outcome is NUM1 points above
the mean (Y_c = NUM1) and time is NUM2 weeks "
intercept 1 time NUM2 time*time NUM2*NUM2
Y_c NUM1 time*Y_c NUM1*NUM2
time*time*Y_c NUM2*NUM2*NUM1 / cl ;
```

Repeated applications of the estimate statement at, for example, every week, $NUM2=(0, 1, \dots, 40)$, give the estimates and confidence limits used to construct Figure 1 of the paper.

Covariates can be included in the model for the exposure $X_{ik} = f_0(t_{ik}) + f_1(t_{ik})Y_i^c + \gamma_C Z_C + \gamma_X Z_X + \delta_{ik}$. These covariates would be included to help reduce the error variance, and therefore possibly reduce the variance of the estimated $f_0(t)$ and $f_1(t)$.

```
model X = time time*time Y_c time*Y_c time*time*Y_c ZC ZX /solution;
```

In the lead example, we included use of lead glazed ceramics as an example. The results are presented in Supplemental Material, Table 4 and Supplemental Material, Figure 6.

If the outcome Y was categorical (e.g., cases and controls), constructing residuals Y^c would not necessarily be straightforward. Instead one can simply use the observed Y . In this case, including covariates Z_C (i.e., confounders) in the exposure model becomes necessary to avoid confounding. The modifications to the code would be to include Y in the `class` statement.

SENSITIVITY ANALYSES

We present results from alternative analyses, namely including child blood lead at 24 months into all analyses in the parent paper, and provide examples of including covariates in exposure model for Methods 3 and 4.

Methods 1 and 2. It is possible to include child blood lead as a confounder or as another window in methods 1 and 2 (Supplemental Material, Tables 1 and 2). Including it as a confounder has some potential advantages and limitations. A potential advantage is that the coefficients for the exposure effects during the prenatal windows can be interpreted as independent effects of prenatal exposure after accounting for early childhood exposure. The primary limitation is that doing this would not always be possible because high correlations between exposures outside the windows of interests and the exposure window being evaluated may actually preclude valid interpretations of regression coefficients (Woodruff et al., 2009). In this case, correlations between prenatal and child blood lead are relatively low (0.17, 0.24, 0.27). In all models shown in Supplemental Material, Tables 1 and 2, child blood lead was not a significant predictor of mental development; however the direction of association was always negative, as would be expected.

Supplemental Material, Table 1: Results of Methods 1 and 2 including child blood lead at 24 months as a confounder

Trimester	Multiple Regression						Multiple Informants Approach (N=169)					
	Simultaneous Adj. ^a			Separate Regressions ^b			GEE			MLE		
	β	95% CI		β	95% CI		β	95% CI		β	95% CI	
1	-5.47	-10.3	-0.67	-3.39	-6.69	-0.1	-3.39	-6.75	-0.04	-4.03	-7.61	-0.45
