**Preliminary Research Case Study Summary**

**Title: Assessing Influence of Confounding Variables in Low Dose Lead Dose Response**

**Version: 1**

**Presented by: Cynthia Van Landingham (cvanlandingham@ramboll.com) and Rosalind A. Schoof (rschoof@ramboll.com)**

**Panel Advisor:**

## 1. Provide a few sentences summarizing the method illustrated by the case study.

This research case study explores possible limitations of current regression models in extrapolation to the low dose region of the dose-response curve, using epidemiological data for lead as an example, due to the existence of unrecognized and uncontrolled confounding. As described by Wilson and Wilson (2016), such confounding may arise "*when the measured association between an exposure variable and an outcome is distorted by an effect of a third variable (called a confounding variable or confounder)*". Wilson and Wilson (2016) report that uncontrolled confounding may be contributing to overestimation of the effects of lead reported by Lanphear et al. (2005), specifically identifying confounding by parental education, intelligence, or household management.

A major limitation in the analysis conducted by Wilson and Wilson (2016) was their inability to assess potential confounders due to lack of access to the original data sets (i.e., the seven longitudinal cohorts). As a result, they only present examples of potential confounding or reverse causation primarily for studies not included by Lanphear et al. (2005). In contrast, Ramboll has access to all the datasets used by Lanphear et al. (2005) as a result of performing additional analyses as published by Crump et al. (2013). In this case study, we present an initial analysis identifying confounding variables based on all seven longitudinal cohorts. We then propose to reanalyze the cohort data using the methods presented in Crump et al. (2013) while also considering the interaction among the identified confounding variables. This analysis, which is an expansion of the original regression method, should permit the determination of whether or not the interactions of these variables with the blood lead levels have a significant effect on the predictions in the low dose range of the dose-response analysis.

## 2. Describe the problem formulation(s) the case study is designed to address. How is the method described in the case useful for addressing the problem formulation?

Prior studies have suggested that the dose-response for lead effects on children's IQ may be supra-linear at blood lead levels less than 10 µg/dL. Such supra-linearity is generally not expected at lower doses for most environmental contaminants and may not be biologically plausible. This case study examines the possible contribution of uncontrolled confounding to the reported nonlinearity of the dose-response.

Lanphear et al. (2005) is an international pooled study of epidemiological data and is one of the key papers indicating that adverse effects due to exposure to lead may occur at levels below those previously regarded as safe. In their analysis, Lanphear et al. (2005) pooled data examining the

association between blood lead concentrations (BPb) in children and measures of their intelligence from seven longitudinal cohorts: Baghurst et al. 1992 (Port Pirie AU), Bellinger et al. 1992 (Boston MA US), Canfield et al 2003 (Rochester NY US), Dietrich et al. 1993 (Cincinnati OH US), Ernhart et al. 1989 (Cleveland OH US), Schnaas et al. 2000 (Mexico City MX), and Wasserman et al 1997 (Yugoslavia). Their overall conclusion was that "*environmental lead exposure in children who have maximal blood lead levels < 7.5 µg/dL is associated with intellectual deficits*." An analysis using the same data as Lanphear et al. (2005) was conducted by Crump et al. (2013) but with differing assumptions: how non-lead variables were controlled, how summary measures of BPb exposure were defined, and which BPb measures and transformations best modeled the data. Crump et al. (2013) concluded that "*there was statistical evidence that the exposure-response is non-linear over the full range of BPb evaluated in these studies, which implies that, for a given increase in blood lead, the associated IQ decrement is greater at lower BPb levels*."

