# Statistics and Policy Decisions: Issues in Statistical Analyses

Helena Chmura Kraemer*

*1116 Forest Avenue, Palo Alto, CA 94301, USA*

**Abstract:** When national policy decisions are to be guided by the results of statistical analyses, it is important, to avoid being misled to look beyond the authors' conclusions and first to assess the study design, measurement and analytic methods, in order to decide whether a study's conclusions rest on a solid foundation. In particular, observational studies must be carefully and critically evaluated. Using a study widely cited concerning the effects of low-level lead exposure and IQ, we illustrate several methodological errors, long known but often ignored. The goal is not to settle the controversies about the effect of lead on IQ, nor to disparage observational studies, for they are the foundation of all studies done to guide policy, but to encourage additional care in the use of such studies to address policy questions.

**Keywords:** Policy decisions, Statistical Significance, Practical or Policy Significance, Methodological Errors, Lead/IQ Association.

## 1. INTRODUCTION

"Do No Harm!" is a principle well understood and accepted when applied to ethical clinical decision-making for individual patients. The same principle also applies to research conclusions that might affect policy decision-making, where an error in design, analysis, or interpretation may harm many thousands of people the researchers themselves have never seen or will see, the target population they are trying to help. To avoid being misled, it is important to look beyond conclusions and critically assess the study design and statistical methods, in order to assess whether a study's conclusions rest on a solid foundation.

We here use the ongoing controversy of the relation of lead to IQ to illustrate some of the red flags that cannot be ignored when making research-based policy decisions based on statistical analyses. Our goal is not to resolve the IQ/lead controversies (e.g., see [1]), for the data needed for such resolution are, we would argue, not yet available. Nor is the goal to discourage observational studies, for such studies are essential in the process of understanding risk. The goal is to encourage careful, valid, critical and candid analysis of such studies, and an awareness of the limitation of such studies as a basis of policy decisions.

## 2. BACKGROUND

Observational studies have long been at the center of the continuing debate concerning the effect of lead on the IQ of children. Minimizing lead sources in the environment is laudable; no one believes that lead is good for child development. Efforts to reduce lead in the environment have been ongoing at least since 1970. In the United States [2] between 1976-1980 and 1988-1991, there was substantial decline in the blood lead levels (BLLs), e.g., for children aged 1-5 years, from an overall mean of 15.0 µg/dL to 3.8 µg/dL. The decrease has largely been attributed to the removal of lead from gasoline and the ban on the use of lead in soldered cans and paint [2]. However, children may still be exposed to lead from a variety of sources such as historical emissions of leaded gasoline or industrial sources present in air, dust, soil and water, lead-based paint in old and deteriorating housing [2], and other sources difficult to regulate. Consequently, current concern centers on the effects of low levels of blood lead (<10 µg/dL) and the question of whether there are tolerable levels of lead exposure, i.e., some level of lead with so little, if any, effect on IQ that the costs or risks of further systematic intervention might exceed its benefit.

A key research study, heavily based on statistical methods, the Lanphear *et al*. study [3] in which 7 observational studies are "pooled" (the "Pooled Study") is often cited by advisory committees for government agencies as compelling evidence of lead effects on IQ at low BLLs (See e.g., Centers for Disease Control and Prevention, Advisory Committee on Childhood Lead Poisoning Prevention Report (2012), United States Environmental Protection Agency (USEPA) National Ambient Air Quality Standards for Lead (2008); USEPA, Integrated Science Assessment for Lead (Draft), (2012), and National Toxicology Program, Monograph on Health Effects on Low-Level Lead (2012)). Should the data in this observational study be the basis for policy decisions?

*Address correspondence to this author at the 1116 Forest Avenue, Palo Alto, CA 94301, USA; Tel: 650 328-7564; E-mail: hckhome@pacbell.net

E-ISSN: 1929-6029/15 © 2015 Lifescience Global

## 3. OBSERVATIONAL STUDIES VERSUS RANDOMIZED CLINICAL TRIALS

The randomized clinical trial (RCT) is generally considered to be the "gold standard" in scientific design to prove causal association. Unlike observational studies, (1) randomization in a well-designed and executed RCT results in two (or more) random samples from the *same* population, thus minimizing the chance that any association seen reflects comparing two (or more) populations differing on more than the specific factor of interest. Observational studies attempt to use mathematical models to "statistically control" for such factors, but the results are much less certain, both because all such factors cannot be known, many are not measured, and the models used are questionable (further discussion below). (2) In a well-done RCT, bias in measuring outcome is minimized by "blinding" outcome assessment; in observational studies, such "blinding' is often difficult to implement. (3) In a well-done RCT, researchers control the protocol of delivery of the treatment and control conditions, and monitor the delivery to ensure "fidelity"; in observational studies, what is experienced is determined by factors outside the control of the researchers and often unknown to them. Not all RCTs are well-done, and there is not, of course, guaranteed infallibility of the conclusions even in well-done RCTs. The CONSORT guidelines [4-8] provide guidance as to which RCTs are more trustworthy, and independent confirmation of the conclusions is still always necessary.