2	1.12	-5.16	1.12	-2.89	-6.76	0.98	-2.89	-6.75	0.96	-2.77	-6.98	1.44
3	1.54	-3.45	1.54	-1.88	-5.46	1.7	-1.88	-5.43	1.67	-1.63	-4.8	1.54
p_{int}	n/a			n/a			0.72			0.12		

^a N=120

^bFor Trimester 1, N=139; Trimester 2, N=159; Trimester 3, N=146

Supplemental Material, Table 2: Including child blood lead at 24 months as another window

Timing	Multiple Regression						Multiple Informants Approach (N=169)					
	Simultaneous Adj. ^a			Separate Regressions ^b			GEE			MLE		
	β	95% CI		β	95% CI		β	95% CI		β	95% CI	
T1	-5.47	-10.3	-0.67	-2.74	-5.78	0.29	-3.14	-6.33	0.06	-4.14	-7.63	-0.65
T2	1.12	-5.16	1.12	-1.37	-4.81	2.07	-3.00	-6.81	0.81	-3.01	-7.22	1.20
T3	1.54	-3.45	1.54	-1.15	-4.20	1.90	-2.70	-5.62	0.22	-1.91	-5.05	1.23
24M	-1.02	-4.27	2.23	-1.19	-3.87	1.49	-1.17	-3.56	1.22	-1.36	-3.57	0.86
p_{int}	n/a			n/a			0.86			0.08		

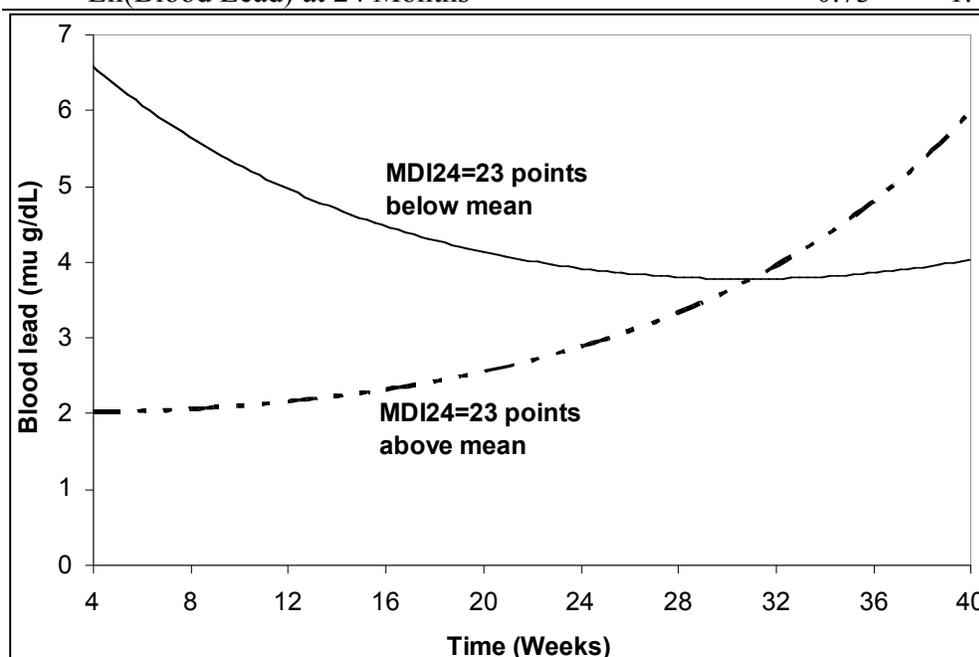
^a N=120

^bFor Trimester 1, N=151; Trimester 2, N=159; Trimester 3, N=146, for 24months, N=169

Method 3. Supplemental Material, Table 3 shows the model parameters from implementing Method 3 where we: 1) included a fixed quadratic effect of time in the exposure model 2) included ceramics use at baseline as a predictor of θ_{0i} , and 3) included child blood lead in outcome model. As compared to the analysis in the main paper, the effect of blood lead level early in pregnancy (θ_{0i}) was slightly weaker ($\beta = -1.93$, $SE=1.14$, $p=0.09$ vs $\beta = -2.11$, $SE=1.08$, $p=0.05$); this reflects that although it is statistically significant, baseline ceramics leads to no improvement in quantifying θ_{0i} . Supplemental Material, Figure 5 shows exposure patterns for two individuals and their outcomes.

Supplemental Material, Table 3. Including fixed quadratic effect of time in exposure model and child blood lead in outcome model.

<u>Exposure Model</u>	Estimate	SE	p
θ_0 (Average Intercept)	2.00	0.07	<.0001
θ_1 (Average change per 12 weeks)	-0.36	0.09	<0.01
θ_2 (Quadratic term)	0.11	0.03	<0.01
γ (Effect of baseline ceramics use on θ_{0i})	0.04	0.02	0.01
Random Intercept SD	0.49		
Random Slope SD	0.19		
Correlation of Random Int. (θ_{0i}) and Slope (θ_{1i})	-0.58		
Residual SD	0.26		
<u>Outcome Model</u>			
Blood lead level at week 7 (θ_{0i} , Random Intercept)	-1.93	1.14	0.09
Changes in blood lead level (θ_{1i} , Random Slope)	1.07	1.61	0.51
Maternal Age (per 5yrs)	2.93	0.79	<0.01
IQ (per 10 pts)	0.68	0.65	0.29
Child's Gender	-5.12	1.70	<0.01
Weight at 24 Months	-2.08	0.94	0.03
Height Z-score at 24 Months	2.72	1.17	0.02
Breast Feeding Duration (per 6mo)	-0.09	0.15	0.55
Ln(Blood Lead) at 24 Months	-0.73	1.40	0.60



Supplemental Material, Figure 5

Offspring MDI and exposure patterns for two participants as estimated from Method 3 including a fixed quadratic effect of time in exposure model.

