Combining Datasets to Predict the Effects of Regulation of Environmental Lead Exposure in Housing Stock

Warren J. Strauss,1 Raymond J. Carroll,2 Steven M. Bortnick,1 John R. Menkedick,1 and Bradley D. Schultz3

1 Department of Statistics and Data Analysis Systems, Battelle
505 King Avenue, Columbus, Ohio 43201-2693, U.S.A.

2 Department of Statistics, Texas A&M University
College Station, Texas 77843-3143, U.S.A.

3 United States Environmental Protection Agency
Office of Pollution Prevention and Toxics (7404)
401 M Street SW, Washington, DC 20460, U.S.A.

Summary

A model for children’s blood-lead concentrations as a function of environmental-lead exposures was developed by combining two nationally representative sources of data which characterize the marginal distributions of blood-lead and environmental-lead, with a third regional dataset which contains joint measures of blood-lead and environmental-lead. The complicating factor addressed in this paper was the fact that methods for assessing environmental-lead were different in the national and regional datasets. Relying upon an assumption of transportability (that although the marginal distributions of blood-lead and environmental-lead may be different between the regional dataset and the nation as a whole, the joint relationship between blood-lead and environmental-lead is the same), the model makes use of a latent variable approach to estimate the joint distribution of blood-lead and environmental-lead nationwide.

Key words and phrases: combining data, environmental statistics, latent variable, lead poisoning, measurement error, residential hazards, transportability.
1. Introduction

Despite dramatic reductions in average blood-lead levels over the past 15 years, lead poisoning continues to be a significant health risk for young children (CDC, 1997). Permanent injuries caused by lead poisoning include cognitive impairments which are likely to affect a child’s development, educational potential, and subsequent ability to function as an adult. With the reductions of lead in air and food, lead in paint, dust and soil have been identified as the principal remaining sources of lead exposure for children. In the U.S., a non-trivial fraction of the housing stock contains significant levels of lead in one or more of the three principal sources, often resulting in a hazardous environment for children. In response to requirements mandated by the Residential Lead-Based Paint Hazard Reduction Act of 1992, the Environmental Protection Agency (EPA) has taken on the responsibility of developing a set of health-based standards for lead in residential environmental media (PROPOSED RULE FR NOTICE). These health-based standards represent levels of lead in paint, dust and soil, above which there may be a significant health risk for children.

Due to the significant public health and economic ramifications of these health-based standards, EPA must also conduct research to demonstrate the costs and benefits associated with different regulatory options. One method for assessing the impact of the health-based standards is to develop a model relating children’s blood-lead concentration to measures of residential lead exposure from paint, dust and soil. This model could then be used to predict the change in the national distribution of blood-lead attributable to changes in residential lead exposure resulting from different standards. The model must be applicable on a national basis. There are two nationally representative surveys with data applicable to this effort. The first survey is the third National Health and Nutrition Examination Survey (NHANES III) which characterizes the national distribution of children’s blood-lead concentrations (CDC, 1997). The second is the U.S. Department of Housing and Urban Development National Survey of Lead-Based Paint in Housing (HUD National Survey) which characterizes the national distribution of residential lead exposure in paint, dust and soil (US
EPA, 1995). Unfortunately, neither of these two nationally representative surveys contain both blood lead (the response variable of interest) and environmental lead (the predictor variables of interest). Data from a third study (The Rochester Lead-in-Dust Study) does contain joint measures of children’s blood-lead and residential lead exposure, but only for the Rochester, New York region (US HUD, 1995). The combination of these three sources of data to develop a model that can be used to predict a national distribution of blood lead as a function of environmental lead levels as measured in the HUD National Survey is the subject of this paper. This model may then be used to predict changes in the national distribution of children’s blood-lead associated with estimated reductions in the national distribution of environmental-lead levels attributable to different regulatory options.

