EPA 540/R-94-039
PB94-963504
OSWER #9285.7-21
December 1994

VALIDATION STRATEGY FOR THE INTEGRATED EXPOSURE UPTAKE BIOKINETIC MODEL FOR LEAD IN CHILDREN

Office of Solid Waste and Emergency Response
U.S. Environmental Protection Agency
Washington, DC 20460

NOTICE

This document provides guidance to EPA staff. It also provides guidance to the public and to the regulated community on how EPA intends to exercise its discretion in implementing the National Contingency Plan. The guidance is designed to implement national policy on these issues. The document does not, however, substitute for EPA's statutes or regulations, nor is it a regulation itself. Thus, it cannot impose legally-binding requirements on EPA, States, or the regulated community, and may not apply to a particular situation based upon the circumstances. EPA may change this guidance in the future, as appropriate.

U.S. ENVIRONMENTAL PROTECTION AGENCY
TECHNICAL REVIEW WORKGROUP FOR LEAD

The Technical Review Workgroup for Lead (TRW) is an interoffice workgroup convened by the U.S. EPA Office of Solid Waste and Emergency Response/Office of Emergency and Remedial Response (OSWER/OERR).

CHAIRPERSON

Region 8
Susan Griffin
Denver, CO

MEMBERS

Region 2
Mark Maddaloni
New York, NY

Region 3
Roy Smith
Philadelphia, PA

Region 5
Patricia VanLeeuwen
Chicago, IL

Region 8
Chris Weis
Denver, CO

NCEA/Washington
Karen Hogan

NCEA/Washington
Paul White

NCEA/Cincinnati
Harlal Choudhury

NCEA/Research Triangle Park
Robert Elias

NCEA/Research Triangle Park
Allan Marcus

ORD/Washington
Barbara Davis

VALIDATION STRATEGY FOR THE INTEGRATED EXPOSURE UPTAKE BIOKINETIC MODEL FOR LEAD IN CHILDREN

Prepared by

THE TECHNICAL REVIEW WORKGROUP FOR LEAD
Office of Emergency and Remedial Response
U.S. Environmental Protection Agency
Washington, D.C. 20460

PREFACE

This document describes the considerations and methods for characterizing the confidence to place in output from the Integrated Exposure Uptake Biokinetic (IEUBK) Model for Lead in Children, version 0.99d. The IEUBK Model has been recommended as a risk assessment tool to support the implementation of the July 14, 1994 Office of Solid Waste and Emergency Response Interim Directive on Revised Soil Lead Guidance for CERCLA Sites and RCRA Facilities.

The development of the model has included the cooperative efforts of several EPA programs over nearly a decade. For the last four years, the development and documentation of the model have been coordinated by the Technical Review Workgroup for Lead, whose members are listed on page iv. The approach recommended in the Validation Strategy has been developed by the Technical Review Workgroup for Lead after considerable debate about the criteria that are useful for determining the validity of a model according to its various applications. It reflects the comments of peer reviewers from within and outside of EPA whose names and affiliations are listed on page v.

This document specifies many aspects in model validation. Several of these aspects of model validation have already been conducted, including comparison with other models, documentation of the Model's scientific basis, code verification, and preliminary empirical comparisons. Additional aspects of model validation are being carried out concurrently with the effort detailed in this document. This document is primarily concerned with empirical comparisons of model predictions with field study data, and should be viewed as a working strategy that can be refined and expanded as new approaches are developed and additional data become available. Comments on the technical content of this document or suggestions for its improvement may be brought to the attention of the Technical Review Workgroup for Lead.

TABLE OF CONTENTS

PEER REVIEWERS

Dr. Steven Bankes
RAND Corporation
Santa Monica, CA

Dr. Dennis Cox
Rice University
Houston, TX

Dr. Buck Grissom
Dr. Steven Haness
Agency for Toxic Substances and Disease Registry
Atlanta, GA

Dr. James Hodges
University of Minnesota
Minneapolis, MN

Dr. Thomas Matte
Centers for Disease Control
Piscataway, NJ

Validation Strategy for the IEUBK Model

1.0 Introduction and Overview

There is a need in decision-making to make a distinction between those levels of environmental lead that could cause children to have blood lead levels that are considered elevated by current standards and those levels at which children are less likely to have elevated blood lead levels. Specifically, one primary goal is to identify those environmental situations in which children have a significant chance of having a blood lead level above a level of concern, currently defined by the Centers for Disease Control to be 10 µg/dL (CDC, 1991). For example, the current EPA/Office of Solid Waste and Emergency Response Soil Lead Guidance aims to limit an individual child's risk of exceeding 10 µg/dL to no more than 5% (USEPA, 1994a).

The Integrated Exposure Uptake Biokinetic (IEUBK) Model for Lead in Children (USEPA, 1994a) enables prediction of children's blood lead levels using information on their multimedia exposure to environmental lead. The Model was developed as an alternative to multiple regression models, which in practice have been difficult to generalize to situations beyond those where the data were specifically collected. This model has been developed using data from many different scientific studies of lead biokinetics, contact rates of children with contaminated media, and data on the presence and behavior of environmental lead. Given estimates of the lead concentrations in environmental media to which children are exposed, the Model provides a central tendency estimate and a probability distribution to characterize predicted blood lead levels.

As should be the case with all model applications, it is important that users of the IEUBK Model understand its fundamental strengths and limitations, the process by which the numerical accuracy of model predictions has been verified, and the extent to which model predictions are supported by comparisons with real-world data. The process of model validation addresses this range of considerations, all of which bear on the level of confidence that users can have in model predictions. In broad terms, an evaluation of model validity therefore includes the following considerations:

(1) The scientific foundations of the model structure. Does the model adequately represent the biological and physical mechanisms of the modeled system? Are these mechanisms understood sufficiently to support modeling?

(2) Adequacy of parameter estimates. How extensive and robust are the data used to estimate model parameters? Does the parameter estimation process require additional assumptions and approximations?

(3) Verification. Are the mathematical relationships posited by the model correctly translated into computer code? Are model inputs free from numerical errors?

(4) Empirical comparisons. What are the opportunities for comparison between model predictions and data, particularly under conditions under which the model will be applied in assessments? Are model predictions in reasonable agreement with relevant experimental and observational data?

