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

1Department of Statistics and Data Analysis Systems, Battelle, 505 King Avenue, Columbus, Ohio 43201-2693, U.S.A.
2Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.
3U.S. Environmental Protection Agency Office of Pollution Prevention and Toxics (7404), 401 M Street SW, Washington, DC 20460, U.S.A.

*email: strauss@battelle.org

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 that characterize the marginal distributions of blood lead and environmental lead with a third regional dataset that contains joint measures of blood lead and environmental lead. The complicating factor addressed in this article was the fact that methods for assessing environmental lead were different in the national and regional datasets. Relying on 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: 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 (Centers for Disease Control, 1997). Permanent injuries caused by lead poisoning include cognitive impairments that 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 United States, a nontrivial 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 (U.S. Environmental Protection Agency, 1998b). 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, the EPA must also conduct research to demonstrate the costs and benefits associated with different regulatory options (U.S. Environmental Protection Agency, 1998a). 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 (Centers for Disease Control, 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 (U.S. Environmental Protection Agency, 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 (U.S. Department of Housing and Urban Development, 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 article. 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 (U.S. Environmental Protection Agency, 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 the Centers for Disease Control's National Center for Health Statistics (NCHS) to trace the health and nutritional status of the noninstitutionalized, civilian U.S. population. To provide for a nationally representative sample and sufficient precision in characterizing key subpopulations, a complex survey design was employed in NHANES III. Approximately 13,000 persons were sampled in NHANES III, phase 2, including approximately 987 one- and two-year-old children. 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 (Centers for Disease Control, 1992).

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 (U.S. Department of Housing and Urban Development, 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 3 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 article, 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 blue 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 provides a measure of the interaction between presence of deteriorating lead-based paint in the child's residence and the child's tendency toward pica or strong mouthing behavior. This variable 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 differences between these distribution parameters across data sources (geometric mean and standard deviation) can be explained by differences in measurement technology (wipe versus blue nozzle) as well as differences in the true
Combining Datasets to Predict Lead Exposure

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>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>

The method 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 is summarized as follows.

Let $Y$ represent the natural log of children’s blood lead levels and let $R$ and $H$ represent the natural log of lead exposure(s) using the method of measurement in the Rochester Study and the HUD National Survey, respectively. Also, let $C$ represent other covariate(s) of interest that are measured similarly in both the Rochester Study and the HUD National Survey (the paint/pica variable). 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). It is assumed that the distribution of $X \mid C$ is the same in Rochester and the HUD National Survey.

3. Assumptions

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

$$Y_{ij} = \alpha Y_{[H,C]} + \beta Y_{[H(C)]} H_i + \beta Y_{[C(H)]} C_i + e_{Y[H,C,i]},$$  \hspace{1cm} (1)

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 United States as a whole and separately in Rochester (i.e., with different multivariate normal [MVN] parameters for the United States and Rochester). While the multivariate normal assumption could not be verified directly due to the above-described limitations on observable data, what is required in our development is that various error distributions be normally distributed. This assumption can be directly verified in Rochester by assessing the distribution of $Y \mid R, C$, as seen in Figure 1. Second, letting $j$ denote region ($=1$ for Rochester, $=2$ for the United States) and $i$ denote study participant, our assumed model is as follows:

$$Y_{ij} = \alpha Y_{[X,C,j]} + \beta Y_{[X(C)]} X_{ij} + \beta Y_{[C(X)]} C_{ij} + e_{Y[X,C,i]},$$  \hspace{1cm} (2a)

Figure 1. Histogram of residuals from a regression of blood lead on environmental lead.
\[ R_{ij} = X_{ij} + \epsilon_{R,X,ij}, \]  
\[ H_{ij} = \alpha_{H|X,j} + \beta_{H|X} X_{ij} + \epsilon_{H,X,ij}, \]  
\[ X_{ij} = \alpha_{X|C,j} + \beta_{X|C} C_{ij} + \epsilon_{X,C,ij}, \]

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 in this analysis, we will use the NHANES III data to allow \( \sigma^2_{Y|X,C} \) to differ between Rochester and the United States (assuming different MVN distributions of \((Y, R, H, X)\) given \( C \) between the United States 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 \( 3 \times 3 \) 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 another. Therefore, \( \Sigma_{X|C} \) will have \( 3 \times 3 \) covariance matrices denoted by \( \Sigma_{R|X}, \Sigma_{H|X}, \) and \( \Sigma_{X|C}, \) respectively, independent of \( j \).  