2. Data and Issues

The primary source of information on environmental lead levels in the national housing stock was the HUD National Survey (US EPA, 1995), an effort to obtain data for estimating the prevalence of lead-based paint and lead contaminated dust and soil in the nation’s housing stock. The HUD National survey, conducted between 1989 and 1990, measured lead levels in paint, dust and soil from 284 privately-owned, occupied housing units. The housing units were selected using a statistically-based sampling design to represent the national housing stock built prior to 1980. Houses built after 1980 were assumed to be free of lead-based paint due to the Consumer Product Safety Commission’s ban on the sale of residential lead-based paint in 1978. Although each of the housing units included in the HUD National Survey was occupied at the time of sampling, information on the blood-lead concentration of resident children was not part of the sampling design and was not included in the survey. Therefore, the HUD National Survey cannot be used on its own to develop a model relating blood-lead to residential lead exposure.

Data from Phase 2 of the third National Health and Nutritional Examination Survey
(NHANES III) can be used to characterize the current national distribution of children’s blood-lead concentrations, however it contains no information on residential lead exposures. NHANES III, Phase 2, conducted from 1991 to 1994, was the seventh in a series of national examination studies conducted by CDC’s National Center for Health Statistics (NCHS) to trace the health and nutritional status of the non-institutionalized, civilian U.S. population. To provide for a nationally representative sample and sufficient precision in characterizing key sub-populations, a complex survey design was employed in NHANES III. Approximately 13,000 persons were sampled in NHANES III, Phase 2, including approximately 987 children one and two years old. As a result, the NHANES III, Phase 2 provides a solid basis for estimating the current national distribution of blood-lead concentrations for children aged 1-2 years.

Unfortunately, a nationally representative source of data with joint measures of blood-lead and residential lead exposures does not exist. However, joint measures of blood-lead and residential-lead were collected in the Rochester Lead-in-Dust Study (US HUD, 1995), a cross-sectional study that recruited 205 children aged 12-31 months in Rochester, New York using a stratified sampling scheme in 1993. The sampling scheme was designed to recruit a high proportion of low income families living in older (pre-1940) housing. Environmental assessment at the primary residence of each recruited child was generally completed within three weeks of the date of blood sample collection, and included samples of paint, dust and soil collected using methods in accordance with a standard lead risk assessment procedure.

A log-linear regression model was developed using the data from the Rochester Lead-in-Dust Study which expressed children’s blood-lead concentrations as a function of the following measures of residential lead exposure: floor dust-lead loading, window sill dust-lead loading, soil-lead concentration, and an indicator of paint/pica hazard.

While the Rochester Lead-in-Dust Study is not nationally representative, the relationships
between children=s blood-lead concentration and measures of residential lead exposure may be considered nationally representative if the exposure and uptake mechanisms, on average, are the same in Rochester and the nation as a whole. While this was not able to be verified empirically, it is a reasonable assumption. Therefore, even though the marginal distributions of children=s blood-lead and residential lead exposure in Rochester are not the same as those in NHANES III or the HUD National Survey, the joint relationships between blood-lead and residential-lead observed in Rochester may serve as a basis for a nationally representative model.

Application of the model relating blood-lead to environmental-lead in Rochester to the nation as a whole would be straightforward, given the assumption discussed above, were it not for the fact that the techniques used to measure lead exposures in the HUD National Survey are different from those used in the Rochester Lead-in-Dust Study. At issue then, in this paper, is how to combine these three sources of data to develop a nationally representative model for the evaluation of EPA=s proposed health-based standards given that some variables have a different interpretation in each of the two studies.