Method 4. Supplemental Material, Table 4 shows parameter estimates from method 4 after including child blood lead at 24 months as an outcome predictor and using time-varying ceramics as a prenatal exposure predictor. Supplemental Material, Figure 6 shows the relative exposure comparing children with low vs. high MDI24 scores. The primary difference is that the time at which exposure differences become non-significant is 18 instead of 17 weeks.

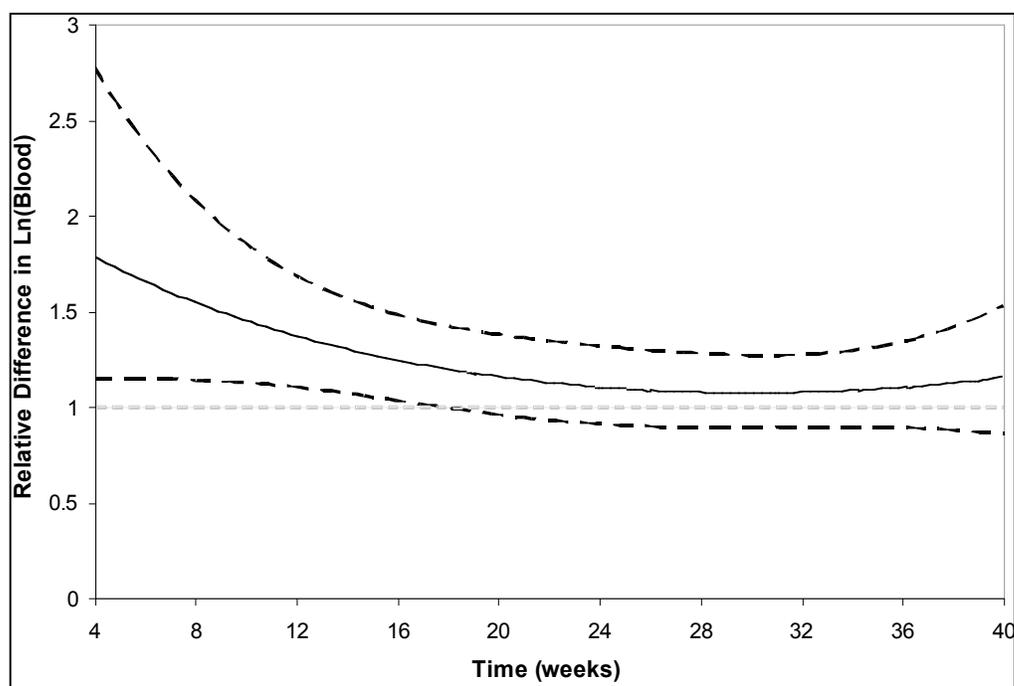
Supplemental Material, Table 4. Model parameter estimates after including time-varying ceramic use as a predictor of prenatal exposure and child's blood lead at 24 months as a predictor of outcome.

<u>Outcome Model</u>	Estimate	SE	p
Maternal Age (5yrs)	2.77	0.81	<0.001
Breast Feeding Duration (6mo)	-0.49	0.91	0.60
Maternal IQ (10 pts)	0.79	0.67	0.24
Child's Gender	-4.23	1.79	0.02
Weight at 24 Months	-1.66	0.68	0.02
Height Z-score at 24 Months	2.58	1.18	0.03
Ln(Blood Lead) at 24 Months	-1.19	1.37	0.38
<u>Exposure Model</u>			
Average exposure pattern, $f_0(t)$			
Intercept, α_{00}	3.00	0.4	<.001
Linear trend, α_{01}	-1.26	0.63	0.046
Quadratic trend, α_{02}	0.44	0.3	0.146
Relationship with MDI, $f_1(t)$ ^a			
Change in the intercept, α_{10}	-0.12	0.04	0.005
Change in linear trend, α_{11}	0.11	0.07	0.098
Change in quadratic trend, α_{12}	-0.04	0.03	0.277
Time-varying Ceramic Use (γ)	0.04	0.01	0.004

^a $H_0: f_1(t)=0$ vs $f_1(t)\neq 0$, $p=0.04$, $H_0: f_1(t)=\text{constant}$ vs $f_1(t)\neq\text{constant}$, $p=0.10$

Supplemental Material, Figure 6

Relative exposure comparing those in the 10th percentile of the MDI distribution to those in the 90th percentile, with 95% point wise confidence intervals. Lower point wise confidence interval crosses 1 at approximately 18 weeks.



REFERENCES

- Horton N.J., Laird N.M., Zahner G.E. (1999) Use of Multiple Informant Data as a Predictor in Psychiatric Epidemiology. *International Journal of Methods in Psychiatric Research* 8:6-18.
- James GM. 2002. Generalized linear models with functional predictors. *J Roy Stat Soc B* 64: 411-432.
- Lin X.H., Ryan L., Sammel M., Zhang D.W., Padungtod C., Xu X.P. (2000) A scaled linear mixed model for multiple outcomes. *Biometrics* 56:593-601.
- Litman H.J., Horton N.J., Hernandez B., Laird N.M. (2007a) Incorporating missingness for estimation of marginal regression models with multiple source predictors. *Statistics in Medicine* 26:1055-1068. DOI: 10.1002/sim.2593.
- Litman H.J., Horton N.J., Hernandez B., Laird N.M. (2007b) Estimation of Marginal Regression Models with Multiple Source Predictors, in: C. R. Rao, et al. (Eds.), *Handbook of Statistics*, Elsevier, San Diego.
- Pepe M.S., Whitaker R.C., Seidel K. (1999) Estimating and comparing univariate associations with application to the prediction of adult obesity. *Statistics in Medicine* 18:163-173.
- Woodruff T.J., Parker J.D., Darrow L.A., Slama R., Bell M.L., Choi H., Glinianaia S., Hoggatt K.J., Karr C.J., Lodbell D.T., Wilhelm M. (2009) Methodological issues in studies of air pollution and reproductive health. *Environmental Research* 109:311-320.

SAS MACROS