4. Parameter Development

In order to fit model (1), we need the joint distribution of \((Y, H)\) given \( C \) in the United States, 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

\[
\begin{bmatrix} Y_{ij} | C_{ij} \\ H_{ij} | C_{ij} \end{bmatrix} \sim \text{MVN} \left\{ \begin{bmatrix} \mu_{Y|C} \\ \mu_{H|C} \end{bmatrix}, \begin{bmatrix} \sigma^2_{Y|C} & \Sigma_{YH|C} \\ \Sigma_{YH|C}^T & \Sigma_{H|C} \end{bmatrix} \right\};
\]

where, using (2a) and (2d),

\[
\mu_{Y|C} = \alpha_{Y|X,C} + \beta_{Y|X} (C_{ij}) (\alpha_{X|C} + \beta_{X|C} C_{ij}) + \beta_{Y|X} C_{ij},
\]

\[
\sigma^2_{Y|C} = \beta_{Y|X} (\Sigma_{X|C} \beta_{Y|X}^T) + \sigma^2_{Y|X,C}.
\]

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

\[
\mu_{H|C} = \alpha_{H|X,j} + \beta_{H|X} (\alpha_{X|C} + \beta_{X|C} C_{ij}),
\]

\[
\Sigma_{H|C} = \beta_{H|X} (\Sigma_{X|C} \beta_{H|X}^T) + \Sigma_{H|X}.
\]

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

\[
\Sigma_{YH|C} = \beta_{Y|X} (\Sigma_{X|C} \beta_{H|X}^T).
\]

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

\[
\mu_{Y|H,C} = \mu_{Y|C} + \Sigma_{YH|C} (H_{ij} - \mu_{H|C})
\]

and variance equal to

\[
\sigma^2_{Y|H,C} = \sigma^2_{Y|C} - \Sigma_{YH|C} \Sigma_{H|C}^{-1} \Sigma_{YH|C}^T.
\]

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

\[
\beta_{Y|H|C} = \beta_{Y|X|C} \Sigma_{X|C} \beta_{H|X}^T \Sigma_{H|C}^{-1},
\]

\[
\beta_{Y|H} = \beta_{Y|X|C} + \beta_{Y|X} \Sigma_{X|C} \Sigma_{H|C}^{-1} \Sigma_{H|X} C_{ij}.
\]

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

\[
\sigma^2_{Y|H,C} = \sigma^2_{Y|C} + \beta_{Y|X} (\Sigma_{X|C} \beta_{Y|X}^T) + \sigma^2_{Y|X,C}.
\]

We now discuss how we estimated the parameters on the right-hand sides of equations (9) and (10). Replicate samples were available (three to six per residential unit) in the Rochester Study and the HUD National Survey, thereby allowing estimation of \( \Sigma_{R|X} \) and \( \Sigma_{H|X} \) (Carroll et al., 1995). Next, the parameters \( \beta_{Y|X|C}, \beta_{Y|X|C}(X), \) 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):

\[
\begin{bmatrix} \beta_{Y|X|C} \\ \beta_{Y|X|C}(X) \end{bmatrix} = \beta_{Y|X,C} = (R^T R - n \Sigma_{R|X})^{-1} R^T Y
\]

and

\[
\sigma^2_{Y|X,C} = \frac{1}{n - p} \left[ (Y - R \beta_{Y|X,C})^T (Y - R \beta_{Y|X,C}) - (n - p) \beta_{Y|X,C} \right] \sigma_{Y|X,C} \Sigma_{R|X}^2 \beta_{Y|X,C}. \]

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 \( \beta_{X|C} \) is the slope of this regression while

\[
\Sigma_{R|X} = \Sigma_{X|C} + \Sigma_{R|X}
\]

is the residual covariance matrix. The estimate of \( \Sigma_{X|C} \) was obtained by subtracting off the above-described estimate of \( \Sigma_{R|X} \) from the estimate of \( \Sigma_{R|X} \).

Next observe that, since \( \beta_{H|X} \) is a diagonal matrix, \( \beta_{H|X} X \) and \( \Sigma_{X|C} \) are diagonalized. Therefore, using the second relationship in (5) along with (13) and since the diagonal
elements of $\beta_{H|X}$ all are nonnegative, $\beta_{H|X}$ was estimated via the square roots of the diagonal elements of $\Sigma_{H|C}^{-1} - \Sigma_{H|C}^{-1}(\Sigma_{R|C} - \Sigma_{R|C})^{-1}$. Alternatively, a different method of moments estimator of $\beta_{H|X}$ is $[\beta_{H|C}^2]^{1/2}$. The former was chosen for this analysis because it takes advantage of the replicate data in $R$ and $H$.