For example, the dust samples on floors and window sills were collected using different measurement technologies in these two studies; by swabbing the surface with a wet wipe in the Rochester Study, and by a low power vacuum in the HUD National Survey, often referred to as the Ablue nozzle@ vacuum method. Another type of measurement difference between the two studies occurred in soil sampling. Although both studies used the same sample collection devices and similar laboratory equipment to analyze the soil samples, the Rochester Study included samples of soil from the perimeter drip-line of the residence, while the HUD National survey included samples of soil from both the drip-line and from other more remote locations in the yard. The paint/pica hazard indicator variable included in the predictive model was assumed to be measured similarly in both the Rochester Study and the HUD National Survey.
Table 1 provides summary statistics for the distributions of environmental lead in floor dust, window sill dust, soil and paint/pica for the Rochester Study and the HUD National Survey. The distribution of wipe sample floor dust-lead loading in the Rochester Study had a geometric mean of 17.9 $\mu$g/ft$^2$ and a geometric standard deviation of 3.2 $\mu$g/ft$^2$. The geometric mean and geometric standard deviation are calculated respectively as the exponentiated mean and standard deviation of the observed natural-log transformed distribution. The corresponding distribution of blue-nozzle dust lead loading observed in the HUD National Survey (before regulation) had a geometric mean of 1.7 $\mu$g/ft$^2$ with a geometric standard deviation of 4.7 $\mu$g/ft$^2$. The differences between these distribution parameters (geometric mean and standard deviation) can be explained by differences in measurement technology (wipe versus blue nozzle), as well as differences in the distribution of environmental lead between the Rochester region and the nation as a whole.

Table 1 Placed Here

Our approach provides a methodology for adjusting the regression model based on data from the Rochester Study to appropriately use environmental lead levels observed in the HUD National Survey as inputs to the model. The adjustment takes into account both systematic differences and differences in error structures between the Rochester predictor variables and the HUD National Survey predictor variables. The method provides a relationship between blood-lead concentration and a set of lead exposure variables and other covariates as they were measured in the HUD National Survey. The method does not provide for a relationship between blood-lead and environmental-lead measured without error, as observed environmental-lead levels will be used as input to the model. The method is summarized as follows.

Let $Y$ represent the natural log of children=s blood-lead levels and let $R$ and $H$ represent lead exposure(s) using the method of measurement in the Rochester Study and the HUD National Survey, respectively. Also, let $C$ represent other covariates of interest that are measured similarly in both the
Rochester Study and the HUD National Survey. We need to regress $Y$ on $(H,C)$, but there are no data to do so directly. Instead, we will link the variables and data sets using a latent variable, $X$, which represents true lead exposures based on the method of measurement used in the Rochester study (representing the true value of the lead exposure as measured by the methods used in the Rochester Study, $X$ may be conceptualized as the average of an infinite number of $R$ samples).

3. Assumptions

Of chief concern is the estimation of the parameters in the following model:

$$Y_i = \alpha_{Y|H,C} + \beta_{Y|H,C} H_i + \beta_{Y|C,H} C_i + e_{Y|H,C},$$

where the errors are independent with mean zero and variance $\sigma^2_{Y|H,C}$.

In order to estimate the parameters in (1), two important assumptions will be made. First, we assume that $(Y,R,H,X)$ given $C$ follows a joint normal distribution, both in the U.S. as a whole, and separately in Rochester (i.e. with different MVN parameters for the U.S. and for Rochester). While the multivariate normal assumption could not be verified directly due to the above described limitations on observable data, this assumption was considered reasonable based on indirect evidence in the form of preliminary analyses (i.e., histograms and normal probability plots) on the available marginal and conditional data, as seen in Figure 1. Second, letting $j$ denote region (=1 for Rochester, =2 for U.S.) and $i$ denote study participant, our assumed model is as follows:

$$Y_{ij} = \alpha_{Y|X,C,j} + \beta_{Y|X,C} X_{ij} + \beta_{Y|X,C} C_{ij} + e_{Y|X,C,ij};$$

$$R_{ij} = X_{ij} + e_{R|X,ij};$$

$$H_{ij} = \alpha_{H|X,j} + \beta_{H|X} X_{ij} + e_{H|X,ij};$$

$$X_{ij} = \alpha_{X|C,j} + \beta_{X|C} C_{ij} + e_{X|C,ij}.$$
where the error terms are independent and normally distributed with mean zero.