```

%macro MultipleInformantMLE(data= , out= , X1=, X2=, X3=, Y=, id=, covariates= );
/*
data= dataset in wide layout
out= name of dataset to store the results
X1= name of variable for exposure at window1
X2= name of variable for exposure at window2
X3= name of variable for exposure at window3
Y= name of variable for outcome
id= person identifier
covariates = list of covariates to adjust for; it must include interactions with
T and main effect of T, literally: eg., covariates = T age*T
The variable T MUST NOT exist in the widelayout dataset, it will get constructed
in this macro.
*/

data ResampledEM;      *need data in long format, similar to Layout 2, but with T as categorical;
  set &data;
  Response = &Y;      T=0;  output;
  Response = &X1;     T=1;  output;
  Response = &X2;     T=2;  output;
  Response = &X3;     T=3;  output;
run;

*run proc mixed to estimate the joint distribution of exposure and outcome;
proc mixed data=ResampledEM method=ml covtest ;
title2 "proc mixed for multiple informant approach";
  class &id T;
  model Response = T &covariates / noint s;
  repeated T /subject=&id type=un ;
  ods output CovParms=CovarP(keep = estimate ) ;
  ods output SolutionF=fixedF(keep = estimate ) ;
run;

*manipulate estimated covariance parameters ;
proc transpose data=CovarP out=transpCovarP;
run;

data transpCovarP;
  set transpCovarP;
  vary=coll;          *residual variance of outcome given covariates only;

```

