NATIONAL BUREAU OF STANDARDS REPORT

10 657

DATA COLLECTION AND ASSIMILATION FOR THE LEAD PAINT POISONING MODEL

U.S. DEPARTMENT OF COMMERCE

NATIONAL BUREAU OF STANDARDS

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards[1] was established by an act of Congress March 3, 1901. The Bureau's overall goal is to strengthen and advance the Nation's science and technology and facilitate their effective application for public benefit. To this end, the Bureau conducts research and provides: (1) a basis for the Nation's physical measurement system, (2) scientific and technological services for industry and government, (3) a technical basis for equity in trade, and (4) technical services to promote public safety. The Bureau consists of the Institute for Basic Standards, the Institute for Materials Research, the Institute for Applied Technology, the Center for Computer Sciences and Technology, and the Office for Information Programs.

THE INSTITUTE FOR BASIC STANDARDS provides the central basis within the United States of a complete and consistent system of physical measurement; coordinates that system with measurement systems of other nations; and furnishes essential services leading to accurate and uniform physical measurements throughout the Nation's scientific community, industry, and commerce. The Institute consists of a Center for Radiation Research, an Office of Measurement Services and the following divisions:

Applied Mathematics—Electricity—Heat—Mechanics—Optical Physics—Linac Radiation[2]—Nuclear Radiation[2]—Applied Radiation[2]—Quantum Electronics[3]—Electromagnetics[3]—Time and Frequency[3]—Laboratory Astrophysics[3]—Cryogenics[3].

THE INSTITUTE FOR MATERIALS RESEARCH conducts materials research leading to improved methods of measurement, standards, and data on the properties of well-characterized materials needed by industry, commerce, educational institutions, and Government; provides advisory and research services to other Government agencies; and develops, produces, and distributes standard reference materials. The Institute consists of the Office of Standard Reference Materials and the following divisions:

Analytical Chemistry—Polymers—Metallurgy—Inorganic Materials—Reactor Radiation—Physical Chemistry.

THE INSTITUTE FOR APPLIED TECHNOLOGY provides technical services to promote the use of available technology and to facilitate technological innovation in industry and Government; cooperates with public and private organizations leading to the development of technological standards (including mandatory safety standards), codes and methods of test; and provides technical advice and services to Government agencies upon request. The Institute also monitors NBS engineering standards activities and provides liaison between NBS and national and international engineering standards bodies. The Institute consists of the following divisions and offices:

Engineering Standards Services—Weights and Measures—Invention and Innovation—Product Evaluation Technology—Building Research—Electronic Technology—Technical Analysis—Measurement Engineering—Office of Fire Programs.

THE CENTER FOR COMPUTER SCIENCES AND TECHNOLOGY conducts research and provides technical services designed to aid Government agencies in improving cost effectiveness in the conduct of their programs through the selection, acquisition, and effective utilization of automatic data processing equipment; and serves as the principal focus within the executive branch for the development of Federal standards for automatic data processing equipment, techniques, and computer languages. The Center consists of the following offices and divisions:

Information Processing Standards—Computer Information—Computer Services—Systems Development—Information Processing Technology.

THE OFFICE FOR INFORMATION PROGRAMS promotes optimum dissemination and accessibility of scientific information generated within NBS and other agencies of the Federal Government; promotes the development of the National Standard Reference Data System and a system of information analysis centers dealing with the broader aspects of the National Measurement System; provides appropriate services to ensure that the NBS staff has optimum accessibility to the scientific information of the world, and directs the public information activities of the Bureau. The Office consists of the following organizational units:

Office of Standard Reference Data—Office of Technical Information and Publications—Library—Office of International Relations.

---

[1] Headquarters and Laboratories at Gaithersburg, Maryland, unless otherwise noted; mailing address Washington, D.C. 20234.
[2] Part of the Center for Radiation Research.
[3] Located at Boulder, Colorado 80302.

NATIONAL BUREAU OF STANDARDS REPORT

NBS PROJECT		NBS REPORT
4314518	December 7, 1971	10 657

DATA COLLECTION AND ASSIMILATION FOR THE LEAD PAINT POISONING MODEL

Milestone 2

by
Judith Gilsinn
Applied Mathematics Division
Lead Paint Poisoning Project
Building Research Division
Institute for Applied Technology
National Bureau of Standards

Sponsored by
Department of Housing and Urban Development

NBS

U.S. DEPARTMENT OF COMMERCE

NATIONAL BUREAU OF STANDARDS

PREFACE

This report can be grouped, conceptually, with 4 other NBS Reports under the blanket title "Pediatric Lead Paint Poisoning in the United States--A Survey with Preliminary Estimates." Under this arrangement, the "parts" of the study would be listed:

Part I NBS Report # 10499 "The Nature of the Lead Paint Poisoning Hazard"

Part II NBS Report # 10657 "Data Collection and Assimilation for the Lead Paint Poisoning Model"

Part III NBS Report # 10653 "Effect of Data Aggregation in Modelling"

Part IV NBS Report # 10654 "A Model to Estimate the Incidence of Lead Paint Poisoning"

Part V NBS Report # 10651 "National Estimates of Lead Based Paint Poisoning of Children (Estimated by Standard Metropolitan Statistical Area)"

These papers were intended as interim progress reports covering work done up to the time of publication. Reports describing validation and refinement of the models and outputs as well as analysis of the outputs will be issued subsequently. A summary report encompassing revision of the current ones and those projected above, under one cover, is anticipated.

ABSTRACT

Criteria are given for the choice of data items and data sources for use in a model to predict the incidence of pediatric lead paint poisoning in the United States. The required environmental and population variables are described, and the sources and availability of these data are compared. Sources of lead poisoning incidence data are also identified and appraised.

