Population-Based Surveillance and Etiological Research of Adverse Reproductive Outcomes and Toxic Wastes

PHASE II

THE EVALUATION AND DEVELOPMENT OF DATABASES ON POTENTIAL ENVIRONMENTAL EXPOSURES

NEW JERSEY DEPARTMENT OF HEALTH
A BETTER STATE OF HEALTH

Jim Florio
Governor

Bruce Siegel, M.D., M.P.H.
Acting State Commissioner of Health

POPULATION-BASED SURVEILLANCE AND ETIOLOGICAL RESEARCH OF ADVERSE REPRODUCTIVE OUTCOMES AND TOXIC WASTES

REPORT ON PHASE II: THE EVALUATION AND DEVELOPMENT OF DATABASES ON POTENTIAL ENVIRONMENTAL EXPOSURES

Frank J. Bove, ScD

March, 1992

Environmental Health Service
Division of Occupational and Environmental Health
New Jersey Department of Health

This report was supported in part by funds from the Comprehensive Environmental Response, Compensation, and Liability Act trust fund through a Cooperative Agreement with the Agency for Toxic Substances and Disease Registry, U.S. Public Health Service.

Project Officer: Larry D. Edmonds
Division of Birth Defects and Developmental Disabilities
National Center for Environmental Health and Injury Control
Centers for Disease Control

PRINCIPAL AUTHORS

Frank J. Bove, ScD, Mark C. Fulcomer, PhD, Judith B. Klotz, DrPH

CO-AUTHORS

Ellen M. Dufficy, BSN, Jorge Esmart, BA, George J. Halpin, MD,
Jonathan E. Savrin, MSPH, Rebecca T. Zagraniski, PhD, Leah Z. Ziskin, MD, MS

ACKNOWLEDGMENTS

The principal authors wish to express gratitude to many people for enabling the completion of this multi-year project and its five reports.

We were very fortunate to have a dedicated staff who contributed to the project over a long period of time: many of these individuals continue to work at the Department, continuing to contribute toward the project's objectives:

> Jorge Esmart, Ellen Dufficy, Marian McElroy, Kay Knoblauch, Barbara Guidici, Debra Dragnosky-Embert, Suzanne Tschachler, and Donna France conducted interviews of the cases and controls; Jorge Esmart, Ellen Dufficy and Mary Knapp compiled, completed, and cleaned the birth outcomes data.
>
> Carmen Pedroza and Jeanette Corbin contributed their secretarial and data entry skills.

Numerous Departmental personnel supported the project through its many phases:

> Elizabeth Shapiro, Pamela Costa and Barbara Kern supported the project through their work on the Birth Defects Registry. Jonathan Savrin assisted in the completion of the reports. Perry Cohn, William Coniglio and Jerald Fagliano contributed technical insights. George Halpin, Susan Lenox-Goldman, Jerald Fagliano, Kathleen Cunningham, Diana Kiel, James Brownlee, Rebecca Zagraniski, and Leah Ziskin provided management support and technical reviews.

This project depended for its success on cooperation and contributions by many individuals outside the Department of Health:

> Barker Hamill and his staff in the Bureau of Safe Drinking Water made it possible to conduct the exposure assessment for the drinking water studies. Robert Tucker and Leslie McGeorge of the Division of Science and Research coordinated the technical reviews and communications within Department of Environmental Protection and Energy. Our able Peer Review Panel was composed of Drs. Howard Kipen, Dirk Moore, Nigel Paneth, and Sherry Selevan. We thank the water companies of northern New Jersey for their consistent cooperation. We are indebted to the hundreds of New Jersey women who shared their time and life experiences with us.

Lastly, this project succeeded because of the constant and creative support of our project officer, Larry Edmonds of the Centers for Disease Control.

POPULATION-BASED SURVEILLANCE AND ETIOLOGICAL
RESEARCH OF ADVERSE REPRODUCTIVE OUTCOMES AND TOXIC WASTES

REPORT ON PHASE II THE EVALUATION AND DEVELOPMENT OF
DATABASES ON POTENTIAL ENVIRONMENTAL EXPOSURES

Frank J. Bove, Sc.D.

EXECUTIVE SUMMARY

This report evaluated the usefulness of computerized environmental databases of the U.S. Environmental Protection Agency (EPA) and New Jersey (NJ) for estimating potential population exposures to environmental pollutants at the municipality level.

The following criteria were among those used to evaluate the databases: practicality of use, the variety of the chemicals reported, the completeness of the database, application of a well-defined set of quality assurance and control (QA/QC) procedures, and adequate editing, management and geocoding.

Most importantly, the sampling data contained in a database had to be appropriate for estimating contamination levels in exposure pathways during the relevant time period for the adverse reproductive outcomes of interest. Sample locations and conditions, the types of contaminants sampled and analyzed, and the analytic detection limits were some of the factors considered in determining whether a database could provide reasonable exposure estimates.

Only three databases contained data on a variety of chemicals documented to be reproductive hazards: EPA's and NJDEP's TRI (Toxic Release Inventory) databases, the NJDEP Pesticide Use Survey and the NJDEP's "A-280" (NJ Safe Drinking Water Act Testing Program For Hazardous Contaminants In Public Water Supplies). Four databases

contained data validated by a standard quality assurance and control program - EPA's SAROAD (Storage and Retrieval of Aerometric Data) and FRDS (Federal Reporting Data System), NJ Trihalomethane Levels in Public Water Systems and A-280. All databases were found, however, to have frequent data entry errors, duplicate entries and inaccurate or missing geocodes.

EPA's STORET (Storage and Retrieval) database was impractical to use because of its immense sample size, very little of which was relevant for estimating pollutant levels in exposure pathways. The sample size of SAROAD database was also immense, but aggregated data on a quarterly basis were available to estimate regional ambient air pollutant levels. Other EPA databases such as the TRI, NEDS (National Emissions Data System) and FRDS had serious data gaps. For example, NEDS excluded industrial plants with relatively small air emissions and TRI excluded certain categories of industrial facilities (e.g., incinerators, electric power plants and plants with relatively small workforces). FRDS reported results only for Maximum Contaminant Level violations, and not all water companies reported their sample results to FRDS.

After review of remedial investigations and health assessments at all NJ National Priority List (NPL) sites, it was concluded that extensive data gaps precluded the development of a NPL site database that would be useful for estimating potential toxic waste exposures to populations.

In general, the environmental databases primarily serve regulatory and administrative purposes. Their usefulness for estimating exposures for epidemiologic studies and ecologic studies at the municipality level is limited.

The project also determined that a geographic information system (GIS) would be useful for population exposure assessment if it could link street addresses with corresponding census blocks and geographic coordinates. In this way, environmental data and census information at the block level could be linked to individual data on disease status.

The incorporation of the US Census TIGER files has given NJDEP's GIS the capability to make such linkages.

CONTENTS

POPULATION-BASED SURVEILLANCE AND ETIOLOGICAL RESEARCH OF ADVERSE REPRODUCTIVE OUTCOMES AND TOXIC WASTES

REPORT ON PHASE II: THE EVALUATION AND DEVELOPMENT OF DATABASES ON POTENTIAL ENVIRONMENTAL EXPOSURES

INTRODUCTION

In 1986, the New Jersey Department of Health (NJDOH) entered into a five-year cooperative agreement with the US Centers for Disease Control (CDC). The goal of the project was to develop and apply appropriate methodology to assess the relationships between adverse reproductive outcomes and population exposures to environmental pollutants, particularly toxic waste site contamination. It was anticipated that the results of the project would be useful for the prevention of adverse reproductive outcomes.

The four main objectives of the project were:

1) To enhance the Department of Health's capability to conduct surveillance and etiological research of birth defects, low birthweight, fetal deaths and infant mortality;

2) To evaluate the appropriateness of available data on environmental pollution for estimating exposures to populations;

3) To perform ecological (correlational) analyses at the municipality level of fetal and infant mortality, birth defects and low birthweight and estimated population exposure to environmental pollutants (using appropriate environmental databases); and

4) To conduct etiological studies of birth defects, fetal deaths and low birthweight and exposures to environmental pollutants.

The project was divided into four "phases" corresponding to these four objectives.