The biological basis of the model and the parameter estimates are described in two other documents, the Guidance Manual for the Integrated Exposure Uptake Biokinetic Model for Lead in Children (1994c) and Technical Support Document: Parameters and Equations Used in the Integrated Exposure Uptake Biokinetic Model for Lead in Children (USEPA, 1994c). Both of these documents are available from the National Technical Information Service. With regard to the third criterion, the coding of model equations has been verified by a separate recoding of the model in another programming language. In addition, independent code verification and validation is currently being conducted and will be described in a forthcoming Technical Memorandum.

This strategy document for the IEUBK Model addresses primarily the last point, comparing Model predictions of blood lead levels with those observed in epidemiologic studies in which environmental lead levels were also characterized. The objective is to describe the Model's performance for specified ranges and combinations of environmental exposures. These results will help IEUBK Model users better understand the strengths and limitations of the IEUBK Model for various applications and identify areas for additional research or model improvement. It should be noted that of these four considerations, the empirical comparisons can be the most telling. If, after extensive empirical testing, statistical evaluation demonstrates that the model predicts the empirical measurements to a degree that is acceptable to the decision maker, that in itself would provide strong support for model applications within the tested range.

There is a substantial body of information on the environmental lead exposures of young children, and a substantial body of work in which both blood lead levels and environmental lead contamination levels have been measured for the same children. This type of information is relevant to the particular applications of the Model in health risk assessment. A full examination of these data should play an important role in evaluating performance of the IEUBK Model. It should also be noted that few other models being applied in environmental health assessment have a similarly extensive and relevant empirical data base for comparison.

The Validation Strategy outlines the data requirements for this type of exercise, and describes the appropriate use of available data in generating and interpreting blood lead

predictions. The statistical methods to be used are primarily descriptive, but formal statistical hypothesis tests can be useful in evaluating the correspondence between distributions of observed and predicted blood lead levels. The document should be viewed as a working strategy that can be refined and expanded as new approaches are developed and additional data become available.

2.0 General Principles of Model Validation and Verification

Before going into the steps to be taken in evaluating the IEUBK Model's performance, it is worthwhile to examine the nature of model validation and verification, apart from the constraints imposed by a particular model structure. The meaning of model validation is evolving. In EPA's recent guidance for conducting peer review of environmental models, several purposes for model development were identified (USEPA, 1994b). These include the use of models as research tools simplification and/or refinement of existing model paradigms or software, their use as screening tools, and the use of models to estimate compliance with regulatory requirements. The criteria for model "validation" must be relevant for the intended application of the model.

First, the process of examining a model's performance by making empirical comparisons with observed data should more correctly be termed confirmation of the model for the conditions defined by the empirical data (Oreskes et al., 1994; see the Appendix for a discussion of some of the issues raised in this paper). Models represent conceptual simplifications of complex biological and physical systems. Model validation should be considered an iterative process that aims to test applications of a model under a variety of conditions (combinations of input variables) that ideally span those conditions where practical applications of the model will be made.

The scope of the conditions under which empirical comparisons are made can be termed the "domain." This is analogous to the mathematical use of the term domain for a function: the set of values of input variables over which the function is applied. By this analogy, each data set with well-characterized inputs represents a separate domain for the model. Multiple comparisons can address the model's agreement with the data. By the same logic, the scope of different model output predictions that is tested against experimental data can be called the "range." Of particular interest are those model features to be utilized in practical applications.

It is convenient to think of the assessment of a model as making comparisons between "certain" experimental data and the uncertain predictions of that model: If the two fail to agree the model is taken to be at fault. This conceptualization, however, usually fails to fit with reality. Experimental data may be in "error" either through statistical fluctuations (which can often be accommodated through the use of appropriate statistical testing) or more problematically, through

systematic biases. In practice, it is not unusual to come across different sets of data that effectively conflict with each other, precluding the possibility that any model can agree with all of the data. With further scientific evaluation, it may or may not prove possible to understand the reasons for the differences between data sets.

It should be noted that while the collection of data relevant to evaluation of the IEUBK Model's performance is quite extensive, it is observational, not experimental in nature. When observational data (such as those obtained in epidemiologic studies) are used in model confirmation, inconsistencies between the modeled and observed systems can be particularly problematic. An observational study does not allow the investigator to fix the values of some variables or isolate processes to be measured quantitatively as can be done in an experiment. In fact, values for some variables may not be ascertainable under field conditions. On the other hand, even if it were acceptable to study children experimentally, that is, by rigorously measuring lead exposure and blood lead levels over full ranges of the lead concentrations and exposures of interest, such data would be unlikely to reflect real-world exposure conditions, due to the practical difficulties in measuring actual exposure.

Another matter concerns the precision of the agreement between model predictions and observed data that should be anticipated. The needed precision in a model also varies with its applications. Therefore, it is most useful if a reasonable standard can be established for the level of agreement that is expected and needed in model comparisons. Such a standard can often be usefully expressed in terms of the absolute or fractional difference between model predictions and observed data that would cause concern for expected model applications.

While statistical tests are useful tools in many model comparison efforts, the standard formulation of tests in terms of a null hypothesis that is to be accepted unless rejected with a certain level of confidence is not fully compatible with the approach to model validation discussed here. When data are sparse, the observation that agreement between the model and the data cannot be statistically rejected should not in itself be taken to represent a successful “validation” of the model. The statistical power to observe a meaningful difference between observed and predicted values may be lacking. On the other hand, when large data sets are available, a slight difference between observations and predictions may be found to be statistically significant, but to have little practical application. Large data sets, however, may make possible distinctions in model performance among subsets of the data, as determined by restricted ranges of the input variables or by other characteristics, such as sex, or type of lead contamination, if the lead contamination is known to differ by neighborhood or some other recorded variable.

3.0 Description of the IEUBK Model

The Integrated Exposure Uptake Biokinetic (IEUBK) Model was designed to provide predictions of the probability of elevated blood lead levels for children. The Model addresses three components of environmental risk assessments 1) the multimedia nature of exposures to lead, 2) lead pharmacokinetics, and 3) significant variability in exposure and risk, through estimation of probability distributions of blood lead levels for children exposed to similar environmental concentrations.