The residual variance of $Y$ given $(X, C)$ is $\sigma^2_{Y|X,C}$ and is assumed to be independent of $j$. Later on in this analysis, we will use the NHANES III data to allow $\sigma^2_{Y|X,C}$ to differ between Rochester and the U.S. (Assuming different MVN distributions of $(Y,R,H,X)$ given $C$ between U.S. and Rochester). Since $(R,H,X)$ are vectors, each consisting of three elements corresponding to the media floor dust, window sill dust and soil, the remaining errors will have 3x3 covariance matrices denoted by $\Sigma_{R|X}$, $\Sigma_{H|X}$ and $\Sigma_{X|C}$, respectively, independent of $j$.

It is assumed that the various processes involved in measuring the different media for lead are independent of one another. Therefore, $\Sigma_{R|X}$ and $\Sigma_{H|X}$ are assumed to be diagonal matrices. For this same reason, the slope parameter $\beta_{H|X}$ in (2c) is assumed to be a diagonal matrix. For example, imprecision or bias in the measurement of lead in floor dust would not impact an observed measurement of lead in soil. Also, since $X$ represents true unobserved measurements and $H$ represents observed measurements, an obvious assumption for the relationship between $H$ and $X$ is that increases in $X$ are associated with increases in $H$. Thus, the diagonal elements of $\beta_{H|X}$ are assumed to be nonnegative. Historic tests on the efficacy of the measuring devices have validated this assumption in practice. Finally, the intercepts are allowed to depend on region ($j$), but the slopes are not, representing a form of the transportability assumption (Carroll, Ruppert, and Stefanski, 1995, pages 10-12). Of course, we observe only $Y_{i2}$ marginally; $Y_{i1}$, $R_{i1}$ and $C_{i1}$ jointly; and $H_{i2}$ and $C_{i2}$ jointly.

4. Parameter Development

In order to fit model (1), we need the joint distribution of $(Y,H)$ given $C$ in the U.S., from which we can construct the conditional distribution of $Y$ given $H$ and $C$, namely model (1). Write the joint distribution of $(Y,H)$ given $C$ as
where, using (2a) and (2d),

\[ \text{and (4)} \]

Next, using (2c) and (2d),

\[ \text{and (5)} \]

Finally, using (2a), (2c) and (2d),

\[ \text{and (6)} \]

From (3), Y given (H,C) is normally distributed with mean equal to

\[ \text{(7)} \]

and variance equal to

\[ \text{(8)} \]

Equations (4), (5) and (6) then can be substituted into (7) to obtain the formulas for the slope parameters in model (1) as

\[ \text{and (9)} \]

Similarly, substituting (4), (5) and (6) into (8) gives the conditional variability of Y given (H,C) as

\[ \text{(10)} \]
We now discuss how we estimated the parameters on the right-hand-sides of equations (9) and (10). First, $\Sigma_{R|X}$ and $\Sigma_{H|X}$ were estimated using variance components models of the individual environmental samples in the Rochester Study and the HUD National Survey, which ultimately were averaged to construct the predictor variables $R$ and $H$ in models (1) and (2a), respectively. For example, when estimating the floor dust-lead loading in Rochester homes and its associated measurement error, the model $R_{ij} = \mu_i + e_{ij}$ was applied; where $R_{ij}$ is the dust-lead level from the $j^{th}$ sample collected in the $i^{th}$ home, $\mu_i$ is the average dust-lead level in the $i^{th}$ home, and $e_{ij}$ is the error left unexplained by the model. It is assumed that the $e_{ij}$ follow a normal distribution with mean zero and variance $\sigma^2_{\text{error}}$, which represents error attributable to spatial, sampling and analytical variability. Thus, $\Sigma_{R|X}$ and $\Sigma_{H|X}$ are diagonal matrices containing estimates of $\sigma^2_{\text{error}}$ from each individual predictor variable represented in $R$ and $H$.