TABLE OF CONTENTS

LIST OF TABLES

DATA COLLECTION AND ASSIMILATION FOR THE LEAD PAINT POISONING MODEL

1. INTRODUCTION

In any project involving the construction of a mathematical model, a major effort must be made to collect and assimilate data to facilitate model development, calibration, validation, and application. Such an effort includes:

a. determination of relevant data items,*

b. determination of the sources of those items,

c. obtaining the data items, and

d. assimilating them into a form to enable further processing.

Once the data have been assimilated, preliminary analysis can be conducted in order to decide which items are most relevant to the process being modelled and which predict best. Further analysis and perhaps modelling may be necessary to fill in any gaps in the existing data.

1.1 Relevant Data

The first step in a data collection effort is the determination of what data items to collect. It may seem at first to be overly pedantic to consider this as a separate step, but if care is not taken at this stage, decisions may be made which are later regretted. It is often better to collect too much data than too little, but there is a practical limit on how much can realistically

*In this report, "data items" will refer to categories of information, while "data" will refer to the numbers in a category.

be assimilated and checked. Deciding which data items to collect involves first learning about the process being modelled in order to see what has an effect on that process. It is at this point that the decision should be made on what the dependent variable of the model should be, i.e., what exactly is being modelled. Relevant literature and conversations with experts in the field can help determine those data items which are being modelled, and may be used to eliminate other items. A preliminary search of data sources may also uncover items which one's intuition may suggest as relating to the process being described. Any previous modelling of the same phenomenon may also provide insight as to the choice of data items to collect.

1.2 Data Sources

Once the kind of data to be collected has been decided upon, one must search for sources of such data items. Of course, the decision on which data items to collect is not entirely independent of which items are available. However, if one is too influenced by sources he knows about, he is unlikely to investigate less obvious sources for a desirable item which is more difficult to acquire. The two major criteria which must be used in assessing the different data sources are the quality and accessibility of the data.

1.2.1 Data Quality

Quality includes accuracy, uniformity, and compatibility with other data sources. In assessing data accuracy it is necessary to know the reputation of the organization collecting the data and to understand their sampling procedures. It has often been said that one can "lie

with statistics," that one can obtain data supporting almost any thesis he wishes. It is therefore necessary to investigate biases in data to eliminate, as much as possible, prejudging the outcome of the modelling process. Another element in assessing the quality of data is the uniformity with which it is collected. In the case of lead poisoning, for instance, it is desirable to have uniformity both with respect to geographical location and socio-economic status, since we must be able to predict for all people in all parts of the country. This ideal is unlikely to be met in full, but the degree to which it can be met by any source will be a major factor in deciding for or against it as a data source. A final aspect of the quality of a data source is its compatibility with data from other sources. This involves, in part, the level of geographical aggregation of the two sources. This problem will be discussed in greater detail in Effect of Data Aggregation in Modelling,* but it will be noted here that the data sources must either be at the same level of aggregation or else be capable of being aggregated or disaggregated, so that all model variables are at the same level. This may require some modelling, particularly if data must be disaggregated.

*NBS Report #10653, Effect of Data Aggregation in Modelling.

1.2.2 Data Accessibility

A second criterion in addition to data quality is the accessibility of the data from any source. It should be noted here that this project has neither the funds nor the time to engage in any major data collection effort. Such efforts require a large staff, many trips into the field, and a great amount of time in both planning and data reduction phases. Since this project is a short term, relatively low budget effort, it was decided to utilize only already existing data sources. This, however, did necessitate a search for those sources and an examination into the real availability of obtaining the information contained in these sources, within the time and budget constraints of the project.

A major consideration of data accessibility is the form that the data are in. Many people assume that because they have a filing cabinet full of papers with numbers on them, they have lots of data. However, if it would take several man-months of work to reduce this file to a manageable form, then these data are not really accessible. To be accessible, data ideally should be in machine readable form (on punched cards or magnetic tape) or else be few enough in number to be reducible to machine readable form with a small effort. One might reasonably expect to code 2000 forms in 2 man weeks, and that is a project one would be willing to undertake, whereas the coding of 200,000 items would be out of the question. Even machine readable data might not be truly available, if it requires either additional hand-coding or very sophisticated data processing techniques to extract the necessary information.

As an example of this, it is necessary for this project to have data broken down by geographical areas. Some sources have such information, but their only geographical reference point is a street address. The state of the art of computer address coding is not very advanced; it has been successfully applied only for a small geographical area. The alternative is to hand code the addresses to our geographical zone system, which is a reasonable procedure if the number to be coded is not too great. Thus, it is necessary to investigate the data sources in detail to ascertain how much work will be necessary to reduce the data to a form which can be used as input to the computer modelling process.

1.3 Obtaining Data

Once the source or sources of data have been chosen one must actually obtain the data. This process is likely to be time consuming and frustrating, but in principle at least it is relatively straightforward, since presumably in evaluating the sources the availability of the data was investigated.

1.4 Data Assimilation

The final stage in a model data collection effort is data assimilation, by which is meant collecting the data in one place and putting it into the final form for which it becomes input into the modelling process. This includes:

1. putting all the data into machine readable form on tape or on punched cards,
2. data processing to select and reformat to facilitate entry into the modelling process,

3. aggregation and disaggregation to ensure all data items appear at the same geographical zonal level,
4. modelling to fill in any gaps in the necessary data, and
5. preliminary analysis to narrow down the number of data items to be used in the model and to better understand the relationships between these variables.

These steps are thus designed to bring all the data together in one place in one form so that they can be examined critically and massaged, if necessary, before it is used directly in a modelling effort.