The Model first estimates a longitudinal exposure pattern, from birth to the age of interest, by using exposure concentrations measured reliably in appropriate settings. The Model accepts exposure data on an annual basis, but allows for entry to characterize a cumulative average exposure for each environmental medium. The Model then predicts a plausible distribution of blood lead for a hypothetical child, or population of children exposed to this inferred exposure pattern. The blood lead distribution incorporates variability associated with repeat sampling, and inter-individual and biological variability, as determined from community blood lead studies of children's residential settings.[1]

Version 0.99d of the IEUBK Model is an expanded version of models used by the USEPA Air and Water programs in support of regulations. The expansions were largely the result of consultation with outside experts and comments made by users of earlier versions of the Model. Two of these versions were reviewed by Science Advisory Boards, and judged to be scientifically sound.

It is beyond the scope of this document to describe the scientific basis of the Model in any great detail. Very briefly, the Model has been based on human data, in order to estimate parameters directly, where possible. The absorption algorithm, was based on data from lead balance and feeding studies in human infants and children. Data from baboon studies were not

---

[1]Consistent with measurements of other metals in tissues of human populations, the distribution of blood lead levels for any relatively homogeneous population closely follows a lognormal distribution (USEPA, 1986). A lognormal distribution is completely specified by its geometric mean and geometric standard deviation.

The standard deviation was estimated for subgroups with relatively homogeneous environmental lead concentrations, minimizing the contribution from the exposure distributions, resulting in a geometric standard deviation of about 1.6 (see the Guidance Manual (USEPA, 1994c) for more details).

The variability in estimates of blood lead levels arises from two factors: variability in environmental concentrations, and the geometric standard deviation reflecting empirical (observed) variability in blood lead levels in children exposed to similar lead concentrations. This could be contrasted with a Monte Carlo simulation in which it would be necessary to estimate population distributions for all input parameters in order to estimate variability in blood leads.

used as the primary basis for any parameter in the IEUBK Model. The baboon data were used to help define a range of plausible values for human children. The exact value of a parameter within that range was selected using comparisons with human data from cross-sectional studies. The compartmental structure was based on earlier models for lead in adults. Some information was derived from field work (e.g., soil ingestion rates in children) and some from surveys (time children spend outside). See the Guidance Manual (USEPA, 1994c), Section 4.6, for a more detailed description of the uptake pathways, and biokinetic compartments and associated lead transfers, and the Technical Support Document (USEPA, 1995) for documentation of the equations and parameter values used and the sources of data considered.

The Model does not aim to reproduce the observed blood lead level for any specific child, because of the practical limitations of exposure characterizations, which are discussed further in Section 4.4. For instance, it is difficult to quantify the influence of mouthing behavior on exposure, especially with existing, data. Most importantly, the Model is not a substitute for medical evaluation of an individual child.

4.0 Analysis of IEUBK Model Performance

This section outlines the steps in using available epidemiologic studies in assessing IEUBK Model performance. First, the objectives to be evaluated are discussed. Second, the attributes of suitable data sets to evaluate the objectives are described. Then, the mechanics of generating Model predictions from suitable data sets are reviewed (see the Guidance Manual for more detail). Last, appropriate analysis methods and interpretation of results are discussed.

4.1 Comparisons to be Made

The general objective to be evaluated in these comparisons is whether the IEUBK Model predictions are consistent with observed blood lead distributions, given adequately characterized exposure profiles. The IEUBK Model produces many testable criteria, such as a geometric mean blood lead and probability distribution corresponding to a multimedia lead exposure combination, and environmental lead concentrations corresponding to a specified uptake pattern and blood lead distribution. The IEUBK Model may be considered more “valid” by some criteria than by others. As explained in the previous section, for example, prediction of a specific child's measured blood lead level is not one of the Model's intended or valid uses. Comparisons will parallel the intended uses of the Model to:

1) Provide a best estimate of the geometric mean blood lead concentration for a hypothetical child aged 0 to 84 months, if he were assumed to reside at a given residence;

2) Provide assistance in estimating blood lead concentrations at undeveloped residential sites that may be developed in the future;

3) Provide a basis for estimating risk of elevated blood lead (i.e., >10 µg/dL) for a hypothetical child of specified age with given site-specific residential lead exposure;

4) Provide a basis for estimating the risk of elevated blood lead concentrations in a given neighborhood by aggregating the individual residential risk estimates;

5) Predict likely changes in risk of elevated blood lead concentrations from exposure to soil, dust, water, or air lead following actions to reduce exposure levels from one or more of these sources;

6) Provide assistance in determining appropriate soil or dust lead target cleanup levels at specific residential sites.

These applications are also described in the Guidance Manual. One other application of the Model, to provide a summary of children's long-term exposure to lead, will not be explicitly pursued in this effort, since none of the available studies were designed to measure children's exposure.

4.2 Data Set Selection

Evaluation of these intended applications, as well as correct application of the Model, clearly depends upon good quality data that describe the actual exposures (“garbage in, garbage out”).

Two general types of epidemiologic studies have been conducted in examining blood lead levels in relation to environmental lead levels, longitudinal and cross-sectional. Longitudinal studies measure the same individuals and their environments periodically over time. These studies should be especially useful, all other factors being equal, since the Model predicts blood lead levels resulting from cumulative exposure to environmental lead. Most commonly, however, we anticipate seeing cross-sectional data sets, those in which environmental and blood lead data were collected at only one time point for each participant. Only data sets which were not used in

calibration will be considered. Data sets will be examined for their demonstration of:

— representativeness of the communities studied, through assessment of statistical study design, including but not limited to such factors as appropriate sample size determination and sampling procedures,
— accepted analytical methods for measuring environmental and blood lead levels; and
— documentation of quality assurance and quality control procedures followed, including reproducibility of environmental and blood lead measurements.

At a minimum, any studies being considered for the comparison exercise should have data of sufficient quality and quantity to adequately characterize the residential home and yard as the exposure unit. This means that for each child in the study, he or she would have had a blood lead taken, and concurrent soil, house dust, and tap water samples collected and analyzed for lead. In addition, information from a demographic/behavioral survey, and interior/exterior paint analysis could be helpful in understanding predictions.

Several additional factors discussed below should influence data set selection; however, the values of the measured environmental and blood lead levels will not be considered at this stage, to minimize/avoid biasing data set selection. First, each study group will be reviewed to assess the extent of missing data, including the extent and any patterns of missing blood and environmental lead data. Data sets with many records missing at least one lead exposure variable will raise concerns about the representativeness of the data for carrying out the empirical comparisons. Data sets missing appreciable amounts of demographic and behavioral information are less desirable than others, and have a lower priority in the overall strategy.