This report focuses on the activities performed in Phase II of the project. Federal and state environmental databases are described and evaluated as to their appropriateness for estimating potential population exposures to environmental pollutants. In addition, the feasibility of developing a database based on data from remedial investigations and health assessments at National Priority List (NPL) "Superfund" sites for the purpose of estimating potential population exposures to toxic waste contaminants is examined.

Phase II had four objectives:

1) To evaluate the usefulness of computerized federal and state environmental databases for estimating population exposures to environmental pollutants at the municipality level for the entire state of New Jersey (NJ);

2) To develop a new database of toxic waste sites on the National Priority List (NPL) based on information contained in their remedial investigations and health assessments, and to evaluate the usefulness of this database for estimating toxic waste exposures to populations;

3) To evaluate the utility of the New Jersey Department of Environmental Protection (NJDEP) Geographic Information System (GIS) for estimating exposures of populations to environmental pollutants; and

4) To define specific populations potentially exposed to environmental pollutants for etiological studies of adverse reproductive outcomes.

To realize these objectives, evaluative criteria were established.

First, a database should be both computerized and practical to use. For example, a database should not contain so much data irrelevant to exposure assessment that massive manipulation would be necessary to obtain useful information from it. In addition, a database should be relatively inexpensive to use. Ideally, it should be easy to download a database to a personal computer.

Second, a database should include information on a sufficient variety of chemicals, i.e., it should include chemicals, for which analytical methods exist, that one could expect to find in the medium (or media) covered by the database, and that are known or suspected to be toxic to human health.

Third, a database should be reasonably complete. For example, it should cover the entire state with minimal data gaps or exclusions. In addition, the sample size should be large enough to evaluate variability over space and time. Ideally, the database would include several samples on each sampling date so that stable estimates of contamination levels could be calculated.

Fourth, the development and maintenance of a database should follow a well-defined set of quality assurance and control (QA/QC) procedures. These procedures would be used routinely to validate the sample data and ensure that data entry errors, duplicate records, errors in geocoding, missing data items and other problems are minimized.

Finally, and most important, the sampling data contained in a database must be appropriate for estimating contamination levels in exposure pathways during the relevant time period for the adverse reproductive outcomes of interest. Exposure pathways include drinking water, indoor and outdoor air, consumption of contaminated food and dermal contact or accidental ingestion of surface soil, surface water, leachate or pesticide residues.

To evaluate the appropriateness of the sample data, factors such as the sample locations and conditions, the types of contaminants sampled and analyzed, and the analytic detection limits should be considered. For example, soil, air and surface water samples taken onsite at an NPL site may be necessary to determine potential exposures to workers onsite, but these samples shed little light on potential exposures to populations residing offsite. Similarly, surface water sample data are uninformative for estimating potential population exposures unless additional information is available on the uses (e.g., recreational) of the surface water represented by the sample.

METHODS

Computerized databases were identified. Those databases that could be easily downloaded to personal computer were obtained from NJDEP and the US Environmental Protection Agency (EPA) and evaluated directly by the project. Evaluations of databases not obtained were based on published reports describing the databases and discussions with researchers in academic, state or federal institutions who had first-hand experience with these databases.

The criteria mentioned in the previous section were applied to each database to assess its suitability for estimating potential population exposures to environmental pollutants. (See Table 1.)

In order to create a database on NPL sites, the project reviewed health assessments and available remedial investigations for all NJ NPL sites. Data abstraction forms were developed and data were entered into a dBase III database file. Frequencies of data gaps were calculated and the overall evaluation of the database was based on the above criteria.

RESULTS

I. FEDERAL DATABASES

A. GEMS (Graphical Exposure Modeling System)

1. Description

The project evaluated the EPA GEMS and the databases to which GEMS provides access - AIRS (Aeromatic Information Retrieval System), STORET (Storage and Retrieval) and FRDS (Federal Reporting Data Systems).

GEMS resides on the VAX computers maintained by the EPA at Research Triangle Park, North Carolina (US EPA, 1986). It is interactive. The user can model the migration of environmental contaminants in various media (e.g., air, groundwater), generate surface water and groundwater maps, display data from meteorological stations, display prevailing wind directions ("wind roses"), plot data overlays and shade areas according to data values. GEMS has the capability to manipulate 1980 census data geographically in order to obtain estimates of population and housing density within various distances (concentric circles) around a site (a latitude/longitude point).

GEMS can link data from different databases and provide overlays of information so that a multi-faceted view of potential exposures to a population is presented. For example, the air model program and the air database can be linked to the population data retrieval program and its census data as well as to the groundwater and surface water databases, and all this information can be presented as overlays on a county map or it can be assigned to a particular latitude and longitude point or zipcode.

However, GEMS is limited by the databases and models it uses. For example, in order to display a wind rose around a toxic waste site, GEMS applies information from the meteorological station nearest to the site. But, it is not unusual for a toxic waste site to have local wind patterns considerably different than those occurring at the nearest (usually quite distant) meteorological station.

Certain databases linked to GEMS, such as the industrial discharge permit databases and the RCRA database, primarily serve management and enforcement purposes and therefore are not useful for exposure assessment purposes. Ambient databases discussed in the next section, such as AIRS, STORET and FRDS, also are of limited use for estimating population exposures because their primary focus is regulatory.

2. Estimating Population Density Using GEMS

In this section, the utility of estimating population and housing densities at various distances from a toxic waste site is assessed. GEMS does not have access to all the census information available at the block level (such as age and sex distribution). Furthermore, during the early part of the past decade, many areas of NJ experienced significant population growth changes so that by 1985 (when the NJ Birth Defects Registry began operation) the 1980 census did not adequately characterize these areas.

The algorithm GEMS uses to calculate population and housing density within concentric circle distances from a latitude and longitude point produces estimates that do not correspond well to estimates made in the field. The algorithm identifies each census block or enumeration district whose centroid point is located within the concentric circle distance requested by the user. Then the algorithm sums the population and housing densities of each of the identified census blocks and enumeration districts.

The results often underestimate the actual population density within a mile of a toxic waste site. For example, at the Helen Kramer Landfill in Mantua, NJ, GEMS estimated no population within a half-mile of the site whereas the remedial investigation (RI) at this site estimated over 300 people living this close to the site.

Similarly, for other NJ NPL sites with RI field estimates of population density such as the Combe Fill South Landfill, Goose Farm, Krysowaty Farm and the GEMS Landfill, GEMS estimated inaccurately that no people resided within one kilometer of these sites. (Actual population densities within one kilometer based on RIs: Combe Fill South Landfill = 170; Goose Farm = 40; Krysowaty Farm = 200; and GEMS Landfill = 480).

However, for the Bog Creek Site, the GEMS estimate of population density within one mile of the site (715) corresponded reasonably well with the RI field estimate (900). For the Caldwell Trucking Site, the GEMS population density estimate within one-half kilometer was much higher than the RI field estimate (1800 vs 1000).

The latitude and longitude points provided in the RI for each of the above sites were entered into GEMS in order to estimate population densities. However, more accurate coordinates for the NPL sites were obtained from the NJDEP Geographical Information System (GIS). To determine whether the source of the disparities between field and GEMS estimates of population density was the algorithm or the latitude and longitude point used for each site, coordinates from both the RI and the GIS were supplied to GEMS and the estimated population densities were compared. The estimates did not differ. Thus, it appears that the source of the disparities is the algorithm used by GEMS.

Summary. GEMS is computerized and practical to use. However, the demographic data available in GEMS is incomplete since only population and housing densities at the block level from the 1980 Census are included. Its algorithm for estimating population and housing densities around a latitude/longitude point appears to be invalid. Its use of meteorologic station information may not be useful to characterize wind patterns at a point source. For these reasons, its use to identify specific populations potentially exposed to pollutants emitted from a point source is severely limited.

B. AIRS (AEROMETRIC INFORMATION RETRIEVAL SYSTEM)

AIRS contains two databases, SAROAD and NEDS, that provide information on the outdoor air exposure pathway.

1. SAROAD (Storage and Retrieval of Aerometric Data)

SAROAD contains air monitoring sample data taken to assess state compliance with National Ambient Air Quality Standards (NAAQS) and the State Implementation Plan (SIP) (US EPA, 1985). NJ has divided itself into nine regions for NAAQS compliance monitoring. Monitoring stations are located at sites which can best characterize each region's air quality. Consequently, monitoring stations are not evenly distributed in the state but are concentrated in the northeast part of the state (Newark-Elizabeth-Jersey City) and in the Camden area (NJDEP, 1989).