```

varx1=col3; varx2=col6; varx3=col10; *residual variances of exposure at each time window given covariates
only;
covyx1=col2; covyx2=col4; covyx3=col7; *covariance of outcome with exposure at each window;
covx1x2=col5; covx1x3=col8; covx2x3=col9; *covariance of exposure between pairs of windows;
corryx1=covyx1/(sqrt(vary)*sqrt(varx1)); *covariate adjusted correlations;
corryx2=covyx2/(sqrt(vary)*sqrt(varx2));
corryx3=covyx3/(sqrt(vary)*sqrt(varx3));
drop _name_ col1-col10;
run;

proc transpose data=fixedF out=transpfixedF;
run;

*manipulate estimated mean parameters ;
data transpfixedF;
  set transpfixedF;
  muy=col1; mux1=col2; mux2=col3; mux3=col4;
  drop _name_ col1-col4;
run;

data &out ;
  merge transpCovarP transpfixedF;

  *exposure regression coefficients for each window;
  beta11=covyx1/varx1;
  beta12=covyx2/varx2;
  beta13=covyx3/varx3;

  *intercepts coefficients for each window;
  beta01=muy - beta11*mux1;
  beta02=muy - beta12*mux2;
  beta03=muy - beta13*mux3;

  *differences in covariate-adjusted outcome-exposure correlation;
  *between pairs of windows;
  CorYX1_CorYX2 = corryx1 - corryx2;
  CorYX1_CorYX3 = corryx1 - corryx3;
  CorYX2_CorYX3 = corryx2 - corryx3;

run;

proc print data=&out;
title2 "Multiple informants estimates for each exposure window";
run;

```

```

proc datasets;
  delete TempOut fixedF transpfixedF CovarP transpCovarP ResampledEM;
  run;

%mend MultipleInformantMLE;

/* The BootstrapForStandardErrors macro employs the resample and MultipleInformantMLE macros */
%macro BootstrapForStandardErrors(BootstrapReplicates=, data= , results=,
                                  N_complete=, N_missing=, NfullSample= , variables=, strata=,
                                  X1=, X2=, X3=, Y=, id=, covariates=);
*initialize dataset for results;
data &results;
set _null_;
run;

*start resampling;
%do run = 0 %to &BootstrapReplicates;

* resample wide dataset, resampled dataset is called ResampledData;
%resample(originaldata=&data, N1=&N_complete, N2=&N_missing, Ntotal=&NfullSample, strata=&strata, variables=&variables,
id=&id)

* run MultipleInformantMLE macro ;
%MultipleInformantMLE(data=ResampledData , out=tempresults , X1=&X1, X2=&X2, X3=&X3, Y=&Y, id=&id,
covariates=&covariates );

*merge previous results to new results;
data &results;
set &results tempresults;
run;

*delete working datasets;
proc datasets ;
delete tempresults ResampledData ;
run;

*clear windows;
dm "out;clear;log;clear;";

%end;
%mend BootstrapForStandardErrors;

```

```

%macro resample(originaldata=, N1=, N2=, Ntotal=, strata=, variables=, id=);
/*This macro resamples data for use in the bootstrap macro; it
assumes data have some missing values in X's and/or Y.
originaldata = data in wide layout;
N1           = number of participants with complete data;
N2           = number of participants with some missing data in X's and/or Y.
Ntotal       = sample size
strata       = name of variable that identifies which observations are complete
              i.e., the indicator variable marked 0 if data has some missing data
              and 1 if complete
variables    = list of variables necessary for the analysis, X's, Y, and Z's ,
              do not include variable with subject's id here
id=         = variable name with subject's id
*/
proc sort data=&originaldata;
  by &strata;
run;

*sample with replacement, maintain % of people with incomplete observations;
proc surveyselect data=&originaldata method = urs sampsize = (&N1 &N2)
  rep=1 out=Resampled ;
  strata &strata;
  id &variables;
run;

data ResampledData ;
  set Resampled ;
  do i = 1 to numberhits;    output;    end;
  drop i;
run;

data fakeids ;
  do i = 1 to &Ntotal; *number of observations with complete covariates;
    &id=i; output;
  end;
  drop i;
run;

data ResampledData ;
  merge ResampledData fakeids ;

```

```

run;

proc datasets ;
  delete Resampled fakeids;
run;
%mend resample;