It will also be necessary to review sampling and analytical protocols in depth in order to note differences in sample collection (for example, wipe vs. vacuum), sample preparation (for example, sieve sizes used for separating soil samples), and analytical methods (for example, atomic absorption or XRF). At this stage, studies must have measurements that the Model can use in its current form, such as, dust lead concentration rather than dust lead loading. These factors may also determine the relative priority of data set consideration in the overall strategy.

Data sets identified to be considered will include a range of environmental conditions, including lead from urban, industrial, mining and smelting sources, such as:

— Three City pre-abatement data from Baltimore, Boston and Cincinnati,
— Boston Women's Hospital (longitudinal)
— Kellogg, ID
— Butte-Silver Bow, MT

— Multi-State study (ATSDR): Palmerton, PA; Granite City, IL; Joplin, MO; Galena, KS
— Cincinnati Longitudinal study
— Bartlesville, OK
— West Dallas, TX
— Leadville, CO
— Rochester, NY
— California (Oakland, Sacramento, Los Angeles)

Similarly extensive data sets, including those from other countries will also be considered as they are made available.

4.3 Use of Data Set to Generate Model Predictions

For the comparison effort, use of the Model involves using environmental lead concentrations for areas in which children are expected to be exposed to lead, corresponding to an individual child. In addition, community-specific information, such as bioavailability of lead compounds found in the community, is especially important.

The default parameters were not intended to be appropriate in every case and should not be expected to correctly predict blood lead concentrations in all situations. Substitution for default Model parameters, however, should be based on other scientific studies of lead-contaminated materials from the site, or on demographic and behavioral studies of potentially exposed children. In addition, this information should be independent of the blood lead data being used for the comparison. Specifically, any regression or structural equation modelling used to estimate these additional parameters, such as bioavailability and absorption, to the observed blood lead levels is inappropriate for generating additional information which would then be used as inputs for Model predictions based on that site.

In some cases, there may have been an earlier study available that identified a major childhood lead exposure problem, and a later follow-up study. The nature of any interventions which may have occurred will be evaluated. If there were no confounding issues identified, one Model testing strategy could be to use the earlier study to estimate the site-specific and general population behavioral parameters that affect lead exposure.[2] The IEUBK Model could then be

[2]The information collected for each study varies, but could optimally include such factors as bioavailibility/absorption of lead by compound in each medium, rate of dirt ingestion, relative exposure to soil and dust, and community-specific diet data (local produce, game and fish consumption, for example). If it appears that these esitimates would not have changed much over the period between the studies, then the earlier estimates could replace the defaults in the later study.

used to predict the blood lead concentrations in the later study using site-specific parameters, if they were measured in the earlier study, together with environmental concentrations and child-specific behavioral or demographic data from the later study. An alternative would be to perform two independent Model comparisons.

An important issue is the inappropriate use of community-wide mean soil and dust concentrations to predict a community geometric mean blood level, as opposed to the recommended approach of using sets of individual child-based environmental measurements to generate a community mean. Since the former approach does not directly take into account the distribution of environmental concentrations, it cannot be expected to describe the blood lead distribution adequately. Such comparisons cannot be expected to provide adequate confirmation or refutation of Model performance. See the Guidance Manual for more discussion of this issue.

Each data set will be reviewed to determine which records have data adequate to characterize exposure. As stated earlier, plausible predictions by the IEUBK Model for a given exposure situation are conditional on the assumption that the significant sources of lead exposure and lead intakes for any child in that situation have been adequately characterized. Ideally, this would involve determining where a child would spend time; collecting soil, house dust, tap water and paint samples from each of those locations; and time-weighting exposure within that exposure unit." Since this is generally impractical, a child's home and yard have been considered in most field studies to be an acceptable surrogate for that exposure unit.

For the purposes of this comparison exercise, there are at least two steps to be taken to distinguish marginally sufficient exposure characterizations. First, records (children) without soil or dust measurements cannot be expected to contain enough information for predicting blood lead levels. These records will be excluded from the analysis. An assessment of the impact on the representativeness of the resulting subsample will be included in each write-up. Second, some data sets include the amount of time on average that a child spends away from home per week, the length of residency at the current address, and the time elapsed since the last renovation in the home. Where this information is available, this will be retained in order to examine any tendency for Model predictions to be systematically less accurate for children with exposures which may have been possibly more variable than other children's. No cut-off will be chosen in advance for the purpose of excluding records; the emphasis will be on describing trends.

Once the set of participants for a particular study with adequate data is identified, the values of the environmental measurements must be available in order to calculate the inputs to the IEUBK Model. At each time point, multiple measurements for an environmental medium could be combined into one composite measurement, usually an arithmetic average if no other weighting scheme is justifiable to reflect a child's activity patterns. In the case of soil, for example, there may be measurements made at several depths; it may be reasonable to exclude some

measurements, and average the others. In general, such decisions will be made on a study by study basis, and where appropriate on a subject by subject basis, depending upon sampling protocols and other available information. All such decision processes will be documented.

The Model will be run for each suitable data set, incorporating any site-specific information that may be available, such as bioavailability/absorption of lead compounds by medium, rate of dirt ingestion, and relative exposure to soil and dust. If no community-specific diet data are available, national averages appropniate.for the time period will be used. There will be no "calibration" or adjusting of fixed (user-inaccessible) Model parameters to agree more closely with measured blood lead levels. In addition, measured blood lead levels will not be examined 11 after the Model predictions have been generated. All the fixed parameters, and Model defaults for user-supplied parameters, have been documented in the Technical Support Document (USEPA, 1994c). All user-supplied parameters for each Model run will be reported.

4.4 Analysis Methods

As described in Section 4.1, comparisons will follow the intended applications of the Model. These can be addressed by considering the following questions:

— Applications 1-2: How do Model predictions of central values (usually geometric means) for blood lead compare with those observed for communities?

— Applications 3-6: More generally, how do Model predictions of population percentiles (such as 95% below 10 µg/dL) compare with observed percentages?

4.4.1 Descriptive Measures

The first question above will be addressed by comparing the geometric mean blood lead level of a set of Model-predicted (geometric mean) blood lead levels for a group of children with the observed geometric mean blood lead level. This will give an overall evaluation of bias in the Model's predictions. In addition, the percent difference between observed and predicted central values will be examined for each data set. Confidence intervals for these statistical measures will be evaluated.