Only half of the counties in NJ have monitoring stations for trace metals (NJDEP, 1989). The bulk of the data are on Clean Air Act "criteria pollutants" (lead, carbon monoxide, ozone, nitrogen dioxide, sulfur dioxide and total suspended and inhalable particulates) with limited data on trace metals, benzo(a)pyrene and extractable organic matter.

The monitoring station data are obtained following a standard QA/QC protocol and covers the years since the 1950s. The sample size for criteria pollutants is immense - 400 million observations from 14,000 monitoring stations nationwide (US EPA, 1985). Thus, it is impractical to manipulate individual observations to obtain the summary information needed for analysis. The same is true for an individual state's database. In NJ, criteria pollutants are continually monitored at 26 stations and data are transmitted every minute, 24 hours per day, to a central computer.

SAROAD routinely provides annual and quarterly summary information on each criteria pollutant for each state and each region within a state. The summary information includes the sample size, the two highest values, number of days above the standards and arithmetic and geometric means.

NJDEP stores all air quality sample data on a central computer. Except for metals (e.g., lead) and particulate organic matter (e.g., Benzo(a)pyrene), NJDEP cannot provide summary data in a computerized database file. A yearly report is published but the data in the report are not computerized in a database file. Therefore, one must obtain summary data from SAROAD.

Summary. SAROAD's sample data are validated by stringent QA/QC procedures and the database is maintained properly. The immense sample size for criteria pollutants makes it impractical to use individual observations, but the database does provide quarterly and annual summary data on a routine basis.

One serious shortcoming of the database is its failure to provide information on a wide variety of chemicals. Data are sparse or nonexistent on trace metals, extractable organic matter and other toxic air pollutants. However, the most serious drawback to the use of SAROAD (or state) data is the uneven distribution of the monitoring stations. For example, three counties in NJ have no stations. This limits the usefulness of the database for characterizing air quality at the municipality level.

2. NEDS (NATIONAL EMISSIONS DATA SYSTEM)

NEDS contains information on each point source's allowable (permitted) emissions as well as its location, processes, production and other characteristics (US EPA, 1985). The pollutants covered are chiefly the criteria pollutants (particulates, sulfur dioxide, nitrogen oxides, carbon monoxide, and total hydrocarbons). Data on other toxic pollutants are sparse. To be included in the database, the point source must be permitted to emit greater than 100 lbs/yr. NEDS was established in 1972 (US EPA, 1985). In NJ, this database has been maintained by NJDEP since 1977 and is called the Air Pollution Enforcement Data Systems (APEDS).

APEDS and NEDS do not reflect the actual emissions of a point source. Often, a source is permitted to emit a much greater amount than it actually emits. This discrepancy is especially great during economic downturns. Currently, APEDS contains data by stack on 13,100 NJ plants (45,000 stacks) including about 5,000 gas stations (Held, 1988).

In 1986, NJDEP initiated the development of an Air Toxics Emission Inventory based on information available in APEDS (Carhart, 1988). The purpose was to expand the focus of APEDS to include toxic air contaminants such as benzene, chlorinated hydrocarbons and trace metals. By utilizing information available from APEDS (e.g., total suspended particles and total hydrocarbons permitted, stack parameters, type of plant), emissions of particular toxic contaminants would be estimated using mass balance or "throughput" modeling.

However, at present there are several problems with NJDEP's Air Toxics Emission Inventory. Numerous data entry errors exist in the coding for municipality of the facility, stack flow rates, raw materials and throughput quantities and other data fields. In addition, some plants do not have emission permits because they were "grandfathered" or have emission rates below the regulatory threshold of 100 lbs/yr. Given these data problems and gaps, NJDEP staff recommend that the database be used to estimate emissions by county rather than by municipality.

<u>Summary</u>. NEDS, APEDS and the Air Toxics Emissions Inventory are practical to use. Although small point source emitters and "grandfathered" plants are not included, NEDS and APEDS cover most point source air emitters. The most serious data gap is the lack of information on toxic air pollutants other than the criteria pollutants.

Another major problem is the lack of quality control in the maintenance of these databases. Data entry errors preclude the use of the databases to characterize total emissions by municipality and may significantly distort county estimates as well.

The protocols used by NJDEP to estimate toxic pollutant emissions from the APEDS plant information are based on EPA documents, but it remains to be seen whether these mass balance or throughput models accurately reflect actual emissions. However, such modeling is an improvement over simply relying on the permit information.

C. TRI (TOXIC RELEASE INVENTORY)

EPA's TRI database, which is now available on NLM's TOXNET, contains estimates provided by each industrial facility of its toxic chemical releases to air, water, land and underground injection as well as off-site waste transfers to publicly owned treatment works, landfills and other sites. This review focused on the use of TRI to estimate population exposures to outdoor air contaminants.

A facility is included in the database if:

1) its Standard Industrial Classification (SIC) code is between 20 through 39;

2) it has 10 or more full-time employees; and

3) it manufactured, processed or otherwise used one or more of the 308 toxic chemicals and 20 chemical categories covered by Section 313 of the Emergency Planning and Community Right-To-Know Act (Title III of the Superfund Amendments and Reauthorization Act) in excess of the specified threshold quantity for the reporting year (US EPA, 1988).

The first survey year was 1987; the threshold quantity for manufacturing was 75,000 lbs, and for "otherwise used", 10,000 lbs.

In NJ, data on 840 facilities were available from the TRI. The compliance rate for reporting is not known. As the threshold quantity for manufacturing has been lowered progressively (1988 - 50,000 lbs; 1989 onward - 25,000 lbs) more facilities have been included in the database. Facilities <u>not</u> included in TRI include gas stations, dry cleaners, incinerators, sewage treatment plants, electric power plants and other combustion sources, warehouses and storage facilities, landfills and hazardous waste sites. These exclusions make the TRI database less comprehensive than the NEDS and APEDS databases.

The greatest advantage TRI has over APEDS and NEDS is the wide variety of chemicals included in the database. On the other hand, a limitation of TRI is that facilities are not required to report toxic byproducts such as dioxin and dibenzofurans.

QA/QC problems exist in the maintenance of the database. The copy available to states was not geocoded by municipality but by zipcode. Spelling errors occurred in the chemical name, facility name and municipality fields. Missing data occurred in the chemical name and CAS number fields. In order to use this database, we corrected the data entry errors and geocoded each facility. Fortunately, the total number of records in the NJ database (3029) was small enough so that it was practical to correct errors and enter the municipality codes.

A major QA/QC problem concerns the self-reporting of emissions by the facilities. Firms are only required to estimate their emissions, except that when actual monitoring has been performed the results must be reported. About 6% of reporting firms actually performed monitoring to estimate emissions. However, it is not known whether standard sampling and analytic procedures were used. About 30% estimated their emissions using mass balance calculations or published emission factors, and these methods have not been standardized. Most firms estimated emissions using "best engineering judgment". The latter is obviously not a standardized procedure, so the accuracy of these estimates is especially suspect.

An additional problem with the quality of the reported estimates is that EPA allows firms to report emissions under 1000 lbs by giving ranges (1 - 499 lbs; 500 - 999 lbs). TRI does not contain any information on peak (short-term) emissions.

Summary. The TRI database is practical to use. It covers a broad spectrum of toxic pollutants. The major data gap concerns the exclusion of small point source emitters such as dry cleaners and gas stations as well as some large point source emitters such as incinerators, power plants and sewage treatment plants. Taken together, these omitted point sources contribute substantially to the total air pollutant load in a municipality.

Serious QA/QC problems exist in the database. Most of the data have not been validated, and there are numerous errors that require manual correction. Given that most of the data are based on "guesstimates", that peak emissions are not provided and that many firms report ranges, using the data for exposure assessment is problematic. It makes little sense to apply air dispersion models to such data or to use the data to assign exposures to individuals in epidemiologic studies. At best, the data can be used in ecologic studies to characterize potential exposures to stack and fugitive industrial toxic air pollutants at the municipality level. However, it must be expected that considerable exposure misclassification at the municipality level will occur.

Consistent with the above evaluation, this project performed ecological analyses using the TRI data to estimate the potential for exposure to air emissions. For all chemicals and for particular chemical classes, stack and fugitive air emissions from all plants in a municipality were summed. These values were compared with municipality rates for various adverse reproductive outcomes in a correlational analysis.