%macro SummaryBootstrap(actualresults= , results= , finalresults= );

proc means data=&results;  ods output summary=resultssummary;  run;

proc transpose data=resultssummary out=transposresultssummary;
run;

proc transpose data=&actualresults  out=transpactual;
run;

data bootmean(rename= (coll=bootmean))
  bootse(rename= (coll=bootse))
  bootmax(rename= (coll=bootmax))
  bootmin(rename= (coll=bootmin));
set transposresultssummary;
if _label_="Mean" then output bootmean;
if _label_="Std Dev" then output bootse;
if _label_="Minimum" then output bootmin;
if _label_="Maximum" then output bootmax;
drop _label_ _name_;
run;

data &finalresults(rename= (coll=Estimate));
merge transpactual bootmean bootse bootmin bootmax;
tstat= coll/bootse;
bootbias= coll-bootmean;
Pvalue = 2*(1-probnorm(abs(tstat)));
NormalBasedCILow= coll- 1.96*bootse ;
NormalBasedCIHigh= coll + 1.96*bootse ;
drop bootmean;
run;

proc print data=&finalresults;
run;

%mend SummaryBootstrap;

```

```

%macro MultipleInformantMLEConstraints(data= , out= , X1=, X2=, X3=, Y=, id=, covariates= ,
                                     type= , LinCovParms= );
/*
data= dataset in wide layout
out= name of dataset to store the results
X1= name of variable for exposure at window1
X2= name of variable for exposure at window2
X3= name of variable for exposure at window3
Y= name of variable for outcome
id= person identifier
covariates = list of covariates to adjust for; it must include interactions with
             T and main effect of T, literally: eg., covariates = T age*T
             The variable T MUST NOT exist in the widelayout dataset, it will get constructed
             in this macro.
type= type of variance for constrained model, usually lin(#), where # is the number of
             covariance parameters
LinCovParms=dataset with constraints for covariance matrix. in the matrix below, the parameters
             correlation between the outcome and exposure at each window (parameter 8) are constrained
             to be equal

data LinCovParms8; *this data needs to be part of the work directory prior to running the macro;
input parm row col1-col4;
datalines ;
1 1 1 0 0 0
2 2 0 1 0 0
3 2 0 0 1 0
3 3 0 1 0 0
4 3 0 0 1 0
5 2 0 0 0 1
5 4 0 1 0 0
6 3 0 0 0 1
6 4 0 0 1 0
7 4 0 0 0 1
8 1 0 1 1 1
8 2 1 0 0 0
8 3 1 0 0 0
8 4 1 0 0 0
;

```

```
run;
```

```
This algorithm is based on the algorithms discussed in
Lin X.H., Ryan L., Sammel M., Zhang D.W., Padungtod C., Xu X.P. (2000)
A scaled linear mixed model for multiple outcomes. Biometrics 56:593-601.
*/
```

```
data ResampledEM;      *need to make data into long format
  set &data;
  Response = &Y;      T=0;  output;
  Response = &X1;     T=1;  output;
  Response = &X2;     T=2;  output;
  Response = &X3;     T=3;  output;
run;

proc sort data= ResampledEM;
  by T;
run;

proc means data=ResampledEM std;
  title "initial standard deviations";
  by T;
  var Response;
  ods output Summary=sumary1;
run;

proc transpose data=sumary1 out=Transposesumary1;
run;

data Transposesumary1;
set Transposesumary1;
if _name_ = "T" then delete;
drop _label_ _name_;
run;

data _null_;
  set Transposesumary1;
  call symput('OutSTD',col1);
  call symput('Exp1STD',col2);
  call symput('Exp2STD',col3);
  call symput('Exp3STD',col4);
run;

data ResampledEM;
```

```

set ResampledEM;
ScaledResponse=.;
if T=0 then ScaledResponse=Response/&OutSTD;
if T=1 then ScaledResponse=Response/&Exp1STD;
if T=2 then ScaledResponse=Response/&Exp2STD;
if T=3 then ScaledResponse=Response/&Exp3STD;
run;

proc sort data= ResampledEM;
  by &id T;
run;