The next question involves characterization of blood lead distributions. Since the Model does not incorporate any exposure variability in its estimation of blood lead levels, each prediction applies only to its set of concentration inputs, one for each child in a data set. The proportion of a population having blood lead levels $\geq$10 µg/dL is estimated using the Model by aggregating the

probabilities of exceeding 10 µg/dL for each child, and averaging them over the entire group. This probability will be compared with the proportion actually observed to have elevated blood lead levels.

Whether or not there is good overall agreement for these criteria just outlined, it will be instructive to examine the correspondence between Model predictions from appropriate studies with observed blood lead levels graphically, illustrating the quantifiable variability in both the observed and predicted blood lead levels. Figure 1 demonstrates the results of Model runs in one community, with Model prediction intervals provided.

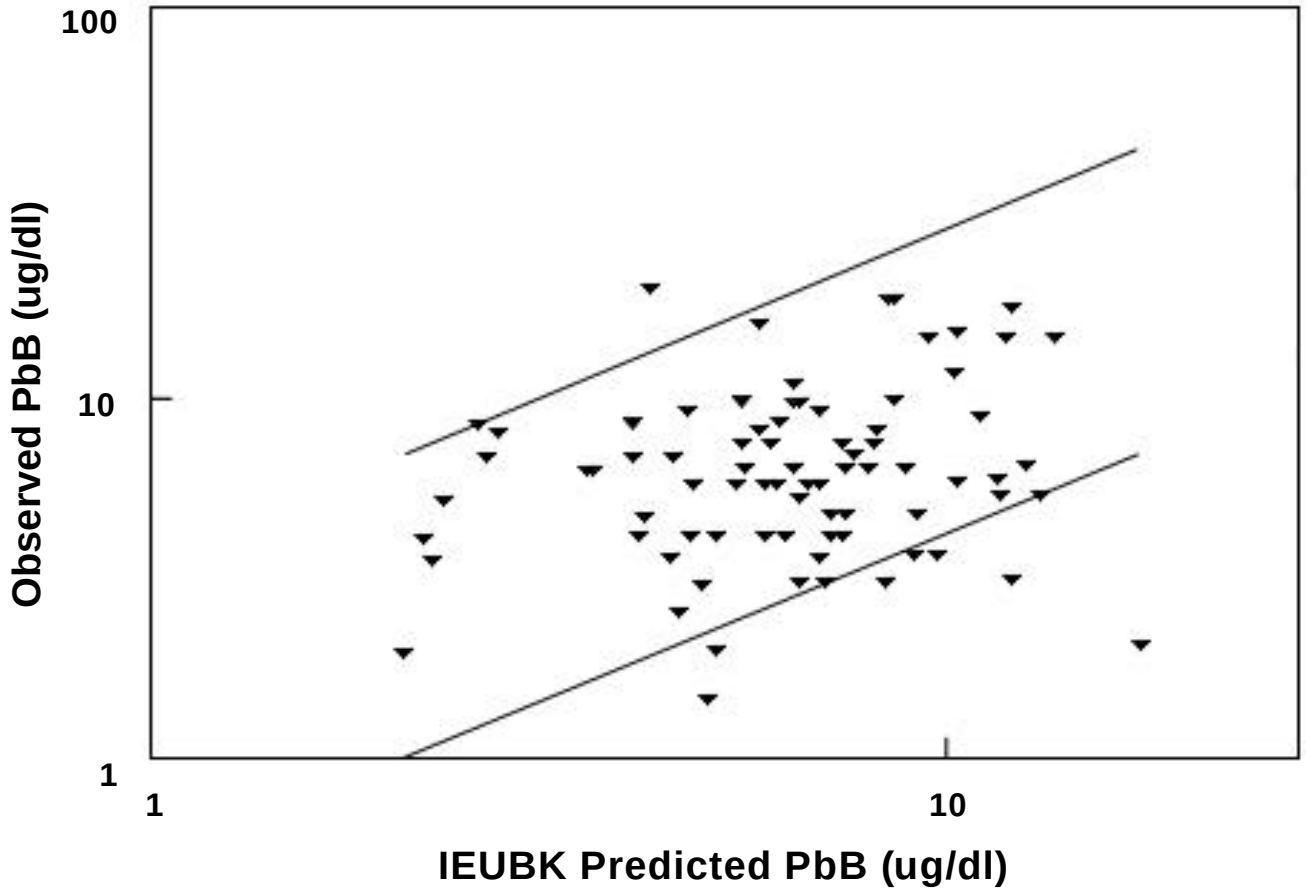
For those data sets with relatively extensive information and adequate sample sizes, subsets will be chosen according to variables which may influence exposure scenario characterizations.

These include, but are not limited to:

— behaviors such as
  - mouthing frequency (times/week),
  - working with lead outside the home or as part of a hobby (yes/no),
  - attendance of the child at a day care center (yes/no-, hours/week);
  - amount of time (per week) spent outdoors, or at various activities;
— neighborhood, age, sex, ethnicity, etc.; and
— presence of lead-based paint, through
  - XRF > 1 Mg/CM^2 and presence of deteriorated paint,
  - renovation work shortly before blood lead measurements, or
  - age of house.

These factors will define subsets that will also be compared graphically, as in Figure 1.

Figure 1. Comparison of blood lead measurements: Observed and IEUBK Predicted (with 95% upper and lower confidence intervals)



4.4.2 Statistical Hypothesis Testing

This section describes considerations in applying statistical tests to Model comparisons. In general, statistical hypothesis testing defines a pass/fail context, implicitly requiring studies designed for the purpose of testing the performance of the model within prespecified limits. Unfortunately, the data available for these comparisons were not collected for the purpose of mechanistic/deterministic model evaluation. As discussed more generally in Section 2.0, the lack of control over the sample sizes may lead to deciding incorrectly that the Model predictions are too different (when the data set is too large) or incorrectly deciding that the Model predictions are not so different, because the particular data set is too small to have sufficient power. In general, statistical similarity or dissimilarity (pass/fail) does not decide Model concordance with observed data, without weighing the importance of such contributing factors as the expected variability in the observed blood and environmental lead measurements, as well as in the Model predictions. In addition, the sample size of each study must be evaluated for use in statistical hypothesis testing. Consequently, it still may be that some of the available data sets will lend themselves to retrospective hypothesis testing, so it will be useful to discuss these factors.