D. STORET (STORAGE AND RETRIEVAL SYSTEM)

STORET is the national ambient water quality system. It contains more than 90 million pieces of data from about 600,000 sampling sites (streams, lakes, ponds, wells, reservoirs, estuaries, oceans, municipal and industrial influents and effluents) throughout the US (US EPA, 1985). STORET is used primarily to monitor compliance with ambient water quality standards and effluent discharge permits. It has also been used to evaluate long-term trends in ambient water quality. Data are available for the years since the early 1960's.

STORET is difficult to use because of its immense size. The sample size varies widely over chemicals and media (e.g., there are very few data on sediment concentrations of many chemicals). Another problem is that most of the database involves information on traditional water parameters (e.g., coliform levels, turbidity, dissolved oxygen, general inorganics, nitrogen and phosphorus) with limited data on priority toxic pollutants such as pesticides, other organics and heavy metals (Staples, 1985).

A major problem with STORET is that quality assurance procedures are not established for data entered into the system (Staples, 1985). Data have been collected by many agencies for various purposes using disparate QA/QC procedures. It is possible that data quality has improved in the 1980's since better laboratory techniques have been available, but no evaluation of data quality has been reported.

Most of the sample data are not relevant for estimating pollutant levels in exposure pathways such as public drinking water supplies or surface water areas where fishing and recreation take place. STORET contains no information on the possible exposure pathways so information from other sources is necessary to evaluate whether any data in STORET can be used to estimate population exposures.

Summary. STORET is impractical to use, its information on toxic pollutants is very limited and the quality of its data is unknown. Worse, most of its data are not relevant for estimating exposures to populations.

E. FRDS (FEDERAL REPORTING DATA SYSTEM)

FRDS contains data by water company purveyors on primary and secondary contaminants in drinking water. These contaminants include inorganics (arsenic, lead, nitrates, mercury, etc.), trihalomethanes and selected pesticides (endrin, lindane, methoxychlor, toxaphene and chlorophenoxys).

Other parameters include the purveyor's number of services (hook-ups) and water source (groundwater or surface water). The database covers the years since 1980.

FRDS is practical to use. However, there are serious sample size limitations and data gaps. Since FRDS' function is primarily regulatory, only violations of the Maximum Contaminant Level (MCL) are recorded. Values under the MCLs are not provided. Most purveyors report only one sample per year for the inorganic and pesticide contaminants and some purveyors do not report at all or report less than annually. The database does not provide information on a variety of chemicals of concern such as solvents and PCBs.

The data are obtained following standard QA/QC procedures but analytic detection limits for the same contaminant vary by purveyor (e.g., one purveyor reported levels of arsenic as "< 0.05 ppm" whereas another reported levels of arsenic as "< 0.002 ppm".)

A hard copy of the FRDS' data on inorganics for NJ was obtained from NJDEP. Numerous data entry errors were found including impossible values and missing "less than" signs.

Although FRDS provides data on an exposure pathway (i.e., drinking water), additional information is needed to estimate exposures to populations. For example, information on a water purveyor's distribution system is necessary to determine the population served by the purveyor. In order to use the inorganic data in our Phase IV four etiological studies, it was necessary to link these data with distribution information obtained from NJDEP files. Since the distribution data are not computerized, the linkage had to be performed manually.

<u>Summary</u>. FRDS is practical to use but has serious data gaps and data entry errors. These problems limit severely the usefulness of the database for characterizing population exposures to public drinking water supplies.

F. HAZARDOUS WASTE DATABASES

The Hazardous Waste Data Management System (HWDMS) contains administrative information on facilities regulated under the Resource Conservation and Recovery Act (RCRA) including location, type of waste processed, compliance enforcement actions and groundwater monitoring actions. Other than location of a facility and its compliance history, there is no information that could be used to evaluate potential exposures to populations.

The Comprehensive Environmental Response, Compensation and Liability Information System (CERCLIS) contains administrative information on toxic waste sites and their location. The database is used to track the progress of these sites through the "Superfund" process. The National Priority List (NPL) Technical Database is maintained by the MITRE Corporation for EPA and contains administrative information on the Superfund sites (MITRE, 1986). Other than location of a site and its hazard ranking score (along with the component scores), there is no information that could be used to evaluate potential exposures to populations. The hazard ranking score (HRS) is described below.

These two databases were obtained as database files from ATSDR (NPL Database) and NJDEP (CERCLIS Database). The project geocoded the databases by county and municipality. After correction of location errors and removal of duplicates, the location of the NPL and CERCLIS sites was used to determine the number of sites per square mile in a municipality for the project's Phase III three ecological (correlational) studies.

Simply knowing the location of a toxic waste site tells us nothing about the possible exposure pathways emanating from the site, the level of contamination in these exposure pathways or the population potentially exposed. In addition, it is often the case in NJ that a landfill site is located at the corner of one municipality while the population potentially exposed resides in the adjacent municipality. Thus, using the town of location as the exposure variable in ecological analyses of landfill sites will incorrectly identify the population potentially exposed.

Given that location of the site is an extremely poor surrogate for potential population exposures to toxic wastes, we decided to evaluate whether the HRS might offer a better exposure surrogate by providing a ranking of each site's exposure potential. If a correlational between the HRS and adverse reproductive outcomes were found, then those sites with high HRS values could be examined further (e.g., reviewing site investigation data or performing site visits) to identify the specific populations potentially exposed.

There are several limitations to the HRS. First, the HRS does not quantify the probability of exposure or its severity. In particular, it does not differentiate between sites with minimal and severe groundwater contamination.

Second, the direction of groundwater movement is not taken into account in the scoring. About 80% of the NPL sites in NJ received a score indicating that contaminants from these sites had entered nearby aquifers which were the sole source of drinking water for nearby communities (i.e., no alternate municipal supplies were available). However, after reviewing remedial investigations at these sites, we found that less than 10% had contaminated public drinking water supplies and less than 25% had contaminated private wells that were not on or immediately adjacent to the sites.

In addition, the HRS scoring for groundwater exposure potential cannot be relied upon to predict future potential exposures. At the time of scoring, little (if any) groundwater samples have been taken at the sites. Hence, the scoring is based on sparse (or no) sampling data.

Third, the risk of air contamination is not considered unless there are sample data indicating that the air has been contaminated. Given that air sampling is virtually never performed before the remedial investigation is begun (and rarely performed during the remedial investigation itself), the HRS does not reflect the potential for population exposures to air contaminants.

Fourth, the scoring for the category "waste quantity" is arbitrary since it is unclear whether the category refers to current or past amounts. In addition, the amount of waste onsite is not indicative of the severity or possibility of exposures to nearby populations. However, the overall HRS is greatly influenced by its score.

Fifth, a large percentage of the information upon which the HRS is based has not been verified. A standard QA/QC procedure for the collection of this information has not been established. Thus, the decision on whether to propose a site for NPL listing is not based on data that has been through a QA/QC process and verified.

Summary. Our conclusion is that the HRS cannot be used as a means of ranking sites for exposure potential and severity.

II. NJDEP DATABASES

In addition to the state versions of the federal databases discussed above, the project has also evaluated databases developed by NJDEP. These include a survey of agricultural pesticide use and public drinking water databases for specific volatile organics and total trihalomethanes. NJDEP's Geographic Information System (GIS) was also evaluated.

A. NJ SAFE DRINKING WATER ACT A-280: A TESTING PROGRAM FOR HAZARDOUS CONTAMINANTS IN PUBLIC WATER SUPPLIES

The A-280 database consists of data from the semi-annual testing of all public water systems for 16 compounds: 14 volatile organics (including trichloroethylene, perchloroethylene and benzene), PCBs and chlordane. Initial testing of public water systems occurred in the winter of 1984-85 (NJDEP, 1986). The first full year of sampling occurred in 1985 (NJDEP, 1987). All public water purveyors must sample their distribution systems twice annually for the 16 contaminants. Only NJDEP-approved laboratories may perform the sample analyses. The compliance rate is better than 90%.