In 1997, Moses and Mosteller pleaded: "Experimentation: just do it!" [9] urging the wider use of well-conducted RCTs as a basis of public health policy decisions. Nevertheless, even today, as noted in the Wall Street Journal (May 4, 2012): "Despite concerns, observational studies have never been more popular", probably because such studies are "easier, cheaper and quicker to do". Ioannidis [10] in a study of highly cited research findings, showed that of the four observational studies there included, the conclusions of three were later contradicted.

Clearly in any particular context, some RCT designs are either not feasible or not ethical: One cannot, for example, randomize families to be exposed to various levels of lead exposure and track development (as one might with laboratory animals) without incurring ethical objections. However, one could randomize families with newborns who already live in at-risk situations where their children would be exposed to lead, to an intervention designed to reduce the lead levels in their environments, versus a "treatment as usual" control group. Randomization ensures that the distribution of IQ potential is similar in the two treatment groups. Longitudinal follow-up would be the same in both groups, "blinded" to group membership. To do such a RCT would eliminate the major problems associated with interpretation of the results from observational studies.

If such a RCT were done, any number of results might be obtained, e.g.:

- It may be that the intervention is ineffective in preserving IQ potential, either because the treatment is ineffective in further reducing lead exposure, or ineffective in reducing BLLs, or that BLL is not a strong causal factor in determining IQ.
- It may be that the intervention is effective in preserving IQ potential, but that the effect size of that increase is outweighed by the costs and risks associated with the intervention. "A nation can be ruined by cleverly crafted short-term solutions to its long-term problems." [11].
- It may be that the intervention is highly effective in preserving IQ, but only for a minority in the population.
- In all these situations, finding this result out in a limited time RCT would prevent implementing policy decisions that are costly, possibly harmful, but ineffective.
- Finally, it may be that the intervention is highly effective for most or all, is cost-effective and without risks of any clinical or practical significance. Those that advocate for policy decisions in the absence of a RCT *assume* this would be the result if a RCT were done. But this is not a foregone conclusion. Lanphear *et al*. [12] recognized this problem, but nonetheless urged that we "acknowledge the limitations of observational epidemiology without prohibiting us from taking action to protect public health. The alternative, to perpetually permit children to be exposed to lead and other emerging toxicants, is both absurd and unacceptable" (p. 197). But what if the actions taken to reduce BLLs were to do more harm than good? Would that not be even more absurd and unacceptable?

There seems to be a feeling that such a RCT is not really necessary, and in some cases, this is correct. However, in what follows we will examine a number of statistical approaches commonly found in observational studies that lead to erroneous conclusions in such studies, using the Pooled Study as an illustration.

## 4. CORRELATES, RISK FACTORS, CAUSAL FACTORS: NOT THE SAME THING!

The Pooled Study was conducted to address questions about lead-associated intellectual deficits at BLLs <10 µg /dL. What does it mean to be "lead-associated"? Only if lead exposure is a causal factor is it appropriate to recommend action to reduce lead exposure as a means of improving IQ.

A *correlate* is a factor that is in some non-specific sense associated with another; a correlation coefficient indicates the strength and direction of the association between two variables, here BLL and IQ. The old adage remains true: one cannot infer causation from correlation. A *risk factor* is a correlate that can be shown to *precede* the other factor in time [13]. A risk factor may or may not be a *causal factor*, i.e., if one were to remove the risk factor, the subsequent outcome may or may not change. A factor might be a risk factor but not a causal factor because it is proxy to another causal factor [14]. For example, high lead exposures seem to be more common in low socio-economic households, households with parents with lower IQ, lower income level, poorer home environment and access to health and educational resources, or other such factors, genetic or environmental, that also influence IQ. It may be that the only reason BLL is correlated with low IQ is because it is yet another indicator of low socio-economic status [15]. Alternatively, it may be that children with lower IQ are more likely to exhibit behaviors that increase exposure to lead, in which case lower IQ may be a risk factor for high lead level rather than vice versa [16]. In either such case, manipulating lead levels may have little or no effect on IQ.

In most cases, demonstrating the required temporal precedence of the risk factor would require a longitudinal study. When the timing of the measurement of the risk factor to the outcome is not clear, then all that can be claimed is that the two are *correlates*, not that one is a risk factor for the other, and assuredly not that one causes the other.

In the Pooled Study, the primary analytic emphasis was on *concurrent* blood lead level (CL). Of the four blood lead concentration measures there considered (concurrent, peak, early childhood and lifetime mean), the only factor satisfying the temporal precedence criterion was early childhood BLL, because it preceded the IQ measure at every site. Lack of distinction between a correlate, a risk factor and a causal factor is a major source of misinterpretation in epidemiological studies. All causal factors are risk factors and all risk factors are correlates, but not vice versa.