Measurement error refers to the difference between a measurement and the actual quantity of interest. It can include technician errors, for instance, but is generally used in broader sense to incorporate all factors which affect precision and accuracy of measurements. Measurement error can affect central tendency estimates, resulting in systematic bias, and in general leads to less precise measurements. What is especially significant about measurement error in this context is that it can be quite large relative to the actual measurements.

In particular, media concentrations measured may not reflect the average concentrations a child contacts. One example is the characterization of a child's indoor lead exposure by dust lead measurements taken right after housecleaning, although the house may be much dustier much of the rest of the time. In addition, blood lead measurements cannot be taken as “gold standards” by which to measure Model performance, because they incorporate some elements of measurement error. Other specific sources of measurement error include:

— Repeat sampling variability - A child's blood lead level should not be expected to be constant within a day, or from day to day. Blood lead concentration is influenced by recent exposure, which is determined by day-to-day variability in activity patterns (e.g., cycles of day care or school activity vs. weekend activity, as affected by weather and seasons, leading to substantial variability in exposures) and timing of exposure relative to eating (leading to variability in lead absorption). Misspecification of environmental lead exposure can result from partial accounting of environmental levels in both time (due to

rapidly varying exposures associated with abatement or frequent moves), or place (due to relatively high or low exposures away from home which have not been accounted for, and heterogeneity of indoor and outdoor environments).

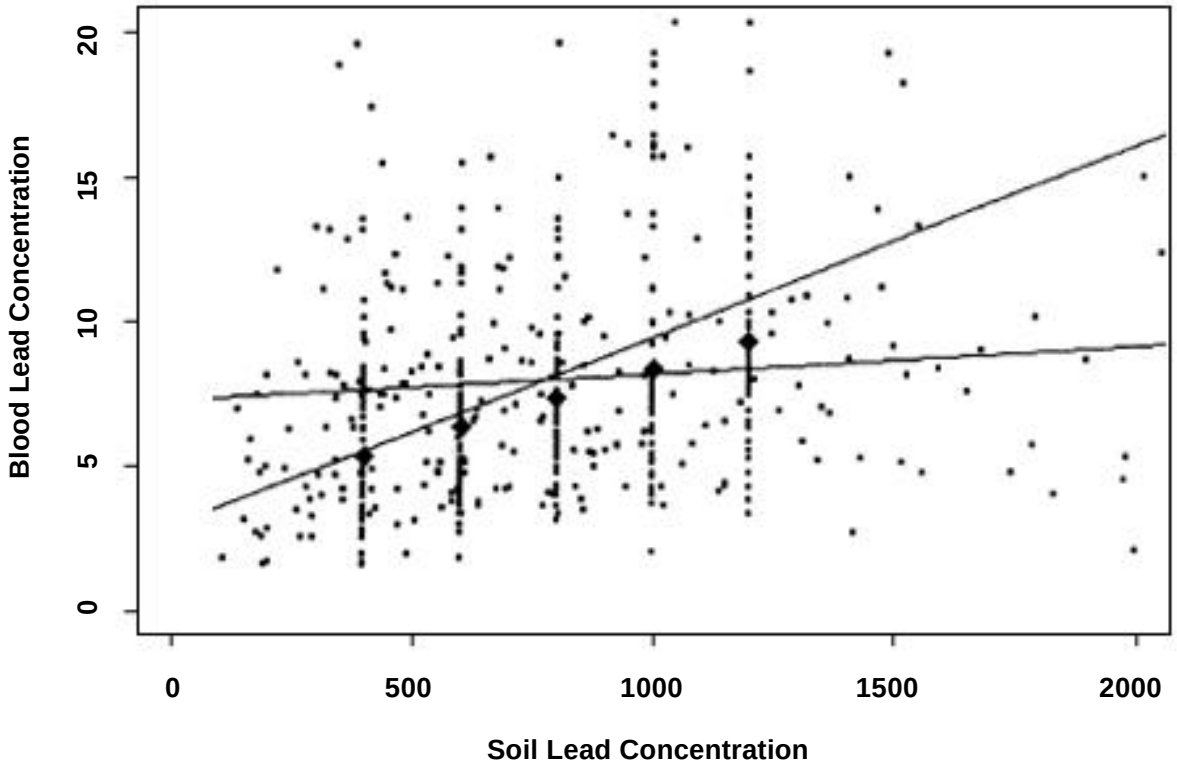
— Analytical variability - Individual blood (and environmental) lead measurements are subject to analytical variability. For example, CDC guidelines suggest that blood lead levels measured by proficient laboratories may vary 4-6 µg/dL from reference values (CDC, 1991). Other information indicates an expected accuracy of ±15%, or ±6 µg/dL, whichever is greater (Weaver, 1994).[3] While within-laboratory variability may currently be relatively low in more proficient laboratories, between-laboratory variability may contribute to some sytematic error or bias. Studies using non-CDC protocols, or studies dating from earlier periods, may differ even more from each other.

Blood lead measurements cannot be taken as "gold standards" by which to judge Model predictions, without acknowledging these sources of measurement error.

Statistical tests are biased by measurement error in both blood and environmental lead concentrations, even if the measurement error is not expected to bias central tendency estimates. Specifically, estimates of slope and correlation coeffecients tend to be lower in the presence of substantial measurement error in the independent variables (Fuller, 1987; Carroll, 1994), as demonstrated in Figure 2. This affects the interpretation of regression models developed from the data sets under consideration, and therefore the use of these models as benchmarks for checking IEUBK Model estimates. Statistical correlation and regression tests of observed and predicted blood lead levels are similarly biased, because the predicted blood lead concentrations acquire substantial measurement error from the input environmental levels. Even informal visual comparison of observed and predicted blood lead, as described in the previous section, is distorted by measurement error. In addition, measurement error can lead to misclassification errors with binary variables, such as the percentage of children with blood lead $\geq$10 µg/dL. This is especially important when blood lead data are subdivided by the value of some exposure index such as soil lead. Statistical tests of goodness of fit can be biased if these misclassification errors are not corrected.

[3]While laboratory methods for determining blood lead levels have been steadily improving, this range is relevant for many data sets available fo rcomparison exercises. The comparisons will use the analytical variability relevant for each data set, to the extent that these data are available.