The database identifies the purveyor, date of the sample, the type of sample (a raw well sample before treatment, a sample at the point where the supply enters the system, or a distribution sample), the contaminants found and the levels detected. Since the purposes of the database are to monitor public water systems and to intervene quickly when contamination levels exceed chemical-specific action levels, NJDEP may require additional sampling if contamination is detected. Some problem systems have been sampled on a monthly basis.

The advantages of the database are the high compliance rate and the mandatory use of certified laboratories that follow standard QA/QC procedures to analyze samples. Most importantly, the types of contaminants analyzed and the rate of sampling make the A-280 data-base unique among other federal and state drinking water databases. For these reasons, this database was used in the project's first Phase IV case-control study as well as in a NJDOH ecological study of leukemia (Fagliano, 1987).

The A-280 database has several disadvantages. First, the data are listed by purveyor, and there is no linkage between purveyor and population served. It was necessary to obtain from NJDEP a "hard copy" list of the purveyors serving each municipality in order to utilize the database in the project's epidemiologic studies.

Second, information on a purveyor's distribution system is not computerized. For example, a purveyor may serve one part of town solely from a surface water source, another part of town solely from groundwater supplies and the rest of the town with a mixture. A purveyor may have a contaminated groundwater supply serving one part of town with the rest of the town served from a relatively clean groundwater supply. In these situations, it is necessary to know which type of supply serves a population in order to estimate drinking water exposures. At present, this information is not computerized and must be obtained from NJDEP files or be requested from purveyors. The process of digitizing this information on NJDEP's GIS is in progress.

Third, for some purveyors only one point sample every six months is available to characterize its entire distribution system. In these situations, exposure estimation requires the assumptions that the purveyor uses only one water source or, if more than one source is used, that there is complete mixing in the distribution system, and that the point in the distribution system sampled is optimal for characterizing the entire system. An additional assumption is that the sample reflects the water quality over a six month period.

Fourth, numerous data entry errors were encountered in the coding of the type of sample. Most often, raw water samples before treatment were miscoded as distribution samples or as samples taken at the point of delivery into the system, resulting in an over-estimation of contamination levels at the tap. After consulting with NJDEP and the purveyors, these errors were identified and corrected.

Fifth, the purveyor performs the routine sampling. Hence, the samples may not be representative of the water supply. NJDEP performs some spot checking, spot sampling and sample validation.

Summary. The A-280 database is practical to use and complete. The sample size for some purveyors is small if only semi-annual sampling is performed. These tend to be the public water systems that are free of contamination.

The variety of chemicals which are analyzed and the strict QA/QC program for sample analysis make the database unique. The database itself requires a quality control program to minimize data entry errors.

The failure to link the database with information on the population served is a serious limitation. The project was required to forge this link by hand for the study area included in the Phase IV etiological studies. An effort is underway using NJDEP's GIS to link the A-280 data, information on the distribution system and information on the population served.

Several assumptions are required in order to use the A-280 database for estimating population exposures. However, in most situations, the assumptions are reasonable and will lead to accurate exposure estimates. For complex systems, it is necessary to obtain information from the purveyors.

For this project, we decided against using the A-280 data in the ecologic studies. One reason was that the data could not be used to characterize drinking water quality in areas of the state primarily served by residential wells. In addition, there were areas of the state in which the distribution systems were so complex that adequate characterization of drinking water quality required supplementing the A-280 data with information from purveyors. It was decided to use the data in the project's epidemiologic studies, focusing on areas primarily served by public systems that were not unduly complex. Additional information was obtained from purveyors.

B. DATABASE ON TRIHALOMETHANE LEVELS IN NJ PUBLIC DRINKING WATER SYSTEMS

NJ public drinking water systems that serve at least 10,000 people are required to test quarterly for total trihalomethanes (TTHM) in their distribution system (i.e., after chlorination). Compliance appears to be complete. Only NJDEP-approved laboratories may perform the sample analyses. Most of the purveyors perform four samples on each quarterly sampling date. The database identifies the purveyor, the date of sample and the levels of TTHM detected. Although the database is computerized on NJDEP's mainframe computer, we obtained a hard copy covering the years 1982-1988.

The TTHM database has similar limitations for estimating contaminant levels in public drinking water supplies as the A-280 database. It must be linked manually to another database supplying the town(s) served by each purveyor. It must be supplemented by distribution system information from

NJDEP and/or the purveyors for those systems utilizing a mixture of groundwater and surface water sources. For exposure estimation purposes, it must be assumed that the samples represent the levels found at the tap. Numerous data entry errors in the sample results, primarily misplaced decimal points, occur in the database. Finally, the purveyors themselves perform the sampling.

Unlike the A-280 database, the TTHM database does not include all the public drinking water systems. Excluded are systems serving less than 10,000 people. However, virtually all of the excluded systems rely on groundwater sources that are relatively free of the organic material necessary for the creation of TTHM. Levels of TTHM in these systems are usually in the minimum detectable range of 1 ppb to 5 ppb.

Summary. The TTHM database is practical to use and sufficiently complete. The sample size is adequate for computing seasonal mean levels of TTHM. The database does not cover a variety of chemicals but, in conjunction with the A-280 database, provides valuable data on potential population exposures to drinking water contamination. The database suffers from inadequate maintenance and its failure to be linked directly with information on the population served. Several assumptions are required in order to use the TTHM database for estimating population exposures. However, in most situations, the assumptions are reasonable and will lead to accurate exposure estimates. The database was used in the project's epidemiologic studies.

C. NJ PESTICIDE USE SURVEY

NJ pesticide regulations require certified applicators to maintain records of pesticide use and to submit that information to NJDEP upon request. In 1986, NJDEP sent survey forms to the 3117 registered applicators in the agricultural community requesting information covering the year 1985 on the pesticides used (active and inert ingredients), the number of acres treated, the crop treated, the method of application (air or ground) and the municipality where the pesticides were applied. The compliance rate was about 95% (Louis, 1989).

The data represented 1721 separate farming operations located in 243 of the 567 municipalities. A total of 176 active ingredients were reported. The project obtained the DBASE III file from NJDEP which was used in the Phase III ecological (correlational) analysis. For all pesticides and for particular pesticide groups, the total amounts of active ingredient applied were calculated for each municipality and compared with municipality rates for adverse reproductive outcomes.

Because it provides information at the municipality level on actual usage for all growers, the NJ survey is the most comprehensive pesticide database in the country. Other states base their estimates of pesticide usage on dealers' sales records or on types of crops grown. A few states survey a percentage of growers in particular areas and the USDA surveys growers of major crops.

However, the NJ survey suffers some of the same limitations as other pesticide surveys. For example, the amount of pesticides used for agricultural purposes in a municipality may bear no relation to the probability or severity of pesticide exposures to populations living in the municipality. Of course, various models could be applied to the data to estimate levels in air, soil, groundwater and surface water but it is not at all evident that the data are appropriate for such extensive modeling or that applying such models to this data would result in more accurate characterization of exposures at the municipality level. In addition, community (as opposed to occupational) exposures to agricultural pesticides are less likely than exposures occurring from household, garden and building applications, community spraying for pests (e.g., gypsy moth and mosquitos) and herbicide spraying on utility and railroad right-of-ways. Finally, the quality of the survey is dependent upon the self-reporting of pesticide applicators.

Summary. The pesticide survey is practical to use. Given the high compliance rate, the database is sufficiently complete. The survey covers only usage in 1985. However, the data are expected to reflect closely the usage over the 1985-1988 period studied by the project since the types and amounts of pesticides used in agriculture in NJ have remained basically unchanged over this period.

The quality of the data is dependent upon industry self-reporting. The data are not directly relevant to an exposure pathway. The use of the database to estimate population exposure at the municipality level requires the assumption that the amount of pesticides applied for agricultural purposes is related to the probability and severity of population exposure in that municipality. Despite these limitations, the database is a vast improvement over pesticide surveys used in previous ecologic and epidemiologic studies.

D. NJDEP'S GEOGRAPHIC INFORMATION SYSTEM (GIS)

The software for NJDEP's GIS is ARC/INFO (Rohardt, 1986). The GIS is similar to EPA's GEMS in that it allows one to manage, analyze, map and report on geographically referenced data. The GIS consists of many layers of cartographic data and corresponding attribute data (e.g., TRI data). All layers are registered to each other and to a common base map - the 1:24,000 scale USGS Topographic Quadrangle.