There is a gray area between what is properly data collection and analysis and what is properly modelling. The two processes are not separate and must be meshed to ensure that the model is developed from an adequate data base. Much can also be learned aobut the process being modelled from a careful analysis of this data base. Some kinds of processes are such that one can discover a functional cause and effect relationship among the variables, and thus build a model before examining the data. However, social processes do not generally fall into this category, and lead paint poisoning is no exception. Although most authors writing about lead paint poisoning have noted characteristics of the environment and population associated with the disease, few of these characteristics, with the exception of the presence of peeling paint, can be shown definitely to have a cause and effect relationship. Therefore, it is necessary to study the data to ascertain which variables and which combinations of variables are most associated with the incidence of lead poisoning. A second reason for analyzing

the data is to discover correlations among the variables. Since most statistical estimation procedures require the variables to be independent, it is desirable to use in the model only variables which are relatively uncorrelated.

Data is used in the modelling process in four ways:

1. for model development,
2. for model calibration,
3. for model validation, and
4. for model application.

The data analyses described above can be used to aid in deciding which variables to include in the model and to help determine the model form. Data is used in the model calibration process for determining the best estimates for model parameters. The model validation process requires comparing the predicted values of the model's dependent variables with actual values. Finally, it is necessary to have values for the independent model variables in order to apply the model. The first three thus require values for both the independent and the dependent variables, while the last requires only the independent variables. Values for more variables are also required during the model development process since the model form has not been established at that time. However, the data collection, assimilation, and analysis process is very much the same for all four of these stages. Since many of the data items required are the same for all four stages, one data search should uncover most sources of data for all stages, although items which are difficult to obtain may require an ongoing effort to improve or locate further sources.

This concludes the general description of the data collection process. The following sections will describe the specific application of this process to the collection and assimilation of data for the lead paint poisoning model.

2. GENERAL DESCRIPTION OF DEMOGRAPHIC DATA REQUIREMENTS

The data requirements for the lead paint poisoning model can be divided into two types 1) demographic and 2) incidence. This division is made primarily because the sources of these two types of data differ, and because incidence data is not needed for prediction, since if it were available for the whole country it would not be necessary to construct a model. Incidence data is required for model development, calibration and validation.

The demographic variables which have been identified with lead paint poisoning can be divided into two types: 1) those concerned with the environment in which the child with lead poisoning lives and which will primarily be associated with housing conditions and 2) those concerned with the socio-economic characteristics of the family of the child. Tables 1 and 2 contain the lists of possible housing and population characteristics which have been associated with pediatric lead poisoning in the literature and in conversations with experts in the field. It can be seen that many of these variables are related. For instance, low income people are more likely to live in housing of poor condition. Therefore, it is probable that only some of these variables will actually appear in any model formulation. However, those that will be the most relevant and the best predictors cannot be known until values for all are in hand and can be analyzed.

TABLE 1: Housing Characteristics Associated with Lead Paint Poisoning

1. Age of housing
2. Condition of housing - dilapidated or deteriorating
3. Value of house
4. Number of renter occupied dwelling units
5. Vacancy rate
6. Number of multiple unit structures

TABLE 2: Population Characteristics Associated with Lead Paint Poisoning

1. Age of children
2. Number of female household heads
3. Population crowding
4. Mobility
5. Income
6. Educational level of head of household
7. Work status of head of household
8. Race
9. Region of birth

As described, some of these variables pertain to individual households, others to area aggregates. The individual household variables can be redefined at area levels when it is desired to incorporate them into area models. Thus, e.g., "Age of housing" may be represented:

mean age of dwelling units in area
mean age of structures in area
median age of dwelling units in area
percentage of dwelling units in area at various age levels

and so forth.

2.1 Environmental Factors

The most obvious pertinent environmental factor is the presence of lead paint in a form that the child can easily ingest. Until the 1940's lead paint was the major good quality interior house paint used in this country. Subsequently, lead paint usage declined as the new latex paints were introduced. In 1955 the American Standards Association (now ANSI) adopted a voluntary standard prohibiting interior house paint from containing more than 1% lead by weight, and requiring a message cautioning against its use on surfaces accessible to children to be printed on the label of any leaded paints. However, a recent New York survey discovered lead in paints being sold for interior use in that city, in August of this year. Lead exterior paint is still sold, and there is little to prevent one from ignoring the warning message and using it on interior surfaces. In addition, children have access to exterior surfaces as well as interior ones and may eat paint chips from railings, porches, garages, exterior window sills and other exterior surfaces. Therefore, some surfaces being painted today may well offer a potential hazard to children in the future. It is impossible to set a cut-off date for which one could say that housing built prior to that time is probably hazardous, and without a major survey of actual housing one can only speculate on the fraction of housing built in any year which has any lead paint in it. Also, although a dwelling unit may be free of lead paint except for one window sill, that surface may have a high enough lead content to kill a child or cause permanent brain damage.