Next, the parameters $\beta_{Y|X(C)}$, $\beta_{Y|C(X)}$, and $\sigma^2_{Y|X,C}$ were estimated via a classical errors-in-variables technique using data from the Rochester Study as follows (Fuller, 1987):

\begin{equation}
\text{(11)}
\end{equation}

and

\begin{equation}
\text{(12)}
\end{equation}

Then, $\Sigma_{H|C}$ was estimated from the residual variance/covariance matrix resulting from a weighted least squares regression of $H$ on $C$ in the HUD National Survey data. The weights in this regression corresponded to the HUD National Survey weights for the number of residential units represented by each individual observation in the dataset, and were used to make the obtained estimates nationally representative.

The parameters $\beta_{X|C}$ and $\Sigma_{X|C}$ were estimated by regressing $R$ on $C$ in the Rochester data, after
noting from (2b) and (2d) that $\beta_{XC}$ is the slope of this regression while

$$\Sigma_{RC} = \Sigma_{XC} + \Sigma_{RX}$$

(13)

is the residual covariance matrix. The estimate of $\Sigma_{XC}$ was obtained by subtracting off the above described estimate of $\Sigma_{RX}$ from the estimate of $\Sigma_{RC}$.

Next, observe that since $\beta_{HX}$ is a diagonal matrix, $\beta_{HX}[\Sigma_{XC}]^D[\beta_{HX}]^T = \beta_{HX}^2[\Sigma_{XC}]^D$, where the notation $[X]^D$, represents the matrix $X$ diagonalized. Therefore, using the second relationship in (5) along with (13), and since the diagonal elements of $\beta_{HX}$ all are nonnegative, $\beta_{HX}$ was estimated via the square roots of the diagonal elements of $[\Sigma_{HC} - \Sigma_{HX}]^D ([\Sigma_{RC} - \Sigma_{RX}]^D)^{-1}$.

The above discussion yields estimates that, when combined according to (9) and (10), give solutions to the parameters $\beta_{Y(H,C)}$, $\beta_{Y(H)}$ and $^2Y_{Y,H,C}$ in model (1). The remaining model (1) parameter to be estimated is the intercept $\alpha_{Y,H,C}$. We used NHANES III data to estimate the intercept $\alpha_{Y,H,C}$ as follows. The overall mean of $Y$, $Y$, was estimated using NHANES III data, and the means of $H$ and $C$, $H$ and $C$, were estimated using the HUD data. In both datasets, weighted means were used in order to make the estimates nationally representative. Then,

$$Y_{Y,H,C} = Y - \beta_{Y(H,C)}H - \beta_{Y(C)}C.$$

(14)

In other words, we calibrated the intercept so that the model=s predicted national mean equals the NHANES III mean, which is consistent with the intent to relate the national distribution of blood-lead to a current national distribution of environmental lead.

Finally, the estimate of $^2Y_{Y,H,C}$ provided above assumes that the parameter $^2Y_{Y,H,C}$ is the same in Rochester and the U.S. This parameter represents the variability in blood-lead concentrations among children exposed to the same environmental conditions. This variability may be attributed to such factors as nutrition, non-residential lead exposure, and a variety of behavior patterns such as play time indoors versus outdoors, cleaning practice, etc. It is reasonable to expect that this variability
may be different in Rochester compared to the entire U.S. Therefore, an alternative estimate, which extends the notion of calibrating the model to NHANES III, allows for $\Theta^2_{Y|X,C}$ to differ between Rochester and the U.S. In this approach, the estimate of $\Theta^2_{Y|H,C}$ is such that the geometric standard deviation of the national distribution of children's blood-lead concentration predicted by the model under pre-regulation environmental conditions matches that of NHANES III. That is,

$$, \tag{15}$$

where $\Theta^2_Y$ is the nationally representative NHANES III variance in blood-lead concentrations on the natural log scale. The $w_i$'s in (15) are the HUD National Survey weights used to make the data nationally representative. Table 3 in the results section below provides estimates based on adjusting the model to match the NHANES III standard deviation, as well as its mean.