## 5. POOLING, MUDDLING, AND META-ANALYSIS

The Pooled Study considered 7 independent, observational studies from 7 different sites, done at different times, following different protocols. The sites varied in virtually every factor studied, including the two primary factors, blood lead index and IQ. (See Crump *et al.* for further discussion of these issues [1]). To treat these 7 studies as if they were replicates, and thus to "pool" their data creates a serious problem.

When one draws multiple samples from the *same* population, it is appropriate to "pool" data, i.e., to treat the entire dataset as one sample in the analysis. However, when each set of data comes from a different population, studied at different ages with different measures, treating the data as if all were drawn from the same population is misleading, a process better called "muddling" rather than "pooling". If all the studies addressed the same research question, the preferred procedure is to estimate the parameter of interest (e.g., a correlation coefficient, a standardized regression coefficient, Cohen's d) in each separate sample, check for homogeneity over the sites, and pool the parameter estimates (not the data) only if there is no heterogeneity, i.e., meta-analysis [17]. If there is heterogeneity and the sites are randomly sampled from some population of sites, one might estimate the mean and standard deviation of the parameter estimates over sites for that population of sites. If there is heterogeneity and the sites are a sample of convenience (as in the Pooled Study), thus not representative of any identifiable population of sites, one might report the results from each site and explore the question as to why sites might differ from one another. It is inappropriate to pool the estimates when heterogeneity is evidenced, or to assume that sites are randomly sampled from a population of sites when that is not so. Finally, it is inappropriate even to use meta-analysis when the various studies address different issues [18], the classical "apples and oranges" problem in meta-analysis [19, 20].

## 6. SIMPSON'S PARADOX, THE ECOLOGICAL FALLACY: PITFALLS OF MUDDLING

The problem of muddling samples from different populations to assess correlation has long been known as generating Simpson's Paradox [3, 21, 22]. The correlation coefficient obtained in the muddled sample is a weighted combination of the ecological correlation between the means of the two variables across the samples and the multiple intra-site correlations [23].

In the Pooled Study, the estimated ecological correlation (that between the site means) between IQ and CL was -.533, i.e., sites with lower IQ means tended to have higher CL means. CL and IQ correlation coefficients for each of the 7 sites in the Pooled Study are presented in Table **1**. The within-site correlations ranged from -.007 at Mexico to -.349 at Rochester, none anywhere near as large as the ecological correlation. The weights in the muddled correlation coefficient reflect the proportion of the total variance coming from within and between samples and thus depend not only on the site differences, but also on the varying sample sizes at the sites. The ecological correlation does not necessarily correspond to that within any site (so interpreting it has been called the Ecological Fallacy). In short, the correlation coefficient obtained from a muddled sample is an estimate of an uninterpretable and meaningless population parameter.

The within-site correlations between CL and IQ are all negative, and range from trivial (Mexico) to moderate (Cleveland, Rochester) [24]. The fact that they are all negative supports the contention that lead is at least a marker for an environment less than optimal for development. Sample sizes ranged from 99 at the Mexico site to 324 at the Port Pirie site. The homogeneity of correlation test was here not statistically significant at the 5% level (ChiSquare test statistic=12.4, p=.053), and the pooled estimate was -.233, suggesting that, in general, CL might account for less than 5% of the variance of IQ, not a strong association. Figure 2 in the Lanphear *et al*. paper [3] (p. 898) conveys the same message with regard to regression coefficients.

## 7. ANALYSIS: A MODEL IS A MODEL IS A MODEL…

"Essentially, all (mathematical) models are wrong, but some are useful [25] (p. 424);… the practical question is how wrong do they have to be to not be useful" [25] (p. 74). All statistical analyses are based on some mathematical model. For a model to be useful, it needs to reflect what is already known about reality (its assumptions) in order to gain further understanding of reality (hypotheses to be tested, parameters to be estimated). Since the conclusions drawn from applying a mathematical model are contingent on the assumptions made, when assumptions are made that do not reflect reality well, the conclusions based on those assumptions may also not reflect reality well. In the Pooled Study, there was extensive mathematical modeling, and some questionable assumptions.

In the Pooled Sample analysis, essentially two models were used: a linear model and a log-linear model. The linear model used in the Pooled Study (in its simplest form) was essentially this:

$$IQ=IQ_0-b\ CL + e,$$

where $IQ_0$ is the average IQ of the population with CL=0, $b\geq0$ indicates how much decrease in IQ would be expected for each unit increase in CL, and e is the deviation of individual subjects' IQs from the expected IQ, which is assumed to be independent of CL. This assumes that, for example, an increase from CL=0 to CL=1 results in the same average decrease in IQ as an increase from CL from 5 to 6 or from 10 to 11, and that IQ approaches zero as CL increases.