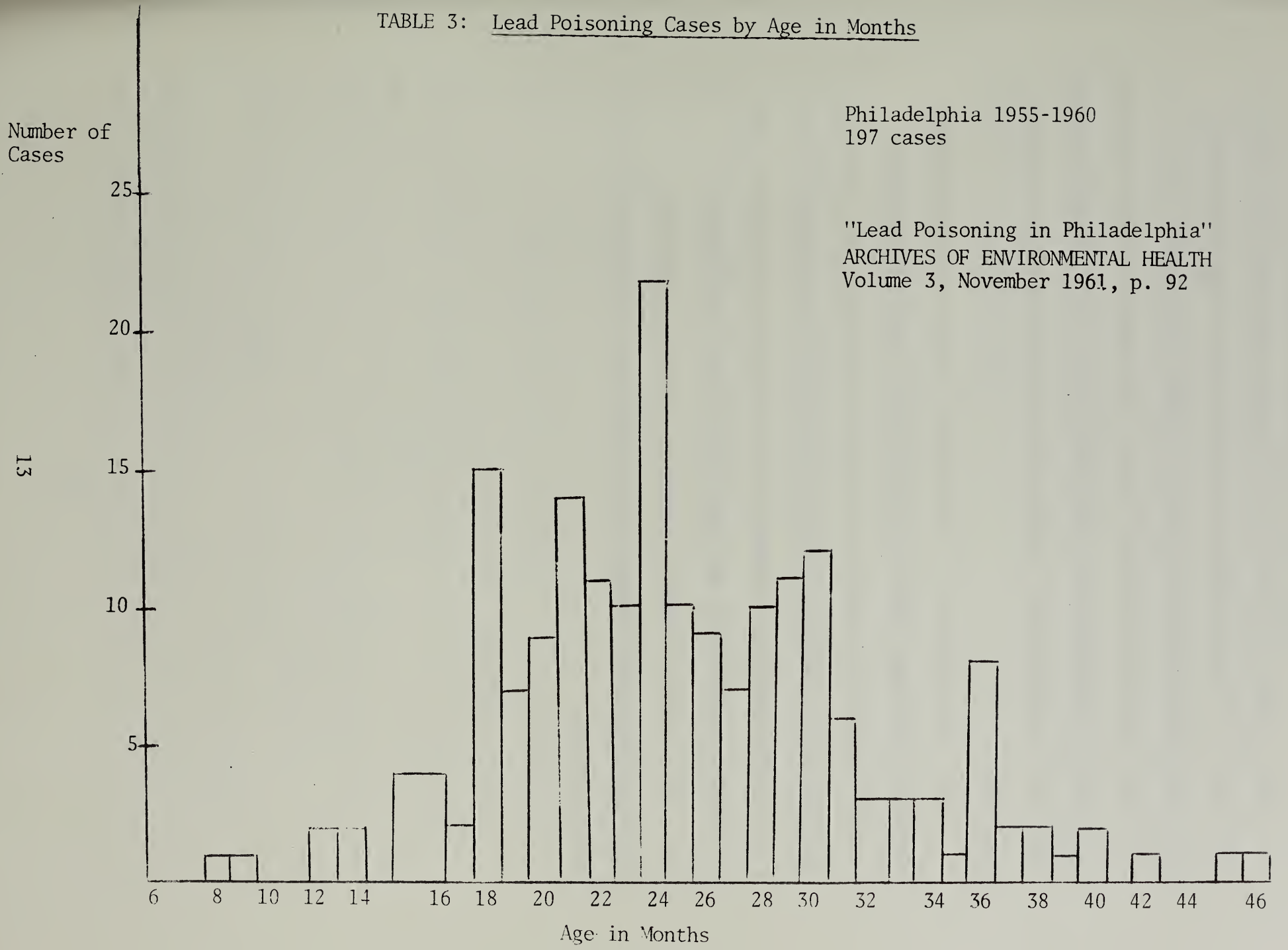
In the absence of more definitive information on the correlation between the age of a house and the presence of lead paint, the condition of the dwelling unit becomes more important. The mere presence of lead paint on a surface does not necessarily offer an immediate hazard to a child, if the surface has been well maintained, the paint is not chipping or peeling, and the surface does not present a chewable edge to the child. Such surfaces are, of course, a potential hazard, since if the dwelling unit is allowed to deteriorate, at some future time peeling and chipping paint may provide a source of lead for a child to ingest. Indeed, several cities with lead poisoning programs cited once good homes, which have been divided into apartment units and which have since deteriorated, as fairly typical lead poisoning environments.

Another characteristic of the housing environment associated with children who get lead paint poisoning, is that most of these children live in rented quarters. A factor here is, of course, the fact that lead poisoning occurs mainly among the poor, who are less able to own their own property and are more likely to live in substandard housing.

2.2 Population Factors

The major population characteristic associated with lead poisoning is the age of the child. All but a small number of reported cases have occurred in children under 6 years old. Table 3 shows that over 85% of lead paint poisoning cases in Phildelphia occurred in children 18 to 35 months old.

TABLE 3: Lead Poisoning Cases by Age in Months



Philadelphia 1955-1960
197 cases

"Lead Poisoning in Philadelphia"
ARCHIVES OF ENVIRONMENTAL HEALTH
Volume 3, November 1961, p. 92

The children who develop lead poisoning come from poor families that have all of the problems of such families. The families have a disproportionate number of female household heads, they live in crowded conditions, they move often, the household head is employed, if at all, in an unskilled or semi-skilled job, and a disproportionate number come from minority groups such as blacks and Spanish speaking Americans.

These characteristics of housing and children have been noted before and have been used by groups that are screening children to locate those with the greatest risk of developing lead poisoning. Some communities which had children with high risk characteristics living in hazardous environments, had no previously reported cases but found lead poisoning when they screened the children for it. Thus, the variables listed in Tables 1 and 2 constitute those demographic data most commonly associated with identified cases of lead poisoning.

3. CHOICE OF DEMOGRAPHIC DATA SOURCE

The major sources of the demographic data listed in Tables 1 and 2 are the U.S. Census and various local sources. These have been judged by the criteria described in Part 1 of this report, and the U.S. Census data has emerged as by far the best source of demographic data for use in a nationwide lead paint poisoning model. The reasoning behind this decision is described below.