proc mixed data=ResampledEM method=ml covtest ;
  title "Starting Values";
  class &id T ;
  model ScaledResponse = T &covariates / noint s;
  repeated T /subject=&id type=un ;
  ods output CovParms=StartCovarP(keep = estimate ) ;
run;

proc transpose data=StartCovarP out=TransposeStartCovarP;
run;

data _null_;
  set TransposeStartCovarP;
  col11=(col2+col4+col7)/3;
  call symput('par1',col1);
  call symput('par2',col2);
  call symput('par3',col3);
  call symput('par4',col4);
  call symput('par5',col5);
  call symput('par6',col6);
  call symput('par7',col7);
  call symput('par8',col8);
  call symput('par9',col9);
  call symput('par10',col10);
  call symput('SigXY',col11);
run;

proc mixed data=ResampledEM method=ml covtest ;
  title "Improve Starting Values";
  class &id T ;
  model ScaledResponse = T &covariates / noint s;
  repeated T /subject=&id type=&type ldata=&LinCovParms;

```

```

    parms (&par1) (&par3) (&par5) (&par6) (&par8) (&par9) (&par10) (&SigXY);
    ods output CovParms=CovarP(keep = estimate ) ;
    ods output SolutionF=fixedF(keep = estimate ) ;
run;

proc transpose data=CovarP out=transpCovarP;
run;

data _NULL_;
set transpCovarP;
  newcol3=col3/(sqrt(col2)*sqrt(col4));
  newcol6=col6/(sqrt(col7)*sqrt(col4));
  newcol5=col5/(sqrt(col2)*sqrt(col7));
  newcol1=col1*(&OutSTD**2);
  newcol2=col2*(&Exp1STD**2);
  newcol4=col4*(&Exp2STD**2);
  newcol7=col7*(&Exp3STD**2);
  call symput('varY',newcol1);
  call symput('varX1',newcol2);
  call symput('varX2',newcol4);
  call symput('varX3',newcol7);
  call symput('X1X2',newcol3);
  call symput('X2X3',newcol6);
  call symput('X1X3',newcol5);
  call symput('ConstCor',col8);
run;

%let i=1;
%do %until(&criteria lt 0.0000001);

data ResampledEM;
  set ResampledEM;
  if T=0 then ScaledResponse=Response/sqrt(&varY);
  if T=1 then ScaledResponse=Response/sqrt(&varX1);
  if T=2 then ScaledResponse=Response/sqrt(&varX2);
  if T=3 then ScaledResponse=Response/sqrt(&varX3);
run;

proc mixed data=ResampledEM method=ml covtest ;
  title "Iteration number &i ";
  class &id T ;
  model ScaledResponse = T &covariates / noint s OUTPRED=predictions_1;

```

```

repeated T /subject=&id type=&type ldata=&LinCovParms;
parms (1) (1) (&X1X2) (1) (&X1X3) (&X2X3) (1) (&ConstCor) / hold=1,2,4,7;
ods output CovParms=CovarP(keep = estimate ) ;
ods output SolutionF=fixedF(keep = estimate ) ;
run;

data predictions_1;
set predictions_1;
NewProd=Resid*ScaledResponse;
run;

proc sort data=predictions_1;
by T;
run;

proc means data=predictions_1;
by T;
Var NewProd;
ods output Summary=sumNewProd(keep = NewProd_Mean);
run;

proc transpose data=sumNewProd out=transpsumNewProd;
run;

data transpsumNewProd;
set transpsumNewProd;
YVariance= col1 * &varY ;
X1Variance= col2 * &varX1 ;
X2Variance= col3 * &varX2;
X3Variance= col4 * &varX3;
keep YVariance X1Variance X2Variance X3Variance;
run;

data _NULL_;
set transpsumNewProd;
call symput('varY' ,YVariance);
call symput('varX1',X1Variance);
call symput('varX2',X2Variance);
call symput('varX3',X3Variance);
run;

proc means data=predictions_1;
Var NewProd;
ods output Summary=NewProd(keep = NewProd_Mean);
run;

```