Using the seven datasets relied upon by Lanphear et al. (2005) and Crump et al. (2013), we used correlation and regression analysis to explore whether any confounding existed between BPb and the covariates previously identified in Crump et al. (2013). Our analysis indicated that six of the variables examined (home score, mother's age, marital status at time of delivery, maternal education, maternal IQ, and ethnicity) are likely or highly likely confounders with the BPb. Four of these variables (maternal IQ, home score, ethnicity, and parental education) were also considered in Wilson and Wilson (2016) as those characteristic variables which may have interaction effects. Birth order, birth weight, gestational age, sex, and alcohol use during pregnancy were not identified as confounders. Confounding effects of the remaining variable, tobacco use during pregnancy, was inconclusive.

In the next phase of our analysis, we plan to include the six variables identified as likely or highly likely to be confounders (i.e., home score, mother's age, marital status at time of delivery, maternal education, maternal IQ, and ethnicity). Modeling of the data will follow the same steps as performed in Crump et al. (2013) using the database as identified in that publication with the inclusion of the interaction terms. Linear, log-linear and log(BPb+1) curves will be fit to the combined studies and used to predict the blood lead levels at which a deficit in IQ can be seen. Models will be fit using all the pertinent measures for blood lead.

Models will start with the basic regression used by Crump et al. (2013), for example:

$$IQ = \beta \times lnc + p1\_i \times site + p2 \times bwgt + p3\_i \times site \times momiq + p4\_i \times site \times medu + p5\_i \times site \times site\_cigs + p6\_i \times site \times site\_alc + p7\_i \times site \times home + p8 \times bo$$

This example shows the regression used by Crump et al. (2013) to estimate the IQ value using the following variables:

Lnc = log of the concurrent blood lead (lnc = log(concurrent + 1))

site = treated as category value so that each of the 7 sites has a unique parameter (p1_i with i = site) associated with it (1=Boston, 2=Cincinnati, 3 = Rochester, 4 = Mexico, 5 = Yugoslavia, 6 = Cleveland and 7 = Port Pirie). This variable serves as a site-specific background IQ level.

Bwgt = birth body weight

Momiq = site specific mother's IQ

Medu = site specific mothers education level

Site_cigs = site specific indicator of cigarettes smoked during pregnancy

Site_alc = site specific indicator of alcohol consumed during pregnancy

Home = site specific HOME scores

BO = birth order
β and Pj or Pj_i_i (j=1,8, and i = 1 to 7) = regression model parameters

To this equation the following variables would be added:

+ p9_i × lnc × site × home + p10 × lnc × mage + p11 × lnc × marital + p12_i × lnc × site × medu +p13_i × lnc × site × momiq + p14 × lnc × ethnicity

Where:
Mage = mother's age
Marital = yes/no was mother married at time of birth
Ethnicity = white or non-white
Pj or Pj_i_i (j=9,14, and i = 1 to 7) = regression model parameters

Each of these additional parameters are confounders with the blood lead variable (lnc). Therefore, when interpreting the relationship between IQ and the blood lead levels, these additional model parameters and the associated regression parameters will be considered in calculating the effect of the blood lead levels on IQ.

The general problem with ultra low dose extrapolation is the possibility of uncertainty in the effect associated with the main exposure variable. As stated by Wilson and Wilson (2016) "(I)nadequate correction of confounding has contributed to incorrect conclusions regarding causality at low levels of lead."

A more detailed description of the initial identification of potential confounding variables and preliminary results can be found in the Appendix.

## 3. Comment on whether the method is general enough to be used directly, or if it can be extrapolated, for application to other chemicals and/or problem formulations. Please explain why or why not.

Meta-analyses of epidemiological studies are increasingly being used to demonstrate low dose adverse effects. In any case where uncontrolled confounding could occur an expansion of typical regression analyses may need to examine the effect of adding additional variables on the exposure parameter. Whenever the outcome variable has multiple factors that directly affect the incidence, and some of those factors also influence the exposure variable, such confounding may occur. For example, as in the current case, mother's IQ may influence both child's IQ and child's blood lead level. Other examples where the measured association between an exposure variable and an outcome at low doses could be distorted by an effect of a third variable include lead exposure and cardiovascular disease, fine particulate matter and cardiovascular disease or pulmonary function, and aggravation of asthma by ozone and fine particulate matter.