**Table 1: Sample Sizes, Arithmetic Means and Standard Deviations for IQ and Concurrent Lead, with the Correlation between them, and that Adjusted for the HOME Score**

| Site | N | IQ | Concurrent Lead (CL) | Correlation IQ versus CL | Adjusted Correlation: IQ versus CL |
|---|---|---|---|---|---|
| Boston | 116 | 116.0 (14.3) | 6.1 (3.7) | -.255 | -.133 |
| Cincinnati | 221 | 87.0 (11.4) | 9.2 (5.3) | -.207 | -.163 |
| Cleveland | 160 | 86.7(16.3) | 15.6 (6.5) | -.328 | -.165 |
| Mexico | 99 | 107.8 (11.0) | 8.2 (4.9) | -.007 | +.027 |
| Port Pirie | 324 | 106.0(13.7) | 13.7 (5.9) | -.247 | -.130 |
| Rochester | 182 | 84.9 (14.4) | 5.1 (3.5) | -.349 | -.270 |
| Yugoslavia | 231 | 74.2 (13.3) | 20.9 (15.5) | -.132 | -.181 |

The Pooled Study also used a log-linear model, essentially:

$$IQ=IQ_0-b \log(CL+1)+e.$$

$IQ_0$ remains the same, but now it is assumed that an increase from CL=0 to CL=1 is equivalent to that from CL=5 to CL=l1, or from CL=10 to CL=21. The model now also assumes that the decrease in IQ as CL increases is most rapid when CL is nearer zero.

The major problem here is that both models *assume* that there is no tolerable level of CL, i.e., no range of CL near zero when IQ remains approximately equal to $IQ_0$. In essence then, the Pooled Study *assumed* the conclusion it wanted to prove.

An alternative model might have been:

$$IQ=IQ_0 +e, \text{ for } CL<c,$$

$$IQ=50+(IQ_0-50)e^{-b(CL-c)}+e, \text{ for } CL\geq c.$$

Then as CL increases, IQ approaches a more reasonable lower limit of 50, not zero, and $0\leq CL\leq c$ indicates the tolerable range of CL. It may well be that fitting such a model to the data at each site would result in an overall estimate of c=0, in which case, as the Pooled Study assumes, there is no support for a tolerable level of CL. But now the hypothesis that there is no tolerable range of CL could be disproved.

When using mathematical models that incorporate many assumptions it is important to first check each assumption carefully against what is known about the situation to assure a reasonable correspondence, and second to distinguish what is assumed from what is to be proven. One cannot assume absence of a tolerable level of CL in order to demonstrate that there is no tolerable level of CL.

## 8. THE PROBLEMS WITH STATISTICAL SIGNIFICANCE

For the last 10-15 years, scholars have emphasized the common misunderstanding of what “statistical significance” means and expressed concern about its misuse [26-31]. Journals have even, on occasion, banned the use of the “p-value” [30, 32, 33]. Generally, a legitimate “statistically significant result” means that the sample size used was sufficiently large to detect some deviation from the null hypothesis. Therefore it is a comment on study design, not on the size or importance of the effect. Consequently, many a “statistically significant result” may be of little practical importance; many a “non-statistically significant result” may be of great importance, but assessed in an inadequately powered study. Consequently, researchers are urged to report effect sizes interpretable to policy makers, with some indication (e.g., a confidence interval) of how precisely the effect size is estimated. Then policy makers could weigh the practical significance of a “statistically significant” effect in deciding whether action should be taken.

A correlation coefficient is one such effect size. As can be seen in Table **1**, the correlation between CL at the various sites and IQ is never large, and even that is perhaps exaggerated because it includes effects of factors, known and unknown, strongly correlated with CL. However, the results are “statistically significant” in all but one site (Mexico) simply because the sample sizes at those sites were large enough to detect some deviation from zero.

An even more serious problem has to do with the legitimacy of the statistical hypothesis-testing done. To be legitimate, the hypothesis tested must be ‘a priori’, i.e., formulated based on rationale and justification that existed *before* the data were accessed. Looking at some or all of the data to develop or modify the hypothesis, then testing that hypothesis on the same data is called “post-hoc” testing. In “post-hoc” testing, inferences based on standard methods are usually incorrect and often exaggerated. In the Pooled Study, some of the hypotheses were developed by examining the data, the most flagrant the selection of CL as the primary outcome because it most strongly supported the authors’ conclusions (p. 896).

The Pooled Study also reported testing for interactions and collinearities. Finding them to be non-statistically significant, it then ignored them for testing or interpretation of the results. If interactions and collinearities are present in the population but ignored in the model, the risk of Type I error (a false positive conclusion), as well as of Type II error (a false negative conclusion), are increased. However, the sample size necessary for adequate power to detect interactions or collinearities of a size that might affect results is generally much larger than that to detect main effects, and the methods for detection are often based on models that may or may not be appropriate to the context. Thus, finding a “non-statistically significant” result is no guarantee that these problems are absent (In the old adage, “Absence of proof is not proof of absence”), and is no justification for ignoring these problems in the models used.