Figure 2. Effect of measurement error in soil lead concentration on soil/blood lead relationship, generated using the IEUBK Model with default settings, evaluated for 24-35 month old children.



The stacked points with diamond central tendency point represent simulated blood lead distributions (with geometric standard deviation 1.6) corresponding to five "true" soil lead concentrations, 400, 600, 800, 1000, and 1200 ppm. The scattered points correspond to the same points as in the "stacks," but have been shifted horizontally with a simulated measurement error distribution (geometric standard deviation 1.9). The line passing nearly through the geometric mean blood lead concentrations is a least squares fit of the non-log transformed blood lead concentrations to the non-log transformed "true" soil lead concentrations. The other (flatter) line corresponds to the fit of the same blood lead concentrations to the soil lead concentrations with measurement error introduced.

Given these considerations, it is essential to allow for them in any statistical analysis of empirical comparisons. Adjustments for measurement errors in predicted values can be estimated. For example, in the regression context this bias amounts to a reduction in the slope and correlation estimates of $\sigma^2_X/(\sigma^2_X + \sigma^2_M)$, where X describes the variance of the "true" environmental measurements, and M describes the measurement error (Fuller, 1987).

The suitability of statistical hypothesis testing will be determined and reported on a study by study basis. This will include consideration of measurement error, as well as appropriate sample size and required agreement. Since this exercise is primarily descriptive, there will be no predetermined level of agreement specified. We are aiming to estimate the level of agreement, under various conditions. Again, this document should be viewed as a working strategy that can be refined and expanded as new approaches are developed and additional data become available. As we gain more experience there will be situations in which it may be valuable to test for similar performance. The following is a discussion of tests which have been considered, and serves as an example of the considerations which will be discussed for each comparison exercise.

The first test under consideration is the paired t-test, which tests the average group difference between two realizations of the same measurement, such as two blood lead measurements for each member of a study population. Let X and Y denote the two measurements, observed and predicted blood leads, respectively. Assumptions involved in this test are that different members of the population are independent, X and Y are not independent and X-Y is normally distributed $(X-Y) \sim N(\mu_{X-Y}, \sigma^2_{X-Y})$. Then the difference X-Y can be tested against the null hypothesis that the mean of X-Y=0, using a t-statistic:

$$\frac{X-Y}{S_{x-y}\sqrt{n}} \sim t_{n-1}$$

If the measurement error in both X and Y is similar and unbiased, the measurement error should cancel out, and not affect the t-test. For the purposes of statistical testing, blood lead data are conveniently described by the lognormal distribution, so the null hypothesis is restated as:

H_O: ln Observed - In Predicted = 0
H_A: ln Observed - In Predicted $\neq$ 0

In order to interpret this in terms of the untransformed measurements, this results in a test of Observed/Predicted = 1.

If it can be assumed that the measured environmental lead levels are a random sample of the environmental conditions, and the blood lead measurements are also a random sample from the community, then the sets of observed and predicted blood lead levels can be considered independent. A regular t-test can then be applied to compare the (geometric) mean observed and predicted blood lead levels. It is known that this assumption is not quite true, yet it is possible that this approach could decrease any bias introduced by measurement error in the environmental lead measurements.

The proportion observed to have elevated blood lead levels can be compared with the proportion predicted to have elevated blood lead levels. This comparison depends upon the assumption made above for using a t-test, that the measured environmental lead levels are a random sample of the environmental conditions, and the blood lead measurements are also a random sample from the community. One test is given by:

$$H_O: p_0 - p_c = 0$$
$$H_A: p_0 - p_c \neq 0$$

where the difference between sample proportions can be assumed to be normally distributed, and the sample estimates can be pooled to estimate a population proportion. This last assumption does not hold because the proportions are different estimates of the same probability of elevated blood lead, while the test was designed to help determine whether two separate samples can be combined into one larger sample. The comparison can be useful, but its significance level will not be accurate. In calculating these proportions, some judgement would be necessary to determine whether a particular sample is large enough that measurement error is unlikely to bias the estimated proportions.

Another approach recommended in model validation literature (for example, Flavelle, 1992) is regression analysis. The goal of regression analysis of observed vs. predicted blood lead levels is ideally to check for a slope of 1 (perfect agreement) and a correlation of 1. The corresponding model is:

$$\ln(\text{Obs}) = a \ \ln(\text{Pred}) + b.$$

Unfortunately, statistical correlation and regression tests are seriously biased by measurement error in both observed and predicted blood lead concentration. As discussed earlier, this bias amounts to a reduction in the slope and correlation estimates of $\sigma^2_X/(\sigma^2_X + \sigma^2_M)$, where X describes the variance of the "true" environmental measurements, and M describes the variance of the measurement error.

4.5 Interpretation of Results

It should be noted that even when a predicted blood lead distribution for a set of exposure inputs seems unlikely to include an observed blood lead level, there may be plausible explanations. More detailed considerations which may impact adequate characterization of exposure conditions and corresponding blood lead levels include, but are not limited to, relative timing of environmental and blood lead sampling, and consideration of relevant occupations, hobbies, house cleanliness, interior/exterior paint condition, children's mouthing behavior, and consumption of imported canned goods, homegrown fruits and vegetables, and game and fish. If data are available, some of these factors can be addressed in IEUBK Model runs (imported canned goods, homegrown food, and game and fish). Otherwise, these impacts should be assessed and noted where appropriate.

Summary statistics describing each data set will be reported to the fullest extent possible within restrictions imposed by confidentiality or other considerations. There will also be a characterization of any differences between the reduced data sets and the original data sets, and a description of the uncertainties inherent in both the Model predictions and the data sets. The suitability of statistical hypothesis testing will be determined and reported on a study by study basis.

We expect that there will be some comparisons in which the predicted and observed blood lead levels agree less well than in others. In some instances, this will be attributable to the data sets' suggesting diverging relationships by themselves. In other instances, the degree or pattern of nonconcordance may suggest further research, such as the need for different characterization of exposure inputs (exterior dust vs. soil, interior dust concentration vs. loading, total amount of dirt ingestion).

The overall approach outlined until now applies primarily to individual data sets. Any conclusions about model adequacy will not rest on the result of applying the Model to any one data set. If the confirmatory exercises are applied to a number of data sets, the resulting conclusion will be derived from a weighting of the strengths and weaknesses of all the exercises, not on simple vote-counting, due to the varying scopes and quality of data collected. There are many possible approaches to combining the result of a sequence of confirmatory exercises. One approach will be to check the consistency of blood lead predictions being higher corresponding to higher environmental concentrations across studies.