The methods used to identify the covariates are general enough that they can be applied to other studies using the study specific variables when covariates are available and confounding with one or more covariates is expected.

## 4. Discuss the overall strengths and weaknesses of the method.

This method provides a powerful approach to critically examine key conclusions about the use of epidemiological data to predict dose-response in the low dose region of the dose-response curve. This is a newly identified potential problem, so additional methodological development may be needed. A key weakness is related to the data needs as described below.

## 5. Outline the minimum data requirements and describe the types of data sets that are needed.

Application of this method requires access to full data sets, and the studies included must have reported robust exposure data with many characteristic variables. In addition, there must be a large number of subjects in the data as there may need to be a number of regression parameters estimated by the regression equations.

## Does your case study:

### A. Describe the dose-response relationship in the dose range relevant to human exposure?

Yes, this method is focused on low dose extrapolation in the range of concentrations most relevant to exposures of sensitive populations. Specifically, this method proposes to incorporate additional factors in such low dose extrapolations to ensure that confounding factors do not result in distortion of the low dose dose-response.

### B. Address human variability and sensitive populations?

Yes, this method relies on epidemiological data that examines a large number of people with variable characteristics and sensitivities. For example, lead is a well-known problem in some populations of more highly exposed children. Our work, and that of the authors of the underlying studies, directly addresses this known sensitive population. Variability is inherently considered due to the need for data from large studies or multiple studies to examine multiple regression parameters.

### C. Address background exposures or responses?

Yes, this method relies on epidemiological data that includes background exposures and responses. In fact, the primary goal of the method is to more robustly account for background exposures and responses in dose-response analyses.

## D. Address incorporation of existing biological understanding of the likely mode of action?

Likely mode of action is not directly considered. Indirectly, alterations to our understanding of the correlation of exposure measures such as blood lead levels with health effects at very low doses may inform arguments about mode of action.

## E. Address other extrapolations, if relevant – insufficient data, including duration extrapolations, interspecies extrapolation?

No other extrapolations are relevant in this instance, as the method is applicable to epidemiological data. Insufficient data are often a problem with epidemiology studies, but this method can be applied after any adjustments for missing data are made. The problem of differences in data collected at different sites can be accommodated by using site specific variables as was done in Crump et al. (2013). The method can also be applied to multiple exposure measures as a means of examining duration directly, for example, Crump et al. (2013) included concurrent BPb, peak BPb, BPb at 24 months, mean lifetime weighted BPb, and early (6 months to 24 months) mean weighted BPb as exposure measures.

## F. Address uncertainty?

The focus of this method is on examining a possible source of uncertainty that has been overlooked in prior studies using regression models with epidemiological data to extrapolate to the low dose region of the dose-response curve. While existing regression models may account for covariation with multiple characteristic variables, they have not accounted for the existence of unrecognized and uncontrolled confounding, where a characteristic variable may distort the measured association between an exposure variable and an outcome. Thus, our model may serve to reduce the uncertainty of the models.

## G. Allow the calculation of risk (probability of response for the endpoint of interest) in the exposed human population?

The focus of this method is on examining previously unrecognized and uncontrolled confounding in prior studies using regression models with epidemiological data to extrapolate to the low dose region of the dose-response curve. As such, this method is expected to increase the accuracy of risk calculations based on low dose dose-response estimates.

## H. Work practically? If the method still requires development, how close is it to practical implementation?

This method is expected to work practically in the specific case to which it is being applied, i.e., for low dose extrapolation of the dose response curve based on a large number of data rich epidemiological studies of lead exposures and IQ. The method is also expected to be generalizable to other chemicals or constituents with a large number of data rich epidemiological studies where the possibility of confounding exists, and the pertinent covariates are available for the study. Determination of whether the method can be incorporated into standard dose-response analyses awaits further development.