Finally, the Introduction [3] ( p. 894) indicates that, in the Pooled Study, the ‘a priori’ hypotheses concerned the effects of BLLs < 10 µg/dL, and were based on the Rochester Longitudinal Study and a reanalysis of the Boston study data. Since the ‘a priori’ hypothesis was based on those studies, it was inappropriate then to include the data from those two studies in any attempt to test that hypothesis. Of the 244 children across all sites with peak BLLs < 10 µg/dL, 144 (59%) were either at the Rochester or Boston sites. There were no children at the Port Pirie site with peak BLLs < 10 µg/dl; this site should have been excluded. The sample sizes at the remaining sites (23, 11, 20, 46) were too small to examine the gradient of IQ with increasing exposure with peak BLLs < 10 µg/dL, particularly when confounding variables were to be considered.

## 9. THE PROBLEM OF CONFOUNDERS

The attempts to “control for” or to “adjust for” certain variables in analysis of observational study results arise from the appropriate concern that the subpopulations exposed to different levels of lead differ on far more than BLL, and thus the apparent correlation between lead level and IQ may actually reflect those factors (confounders) rather than lead level itself (a concern largely mitigated, but not removed, by randomization in a RCT). If so, efforts to reduce lead levels that do not also change confounders could be doomed to failure. However, in an observational study, one never knows all the relevant confounders, only those recognized and measured at or before the time of measurement of the risk factor of interest.

Moreover, the problem is not only with unrecognized confounders. For example, the site differences seen in Table **1** were not completely ignored in the Pooled Study, but were dealt with by “controlling for” site effects in the analysis, but it was assumed that there were no site by CL interactions on IQ, an assumption at least questionable given the site differences seen in Table **1**.

When there are suspected confounding variables (site as well as other factors such as the HOME score, maternal IQ, etc.), researchers often include those factors in a linear model and claim that they have “controlled for…” those factors, suggesting that what results represents a “purer” estimate of the causal effect of lead for all in the population. To illustrate such an analysis here, in Table **1**, the correlations between CL and IQ adjusted for the HOME score are also presented. Those correlations are substantially reduced and range from r=+.027 at the Mexico site, to r=-.270 at the Rochester site. Once again no statistically significant heterogeneity of correlation was detected across sites (Chi Square statistic=6.2, p=.396), with a pooled value of -.157, thus suggesting that CL now accounts for perhaps 2.5% of the variance in IQ.

The difficulty is that the methods used to “adjust” are often based (as was the partial correlation coefficient reported above) on the assumption that the association between lead and IQ is linear (or at least monotonic) and the same regardless of the level of the other factors, i.e., that there is no interactive effect of lead and the confounder on IQ. For example, the Home score is the strongest correlate of IQ in every US site, “an index that reflects the quality and quantity of emotional and cognitive stimulation in the home environment” [3] itself quite strongly correlated with Maternal IQ. If one could stratify the population at each site by their HOME scores, the patterns of IQ means as a function of lead level within each stratum should be monotonic and parallel each other at every site, at least at the population level. Such an assumption is often untrue.

To illustrate this point, we attempted to so stratify on the HOME score and on ranges of CL, examining only those cells with minimally 5 subjects. For no HOME stratum could we see the full pattern of mean IQ as a function of CL across all 7 sites (evidence of collinearities). The closest was when HOME>35 (the most advantaged group), where the full pattern is seen for 5 of the 7 sites, shown in Figure **1**. If these results were to be believed (and they should not be!), it would appear that the highest IQ is achieved when the concurrent lead level is between 5 and 7.5 µg/dL, suggesting that a little lead might help IQ! Why should we not believe these results? This harkens back to the issue of hypothesis-generating versus hypothesis-testing. There was no ‘a priori’ theoretical rationale or empirical justification for a hypothesis that CL between 5 and 7.5 µg/dL is best for IQ development. This result was found in exploration, and should be given credibility only if the hypothesis can be tested and supported in *independent* studies. We would venture to guess that this result would not be confirmed.

In attempting this stratification, it was noted that at the Boston site, only 1 (0.9% of the total) participant had a HOME score< 35 and that one child had a