The standard errors of the parameter estimates for model (1) that are given in Table 2 below were estimated from bootstrap samples of the Rochester data set. See (Efron and Tibshirani, 1993) for further details on the bootstrap method of estimating standard errors.

5. Results

Table 2 compares the parameter estimates of a model that used the Rochester Study data only versus the Combined Datasets model given by (1). Keep in mind that the explanatory variables are different for the two models, based on differences in sampling and measurement methods associated with the dust and soil predictor variables. The parameter estimates for the Rochester Study model are based on a log-linear regression model of the observed data in the Rochester Study. The parameter estimates for the Combined Datasets model are derived from the estimation procedure detailed in sections 3-5, and correspond to the environmental predictor variables as they were measured in the pre-regulation HUD National Survey. In addition, the intercept and error attributable to random variability in this table are calibrated so that the geometric mean and standard deviation
of the predicted national distribution of blood-lead concentrations matches that of NHANES III when the Combined Datasets model is applied to the pre-regulation data from the HUD National Survey.

**Table 2 Placed Here**

Table 2 indicates that adjusting the Rochester Study model to be nationally representative (i.e. developing the Combined Datasets model) results in slope estimates much closer to zero for all three measurement error adjusted covariates (floor dust-lead loading, window sill dust-lead loading, and soil-lead concentration). In contrast, the Combined Datasets model has a much higher estimate of the intercept and slightly higher estimates of the error attributable to random variability and of the one covariate in the model not adjusted for measurement error, namely interior paint/pica hazard.

**Application of the Combined Datasets Model**

In order to predict a national distribution of blood-lead concentrations, the Combined Datasets model was applied to observed environmental-lead levels and estimated post-regulation modified environmental lead-levels from the HUD National Survey. To construct the distribution of post-regulation environmental-lead levels, observed levels of lead in environmental variables in the HUD National Survey were compared to proposed health based standards (e.g. 100 μg/ft² for floor dust-lead loading, 2000 μg/g for soil removal). Changes (motivated by the regulation) to the environmental-lead levels were assumed to occur in the HUD National Survey residential units that had levels of lead in environmental variables that were above the proposed standards. If a change was triggered, post-regulation lead levels for homes that were above the standard were adjusted downward. The model was used to predict a geometric mean blood-lead concentration associated with each residential unit in the HUD National Survey. Observed environmental-lead levels and estimated post-regulation modified environmental lead-levels at each HUD National Survey residential unit were used as input to the model to construct the pre- and post- regulatory geometric mean blood-lead level associated with each residential unit in the HUD National Survey. A log-normal distribution of blood-lead concentrations was then constructed for each residential unit, with
the predicted geometric mean, and a geometric standard deviation of 1.994 (determined by exponentiating the error attributable to random variability from the Combined Datasets model). The predicted national distribution of children’s blood-lead concentrations was then constructed using a weighted average (with weights from the HUD National Survey) of the resulting log-normal distribution associated with each residential unit.

Of primary interest to EPA was estimation of the percent of children nationwide whose blood-lead concentration would equal or exceed certain health thresholds (e.g. 10, 20, 30 μg/dL). The estimated percentage of children exceeding these thresholds were called exceedance percentiles. Table 3 displays the geometric mean, geometric standard deviation, and three exceedance percentiles of interest for the pre-regulation national blood-lead distribution as estimated by NHANES III, the model based on the Rochester Study data only, and the Combined Datasets model, and one post-regulation distribution estimated using the Combined Datasets model.