3.1 Uniformity

A main deficiency of local data is its lack of uniformity across the nation. The quality and availability of data generally depends on the city size and/or the abilities of a department head. Large cities have the funds and staff to collect, maintain and analyze data. Pilot programs in some cities have led to the collection and processing of some information. However, in general, the kinds and quality of the data retained by different localities vary markedly from place to place and data from one city is not necessarily likely to be compatible with data from another. The size of local zonal systems will very likely vary. (In fact, the zonal systems used by different departments such as health and police within the same city are likely to be different.) The type and definition of the data collected may vary. For example, one jurisdiction may date the age of a structure from the time it was first assessed, while another may use the date of issuance of a building permit.

It is one of the goals of the U.S. Census Bureau of collect and make data available on a nationwide basis so that different areas of the country may be compared. There may be some uncertainty that the

Census Bureau has been able to achieve this goal, but for the most part the Census data is much more uniform than any other alternative.

3.2 Accuracy

In comparing the U.S. Census data with local sources on the basis of accuracy, local sources in some areas may be better. However, again because of the lack of nationwide uniformity, some local sources may be less accurate. In fact, many municipalities themselves depend on Census statistics. The main area in which local data would be most useful and most accurate is in the area of housing statistics, since most local governments must assess property to obtain property taxes. Also many local building departments have a great deal of information in the form of experienced and knowledgeable personnel, but this is not codified in any form and cannot truly be called data. City personnel may, for instance, be able to point out dilapidated units, but until the address and number of dwelling units in each such structure is actually listed, that information is not available data. Thus, although local municipalities may provide a potential data source which might in some cases be more accurate, more current and better suited to our needs than the U.S. Census, the lack of uniformity and the problems of obtaining such data again lead us to the U.S. Census as the better data source.

3.3 Accessibility

Another consideration upon which the decision to use U.S. Census data was based is the accessibility of the data. As described above, much of the data potentially available from local sources is not in

a form in which they can be processed. Some are not even written down and what is written down must be extracted, aggregated by some zonal system, and put into machine readable form. Few cities, except large ones such as New York, have such data on tape or punched cards, and even those would require some hand coding.

The main drawback to using local sources for data is the sheer magnitude of the effort required to obtain information from all of the major cities in the country. This project lacks both the funds and, more importantly, the time to engage in such an effort. Local sources are, however, being consulted in the cities of New Haven, Connecticut, and Aurora, Illinois, which are principal sources of incidence data that will be used for model development and validation. The city of New Haven has provided us with a map showing current areas of urban blight. This has been used to aid in modelling current unsound housing figures.

Although the 1970 Census was conducted a year and a half ago, not all of the data is currently available. Therefore, it has been necessary to investigate 3 different sets of census data: the 1960 Census, the 1970 Census First Count data, and 1970 Census Fourth Count data.

3.3.1 1960 Census Data

The 1960 data are already available, tested, and known to be reliable at different levels of aggregation. However, these data are more than 10 years old and urban renewal, changing patterns of residence, and land use differences have all contributed to the present inaccuracies of the data. It is desirable, therefore,

to use more current data wherever possible. Be that as it may, one variable listed in Table 1, namely condition of housing, will not be available to this project from the 1970 Census. In the interest of uniformity, the Census Bureau eliminated this variable from the 1970 Census, because it required a subjective decision by the enumerator. Thus, without further modelling, the 1960 Census is the sole source of this data item. In addition, because of delayed publication of the 1970 Fourth Count data, the 1960 Census will be the only source of several other variables listed in Table 2. The 1960 Census data appear in books available from the Commerce Library and on tape. Because of the difficulties associated with extracting the desired data from the tapes and the number of tapes involved, it has been decided to hand code the data items from the 1960 Census as required.

3.3.2 1970 First Count Census Data

1970 First Count Census data is currently available both from the Census Bureau and from private vendors who have processed it for easier use. We have obtained 1970 First Count data and have reformatted it for our own use for the selected areas required by the modelling process. The major limitation of the First Count data is that certain key items such as age and condition of housing, income, educational level, work status, mobility, and region of birth are not included. As indicated before, the condition of housing will not be available for 1970 at all since that data item was not collected.

3.3.3 1970 Fourth Count Census Data

1970 Fourth Count Census data will contain all the data items in Tables 1 and 2 except the condition of housing. However, Fourth Count data is not yet available, although it was orginally slated for release last August. Current scheduling calls for the release of this data in the Spring of 1972. Thus, it will not be available in time to be used for the January report to Congress, but may possibly, if released early in the Spring, be useful for the final report and modelling effort. The Census Bureau has developed a model for predicting housing condition as a function of Fourth Count variables. This model may be particularly useful when predicting lead poisoning incidence for areas in a city which have changed greatly in the last 10 years because of urban renewal projects and changing land use patterns.

Because of data unavailability, it has been necessary to make use of a combination of 1960 data and 1970 First Count, using the more current data whenever possible. Table 4 lists the Census data categories, the required data in each category, the current status of that data, and problems with that category. Data which is "on cards" is on punched cards which have been used; data on tape has been reformatted to permit easier handling on the NBS computer.