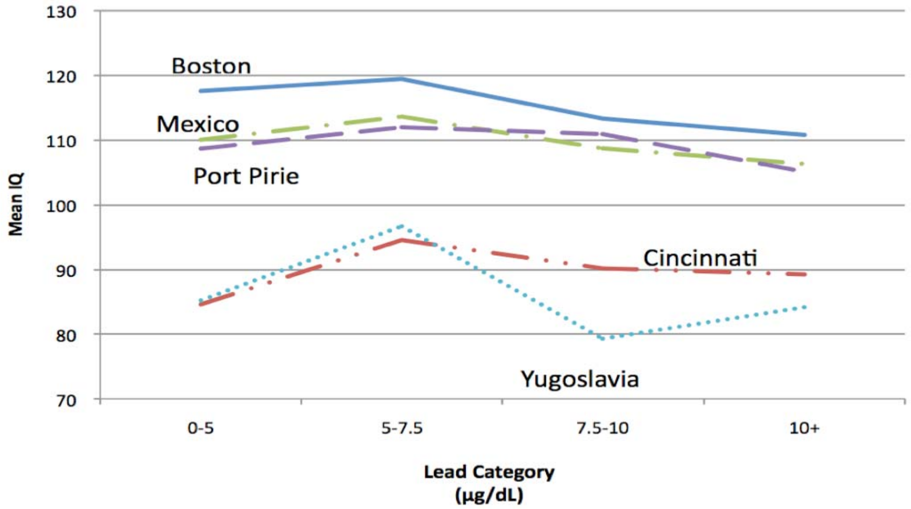

**Figure 1:** HOME score > 35: IQ means by concurrent lead category and site.

CL≥10. At Port Pirie, only 25 (7.7% of the total) had a HOME score < 35 and all had CL≥10. In contrast, Rochester had 167 participants (91.8% of the total) with a HOME score < 35, and of those only 10.2% had CL≥10. Thus, the data of the Pooled Study continue to indicate substantial interactions and collinearities among Site, CL and HOME.

In the Pooled Study, consider only two of the possible "confounders": the HOME score and Maternal IQ. These two factors are correlated with each other (correlations ranging from .292 at Cincinnati to .458 at Boston) (see Table **2**), since both are indicators of socio-economic status. Both are also correlated with CL, at some sites, quite strongly. Finally, the strongest correlation with IQ in these studies was not with CL, but either HOME or Maternal IQ. In short, efforts to further control lead levels without concomitant efforts to remove other disadvantages of low socio-economic status, and perhaps even genetic effects, may not be fruitful, and may subject already stressed families to even further stress.

## 10. DISCUSSION

There are a number of points, all long known but worth reinforcing, that are here made about the general statistical approaches to evaluation of a risk factor:

- Causal inferences drawn from observational studies should always be made in very tenuous fashion. When comparing a sample from populations at different exposure levels, there is always the possibility that these populations differ on far more than the exposure of interest. There are statistical methods (e.g. "adjusting", propensity analysis, [34]) that, carefully used, can reduce the possibility of falsely attributing to the risk factor of interest that which is caused by other related factors. However, since there are always factors confounding the association between the risk factor of interest, some unknown or unmeasured, there must always remain some reasonable doubt. Nevertheless, because RCTs are difficult and expensive,

**Table 2: Correlation among Confounded Factors in the Full Sample and the Number and Percentage of that Full Sample with CL<10 µg/dL, and with Peak Lead Level <10 µg/dL**

| Site | HOME<br>vs. CL | Maternal IQ<br>vs. CL | HOME<br>vs. Maternal IQ | CL<10 µg/dL | Peak Lead <10 µg/dL |
|---|---|---|---|---|---|
| Boston | -.351 | -.115 | .458 | 96 (82.8%) | 41 (35.3%) |
| Cincinnati | -.211 | -.196 | .292 | 146 (66.2%) | 23 (10.4%) |
| Cleveland | -.414 | -.369 | .430 | 32 (20.0%) | 11 (6.9%) |
| Mexico | -.143 | -.214 | .370 | 70 (70.7%) | 20 (20.2%) |
| Port Pirie | -.399 | -.318 | .406 | 88 (27.2%) | 0 (0.0%) |
| Rochester | -.344 | -.270 | .343 | 165 (90.7%) | 103 (56.6%) |
| Yugoslavia | +.028 | +.083 | .403 | 91 (39.4%) | 46 (19.9%) |

observational studies can constitute a valuable first approach to understanding the association between a risk factor of interest and an outcome. However, these observational studies should be carefully designed, conducted, analyzed, and judiciously interpreted. If, in such studies, it can be documented that the association between the risk factor and outcome is strong (not merely statistically significant), and remains strong even after confounders are carefully explored, a RCT may not be necessary. But when the association is weak (even if statistically significant) and becomes substantially weaker after confounders are explored, it is risky to base policy decision on such data.