Eventually, NJDEP plans to centralize all its databases in its GIS. Recently, the US Census TIGER (Topologically Integrated Geographic Encoding and Referencing system) files containing geographic coordinates for streets, addresses and census boundaries were added to the GIS. With the TIGER files, the GIS has the capability to match a street address with its census block and state plane coordinates.

In theory, linking these databases in the GIS would allow, for example, one to determine automatically the rates of congenital anomalies, prematurity and low birthweight at various distances from a hazardous waste site. Unfortunately, NJDOH does not computerize the street address of the mother's residence at time of birth from the birth certificate, so a considerable amount of manual work is still necessary.

The GIS is most appropriate for exposure assessment purposes and to delineate populations for etiological study. However, its use for the evaluation of geographic disease clusters is problematic. For example, simply locating cases of a specific disease on a GIS map and measuring their distances from a pollution source, ignores variations in the geographic distribution of populations (Bithell, 1989). In one situation, cases of cancer were found near a toxic waste site in the path of the prevailing winds and no cases were found on the opposite side of the site. It was later realized that the latter side of the site was almost entirely farmland with no population.

When the 1990 census data becomes available on the GIS it will be possible, using the TIGER files, to take into account the geographic distribution of populations. For adverse reproductive outcomes, birth certificate data could be used to determine the geographic distribution of birth populations. Disease rates could be obtained at the census block level. In situations where the census block covers a relatively large area, this approach may still be inadequate to account properly for variations in the geographic distribution of populations.

Another approach to this problem is the use of density equalizing map software (Selvin, 1987) that can transform maps so that population is distributed uniformly. This software is not yet available for use in the GIS. However, this software creates problems in interpretation. Once a map has been transformed, measuring distances of cases from a pollution source become meaningless since the map's space has been expanded or contracted to produce a map of uniform population density. On the one hand, because the space dimension is distorted, a birth defect case in a sparsely populated area may appear on the density equalized map to reside adjacent to a toxic waste site even though the case's home is actually several miles away. On the other hand, a birth defect case living adjacent to the site but in a densely populated area may be shown on the map to be residing several miles away from the site.

Maps of the distribution systems of public drinking water purveyors are available on the GIS. With databases such as A-280, TRI and the Pesticide Survey available on the GIS along with the TIGER files, it is becoming possible to display a multi-faceted picture of the levels of contaminants within the various possible exposure pathways and to define potentially exposed populations, automatically at the census block level.

Summary. The GIS would be useful for population exposure assessment if it could link street addresses with corresponding census blocks and geographic coordinates. In this way, environmental data and census information at the block level could be linked to individual data on disease status. The incorporation of the US Census TIGER files has given NJDEP's GIS the capability to make such linkages.

III. DEVELOPMENT OF A NJDOH NPL DATABASE

A. INTRODUCTION AND SUMMARY OF FINDINGS

Our objective in developing a database for NPL ("Superfund") sites was to create a database useful for characterizing potential or actual exposures to populations from NPL sites. The database was developed from information contained in Remedial Investigations (RIs) and other sources such as ATSDR and NJDOH Health Assessments, Interim Reports or Preliminary Assessments, Remedial Action Master Plans (RAMPs), and available sample data from USEPA and NJDEP.

The database included only types of information contained in the RIs and other sources deemed relevant for characterizing exposures to populations. Such information included demographic data, site history, sample data that characterized a complete exposure pathway and exposure/risk assessment information.

The database has three parts: 1) a general assessment of the site; 2) the levels of contaminants within each exposure pathway of the site; and 3) the size of the potentially exposed population within each town affected by the site. It was originally planned to use the database to develop more refined variables than the mere location of the site in order to estimate potential population exposures to NPL sites at the municipality level. These variables would then be used in the Phase III ecological (correlational) analysis. In addition, the database would be used to define populations potentially exposed to NPL sites for epidemiologic research.

After review of all the information available to us on the NPL sites, we concluded that our objective for this database could not be met. For all sites, immense data gaps existed in the demographic information. For most sites, data on the levels of contaminants in the exposure pathways were either sparse or non-existent. For a few sites, the sample data failed to pass QA/QC evaluations. For most of the recently proposed NPL sites, a preliminary investigation had not been completed. For these reasons, the only completed aspect of the database was the general assessment of the sites. Therefore, the database was not used in Phase III or Phase IV of the project.

B. CRITERIA FOR INCLUSION OF DATA IN THE NPL DATABASE

A key criterion for including sample data and other information in the NPL database was the quality of the data. Sample data had to be obtained and analyzed following standard QA/QC procedures. Each sample had to be analyzed for all substances appropriate to the history of dumping at the site, and the analytical detection limits had to be sufficiently low so that proper estimates of potential exposure could be made.

Another key criterion for including sample data in the database was the sample's representativeness, i.e., whether the sample provided representative information on a potential exposure pathway. Only sample data relevant for characterizing an exposure pathway would be included in the database. In order to determine its representativeness, a sample's location and the conditions under which it was obtained had to be evaluated. For example, in lieu of private well or public well data, only those groundwater samples in the vicinity of private wells and/or public wells provided relevant information on the drinking water exposure pathway. (Defining "vicinity" in a particular instance is a subjective decision that requires documentation). Onsite surface soil, air and surface water samples were relevant if the public had access to a site and/or workers onsite were at risk of exposure. Offsite air, water and surface soil samples were relevant only if they were taken in areas in use by the public.

The project also evaluated the quality of the data sources - the RI, the health assessment and the preliminary assessment. Criteria used to evaluate these data sources included: 1) the adequacy of the demographic information presented, 2) the completeness of information on potable water use in the vicinity of a site (including number of private wells and/or public wells and their depths, and the distribution of public water), 3) the availability of offsite air and surface soil data in the vicinity of homes, and 4) the adequacy of data on the surface water exposure pathway (including surface water and sediment samples in waters used for fishing and/or recreation, fish samples and shoreline/beach soil samples).

C. DESCRIPTION OF THE NJDOH NPL DATABASE

The database was planned to have three parts. The first part was the general assessment of the site. It contained our evaluation of the quality of information for each exposure pathway. Also included were data on demographics, site accessibility, number of private wells contaminated and whether public water systems were contaminated. Finally, it included an assessment of whether any exposure pathways emanating from the site were "complete" (i.e., potential population exposures existed) at any time during the period 1985 to 1988. A sample of the data entry form used for the general assessment is provided as Appendix A.

The second part of the database included the maximum levels in each of the site's exposure pathways of the 128 chemicals listed on EPA's hazardous substance list. A sample of the data entry form used for the exposure pathway sampling results is provided as Appendix B.

The third part of the database defined, for each municipality affected by the site, the size of the population potentially exposed to any of the site's exposure pathways. A sample of the data entry form used for the population exposure assessment is provided as Appendix C.

An additional database was created containing a narrative summary of the information contained in each RI reviewed.

D. GENERAL ASSESSMENT OF THE DATA AND THE DATA SOURCES

NJ/ATSDR Health Assessments for 96 NPL sites in NJ were reviewed. For 15 of these sites, the preliminary remedial investigations were available and reviewed (see Table 2). For 45 sites, the complete remedial investigations were available and reviewed. The remaining 36 sites were in an early stage of investigation with only preliminary information (i.e., sample data used for the HRS, a brief site description and history, and/or the Remedial Action Master Plan) available for review.

Thirty-four (35.4%) of the sites had sufficient information to assess population exposures (see Table 3). At 36 sites, no exposure assessment had been completed since the investigations were at a very preliminary stage. The most common data gap was information on air exposure pathways. Air sample data that passed QA/QC were available for only 15 sites and sixty-one (63.5%) of the sites had no air data whatsoever. In addition, most of the sites had insufficient data on groundwater migration (59.4%) and drinking water from private wells or public systems (58.3%). Few sites had data relevant to estimate exposures due to accidental ingestion and/or dermal contact with surface water and surface soil. For 38 (39.6%) sites, the only information on the population at risk was the number of people residing within three miles of the site.

For 39 (40.6%) sites, there was no information about whether trespassing occurred onsite and 12 (12.5%) sites had no information on site accessibility by the public (see Table 4.)

For each site, we examined the available data for the site's possible exposure pathways and assessed the likelihood that population exposures to any of these pathways occurred over the 1985-1988 period studied by the project. If population exposures to an exposure pathway were likely, the pathway was considered "complete". The results of this assessment are presented in Table 5.