**Table 3 Placed Here**

Table 3 highlights several results of interest, including the following:

1. The estimated proportions of pre-regulation blood-lead concentrations greater than or equal to 10, 20 or 30 μg/dL using the Combined Datasets model predictions are quite similar to the corresponding proportions estimated by NHANES III.
2. The Combined Datasets model predicts about a 15%, 30% and 40% reduction in the proportions of blood-lead concentrations at or above 10, 20 or 30 μg/dL, respectively, due to the estimated effect of the implementation of the set of health-based standards used as an example in Table 3.

Note that Table 3 presents results for only one postulated post-regulation scenario. The Combined Datasets model was used by EPA to estimate the effect of many different scenarios and to assess the
relative risk reduction with respect to the costs of implementation for the various different scenarios.

6. Discussion of Other Statistical Methods

In building parameter estimates for model (1), we used a latent variable model (2a) - (2d). While the parameter estimation technique was the method of moments, to infer (1) from (2a) - (2d) requires assumptions of (log)normality. An alternate approach is to use likelihood-type analysis based on such parametric assumptions. To do this, one must account for the complex survey structure of NHANES III and the HUD National Survey. One possibility, which we are currently investigating, is to use a weighted likelihood approach, i.e, weight the log-likelihood in these surveys by the sampling weights. The approach is effectively a weighted estimating equation approach. Potentially, using ideas of Robins, et al. (1994), such an approach might lead to more efficient parameter estimates than the method of moments.

ACKNOWLEDGEMENTS

This research was supported by Contract No. 68-D5-008 through the U.S. Environmental Protection Agency. Carroll=s research was also supported in part by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M University Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30 - E509106).
REFERENCES


Table 1  Summary Statistics for the Marginal Distributions of Environmental Lead in the Regional Data Set and in the HUD National Survey.

<table>
<thead>
<tr>
<th>Environmental Media</th>
<th>Distributional Parameter</th>
<th>Regional Data (Rochester)</th>
<th>HUD National Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Floor Dust-Lead Loading</td>
<td>Measurement Technology</td>
<td>Wipe</td>
<td>Blue Nozzle Vacuum</td>
</tr>
<tr>
<td></td>
<td>Geometric Mean</td>
<td>17.9</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>Geometric Standard Deviation</td>
<td>3.2</td>
<td>4.7</td>
</tr>
<tr>
<td>Window Sill Dust-Lead Loading</td>
<td>Measurement Technology</td>
<td>Wipe</td>
<td>Blue Nozzle Vacuum</td>
</tr>
<tr>
<td></td>
<td>Geometric Mean</td>
<td>195.6</td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td>Geometric Standard Deviation</td>
<td>3.9</td>
<td>10.0</td>
</tr>
<tr>
<td>Soil-Lead Concentration</td>
<td>Location of Sample Collection</td>
<td>Drip-Line</td>
<td>Drip-Line and Remote</td>
</tr>
<tr>
<td></td>
<td>Geometric Mean</td>
<td>729.2</td>
<td>61.9</td>
</tr>
<tr>
<td></td>
<td>Geometric Standard Deviation</td>
<td>3.6</td>
<td>4.5</td>
</tr>
<tr>
<td>Indicator of Interior Paint / Pica Hazard</td>
<td>Measurement Technology</td>
<td>Portable XRF</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% with Paint/Pica Hazard</td>
<td>9.5%</td>
<td>1.36%</td>
</tr>
</tbody>
</table>
Table 2. Parameter Estimates and (Associated Standard Errors) for the Model Based on the Rochester Study Data and the Combined Datasets Model.