- It is inappropriate to assume that a conclusion is true, and then, if the data do not contradict that assumption, to claim that the conclusion has been shown to be true.
- It is important to distinguish correlates from risk factors from causal factors. All causal factors are risk factors, and all risk factors are correlates, but not vice versa.
- When dealing with samples from different populations addressing the same research question (e.g., different studies or sites), the preferred method is a meta-analytic approach. In the absence of heterogeneity among the populations in the parameter estimated, the meta-analysis will yield essentially the same results as would the pooled analysis, but where there is heterogeneity, the results obtained by applying these two approaches may be drastically different.
- In any study, observational or RCT, the primary focus of attention should be on estimation of interpretable effect sizes and their confidence intervals, either in addition to, or in place of, p-values. Then the issue of clinical or policy significance of a statistically significant result is based on assessing the magnitude of the effect size, the precision with which it is estimated, and, most importantly, consideration of the impact of such an effect on the population in question. It would be useful to estimate how many families currently lead-exposed would require intervention (at what cost per family and at what burden to those families) in order to increase the IQ of the children by a sufficient number of points to make a difference in their lives. Generally a 2-3 point difference in IQ will make very little difference.
- Mathematical models play an important role in all statistical analyses. However, it is the responsibility of the analysts to check that all assumptions made by the model are reasonable in the context in which they propose to use it, and to avoid assuming the desired conclusion. Nikola Tesla early in the 20th century is quoted as saying [35]: “Today's scientists have substituted mathematics for experiments, and they wander off through equation after equation, and eventually build a structure which has no relation to reality.” That is even a greater danger today, with the readily available computer programs that easily fit very complex models to any dataset.
- A distinction must be made between exploratory studies meant to generate hypotheses to be tested in future independent studies (and to provide empirical justification for those hypotheses as well as information on how best to design such studies), and a hypothesis-testing study. Whether observational or RCT, hypothesis-testing studies must have certain “a priori” hypotheses, with the appropriate rationale and justification; a design appropriate to those hypotheses; an analytic approach set up *before* the data are accessed; and should be adequately powered to test those hypotheses. When that hypothesis testing is completed, it would be wasteful not to explore the data further, both to provide greater insight into the conclusions drawn on the primary hypotheses as well as to generate hypotheses that might broaden or deepen understanding of the issues to be tested in future studies. However, the researchers should not test hypotheses on the same data that generated them, or present the hypotheses so generated as “conclusions” before independent validation.

We propose that the evidence Lanphear *et al*. presented to support their conclusion is not convincing, and should *not* be used to guide policy decisions. In 1994, Pocock and Smith [17], using meta-analysis, pointed out that: “While low level lead exposure may cause a small IQ deficit, other explanations need consideration…. Even if moderate increases in body lead burden adversely affect IQ, a threshold below which there is negligible influence cannot currently be determined. Because of these uncertainties, the degree of public health priority that should be devoted to detecting and reducing moderate increases in children's blood lead, compared with other important social detriments that impede children's development, needs careful consideration." (p. 1189). Almost 20 years later, that conclusion has not changed.

In general, when considering whether to base policy decisions on observational studies, great care must be taken to consider the sampling, measurement, design and analytic decisions made in the study and the impact of those decisions on the credibility of results. Even then, the emphasis should be on effect sizes interpretable in terms of impact on society if policy actions are taken, rather than on statistical significance, and due consideration should be given to the harm that could result from unnecessary or ill-considered interventions. In general, it is advisable that policymakers asking for evalutions and conclusions from research studies consult a panel of experts *not* involved in producing those studies to ensure objective such evaluations.

## REFERENCES

[1] Crump KS, Van Landingham C, Bowers TS, Dexter C, Chyandalia JK. 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. Critical Reviews in Toxicology 2013; 45(9): 785-99. http://dx.doi.org/10.3109/10408444.2013.832726

[2] Pirkle JL, Brody DJ, Gunter EW, Kramer RA, Paschal DC, Flegal KM, *et al*. The Decline in Blood Lead Levels in the United States: The National Health and Nutrition Examinations Surveys (NHANES) Journal of the American Medical Association 1994; 272(4): 284-91. http://dx.doi.org/10.1001/jama.1994.03520040046039

[3] Lanphear BP, Hornung R, Khoury J, Yolton K, Baghurst P, Bellinger DC, *et al*. Low-Level Environmental Lead Exposure and Chidlren's Intellectual Function: An International Pooled Analysis. Environmental Health Perspectives 2005; 113(7): 894-9. http://dx.doi.org/10.1289/ehp.7688

[4] Schulz KF, Altman DG, Moher D, Consort_Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. British Medical Journal 2010; 340: 698-702. http://dx.doi.org/10.1136/bmj.c332

[5] Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW. Reporting of Noninferiority and Equivalence Randomized Trials: An Extension of the CONSORT Statement. Journal of the American Medical Association 2006; 295(10): 1152-60. http://dx.doi.org/10.1001/jama.295.10.1152

[6] Altman DG, Schulz KF, Hoher D, Egger M, Davidoff F, Elbourne D, *et al*. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Annals of Internal Medicine 2001; 134(8): 663-94. http://dx.doi.org/10.7326/0003-4819-134-8-200104170-00012

[7] Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, *et al*. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. Journal of the American Medical Association 1999; 276: 637-9. http://dx.doi.org/10.1001/jama.1996.03540080059030

[8] Rennie D. How to report randomized controlled trials: The CONSORT Statement. Journal of the American Medical Association 1996; 276(8): 649. http://dx.doi.org/10.1001/jama.1996.03540080071033

[9] Moses LE, Mosteller F. Experimentation: Just do it! 1995.

[10] Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. Journal of the American Medical Association 2005; 294(2): 218-28. http://dx.doi.org/10.1001/jama.294.2.218

[11] Keyfitz N, editor. Why forecasts fail and policies are often frustrated. Oxford: Clarendon Press; 1997.