For 40 (41.7%) sites, we were unable to assess whether a complete air exposure pathway existed. For 9 (9.4%) sites, it appeared likely from either the sample data or community complaints that a complete air exposure pathway existed. For the remaining sites, it appeared unlikely that a complete air exposure pathway existed, although for many of these sites the air data were poor or non-existent. (Many of these sites were in isolated areas.)

After review of the remedial investigations and health assessments, the project determined that the majority of the sites did not contaminate public water systems (86.5%) or private wells (61.5%). For 25 (26%) sites, the information was inadequate to assess whether private wells were being affected. For 9 (9.4%) sites, there was no information on whether public systems were being impacted. A majority of the sites did not have complete surface soil (58.3%) or surface water (62.5%) exposure pathways. For 27 (28.1%) sites, information was inadequate to assess whether the surface soil and surface water exposure pathways were complete.

In general, the information available on NPL sites from which to assess potential population exposures is of poor quality. The majority of the sites in NJ have not been adequately investigated (See Table 3.) In particular, information on demographics and air quality near these sites is so poor that for virtually all sites an exposure assessment cannot be done with any confidence. It is clear that the exposure and health assessments must be given much higher priority during the remedial investigations.

Given the paucity of data on exposure pathways and populations-at-risk that was available to us, the project decided to complete only the general assessment part of the database. The database containing the narrative summaries of the RIs was also completed.

E. DEVELOPING A DATABASE OF SAMPLE RESULTS FROM REMEDIAL INVESTIGATIONS THAT IS USEFUL FOR EXPOSURE ASSESSMENT

In the process of considering the feasibility of developing a database of sample data from the RIs, several issues emerged. First, there was the question of what sample data to include. As stated above, we decided to include only data representing contaminant levels in complete exposure pathways since the database was to be used for the purpose of exposure assessment. The database would include all 128 chemicals in EPA's Hazardous Substance List (HSL).

Immediately, difficulties were encountered. The exposure pathways themselves were usually poorly defined so it was not obvious which sample data represented the potential exposure. Often the samples were widely scattered and their results varied greatly. In many instances, samples taken in the areas of heavy contamination were invalidated for QA/QC reasons. Finally, some samples were not analyzed for all the HSL chemicals.

In general, the RIs contained few samples that were reliable, validated and complete and that represented contaminant levels in complete exposure pathways. The reason for this is that the primary purpose of sampling at an NPL site has been to determine the extent of environmental contamination in order to plan for eventual site clean-up. To date, exposure assessment has been of secondary consideration in the site investigation. Most sampling tended to be irrelevant to exposure assessment, and the choice of sampling locations (&/or times) was not based on exposure assessment criteria.

Another issue in the development of the database was whether to report all sample data relevant to a complete exposure pathway or use a summary measure such as maximum or geometric mean concentration. Recording all the sample data would simply overwhelm the database and those using it.Therefore, it was decided that a summary measure would be used to characterize an exposure pathway.

The choice of a summary measure depends on a number of factors including the outcome of interest, the effects of the contaminant(s) of interest and the number of samples available. Our decision was to use the maximum concentration found among the samples that were relevant to a particular exposure pathway. One reason was that rarely were there more than a few samples available in an exposure pathway from which to estimate a mean or median.

Since exposure assessment was of secondary concern in the RI, a second reason for using the maximum concentration was that sampling locations were often not optimal for representing contaminant levels at the point of exposure. Thus it was conceivable that the maximum value might actually better reflect the intensity of exposure than the mean value.

Given the problematic nature of the data, it was not clear that the geometric mean, arithmetic mean or the median had any advantage over the maximum value. Use of the maximum value also required less staff time and effort. Finally, if we were using present sample data to estimate historic exposure (when contamination was most often worse), the maximum concentration might better reflect past exposure than other summary measures. Even if the mean concentration best reflected the levels at the time of sample, it could seriously underestimate the levels at which the population was exposed in the past.

A final issue in the development of the database concerned off-site contamination that the RI claimed to be unlikely from the NPL site. We decided to include such data since: 1) the purpose of the database was to include all information in the RI useful to assess exposures to a population from environmental contamination even if there was some question as to whether it was site-related, and 2) most often, it was unclear whether the site was responsible or not for the contamination. Misclassifying the site as responsible was not an important problem for exposure surveillance or for the correlational/ecological studies. If an epidemiologic study was contemplated for the site and its surrounding area, then an attempt would be made to determine the source of the contamination.

CONCLUSION

There is no published report by either federal or state agencies that evaluates the use of environmental databases for surveillance and epidemiologic research. The need for such an evaluation is growing as progressively more states develop surveillance systems for cancer and/or birth defects and attempt to use these systems to explore causal factors in the environment. Since state public health agencies have limited resources to perform sampling to estimate population exposures to environmental contaminants, their attention has focused increasingly on the use of environmental databases to generate hypotheses and prioritize future research directions (Zagraniski, 1989).

One objective of Phase II of the project was to address this need by preparing a report evaluating the usefulness of computerized databases for estimating exposures of populations to environmental pollutants. The evaluation required the development of criteria to judge the appropriateness of the databases for surveillance and epidemiologic research. We are aware of no published standard set of criteria for this purpose, but a set of criteria has been proposed (Zagraniski, 1986).

Except for STORET, all the databases evaluated were practical to use. However, most databases failed one or more of the other evaluative criteria. For example, most databases covered only a limited number of chemicals. However, the TRI database and NJDEP's drinking water databases and pesticide survey covered a wide variety of chemicals. Several databases lacked completeness due to exclusions and other missing data problems. These included FRDS, NEDS, TRI, and NJDEP's Trihalomethane database. Several databases had no standard QA/QC procedures to validate the data. These included TRI, STORET and NJDEP's pesticide survey. Finally, most databases contained data entry errors and duplicate records, and the coding of county/municipality was inaccurate or nonexistent.

After review of available federal and state environmental databases, the project decided that EPA's TRI air emissions database and NJDEP's Pesticide Survey were most suitable to for exploration in the Phase III ecological (correlational) analyses, despite the limitations of these databases described above. The locations of the NPL sites and the CERCLIS sites were used in the Phase III analyses as well. NJDEP's A-280 and Trihalomethane databases were found suitable for the Phase IV etiological studies.

Since none of the environmental databases were designed for the purpose of exposure assessment, it was expected that their usefulness in epidemiologic research would be limited. Only the drinking water databases were appropriate to characterize exposures at the level of individuals. However, even for these, exposure estimation was not straightforward but required numerous assumptions. For example, the assumption was regularly made that a sample in one part of a company's system was representative of the entire system since no other data were available. Infrequent sampling meant that exposure over a considerable period of time was estimated by a single sample.

Certainly, it is appropriate to use several of the other environmental databases to provide a crude picture of potential exposures at the municipality level. However, their use in ecologic studies requires rather shaky assumptions (and even "leaps of faith"), making interpretation of study findings difficult. Nevertheless, it is most often the only data available to environmental epidemiologists. It is possible that findings from ecologic studies utilizing these databases might suggest directions for further research. This has been the case for studies utilizing the drinking water databases and SAROAD. None of the other databases have been used in ecologic or epidemiologic studies.

A second objective of Phase II of the project was to develop a new NPL database that would be useful for estimating toxic waste exposures to populations. MITRE Corporation had developed a NPL database derived from initial site investigations used to determine the HRS for each site. However, access to the database was restricted and MITRE warned against relying on the database for exposure assessment purposes (MITRE, 1986). We decided to base our NPL database on the remedial investigation and health assessment performed at each site.

After review of remedial investigations and/or health assessments at 96 NPL sites in NJ, we concluded that a NPL database would not be useful for estimating potential exposures to populations. In particular, we concluded that developing a database containing the results of sampling performed in the investigations would require a considerable investment of staff time with no apparent benefit. In general, the remedial investigations contained few samples that were reliable, validated and complete and that could be used to estimate exposures to populations. These investigations were primarily concerned with determining the extent of environmental contamination in order to plan for eventual site clean-up, and exposure assessment was only of secondary importance. Most sampling tended to be irrelevant to exposure assessment, and the choice of sampling locations (&/or times) was not based on exposure assessment criteria.