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Rochester Study Model</th>
<th></th>
<th>Combined Datasets Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter</td>
<td>Estimate</td>
<td>Parameter</td>
<td>Estimate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(S.E.)</td>
<td></td>
<td>(S.E.)</td>
</tr>
<tr>
<td>Intercept</td>
<td>$\alpha_{Y</td>
<td>R_1,R_2,R_1,C}$</td>
<td>0.418 (0.240)</td>
<td>$\alpha_{Y</td>
</tr>
<tr>
<td>Area-Weighted Arithmetic Mean</td>
<td>$\beta_{Y</td>
<td>R_1(C)}$</td>
<td>0.066 (0.040)</td>
<td>$\beta_{Y</td>
</tr>
<tr>
<td>Floor Dust-Lead Loading</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area-Weighted Arithmetic Mean</td>
<td>$\beta_{Y</td>
<td>R_2(C)}$</td>
<td>0.087 (0.036)</td>
<td>$\beta_{Y</td>
</tr>
<tr>
<td>Window Sill Dust-Lead Loading</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soil-Lead Concentration</td>
<td>$\beta_{Y</td>
<td>R_1(C)}$</td>
<td>0.114 (0.035)</td>
<td>$\beta_{Y</td>
</tr>
<tr>
<td>Indicator of Interior Paint/Pica</td>
<td>$\beta_{Y</td>
<td>C(R_1,R_2,R_1)}$</td>
<td>0.248 (0.100)</td>
<td>$\beta_{Y</td>
</tr>
<tr>
<td>Error Attributable to Random Variability</td>
<td>$\sigma_{Y</td>
<td>R_1,R_2,R_1,C}$</td>
<td>0.562</td>
<td>$\sigma_{Y</td>
</tr>
</tbody>
</table>

1 Standard errors associated with parameter estimates from the Combined Datasets Model were calculate via non-parametric bootstrap of data from the Rochester Study.
Table 3. Predicted National Distribution Characteristics for NHANES III, the Rochester Study Model, and the Combined Datasets Model.

<table>
<thead>
<tr>
<th>Predicted Model Results</th>
<th>Parameter</th>
<th>Pre-Regulation Blood-lead Levels</th>
<th>Post-Regulation Blood-lead Levels²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NHANES III</td>
<td>Rochester Study</td>
</tr>
<tr>
<td>National Geometric Mean</td>
<td>( \mu_p )</td>
<td>3.14</td>
<td>6.36</td>
</tr>
<tr>
<td>National Geometric Standard Deviation</td>
<td>( \sigma_p )</td>
<td>2.09</td>
<td>1.85</td>
</tr>
<tr>
<td>Estimated Distribution Exceedance Percentiles</td>
<td>% ( \geq 10 \mu g/dL )</td>
<td>5.88%</td>
<td>22.90%</td>
</tr>
<tr>
<td></td>
<td>% ( \geq 20 \mu g/dL )</td>
<td>0.43%</td>
<td>2.90%</td>
</tr>
<tr>
<td></td>
<td>% ( \geq 30 \mu g/dL )</td>
<td>0.07%</td>
<td>1.00%</td>
</tr>
</tbody>
</table>

¹ The estimates from the Combined Datasets model are those that would be obtained when adjusting the model to produce a pre-regulation estimated national geometric standard deviation equal to that of NHANES III. This adjustment was not applied in the development of the model for the assessment of the implementation of EPA health-based standards.

² The proposed health based standards used to calculate this example of a post regulation distribution were 100 \( \mu g/ft^2 \) for floor dust-lead loading (on a wipe scale), 500 \( \mu g/ft^2 \) for window sill dust-lead loading (on a wipe scale), 2000 \( \mu g/g \) for soil removal, 5 ft² damaged lead-based paint (LBP) for paint repair, and 20 ft² damaged LBP for paint abatement. For homes which exceeded a standard, the estimated effect of the regulation was to reduce the environmental level(s) which exceeded the standard to: 40 \( \mu g/ft^2 \) for floor dust-lead loading (on a wipe scale), 100 \( \mu g/ft^2 \) for window sill dust-lead loading (on a wipe scale), 150 \( \mu g/g \) for soil, and 0 ft² damaged LBP.
Figure 1  Joint distribution of blood-lead and environmental-lead observed in the three data sources.