[12] Lanphear BP, Hornung R, Khoury J, Dietrich KN, D.A. C-S, Canfiled RL. The Conundrum of unmeasured confounding: "can some of the detrimental neurodevelopmental effects attributed to lead be due to pesticides? by Brian Gulson". Sci Total Environ 2008; 396(2-3): 196-200. http://dx.doi.org/10.1016/j.scitotenv.2008.01.039

[13] Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. Coming to Terms with the Terms of Risk. Archives of General Psychiatry 1997; 54: 337-43. http://dx.doi.org/10.1001/archpsyc.1997.01830160065009

[14] Kraemer HC, Stice E, Kazdin A, Kupfer D. How do risk factors work together to produce an outcome? Mediators, Moderators, Independent, Overlapping and Proxy Risk Factors. The American Journal of Psychiatry 2001; 158: 848-56. http://dx.doi.org/10.1176/appi.ajp.158.6.848

[15] Smith MA, Grant LD, Sors AI. Lead Exposure and Child Development Dordrecht/Boston/London Kluwer Academic Publishers 1989.

[16] Shannon M, Graef JW. Lead Intoxication in Children with Pervasive Developmnetal Disorders Clinical Toxicology 1996; 34(2): 177-81.

[17] Pocock SJ, Smith M, Baghurst P. Environmental lead and children's intelligence: a systematic review of the epidemiological evidence. . British Journal of Medicine 1994; 309: 11898-1197. http://dx.doi.org/10.1136/bmj.309.6963.1189

[18] Thacker SB, Hoffman DA, Smith J, Steinberg K, Zack M. Effect of Low-level Body Burdens of Lead on the Mental Development of children: Limitations of Meta-analysis in a Review of Longitudinal Data. Archives of Environmental Health 1992; 47(5): 336-46. http://dx.doi.org/10.1080/00039896.1992.9938372

[19] Wortman PM. Judging Research Quality. In: Cooper H, Hedges LV, editors. The Handbook of Research Synthesis. New York: Russell Sage Foundation; 1994. p. 97-109.

[20] Hall JA, Tickle-Degnen L, Rosenthal R, Mosteller F. Hypotheses and Problems in Research Synthesis. In: Cooper H, Hedges LV, editors. The Handbook of Research Synthesis. New York: Russel Sage Foundation; 1994. p. 17-28.

[21] Simpson EH. The Interpretation of Interaction in Contingency Tables. Journal Of The Royal Statistical Society, SerB 1951; 13: 238-41.

[22] Samuels ML. Simpson's Paradox and Related Phenomena. Journal of the American Statistical Association 1951; 88: 81-8.

[23] Kraemer HC. Individual and Ecological Correlation in a General Context: Investigation of Testosterone and Orgasmic Frequency in the Human Male. Behavioral Science 1978; 23: 67-72. http://dx.doi.org/10.1002/bs.3830230203

[24] Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

[25] Box GEP, Draper NR. Empirical Model-Building and Response Surfaces. New York, NY: John Wiley & Sons, Inc.; 1986.

[26] Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychological Methods 2000; 5(2): 241-301. http://dx.doi.org/10.1037/1082-989X.5.2.241

[27] Krantz DH. The null hypothesis testing controversy in psychology. Journal of the American Statistical Association 1999; 44(448): 1372-81. http://dx.doi.org/10.1080/01621459.1999.10473888

[28] Thompson B. Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. Educational Psychology Review 1999; 11: 157-69. http://dx.doi.org/10.1023/A:1022028509820

[29] Wilkinson L, The_Task_Force_on_Statistical_Inference. Statistical Methods in Psychology Journals: Guidelines and Explanations. American Psychologist 1999; 54: 594-604. http://dx.doi.org/10.1037/0003-066X.54.8.594

[30] Shrout PE. Should significance tests be banned? Introduction to a special section exploring the pros and cons. Psychological Science 1997; 8(1): 1-2. http://dx.doi.org/10.1111/j.1467-9280.1997.tb00533.x

[31] Hunter JE. Needed: A ban on the significance test. Psychological Science 1997; 8(1): 3-7. http://dx.doi.org/10.1111/j.1467-9280.1997.tb00534.x

[32] Trafimow D, Marks M. Editorial. Basic and Applied Social Psychology 2015; 37: 1-3. http://dx.doi.org/10.1080/01973533.2015.1012991

[33] Trafimow D. Editorial. Basic and Applied Social Psychology 2014; 36(1): 1-2. http://dx.doi.org/10.1080/01973533.2014.865505

[34] Rutter M. Epidemiological methods to tackle causal questions. International Journal of Epidemiology 2009; 38: 3-6. http://dx.doi.org/10.1093/ije/dyn253

[35] Tesla N. BrainyQuote.com 2015 [February 26,2015 ]. Available from: http://www.brainyquote/quotes/authros/n/nikola_tesla.html

Received on 16-02-2015 Accepted on 24-03-2015 Published on 21-05-2015

http://dx.doi.org/10.6000/1929-6029.2015.04.02.1