Since the health assessments were based on the sampling performed during the remedial investigations, they also did not prove useful for estimating population exposures. It is obvious that major changes are necessary in the way remedial investigations are performed so that exposure assessment is given a higher priority.

A third objective of Phase II of the project was to assess the utility of NJDEP's GIS for characterizing exposures to populations. GIS technology has been used in the past to estimate exposures to soils contaminated by a lead smelter (von Lindern, 1986). Our conclusion was that NJDEP's GIS holds much promise for future environmental exposure assessment work.

REFERENCES

Bithell JF and Stone RA: On statistical methods for analyzing the geographical distribution of cancer cases near nuclear installations. J Epidemiol Comm Health 1989; 43:79-85.

Carhart BS, et. al.: Refining Air Toxics Emissions Inventories:A New Jersey Case Study. Presentation at the 81st Annual Meeting of APCA. June 1988.

Fagliano J, et. al.: Drinking Water Contamination and the Incidence of Leukemia: An Ecologic Study. New Jersey Department of Health, Division of Occupational and Environmental Health, December 1987.

Held JL, Opperman AJ, Laznow J: Maturing of New Jersey's Air Toxic Program. Presentation at the 81st Annual Meeting of APCA. June 1988.

Louis JB, Robson MG and Hamilton GC: New Jersey Pesticide Use Survey. Presented at a conference sponsored by the Virginia Water Resources Research Center. Richmond; March 1989.

MITRE Corporation: National Priorities List Technical Data Base. Draft. February 1986.

NJDEP, Office of Science and Research and Division of Water Resources: Final Report - Results of Initial Testing for Hazardous Contaminants in Public Water Supplies Under Assembly Bill A-280 Through January 9, 1985. February 1986.

NJDEP, Office of Science and Research and Bureau of Safe Drinking Water: Final Report - Results of Testing for Hazardous Contaminants in Public Water Supplies Under Assembly Bill A-280 Through December 31, 1985. March 1987.

NJDEP, Division of Environmental Quality: 1988 Air Quality Report. July 1989.

Rohardt S, et. al.: Development of a Geographic Database for the Department of Environmental Protection: Preliminary Assessment of Divisional Requirements for Geographic Information. NJDEP, Office of Science and Research. March 1986.

Selvin S, et. al.: Spatial Distribution of Disease: Three Case Studies. JNCI 1987; 79:417-423.

Staples CA, Werner AF, Hoogheem TJ: Assessment of Priority Pollutant Concentrations in the United States Using STORET Database. Environ Tox Chem 4, 1985:131-142.

U.S. EPA, Office of Toxic Substances: Methods For Assessing Exposure to Chemical Substances in the Ambient Environment. July 1985.

U.S. EPA, Office of Toxic Substances: Graphical Exposure Modeling System (GEMS) User's Guide, Draft. April, 1986.

U.S. EPA, Office of Toxic Substances: Toxic Chemical Release Inventory Reporting Form R and Instructions. March 1988.

von Lindern IH and von Braun MC: The Use of Geographic Information Systems as an Interdisciplinary Tool in Smelter Site Remediations. Presentation at The 7th National Conference On Management Of Uncontrolled Hazardous Waste Sites. December 1986.

Zagraniski RT: "Environmental Data Sources For Surveillance of Environmental Reproductive Hazards". Presented at the Second National Environmental Health Conference sponsored by CDC, ATSDR and ASTHO. Denver; February 1986.

Zagraniski RT: "Use of Environmental Data Bases". Presentation at the Association of State and Territorial Health Risk Assessment (ATSDR) Workshop on Environmental Data Bases. San Antonio; June 1989.

TABLES

TABLE 1

AVAILABLE DATABASES AND EVALUATION CRITERIA

DATABASE	PRACTICAL TO USE	INCLUDES VARIETY OF CHEMICALS	COMPLETE	STANDARD QA/QC	DATABASE MAINTAINED PROPERLY	USEFUL FOR EXPOSURE ASSESSMENT
Federal Databases:						
GEMS	X					
SAROAD	X			X	X	
NEDS/ APEDS	X					
TRI	X	X				X
STORET			X			
FRDS	X			X		X
CERCLIS	X		X		X	
MITRE-NPL DB	X		X			
State Databases:						
NJDEP-A280	X	X	X	X		X
NJDEP-TTHM	X			X		X
NJDEP PESTICIDE SURVEY	X	X	X			X

NJDOH, 1992

TABLE 2

STAGE OF REMEDIAL INVESTIGATION (RI) COMPLETED FOR 96 NPL SITES IN NEW JERSEY (as of 1988)

	PHASE 2 (completed RI)	PHASE 1 (preliminary RI)	PRE-RI
NPL Sites	45	15	36
Percent of Total	47%	16%	37%

NJDOH, 1992

TABLE 3

PERCENT OF 96 NPL SITES IN NEW JERSEY BY
DATA QUALITY OF REMEDIAL INVESTIGATIONS (as of 1988)

	ONSITE/ OFFSITE AIR	OFFSITE GROUND WATER	DRINKING WATER DATA	EXPOSURE ASSESSMENT	DATA ON POPULATION AT-RISK
ADEQUATE	15	40	42	35	60
POOR	21	52	38	65	40
NO DATA	63	7	20	0	0

NJDOH, 1992

TABLE 4

REMEDIAL INVESTIGATION INFORMATION ON SITE ACCESSIBILITY AND TRESPASSING ON 96 NPL SITES IN NEW JERSEY (as of 1988)

	NUMBER OF SITES	PERCENT OF TOTAL
SITE ACCESSIBILITY: Completely Fenced	24	25
Partial Restriction	21	22
No Restriction	40	41
No Information	12	12
ON-SITE TRESPASSING: Yes	31	31
No	28	28
ND	40	41

ND = Could Not Be Determined
NJDOH, 1992

TABLE 5

PERCENT OF 96 NPL SITES IN NEW JERSEY
WITH COMPLETE EXPOSURE PATHWAYS (as of 1988) *

	AIR	PRIVATE WELLS	PUBLIC DRINKING WATER	OFFSITE SOIL	OFFSITE SURFACE WATER
YES	9	12	4	13	9
NO	49	61	86	58	62
ND	42	26	9	28	28

* = Based on RIs, HAs, and other information
ND = Could not be determined
NJDOH, 1992

APPENDICES

Major exposure pathways identified by RI risk assessment:

Pathway (0=n; 1=y; 8=N/A; 9=no info)	# potentially exposed
1. air __	1. ________
2. SW/sediments __	2. ________
3. soil __	3. ________
4. private wells __	4. ________
5. public water __	5. ________

	Population	Housing
0.25 mi		
0.5 mi		
1.0 mi		
1.5 mi		
2.0 mi		

SUBSTANCES NOT LISTED ON THE MEDIA SPECIFIC FORM

SUBSTANCE	PRESENT ONSITE	PRESENT OFFSITE	MEDIA CODE	MAX CONC.*
Radiation:				
alpha				
beta				
gamma				
Pesticide(s):				
1.				
2.				
3.				
Other:**				
1.				
2.				
3.				

* (ppm). Maximum concentration in an exposure pathway.

** Substances important to the exposure assessment (e.g., chemicals unique to a site).

Appendix B

: 1=ABSENT; 2=BELOW METHOD DETECTION; 3=PRESENT, QUALIFIED; 4=PRESENT, ESTIMATED; 5=PRESENT; 9=NOT TESTED/REJECTED

ALL CONCENTRATIONS ([ ]) ARE IN PARTS PER MILLION

ICAL NAME	SURFACE WATER ON ?|[ ]	SURFACE WATER OFF ?|[ ]	GROUNDWATER ON ?|[ ]	GROUNDWATER OFF ?|[ ]	AIR ON ?|[ ]	AIR OFF ?|[ ]	DRINKING WATER PRI ?|[ ]	DRINKING WATER PUB ?|[ ]	SEDIMENT ON ?|[ ]	SEDIMENT OFF ?|[ ]	SURFACE SOIL ON ?|[ ]	SURFACE SOIL OFF ?|[ ]
methylene chl												
PCE												
TCA (1,1,1)												
TCE												
vinyl chlor												
chlorobenzene												
1,2-DCB												
1,3-DCB												
1,4-DCB												
1,2,4-TCB												
hexchlbenzene												
benzene												
toluene												
styrene												
xylenes												
benzoic acid												
benzyl alcohol												
phenol												