## REFERENCES


Baghurst PA, McMichael AJ, Wigg NR, et al. (1992). Environmental exposure to lead and children's intelligence at the age of seven years. The Port Pirie Cohort Study. N Engl J Med, 327, 1279–84.

Bellinger DC, Stiles KM, Needleman HL. (1992). Low-level lead exposure, intelligence and academic achievement: a longterm follow-up study. Pediatrics, 90, 855–61.

Canfield RL, Henderson CR, Cory-Slechta DA, et al. (2003). Intellectual impairment in children with blood lead concentrations below 10 micrograms per deciliter. N Engl J Med, 348, 1517–26.

Crump KS, Landingham CV, Bowers TS, Cahoy D, Chandalia JK. 2013. A statistical reevaluation of the data used in the Lanphear et al. (2005) pooled-analysis that related low levels of blood lead to intellectual deficits in children. Crit Rev Toxicol. 43:785–799.

Dietrich KN, Berger OG, Succop PA, et al. (1993). The developmental consequences of low to moderate prenatal and postnatal lead exposure: intellectual attainment in the Cincinnati Lead Study Cohort following school entry. Neurotoxicol Teratol, 15, 37–44.

Ernhart CB, Morrow-Tlucak M, Wolf AW, et al. (1989). Low level lead exposure in the prenatal and early preschool periods: intelligence prior to school entry. Neurotoxicol Teratol, 11, 161–70.

Lanphear BP, Hornung R, Khoury J, et al. (2005). Low-level environmental lead exposure and children's intellectual function: an international pooled analysis. Environ Health Perspect, 113, 894–9.

Schnaas L, Rothenberg SJ, Perroni E, et al. (2000). Temporal pattern in the effect of postnatal blood lead level on intellectual development of young children. Neurotoxicol Teratol, 22, 805–10.

Wasserman GA, Liu X, Lolacono NJ, et al. (1997). Lead exposure and intelligence in 7-year-old children: the Yugoslavia Prospective Study. Environ Health Perspect, 105, 956–62.

Wilson, I.H.; Wilson, S.B. Confounding and causation in the epidemiology of lead. Int. J. Environ. Res. Public Health 2016, 1–16.

## APPENDIX

## DATA

Each of the studies used by Lanphear et al. (2005) and Crump et al. (2013) followed a cohort of children from birth and measured levels of lead in their blood at certain defined times. From time to time their intellectual development was measured along with other variables that might affect or correlate with that development. We used the BPb levels and twelve characteristic variables (home score, mother's age, marital status at time of delivery, maternal education, maternal IQ, ethnicity, birth order, birth weight, gestational age, sex, alcohol use during pregnancy, and tobacco use during pregnancy) reported in the cohort studies to determine if any of the characteristic variables could potentially confound the results reported in Lanphear et al. (2005) or Crump et al. (2013). Several of the characteristic variables (home score, maternal education, maternal IQ, maternal alcohol use, and maternal smoking) have been defined as site-specific due to being defined or measured in different ways in different studies (Crump et al. 2013).

## METHODS

### Correlation

A correlation analysis was conducted to identify which of the characteristic variables, if any, were significantly correlated with both the reported IQ of the children and the BPb concentrations. Being correlated with both provides an indication that the characteristic variable has an effect on both the final outcome (IQ) and the expected cause of the final outcome (BPb). The Spearman and Pearson correlation procedures provided in SAS were used to evaluate whether correlation exists between the characteristic variables and the exposure parameter variables. Eleven exposure variables were evaluated (IQ, concurrent BPb, peak BPb, lifetime weighted mean BPb, BPb at 24 months of age, early weighted mean BPb, and the natural logs of the BPb values).