```

data _NULL_;
  set NewProd;
  criterion = abs(NewProd_Mean - 1);
  call symput('criterion',ROUND(criterion, 0.000000001));
run;

proc print;
  title "The criterion is: &criterion";
run;

%let i= %eval(&i + 1);
%end;

proc print data=transpsumNewProd;
  title "Converged in &i iterations. The final criterion is: &criterion";
  title2 "The error variances are:";
run;

proc transpose data=CovarP out=transpCovarP;
run;

data transpCovarP;
  set transpCovarP;

  vary=&varY;  varx1=&varX1; varx2=&varX2; varx3=&varX3;  corryx =col8;

  covx1x2=col3*sqrt(varx1)*sqrt(varx2);
  covx1x3=col5*sqrt(varx1)*sqrt(varx3);
  covx2x3=col6*sqrt(varx2)*sqrt(varx3);

  covyx1=corryx*sqrt(vary)*sqrt(varx1);
  covyx2=corryx*sqrt(vary)*sqrt(varx2);
  covyx3=corryx*sqrt(vary)*sqrt(varx3);

  drop _name_ col1-col8;  run;

data _null_;
  set transpCovarP;
  call symput('covx1x2',covx1x2);
  call symput('covx1x3',covx1x3);
  call symput('covx2x3',covx2x3);
  call symput('covyx1',covyx1);
  call symput('covyx2',covyx2);
  call symput('covyx3',covyx3);
run;

```

```

proc mixed data=ResampledEM method=ml ;
  title "Final Likelihood (constrained)";
  class &id T ;
  model Response = T &covariates / noint s ;
  repeated T /subject=&id type=un;
  parms (&varY)
        (&covyx1) (&varX1)
        (&covyx2) (&covx1x2) (&varx2)
        (&covyx3) (&covx1x3) (&covx2x3) (&varx3) / hold=1,2,3,4,5,6,7,8,9,10;
  ods output SolutionF=redonefixedeffects(keep = estimate ) ;
  ods output FitStatistics=FitConstrained;

run;

proc mixed data=ResampledEM method=ml ;
  title "Final Likelihood (un constrained)";
  class &id T ;
  model Response = T &covariates / noint s ;
  repeated T /subject= &id type=un;
  ods output FitStatistics=FitUnConstrained;

run;

data CompareFit;
  merge FitConstrained(rename= value=ConstrainedModel) FitUnConstrained(rename= value=UNConstrainedModel);
  LRTStat= .;
  if Descr="-2 Log Likelihood" then LRTStat= ConstrainedModel - UNConstrainedModel ;
  pvalue=1-probchi(LRTStat,2);

run;

proc print data=comparefit;
  title "Constrained vs Unconstrained model";
  title2 "test of equal association between exposure and outcome across windows";
run;

data _null_;
  set CompareFit;
  if Descr ne "-2 Log Likelihood" then delete;
  call symput('LRTStat',LRTStat);
  call symput('LRTpvalue',pvalue);
  run;

proc transpose data=fixedF out=transpfixedF;
  run;

  data transpfixedF;

```

```

set transpfixedF;
  muy=col1*sqrt(&varY);
  mux1=col2*sqrt(&varX1);
  mux2=col3*sqrt(&varx2);
  mux3=col4*sqrt(&varx3);
  drop _name_ col1-col4;
run;

data &out;
  merge transpCovarP transpfixedF;
  beta11=covyx1/varx1;
  beta12=covyx2/varx2;
  beta13=covyx3/varx3;

  beta01=muy - beta11*mux1;
  beta02=muy - beta12*mux2;
  beta03=muy - beta13*mux3;

  col5=col5*sqrt(&varY);
  col6=col6*sqrt(&varY);
  col7=col7*sqrt(&varY);
  col8=col8*sqrt(&varY);
  col9=col9*sqrt(&varY);
  col10=col10*sqrt(&varY);
  col11=col11*sqrt(&varY);

  LRTStat=&LRTStat;
  LRTpvalue=&LRTPvalue;
run;

proc datasets;
  delete CovarP fixedF transpfixedF transpCovarP ResampledEM;
run;

%mend MultipleInformantMLEConstraints;

```