3.4 Choice of Zonal System

The decision to use Census data includes also the decision to use the Census zonal system. This system has many levels of aggregation, but the primary ones for our purposes are the Census block group, the

TABLE 4: Census Data Status

Data Item	Availability	Problems
1. 1960 Census		
A. Census tract level		10 years old
1. New Haven	on cards	
2. Aurora, Ill.	on cards	
3. Chicago	on cards	
B. SMSA level (whole U.S.)	on cards	
2. 1970 First Count		
A. Census tract level		Does not contain age or condition of housing, income, education, or work status
1. New Haven	on cards	
2. Aurora, Ill.	on cards	
3. Chicago	on tape	
4. New York City	on tape	
B. SMSA level (whole U.S.)	on cards	
C. County level (whole U.S.)	on tape	
3. 1970 Fourth Count	not yet	May not be available in time, will not contain condition of housing

Census tract, and the Standard Metropolitan Statistical Area (SMSA). There are about 250,000 Census block groups in the country, 50,000 Census tracts and 243 SMSA's. There are about 3,000 to 10,000 people in a Census tract, and 4 to 6 block groups per census tract. The New York SMSA has over 2,000 Census tracts in it. It was decided to use Census tract level data for model development and calibration because 1) incidence data was available only for a few cities so that it was necessary to use a level of aggregation less than a whole city, and 2) census tracts are more easily available, there is a relatively small number of them, and the data set is more manageable than if block groupings were to be used. In the future it may be desirable to make use of block groupings, since because they are smaller, they should be more homogeneous. On the other hand it was decided to predict for the entire United States on the SMSA level. This can be done in two ways: 1) predict for each Census tract on the basis of the characteristics of that tract and aggregate to the whole SMSA or 2) predict for the whole SMSA on the basis of aggregate and averaged variables. It was decided to take the second choice, because of the volume and cost of obtaining data for the first procedure. A third alternative is to predict for each county in the country. This was rejected for three reasons: 1) There are over 3,000 counties which is a large volume of 1960 data to code and display effectively. 2) Different states have quite different size counties. 3) All available incidence data is for urban populations, and it is believed that the model will not predict as well for rural as for urban areas. Many counties are

rural, but SMSA's, by definition are urban and suburban.

There are problems of consistency between 1960 and 1970 Census at all levels of aggregation. Counties in Virginia do not remain the same, since cities can become independent of counties. Census tracts are split, sometimes into several pieces, as areas grow. Even the block groupings change with population shifts. In 1960 there were fewer than 200 SMSA's, but in 1970 there are 243. Some of these SMSA's are new, but in at least one case 2 formerly separate ones (Greensboro-High Point and Winston Salem) have been combined. Thus, special care must be taken whenever 1960 and 1970 data are combined.

The major criteria involved in the decision as to the choice of a demographic data source have been the time for locating, obtaining, and assimilating data from any source, the uniformity and applicability of the data, and its availability. The final source is a combination of 1960 Census data and 1970 First Count Census data at the census tract level for selected cities and the SMSA level for the country.

4. INCIDENCE DATA

In addition to the demographic data needed for the modelling process, data on the incidence of lead poisoning are also needed. It became obvious at the outset of this project that the definition of a lead poisoning "case" was unclear and differed from one community to another and from one doctor to another. Different cities use different tests to determine which children have lead poisoning, and these tests have differing degrees of reliability. Therefore, the Surgeon General's recommendation of a lead content of 40 μg per 100 ml of whole blood is being used as a precise definition of lead poisoning incidence as it relates to the model and to the data to be obtained. Thus, the incidence data required are the number of children screened in each zone and the number of those screened with blood lead levels greater than 40 μg/100 ml. It would be more desirable, in fact if the whole blood lead level distribution were available so that one could estimate the number over any desired limit value.

Several criteria were used in evaluating various sources of incidence data.

1. The data should be acquired from an area with a mass screening program,
2. The screening program should be as unbiased as possible,
3. The data must be able to be readily aggregated by the geographical zonal system chosen for the modelling process,
4. Data should be determined by a blood lead level test.

4.1 Mass Screening

Data should be obtained from a mass screening program in order to eliminate as much bias as possible and to minimize the difficulties in factoring screening to the whole population. Many testing programs around the country have only screened those children whose parents brought them into a clinic or hospital. They have no community outreach programs to go out and contact children to be tested. Children who are brought in to well baby clinics are apt to be those whose parents are most concerned about them, and who are most able to spend the time, effort and money to bring the child into a central facility. Thus, one might expect such children to suffer from a lower incidence of lead poisoning. On the other hand, a child already exhibiting the symptoms of lead poisoning would be very likely to be taken into a clinic or hospital because the child is sick. Thus, a program which lacks community outreach introduces a bias, but it is difficult to decide whether the bias is in favor of higher or lower incidence rates, and it is quite impossible to know the size of the bias. Only testing a large proportion of all children in an area and in particular testing those children who have little or no contact with the organized medical establishment can lessen possible bias.

In addition to reducing the bias, large screening programs reduce some of the problems associated with factoring a screening sample to the whole population of children. If one has the number of children screened and the fraction of those with lead poisoning, one must still estimate the total number of children with the disease.

This estimation process is called factoring. In programs which screen children only in a hospital, for instance, some area may be represented by only a small number of children, a high fraction of which have elevated blood lead levels. This fraction need not reflect the true incidence of the area because it is based on a small biased sample. In addition, lead poisoning is a phenomenon which, prevalent at a high enough level to be of epidemic proportions, is found in significantly less than half the children tested. A small sample might miss it entirely, and certainly if one wanted to estimate the fraction to two digits, one ought to screen at least 100 children in each zone. Thus, a mass screening effort is necessary for an accurate predictive model and zones for which the number screened is too small, should be dropped from the data list.

4.2 Bias

Bias is, however, present to some degree in all existing programs. Since time and budgetary constraints do not permit this project to conduct screening, it is necessary to obtain data from already existing sources. The primary purpose of existing screening programs is to find as many lead-poisoned children as possible, rather than to provide data for a lead poisoning model. Therefore, they focus their screening efforts on those areas of the city or state whose characteristics, as described in Tables 1 and 2 above, identify them as high risk areas. Thus, we cannot hope to obtain data free from such a bias toward high risk areas. However, it is desirable that data be as free from other known biases as possible.