For those variables identified as having continuous results (e.g., birth order, birth weight, gestational age, home score, mother's age, maternal education, and maternal IQ) the Pearson correlation was applied while for those having categorical results (e.g., marital status, race, and sex) the Spearman correlation was applied. Different types of responses were specified for the alcohol and tobacco use during pregnancy variables. In these cases, the Pearson or Spearman correlation was used where appropriate for each of the individual locations.

### Regression Modeling

In addition, a second analysis was conducted using the SAS multi linear regression procedure, PROC GLM to identify confounding variables. Using PROC GLM we evaluated the association between a given exposure variable and the outcome and the effect on that association when an additional independent variable is added to the regression. Our regression model consisted of the dependent variable, child's IQ (iq) and the independent variable natural log of concurrent lead (lnc[1]). Therefore, our initial model would be:

$$iq = b_0 + b_1 lnc$$

[1] Both Lanphear et al. (2005) and Crump et al. (2013) identified concurrent lead as the best statistical descriptor of the exposure response. Crump et al. (2013) states "this analysis, and particularly the results after eliminating influential points, supports the choice of concurrent BPb as providing the best description of the exposure-response curve."

where $b_0$ is the variable representing the intercept and $b_1$ is the estimated regression coefficient quantifying the association between iq and lnc. The resulting $b_1$ was then compared to a $\hat{b}_1$ produced when each of the characteristic variables were individually included in the model, as an example:

$$iq = b_0 + \hat{b}_1\, lnc + b_2 bwgt.$$

If the percent change between the $b_1$ and $\hat{b}_1$ estimates is greater than ±10% then that characteristic variable is considered to be a confounder. Note that those characteristic variables considered to be site-specific are evaluated by combining the site variable (location) with the characteristic variable, as an example:

$$iq = b_0 + \hat{b}_1 lnc + b_2 site * home.$$

## PRELIMINARY RESULTS

The variables ethnicity (race), home score (home), marital status (marital), mother's age (mage), mother's education (medu), and mother's IQ (momiq) were identified as potential confounders due to their significant correlation with both the child's IQ and natural log of concurrent lead (lnc) concentrations. When evaluated using regression modeling, the characteristic variables ethnicity (race), home score (home), marital status (marital), mother's education (medu), and mother's IQ (momiq), were identified as confounders. Table 1 provides a summary of the confounder identification using the following designations:

- Unlikely – variables that were not identified in the regression analysis as resulting in a percent change of greater than ±10% in the beta estimate but which were identified as having correlation for specific-sites in the correlation analysis.
- Likely – variables that were either identified as potential confounders in the correlation analysis or identified in the regression analysis as resulting in a percent change of greater than ±10% in the beta estimate, but not both.
- Highly Likely – variables both identified as potential confounders in the correlation analysis and identified in the regression analysis as resulting in a percent change of greater than ±10% in the beta estimate.

Wilson and Wilson (2016) considered maternal IQ, home score, race, and parental education as those characteristic variables which may have interaction effects. This analysis confirms that all those variables can be potential confounders using the data from the Lanphear et al. (2005) and Crump et al. (2013) analyses. Therefore, any reanalysis of the cohort data using the method's presented in Crump et al. (2013) should consider the interactions among the confounding variables identified in this analysis, particularly ethnicity, home score, marital status, mother's education, and mother's IQ as these were identified as potential confounders in both the correlation analysis and regression modeling.

**Table 1. Characteristic variables confounder identification based on both correlation and regression analyses**

| Variable Name | Description | Confounder Identification |
|---|---|---|
| bo | Birth order | No |
| bwgt | Birth weight | No |
| gage | Gestational age | No |
| home | Home score with fewest missing | Highly Likely |
| mage | Mother's age | Likely |
| marital | Marital status at delivery | Highly Likely |
| medu | Maternal education | Highly Likely |
| momiq | Maternal IQ | Highly Likely |
| race | Ethnicity | Likely |
| sex | Gender of child | No |
| site_alc | Alcohol use during pregnancy | No |
| site_cigs | Tobacco use during pregnancy | Unlikely |