The above discussion yields estimates that, when combined according to (9) and (10), give solutions to the parameters $\beta_{Y|H(C)}$, $\beta_{Y|H(C)}$, and $\sigma_{Y|H,C}^2$ 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 $Y$, $\mu_Y$, was estimated using NHANES III data, and the means of $H$ and $C$, $\mu_H$ and $\mu_C$, respectively, were estimated using the HUD data. In both datasets, weighted means and variances were used in order to make the estimates nationally representative. Then

$$\alpha_{Y|H,C} = \mu_Y - \beta_{Y|H(C)} \mu_H - \beta_{Y|C(H)} \mu_C. \quad (14)$$

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 $\sigma_{Y|H,C}^2$ provided above assumes that the parameter $\sigma_{Y|X,C}^2$ is the same in Rochester and the United States. 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, nonresidential lead exposure, and a variety of behavior patterns such as play time indoors versus outdoors, cleaning practices, etc. It is reasonable to expect that this variability may differ in Rochester compared to the entire United States. Therefore, an alternative estimate, which extends the notion of calibrating the model to NHANES III, allows for $\sigma_{Y|X,C}^2$ to differ between Rochester and the United States. In this approach, the estimate of $\sigma_{Y|H,C}^2$ is such that the geometric standard deviation of the national distribution of children’s blood lead concentration predicted by the model under preregulation environmental conditions matches that of NHANES III, i.e.,

$$\sigma_{Y|H,C}^2 = \sigma_Y^2 - \sum_{i=1}^n w_i \left[ \beta_{Y|H(C)} (H_i - \mu_H) + \beta_{Y|C(H)} (C_i - \mu_C) \right]^2, \quad (15)$$

where $\sigma_Y^2$ 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 2 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 3 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 3 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 preregulation HUD National Survey. In addition, the intercept and error attributable to random variability in this table are calibrated so that the geo-

<p>| Table 2 Predicted national distribution characteristics for NHANES III, the Rochester Study model, and the combined datasets model |
|-----------------|-----------------|-----------------|-----------------|
|                 | Preregulation blood lead levels | Postregulation blood lead levels $^a$ | Combined datasets model |</p>
<table>
<thead>
<tr>
<th>Predicted model results</th>
<th>Parameter</th>
<th>NHANES III</th>
<th>Rochester study</th>
<th>Combined datasets model $^b$</th>
<th>Combined datasets model</th>
</tr>
</thead>
<tbody>
<tr>
<td>National geometric mean</td>
<td>$\mu_P$</td>
<td>3.14</td>
<td>6.36</td>
<td>3.14</td>
<td>3.03</td>
</tr>
<tr>
<td>National geometric standard deviation</td>
<td>$\sigma_P$</td>
<td>2.09</td>
<td>1.85</td>
<td>2.09</td>
<td>2.05</td>
</tr>
<tr>
<td>Estimated distribution</td>
<td>% $\geq 10 \mu g/dl$</td>
<td>5.88%</td>
<td>22.90%</td>
<td>5.88%</td>
<td>4.91%</td>
</tr>
<tr>
<td>exceedance percentiles</td>
<td>% $\geq 20 \mu g/dl$</td>
<td>0.43%</td>
<td>2.90%</td>
<td>0.64%</td>
<td>0.45%</td>
</tr>
<tr>
<td>(% of population $\geq$</td>
<td>% $\geq 30 \mu g/dl$</td>
<td>0.07%</td>
<td>1.00%</td>
<td>0.12%</td>
<td>0.07%</td>
</tr>
</tbody>
</table>

$^a$ The proposed health-based standards used to calculate this example of a postregulation 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 $ft^2$ damaged lead-based paint (LBP) for paint repair, and 20 $ft^2$ damaged LBP for paint abatement. For homes that exceeded a standard, the estimated effect of the regulation was to reduce the environmental level(s) that 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 5 $ft^2$ damaged LBP.

$^b$ The estimates from the combined datasets model are those that would be obtained when adjusting the model to produce a preregulation 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.
metric 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 preregulation data from the HUD National Survey.

Table 3 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 postregulation modified environmental lead levels from the HUD National Survey. To construct the distribution of postregulation environmental lead levels, observed levels of lead in environmental variables in the HUD National Survey were compared with 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, postregulation 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. 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 the 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 percentages of children exceeding these thresholds were called exceedance percentiles. Table 2 displays the geometric mean, geometric standard deviation, and three exceedance percentiles of interest for the preregulation 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 postregulation distribution estimated using the combined datasets model.