4.3 Zonal System Coding

Another criterion, which has in fact proved to be a major limiting factor in the choice of data sources, is that it is necessary that the form of the data be such as to facilitate aggregation by a zonal system. Few of the data sources have any zonal system classification built into their data; most just record the child's street address. Since, as previously mentioned, the state-of-the-art of computer address coding is quite rudimentary, any coding of zones from addresses must be done by hand. A partial solution to this problem is to sample from the complete data base, taking a small enough sample so that it can be hand coded within project time and budget constraints. There is, however, the difficulty of deciding how representative a small sample is, and this can lead to problems in factoring the sample to the whole population. The ideal data source would be a small city in which only a relatively small (in absolute) number of children need be screened in order to account for a large percentage of the total number of children.

Two points must be made here. The first is that it is necessary to code both the addresses of children who have elevated blood lead levels and those who do not. One must have the number of children screened in order to be able to factor up the number of children with elevated blood lead levels to those which might be expected if all children had been screened. Thus, if 25 children out of the 100 screened had elevated blood lead levels and there were 500 children in the zone, one would expect 125 children in the zone to have elevated blood lead levels. (This of course assumes the 100

children screened were representative of all children in the zone.) It would not be enough to just know that 25 children in the zone had elevated blood lead levels, since the incidence rate could be anything from 5 to 100 percent. A second suggestion has been to code only data from a few zones. The difficulty with this alternative is that one cannot know the zone before the address is coded, so that all must be coded to know which to pick and which to discard. However, if there is other information, such as the drawing site, which can be used to eliminate a large proportion of the data such a plan might be feasible.

4.4 Test Used to Screen

The final criterion on which to base the choice of sources for incidence data is a medical one. The data must be based on blood lead level determinations. Other methods are less reliable; in particular they result in too many false negatives and thus could lead to under predication by as much as 30 percent. The blood lead determination is the test recommended by the Surgeon General. Recent efforts have been made to test out a micro-blood test, but until this has been proven, the standard blood lead test requiring 5 ml of blood is generally considered the most reliable.

4.5 Sources

Table 5 lists known programs, their sizes, known biases, and test methods. Question marks indicate that the particular item is not known. The major mass screening programs are in Chicago and New York. Philadelphia is currently screening in model cities areas but is not using the standard blood lead determination test. In

TABLE 5: Lead Poisoning Screening Programs

	SIZE	BIAS	TEST
Aurora, Ill.	1478 (7/71-now)	mass screening	blood
Baltimore BCEM	small 50-100 in each of 30 cities	hospitals selected areas	blood micro
Boston	small	community groups	micro
Connecticut	New Haven 1897 Hartford 147 Waterbury 500 Bridgeport 2 cases New London 1 case Stanford 130	some screening but through clinics etc.	ALA urine
Chicago	28008 (1967) 40785 (1968) 47527 (1969) 44347 (1970) 28973 (1971)	mass screening	blood
Cincinnati	small	hospitals	?
Cleveland	planning	?	?
Delaware	Dover 3 cases Wilmington 6 cases	no mass screening	?
Denver	none	-----	-----
Detroit	starting	?	?

	SIZE	BIAS	TEST
Illinois	Aurora 449 Springfield 670 Peoria 387 E. St. Louis 376 Decatur 793 Joliet 383 Rock Island 285 E. Moline 293 Robbins 103 Harvey 226 Carbondale 264	selected areas	blood
Kansas City, Mo.	planning	hospital based	?
Louisville	2 cases	no screening	?
Merimac Valley, Mass.	small	?	?
Milwaukee	none	------	------
Minneapolis	small	student committee organized	ALA urine
Nassau Co. N.Y.	3 cases	?	?
Newark	1973 (1/69-3/70)	selected areas	ALA urine
New Haven	1897 (1969) 1990 (1970) 364 (1/71-6/71)	mass screening in 1970, but only clinics and hospitals in 1971	blood (some only ALA urine)
New Orleans	300	health centers & small screening	micro
New York	80,000 per year (1970-71)	mass screening	blood
New York State	reportable (see Nassau Co.)	-----	----

	SIZE	BIAS	TEST
Norfolk	1,200	well baby clinics	micro
Oklahoma City	4 cases	?	?
Philadelphia	5,000 to be screened (2/71-6/71)	model cities areas	blood enzyme
Portland, Me.	1,000 (1970)	small mass screening, will have in well-baby clinics	ALA urine
Rhode Island	Providence 2,600	head start, clinics	hair sample, if positive, take blood
Rochester, N.Y.	small	hospital and community groups	?
San Francisco	small (1,000)	?	?
St. Louis	1847 (since 6/28/70)	well baby clinics and community groups	micro
Washington, D.C.	small (6/70-808)	hospitals, UPO centers and well baby clinics	808 micro, blood in hospitals

1970 New Haven conducted a mass screening program, but because of lack of funds this program was not continued in 1971. Since July 22, 1971, Aurora, Illinois, has screened over 1400 children out of about 10,000 children in the age group 1-6, of which 2,000 are estimated as high risk children. These five programs are the only ones which can be described as mass screening programs. Two other programs should be noted here, because although neither is a mass screening program in the sense of screening a large proportion of the children in an area, both are attempting to screen children from different cities to ascertain the geographical range of lead poisoning. These two programs are the Illinois screening of 10 Illinois cities and HEW's Bureau of Community and Environmental Management's (BCEM) program of screening in 30 U.S. cities spread throughout the country.