As mentioned earlier, there are many community-specific characteristics which may be difficult to quantify—such as the impact of varying lead species on absorption processes, and the effect of different cultural practices on children's activity patterns; and study-specific characteristics, such as seasonal variability and interlaboratory analytic variability. For these

reasons, the overall approach to be taken will be first to treat each data set separately, and then determine whether the Model is relatively more or less predictive for categories across data sets, such as by age groups.

5.0 Next Steps

It is anticipated that this will be a continuing process, with results of additional comparisons issued periodically. The first set of empirical comparisons will be submitted for publication in a peer-reviewed journal in the first half of 1995. In addition, a summary of these results will be presented in a poster at the 1995 Society of Toxicology meeting. As mentioned above, some comparisons may suggest options for further development, such as alternate characterization of exposure inputs (exterior dust vs. soil, interior dust concentration vs. loading, total amount of dirt ingestion).

In addition, it is anticipated that a workshop to discuss validation issues will be scheduled in late 1995.

6.0 References

CDC. October 1991. Centers for Disease Control. Preventing Lead Poisoning in Young Children. Atlanta, GA: CDC, US Department of Health and Human Services.

Carroll RJ, LA Stefanski. 1994. Measurement Error, Instrumental Variables and Corrections for Attenuation with Applications to Meta-Analyses. Statistics in Medicine. 13: 1265-1282.

Flavelle, P. 1992. A quantitative measure of model validation and its potential use for regulatory purposes. Advances in Water Resources. 15: 5-13.

Fuller WA. 1987. Measurement Error Models. New York: John Wiley & Sons.

Oreskes N, K Shrader-Frechette, K Belitz. 1994. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. Science. 263: 641-646.

USEPA, 1994a. Integrated Exposure Uptake Biokinetic Model for Lead in Children. (IEUBK), Version 0.99d. Washington, DC. PB94-501517.

USEPA, 1994b. Agency Guidance for Conducting External Peer Review of Environmental Regulatory Modeling. Agency Task Force on Environmental Regulatory Modeling.

USEPA, 1994c. "Guidance Manual for the Integrated Exposure Uptake Biokinetic Model for Lead in Children." Washington, D.C. EPA/540/R-93/081, PB93-963510.

USEPA, 1995. "Technical Support Document: Parameters and Equations Used in the Integrated Exposure Uptake Biokinetic Model for Lead in Children (v 0.99d)." EPA 540/R-94/040, PB94-963505.

Weaver, VM, 1994. "Occupational Lead Exposure: Health Effects and Medical Monitoring." Presentation at American Industrial Hygiene Association and The Johns Hopkins University Professional Development Conference.

APPENDIX

A recent perspective on the evaluation of scientific models is provided by Oreskes et al. (1994). These authors present an analysis of logical, philosophical, and some practical problems in evaluating models. A central argument of the paper addresses the impossibility of "verifying" models: Verify is taken to mean "demonstrate the truth of." The authors make the point that it is not possible to conclusively demonstrate the truth of any scientific proposition. This corresponds to the statement that a scientific hypothesis can be disproved by experimentation, but not conclusively proved, because there will always be different experiments that could be done to test different applications of a hypothesis. This argument is applied to models: "The more complex the hypothesis the more obvious this conclusion becomes. Numerical models are a form of highly complex scientific hypotheses."

Given the futility of "verification," Oreskes et al. discuss approaches to the "confirmation" of models. The term confirmation is used to describe the situation where observations are found to agree with model predictions, and it is noted that "confirming observations do not demonstrate the veracity of a model or hypothesis, they only support its probability." Furthermore, "the greater the number and diversity of confirming observations, the more probable it is that the conceptualization embodied in the model is not flawed."

The authors also discuss several more specific issues that are encountered in the evaluation of numerical models:

— Models are simplified descriptions of reality, discrepancies from model predictions can be expected if one examines the behavior of a real system in enough detail.

— Model parameters will themselves be established with error and their estimation may involve a number of secondary assumptions that themselves may be open to doubt.

— Measurements of both independent and dependent variables, are not themselves straight forward error free observations, but typically involve inferences and assumptions. These "embedded assumptions may themselves be subject to experimental test, but their truth cannot (for the reasons mentioned above) be unequivocally established. When model comparisons fail, the problem may be with the model, the auxiliary hypothesis, or both. Similarly, an apparently successful confirmation could result if model errors and errors in auxiliary hypotheses canceled each other.

— Models are often calibrated: The model is adjusted so that predictions better correspond to results in studies where both independent and dependent variables were established. If, then, comparison with additional data sets confirm the model predictions the authors suggest the term “empirically adequate.” However, based on their experience with earth sciences models, Oreskes et al. indicate a general presumption that additional “calibration” will be needed when a model is applied to new data sets. If calibration is needed to “force” model agreement with each data set, this would clearly detract from model confirmation. It is noted that an “empirically adequate” model may still fail when extrapolated over much larger time frames or to future conditions that are not similar to those for which the model was confirmed.

The approach to model confirmation presented in the Oreskes paper is consistent with the emphasis in this strategy document on the utility of empirical comparisons of IEUBK model predictions with observational data, This strategy does not anticipate that empirical comparisons can provide conclusive “verification” of the model, but rather, they can contribute to an overall evaluation of the credibility of model predictions.

The attention that Oreskes et al. give to the “auxiliary hypothesis” is also relevant to the task of making empirical comparisons the IEUBK model predictions. Assumptions that arise in comparing IEUBK model predictions to observational data are discussed at some length in this strategy.

In a final section of their paper, “Then what good are models?”, Oreskes et al. offer the opinion that “...the primary value of models is heuristic: models are representations, useful for guiding further study but not susceptible to proof.” However, it is important to recall at this point that the authors' conclusions about the non-verifiable nature of models was presented as a special case of the non-verifiable nature of scientific theories and hypotheses in general. In many fields, a major application of both general theories and specialized models is to make predictions in support of practical actions. The strength of the scientific foundations of a model and the accumulated observational evidence confirming, or conflicting with, model predictions must be weighed when considering practical, predictive applications. This evaluation cannot be made in isolation, the relative merits and confirming evidence for alternate approaches to support decisions must also be considered.