<table>
<thead>
<tr>
<th>Variable description</th>
<th>Rochester Study model</th>
<th>Combined datasets model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>α_{Y</td>
<td>R_1,R_2,R_3,C}</td>
</tr>
<tr>
<td>Area-weighted arithmetic mean floor-dust lead loading</td>
<td>β_{Y</td>
<td>R_1(C)}</td>
</tr>
<tr>
<td>Area-weighted arithmetic mean window sill-dust lead loading</td>
<td>β_{Y</td>
<td>R_2(C)}</td>
</tr>
<tr>
<td>Soil lead concentration</td>
<td>β_{Y</td>
<td>R_3(C)}</td>
</tr>
<tr>
<td>Indicator of interior paint/pica hazard</td>
<td>β_{Y</td>
<td>C(R_1,R_2,R_3)}</td>
</tr>
<tr>
<td>Error attributable to random variability</td>
<td>σ_{Y</td>
<td>R_1,R_2,R_3,C}</td>
</tr>
<tr>
<td></td>
<td>α_{Y</td>
<td>H_1,H_2,H_3,C}</td>
</tr>
<tr>
<td></td>
<td>β_{Y</td>
<td>H_1(C)}</td>
</tr>
<tr>
<td></td>
<td>β_{Y</td>
<td>H_2(C)}</td>
</tr>
<tr>
<td></td>
<td>β_{Y</td>
<td>H_3(C)}</td>
</tr>
<tr>
<td></td>
<td>β_{Y</td>
<td>C(H_1,H_2,H_3)}</td>
</tr>
<tr>
<td></td>
<td>σ_{Y</td>
<td>H_1,H_2,H_3,C}</td>
</tr>
</tbody>
</table>

Note that Table 2 presents results for only one postulated postregulation scenario. The combined datasets model was...
used by the 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 scenarios.

6. Discussion

Although the purpose of this article is to demonstrate a methodology for combining data sets, ultimately the model is intended to predict a national distribution of children's blood lead. Therefore, certain caveats associated with estimating model parameters should be recognized. For example, since the Rochester Study sampled only households with low income and built before 1940, $\sum X|C$ estimated using Rochester data may be too small due to homogeneity of lead media within the particular subpopulation of sampled housing or too large due to environmental characteristics unique to this subpopulation. Further, HUD sampling weights in (15) reflect all houses built before 1980, and therefore $\sigma^2_{Y|H,C}$ may be underestimated. Finally, estimating the standard errors of model parameter estimates was difficult because multiple data sets were combined and NHANES III data are based on a complex sampling scheme. Since the model was essentially based on Rochester data but calibrated according to HUD and NHANES III data, standard errors were estimated by bootstrapping the Rochester data only. This may result in underestimation due to bootstrapping only the Rochester data, effectively treating HUD and NHANES III parameter estimates as known.

In building parameter estimates for model (1), we used a latent variable model (2a)–(2d). The parameter estimation technique was the method of moments, which is potentially an inefficient method for estimating model parameters. To infer (1) from (2a)–(2d) requires assumptions of normality for $Y \mid R, C$ and $Y \mid H, C$ and transportability of the regression relationship from Rochester to the nation as a whole. 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 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).

Résumé

Un modèle liant la concentration en plomb dans le sang des enfants et l'exposition au plomb présent dans l'environnement a été développé en combinant deux sources de données représentatives caractérisant les distributions marginales du plomb dans le sang et dans l'environnement à l'échelle nationale et une troisième source de données régionale concernant des mesures conjointes du taux de plomb dans le sang et dans l'environnement. Le facteur de complication pris en compte dans cet article concerne le fait que les méthodes d'évaluation du plomb dans l'environnement sont différentes entre les sources de données nationale et régionale. Basée sur une hypothèse de transportabilité (on suppose que bien que les distributions marginales du taux de plomb dans le sang et dans l'environnement puissent être différentes entre les données régionale et nationale, la relation conjointe entre le taux de plomb dans le sang et dans l'environnement est la même aux 2 échelles), le modèle utilise une approche par variables latentes pour estimer la distribution conjointe des taux de plomb dans le sang et l'environnement à l'échelle nationale.

References


U.S. Environmental Protection Agency. (1998a). Risk analy-
sis to support standards for lead in paint, dust and soil.


Received May 1998. Revised August 2000.
Accepted September 2000.