Table 6 lists the major incidence data sources and problems associated with each. The screening programs in New York and Chicago have a great deal of data potentially available. However, the problem of address coding these data precludes their immediate use. The Chicago data is not in machine readable form and would have to be coded, but New York has put its case and screening data on tapes. Philadelphia's program is new and is not using the standard blood lead determination. The two programs, therefore, which offer greatest possibility of use are those in New Haven and Aurora. The New Haven data are immediately available, and there is a published report containing the number of children screened and the number of elevated blood lead levels (greater than 40 μg /100 ml) for each of the census tracts in that city. Aurora is also a good potential data source,

TABLE 6: Incidence Data Sources and Problems

SOURCE	AVAILABILITY	PROBLEMS
1. New Haven	in use	
2. New York	on tape in N.Y.	Only has addresses, large amount to process
3. Chicago	on file cards in Chicago	Large number screened, needs to be put on tape, will cost 11¢/child
4. Aurora, Ill.	on forms, can photo reproduce	Will probably need some geographical coding but only 2000 records
5. BCEM	5 cities on hand prelim.	Only has distribution of blood leads, no geographic information, small sample
6. State of Illinois	6 cities on hand	Only has distribution of blood leads
7. Philadelphia, Baltimore, Washington	Forms, etc.	Too small samples, not mass screening, not representative of entire city.

since although a large percentage of the children at risk have been screened, this represents a relatively small absolute number of children and one which can be hand coded within the project time frame and budget limitations. The OEO office in Aurora, which has been directing the screening program there, has been very cooperative, is currently coding the screening results. This source will, in the near future make these available aggregated by census tract.

5. FUTURE DATA COLLECTION

There are still some deficiencies in the data sets described above. Future data collection efforts will be aimed at alleviating these deficiencies as well as obtaining additional data for a more detailed model.

Although the main lack of data lies in the incidence data area, there is still some deficiency in the Census demographic data. The major problem here is the late release of the 1970 Census Fourth Count data. If the Fourth Count data becomes available for any part of the country, it will be obtained and assimilated for use in the modelling process as soon after its release as possible.

5.1 Census Data

Further Census First Count data may be desired for future model versions, in particular data at the block group level for a more detailed model. The major thrust for the first phase modelling effort has been toward developing a model to predict the national lead poisoning incidence. It is expected that future modelling efforts may be directed toward developing a model for use by a city to determine the geographical extent and level of lead poisoning throughout the city. It is undesirable to use census tracts because they are too large and inhomogenous to predict well for this effort. Therefore, it will be necessary to obtain Census data at the block level of aggregation for a selected set of cities for the future model process. Incidence data will also have to be obtained at this level for model development, calibration, and validation.

5.2 Existing Incidence Data

Deficiencies in incidence data can be remedied in two ways: 1) by utilizing existing data sources which have too much information to be assimilated in the short term, and 2) by designing a data collection effort to go out and screen selected small groups of children. Neither of these methods have been used for the first phase because of the time and cost involved. However, since the major difficulty in model development during the first phase has been the lack of incidence data, a heavy push toward obtaining extra data has a high expectation of producing a much better model and is particularly necessary for applying the model to geographical areas at lower levels of aggregation.

There are three possible existing sources which may be utilized in a further data collection effort; New York, Chicago, and Aurora. New York's screening data is currently on magnetic tape, but the volume which would have to be address coded precluded its use for the first phase of modelling. One way of possibly avoiding this large volume is to sample from the data base. Since the Lead Poisoning Control Bureau of the City of New York has been very cooperative, this procedure seems to offer much promise, although the difficulties of factoring a 1 or 2 percent sample to the total population must be investigated. Chicago's data is not in machine readable form. A computer firm in Chicago has submitted a proposal to the City of Chicago to put the screening data on tape, including address coding, for about 11 cents a child plus some fixed costs. This amounts to about $25,000 for all four years' data, but the

firm has estimated the cost for one years' data at about $8,000. This cost does not seem exhorbitant and would free project staff somewhat from having to supervise the coding effort. A third source of extra data is the Aurora screening effort which is being used for validation. Aurora has records containing information about each child screened. If proprietary and invasion of privacy considerations can be worked out between the Aurora OEO office and ourselves, we may be able to obtain this extra information which could be used to discriminate between the characteristics of the lead poisoned child and one who is healthy.

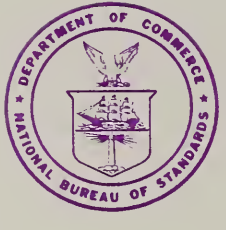
5.3 Collection of Extra Incidence Data

Even if all three sources of further data were to be exploited, there are still some data deficiencies remaining. The first of these is the lack of information about blood lead levels of children who are not living in the environment described in section 2 of this report. In particular, it would be desirable to screen a sample of white middle-class children living in older, but well kept homes. In addition, it would be helpful to be able to establish a correlation between the age of the house and the presence of lead paint. This latter would entail a housing screening program.

A second effort at obtaining more screening data would be to screen a large percentage of children in selected block groups in several cities. Some of these block groups might be chosen at random in order to obtain some idea of the blood lead levels of a normal urban child. Others could be chosen more specifically to study such things as regional effects. However, the main reason for

for using block groups is to be able to obtain screening data associated with areas which are more homogeneous with respect to socio-economic characteristics. If the block groups are in different cities they may be more representative of the nation as a whole, than if they were all in the same city.

Thus, future data collection efforts may be broken down into three areas: 1) obtain more Census data as needed, 2) try to utilize existing sources which could not be used in the first phase, and 3) request small, well designed screening efforts to fill in gaps in existing data.


DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS